AISB/IACAP World Congress 2012

Birmingham, UK, 2-6 July 2012

LINGUISTIC AND COGNITIVE APPROACHES TO DIALOGUE AGENTS

Rafal Rzepka, Michal Ptaszynski and Pawel Dybala

(Editors)

Part of

Published by
The Society for the Study of
Artificial Intelligence and
Simulation of Behaviour

http://www.aisb.org.uk

ISBN 978-1-908187-16-1

Foreword from the Congress Chairs

For the Turing year 2012, AISB (The Society for the Study of Artificial Intelligence and Simulation of Behaviour) and IACAP (The International Association for Computing and Philosophy) merged their annual symposia/conferences to form the AISB/IACAP World Congress. The congress took place 2–6 July 2012 at the University of Birmingham, UK.

The Congress was inspired by a desire to honour Alan Turing, and by the broad and deep significance of Turing's work to AI, the philosophical ramifications of computing, and philosophy and computing more generally. The Congress was one of the events forming the Alan Turing Year.

The Congress consisted mainly of a number of collocated Symposia on specific research areas, together with six invited Plenary Talks. All papers other than the Plenaries were given within Symposia. This format is perfect for encouraging new dialogue and collaboration both within and between research areas.

This volume forms the proceedings of one of the component symposia. We are most grateful to the organizers of the Symposium for their hard work in creating it, attracting papers, doing the necessary reviewing, defining an exciting programme for the symposium, and compiling this volume. We also thank them for their flexibility and patience concerning the complex matter of fitting all the symposia and other events into the Congress week.

John Barnden (Computer Science, University of Birmingham)
Programme Co-Chair and AISB Vice-Chair
Anthony Beavers (University of Evansville, Indiana, USA)
Programme Co-Chair and IACAP President
Manfred Kerber (Computer Science, University of Birmingham)
Local Arrangements Chair

Foreword from the Symposium Organizer

Five decades of failure to pass the Turing test by computers lead us to rethink previous approaches, lean towards new technologies and knowledge sources, and combine them with advances in philosophy, linguistics and cognitive science. We stress the fact that the age of information explosion gives us a whole new spectrum of possibilities for creating an intelligent machine. Many marvelous ideas of the dawn of Artificial Intelligence research faced problems of exceptions and the impossibility of manual input of all needed knowledge, but today we have vast amounts of data from sensors and text so that we can rethink classical AI methods and approaches.

The increased use of WWW, RFID, Bluetooth, etc. could allow us to determine standard human behaviors, emotions or even moral reasoning according to the Wisdom of Crowds hypothesis. Collective input data can also help to retrieve knowledge about the physical world we live in. By combining Natural Language Processing methods with cognitive approaches and philosophy of mind, we can discover a new range of intelligent systems that understand us, our environment and our feelings. In this context, we see a role for NLP and cognitive approaches to play in developing a new generation of user-friendly, (also ethically) safe systems that, through interaction with the user and the world, can learn how to reason, behave or speak naturally.

We are interested in original papers on systems and ideas for systems that use common sense knowledge and reasoning, affective computing, cognitive methods, learning from broad sets of data and acquiring knowledge, or language and user preferences.

The symposium intends to spark an interdisciplinary discussion on joining forces to return AI to its original, broader and deeper goals which are currently represented by AGI - Artificial General Intelligence.

Rafal Rzepka (Hokkaido University, Japan)
Symposium Organizer

Organizer:

Rafal Rzepka (Hokkaido University)

Editors:

Rafal Rzepka (Hokkaido University), Michal Ptaszynski (Hokkai-Gakuen University) and Pawel Dybala (Otaru University of Commerce)

The Program Committee:

Aladdin Ayesh, De Montfort University, UK
Kenji Araki, Hokkaido University, Japan
Eleanor Clark, Hokkaido University
Haris Dindo, University of Palermo, Italy
Pawel Dybala, Otaru University of Commerce, Japan
Ben Groetzel, Novamente, USA
Yasutomo Kimura, Otaru University of Commerce, Japan
Fumito Masui, Kitami Institute fo Technology, Japan
Koji Murakami, Rakuten, USA
Michal Ptaszynski, Hokkai-Gakuen University, Japan
Tyson Roberts, Google, Japan
Marcin Skowron, Austrian Research Institute of Artificial Intelligence, Austria
Masato Tokuhisa, Tottori University, Japan
Zygmunt Vetulani, Adam Mickiewicz University, Poland

The website of our symposium:
http://arakilab.media.eng.hokudai.ac.jp/Turing

Cite as:
Rzepka, R. Ptaszynski, M. and Dybala, P. (eds.) (2012), Linguistic and Cognitive Approaches To Dialogue Agents (AISB/IACAP Symposium).

Lastname, Firstname (2012), 'Paper Title', in Rzepka R., Ptaszynski, M. and Dybala, P. (eds.), Linguistic and Cognitive Approaches To Dialogue Agents (AISB/IACAP Symposium), xx-xx (2012).

Table of Contents

Developing Embodied Multisensory Dialogue Agents

Michał B. Paradowski[1]

Abstract. A few decades of work in the AI field have focused efforts on developing a new generation of systems which can acquire knowledge via interaction with the world. Yet, until very recently, most such attempts were underpinned by research which predominantly regarded linguistic phenomena as separated from the brain and body. This could lead one into believing that to emulate linguistic behaviour, it suffices to develop 'software' operating on abstract representations that will work on any computational machine. This picture is inaccurate for several reasons, which are elucidated in this paper and extend beyond sensorimotor and semantic resonance. Beginning with a review of research, I list several heterogeneous arguments against disembodied language, in an attempt to draw conclusions for developing embodied multisensory agents which communicate verbally and non-verbally with their environment.

Without taking into account both the architecture of the human brain, and embodiment, it is unrealistic to replicate accurately the processes which take place during language acquisition, comprehension, production, or during non-linguistic actions. While robots are far from isomorphic with humans, they could benefit from strengthened associative connections in the optimization of their processes and their reactivity and sensitivity to environmental stimuli, and in situated human-machine interaction. The concept of multisensory integration should be extended to cover linguistic input and the complementary information combined from temporally coincident sensory impressions.

Keywords: embodiment, sensorimotor resonance, semantic resonance, language, multisensory integration, robotics

1 INTRODUCTION

… His eyes only see
His ears only hear …
—Wisława Szymborska *No End of Fun* (1967)

In the 'traditional' view, going back to René Descartes, cognition has been seen as manipulation of symbolic, mental representations, with the brain conceived of as an input-output processor, a problem-solving device running abstract, generalised computational programs which enable us to process incoming data into a perception/interpretation of the outside world. This 'software', separate from the body, was equated with the mind, while the body was regarded as an output system attached to the cognitive processing system, with similar tasks achieved by applying the same underlying motor program to different effectors:

> *magnam esse differentiam inter mentem & corpus, in eo quňd corpus ex naturâ suâ sit semper divisibile, mens autem plane indivisibilis … mentem a corpore omnino esse diversam.*[2]
> —Descartes (1641) *Meditationes de prima philosophia VI*;19

The information-processing approach or computer metaphor has become further entrenched over the latter half of the previous century due to the adoption of the digital computer as the platform to run the symbolic computations (Hoffmann *et al.*, n.d.).

However, this dualist perspective has been increasingly challenged, beginning with Edmund Husserl, Martin Heidegger, John Dewey, and Maurice Merleau-Ponty, and it is today widely acknowledged that perception and cognition are grounded in bodily experience. The brain is not the sole problem-solving resource we have at our disposal; the organiser/filtering machine is the *body-en-total*. Heuristics depend on our physiology; cognition is not only influenced and biased by states of the body, but crucial to it are also the rest of the body beyond the brain, as well as the environment.

Until very recently, most language research has, in a Cartesian manner, traditionally regarded linguistic phenomena as internal, mental, isolationist and amodal (that is, separate and independent from perception, action and emotion systems, and the body); a view endorsed in psychology (e.g. Geschwind 1970; Kintsch 1998), philosophy (e.g. Katz & Fodor 1963; Fodor 1983), and linguistics (e.g. early Chomsky – 1957, 1975; Nowak *et al.* 2002; Jackendoff 2002)[3]. For instance, Chomsky's most seminal theories were based on mathematical formalism and saw language as governed by a context-free grammar extended with transformational rules operating on (non-semantic) symbol strings and complemented by morphophonemic rules, with autonomous syntax at the core of the theory of language. The reason why his views for a long time did not go beyond such a perspective should not come as a surprise. His *Syntactic Structures*, which became a revolutionary and foundational work[4] in linguistics, grew out of a series of lecture notes for an audience of undergrad (mainly electrical engineering and maths) students at the MIT.[5] Also, Chomsky's ideas were born at the

[1] Inst. of Applied Linguistics, Univ. of Warsaw, ul. Browarna 8/10, PL-00-311 Warsaw, Poland. E-mail: michal.paradowski@uw.edu.pl.

[2] "There is a great difference between mind and body, inasmuch as body is by nature always divisible, and the mind is entirely indivisible. […] the mind or soul of man is entirely different from the body."

[3] A *votum separatum* in this domain is the field of biolinguistics, which hypothesizes a strong genetic (or neurobiological) endowment for language (UG) and determination of its structure (e.g. postulating selectional—i.e. evolutionary fitness—advantages), treating the language faculty on a par with other biological systems (see e.g. Meader & Muyskens 1950; Lenneberg 1967; Piatelli-Palmarini 1989; Hauser *et al.* 2002; Chomsky 2005; Di Sciullo & Boeckx 2011).

[4] Ranked #1 on the list of the one hundred most influential works in cognitive science from the 20th century, selected by the University of Minnesota Center for Cognitive Sciences; http://www.cogsci.umn.edu/OLD/calendar/past_events/millennium/final.html

[5] Before that, at the University of Pennsylvania, Chomsky studied logic and foundations of mathematics.

same time as the establishment of computer science as a distinct academic discipline, the beginnings of computational linguistics, and the founding of AI research, which all shared the dominant idea that thought can be described with formal logic.

The generative school inspired several decades of linguistic thought, and even theories trying to modify or undermine its tenets were still relying on the underlying view of language as a system manipulating abstract symbols. This dualistic view could lead one into believing that in order to credibly emulate linguistic behaviour, it suffices to develop 'software' operating on (i.e. applying combinatorial rules such as Merge and Move to) abstract representations[6] that will work on any computational machine, and that its operations will be implementation-independent, functioning identically regardless of the physical hardware.

2 EMBODIED LANGUAGE IN HUMANS

to turn print into exciting situations in their skulls
—Kurt Vonnegut *Slaughterhouse-Five* (1969:205)

The dualistic approach just outlined above works to some extent in statistical machine translation, automatic text indexing and retrieval (think e.g. search engines), natural-language interfaces or dialogue systems, but if the system to be developed is to truly mimic human behaviour, the disembodied picture is not very accurate for several reasons. One may be doubtful about modularity and the existence of a specifically dedicated innate language acquisition device, but must still take into account the following phenomena and theoretical developments:

1. **lateralization** and **localization** of the language faculty in the brain. Linguistic capabilities have been shown to be limited to certain areas of the cerebrum, as evidenced primarily by various language disorders:[7]
   - receptive aphasia, commonly known as Wernicke's aphasia (Wernicke 1874): damage to the medial temporal lobe destroying local language regions and cutting them off from most of the occipital, temporal and parietal regions (cf. e.g. Price 2000; Bookheimer 2002; Damasio *et al.* 2004);
   - expressive aphasia (aka Broca's or agrammatic aphasia; Broca 1861);
   - abnormal language developed in individuals with the left hemisphere removed (Dennis & Whitaker 1976);[8]
   - Specific Language Impairment (SLI), which is unrelated to other developmental disorders, mental retardation, brain injury, or deafness (e.g. Joanisse & Seidenberg 1998; Bishop & Snowling 2004; Archibald & Gathercole 2006);
   - other cases of people with normal nonverbal abilities but impaired language, and 'normal' language but cognitive deficits (cf. the classic case studies of individuals with incommensurable linguistic and cognitive capacities: Genie (Curtiss 1981), Laura (Yamada 1990), Clive (Smith 1989), or Christopher (Smith *et al.* 1993)[9].

While these deficits cannot straightforwardly be taken as proof of the *modularity* of language (*cf.* e.g. Calabrette *et al.* 2003; Fodor 2005), they do point to *localisation* of language processes;

2. embodiment of language in neuronal circuitry. FMRI studies have shown '**activation**' of certain brain areas involved in **language processing** (e.g. Osterhout 1997; Hagoort *et al.* 1999; Embick *et al.* 2000; Horwitz *et al.* 2003; Pulvermüller & Assadollahi 2007), with different levels of language processing identified in specific regions, e.g. loci of syntax mainly in left-perisylvian language regions, especially Broca's and Wernicke's areas, but also adjacent neocortical areas, the insula, and subcortical structures including basal ganglia (*cf.* e.g. Ullman 2001; Grodzinsky & Friederici 2006), or phonology in the superior temporal sulcus and anterior superior temporal cortex (*cf.* e.g. Diesch *et al.* 1996; Obleser *et al.* 2006; Uppenkamp *et al.* 2006);
3. genetic influence on language. While mutations of the Foxhead box protein 2 (**FOXP2** gene), deemed to cause a severe speech and language disorder (e.g. Lai *et al.* 2001; Vernes *et al.* 2008; Fisher & Scharff 2009),were initially taken as evidence for a 'language gene', it was later discovered that the protein impacts a wide range of phenotypic features all over the body (including facial motor control) and that the impairments of the family affected with the mutation went beyond language to other cognitive capacities. It is now more believed that it is networks of gene interactions rather than individual genes that have an influence on language (Knopka *et al.* 2009), but the neurobiological influence is there;
4. many Universal Grammar-based constraints now being reinterpreted as **learning** and **processing constraints**. That is, the difficulty in the acquisition of certain aspects of language are being accounted for by their complexity, the computational load under which the user/learner operates, his/her memory and attention limitations, or ease of access to

[6] Understood as terminal symbols, which can—subsequently or concurrently—be equipped with referential, meaning-bearing properties.

[7] Theoretically, an injury disrupting the system's functioning may only show the involvement of the affected region, not that the whole functionality was due to that region. However, interestingly, not only spoken, but also sing language is left-lateralised (with use of classical language areas—e.g. Broca's (Horwitz *et al.* 2003)—in sentence processing and LH damage associated with lexical comprehension, with a difference in more posterior activation in areas responsible for processing vision and movement; Woll 2012). While signing patients with *RH* damage perform within the normal range on language tests, with the exception of tests of locative sentence comprehension, these problems appear to mean not linguistic malfunction *per se*, but an indirect consequence of more general cognitive deficits: in areas such as classifiers, spatial verbs, and grammar relying on space, sign language processing is reliant on visuospatial cognition (*ibid.*; Woll & Morgan 2012).

[8] Although one must be cautious about the conclusions since the cortical development in the subjects of the study was not normal in the first place (Chomsky 1980:264).

[9] In short, Christopher was able to acquire natural languages (with great aptitude, too, especially regarding morphology), but not ones violating the constraints of Universal Grammar. (The picture is more complex, but does not invalidate the basic claim.) But see e.g. Karmiloff-Smith (1998), Johnson *et al.* (1999), or Elsabbagh & Karmiloff-Smith (2006) for reports on Williams syndrome questioning evidence for a clear-cut dissociation of innate mechanisms for language. While the syndrome was originally postulated as characterised by preserved language in the presence of marked visual-spatial impairments, hence as evidence for modularity (*cf.* e.g. Bellugi *et al.* 1988, 1994), it was subsequently observed that actually language is not wholly intact (e.g. involving prepositional errors; Rubba & Klima 1991, Capirci *et al.* 1996, Volterra *et al.* 1996; Karmiloff-Smith *et al.* 2003; Woll 2012).

representations (*cf.* e.g. Wakabayashi 2002; Van Hell & De Groot 1998; Wątorek 2008);

5. maturation and the **critical/sensitive period**[10] (but consider e.g. Marinova-Todd *et al.* 2000 for a contradictory view);
6. the Chomskyan **competence** vs. **performance** distinction (Chomsky 1965)[11], explaining mistakes in (originally native) language users' output (i.e., their actual deployment of the linguistic capacity) attributable to such psychosomatic states and factors affecting them as fatigue, tedium, intoxication, drugs, sudden changes of mind, haste, inattention, or external distractions;
7. interaction between (context-bound) language comprehension and production, and sensorimotor activation, manifested in both directions by:[12]
   - **motor resonance** observed in linguistic (Lakoff & Johnson 1980; Lakoff 1987), behavioural (primarily with priming[13] modulating motor performance; e.g. Tanenhaus *et al.* 1995; Gentilucci *et al.* 2000; Spivey *et al.* 2001; Glenberg & Kaschak 2002; Glover *et al.* 2004; Buccino *et al.* 2005; Boulenger *et al.* 2008; Nazir *et al.* 2008; Frak *et al.* 2010; for grammar *cf.* Madden & Zwaan 2003; Bergen & Wheeler 2010), neuroimaging and TMS studies[14] (e.g. Zatorre *et al.* 1992; Fadiga *et al.* 2002; Tettamanti *et al.* 2008; Fischer & Zwaan 2008; Kemmerer *et al.* 2008; Boulenger *et al.* 2009; Willems *et al.* 2010; for activation in visual areas *cf.* Martin *et al.* 1996; Pulvermüller & Hauk 2006; Simmons *et al.* 2007; in the olfactory cortex *cf.* González *et al.* 2006);
   - **semantic resonance** (brain language areas getting activated during sensorimotor action; Bonda *et al.* 1994; Pulvermüller *et al.* 2005; Rueschemeyer *et al.* 2010);[15]
   - verbalization of memory facilitated when assuming the original body position during recall (Dijkstra *et al.* 2007)[16], linguistic tasks expedited when accompanied by action (Rieser *et al.* 1994), and sensorimotor experiences intertwined with cognition in episodic memory (Pfeifer 2011);
   - faster comprehension of depictions of spatial associations than of descriptions of spatial *dis*sociations[17] (Glenberg *et al.* 1987); speedier recognition of words with 'body-object interaction' than of ones without (Siakaluk *et al.* 2008);
   - semantic interference and facilitation in the Stroop effect (longer RTs needed to name colour names written in incongruent ink hue; Jaensch 1929; Stroop 1935);
   - clinical studies indicating that processing of action concepts degrades if action- or vision-related brain areas are lesioned in motor neuron diseases (Damasio *et al.* 1996; Bak *et al.* 2001; Neininger & Pulvermüller 2003) and semantic dementia (Pulvermüller *et al.* 2010);
   - comprehension of action words deteriorating after loss of procedural knowledge (*cf.* Boulenger *et al.* 2008 on Parkinson's disease patients; also Bak *et al.* 2006);
8. parallel emergence of speech and gesture in infancy (Iverson & Thelen 1999);
9. co-speech gesture reducing cognitive load (Goldin-Meadow *et al.* 2001), and indications of a dual-task advantage for bimodal (signed-spoken) language production (i.e., production of code-blends, with elements of the signed and spoken languages appearing simultaneously; Kaufmann & Kaul 2012); or
10. Conceptual Blending theory (Fauconnier & Turner 2002) explaining language creativity as a semantic process operating on the output of perception and interaction with the world to create new structures.

Thus, independently of theoretical persuasion, without taking into account both the architecture of the human brain, and

[10] The Critical Period Hypothesis (or its idea), proposed by Penfield and Roberts (1959), posits the existence of an ideal window of time during which genetically endowed language acquisition can—given adequate stimuli—take place spontaneously, relatively effortlessly, and characteristically meeting a high degree of success, after which acquiring a language naturally, automatically and with complete ultimate attainment becomes impossible. "The earlier the better" rule of thumb captures the negative correlation between the age of acquisition onset and subsequent asymptotic attainment. Most evidence to support the claim was supplied by Eric Lenneberg (1967) in his *Biological Foundations of Language*. While the existence of a critical period is widely accepted where first language acquisition is concerned, attempts to extend it to second language acquisition still arouse a good deal of contention (for instance, Lamendella (1977) suggested the term 'sensitive period' to emphasise the fact that acquisition may be more efficient during childhood, but not restricted to that period).

[11] The distinction can be considered on the example of any organic system: "Studies of the digestive system, for example, distinguish between its structural properties and what it is doing after you ate a sandwich" (Noam Chomsky, p.c., 8 Nov 2011), and can actually be traced back to the classic Aristotelian dichotomy between δύναμις (potentiality) and ἐνέργεια (actuality).

[12] This seems to be a reflection of a more general phenomenon where "there is no animal in which there is known to be a complete segregation of sensory processing" (Stein *et al.* 1996:497).

[13] E.g. in the form of mention of tool and action concepts.

[14] Somewhat importantly, motor resonance was not observed when the stimuli were used in idiomatic contexts (Rueschemeyer *et al.* 2010a) or metaphorical ones. Regarding the latter, Raposo *et al.* (2009) found activity in the pre- and motor cortex for literal-only usages of arm- and leg-related Vs, while Bergen *et al.* (2007) likewise demonstrated that visual imagery is triggered in sentence comprehension tasks (where general words of motion were employed) only where the utterances have literal spatial meaning. However, the picture is not completely clear-cut. This year, Lacey *et al.* (2012) showed that textural metaphors do activate parietal operculum regions important to the sense of touch. To explain this discrepancy, one could posit a qualitative difference between 'directly' embodied sensory experiences (e.g. texture or temperature) and more 'indirect' ones such as those grounded in visual perception. The former are more 'primary':

i)sensed earliest – already in the womb, tactition being the first sense that begins to develop before 8 weeks gestational age together with the emergence of the nervous system (Montagu 1978), before taste and smell (14 weeks g.a.), hearing (16 weeks g.a.; Shahidullah & Hepper 1992) or vision (week 18 onwards),

ii)available in more 'primitive' organisms without vision or hearing,

iii)perceptible during half-sleep, and

iv)impacting our bodily functioning more strongly (the somatic reaction to extremely high or low temperatures, pressure or skin irritation is more likely to be stronger than e.g. to an unpleasant sight or sound).

This might account for the lack of activation in visual cortical areas.

[15] But see e.g. Bedny *et al.* (2008), Postle *et al.* (2008), or Kemmerer & Gonzalez-Castillo (2010) for opposing views.

[16] This conviction can also be found in 'folk wisdom'. For instance, in one episode of a Malaysian edutainment program for children which I was consulting on for a European broadcaster, a monkey was hanging upside down because that was the position in which she last saw her orange juice.

[17] I.e. texts describing an event in which the main character was spatially dissociated from a target object, e.g.:

> John was preparing for a marathon in August. After doing a few warm-up exercises, he *took off* his sweatshirt and went jogging. (emph.added)

embodiment—the interaction of the language faculty with the sensory apparatus and motor system—it is unrealistic to replicate accurately the processes which take place during language acquisition, comprehension, or production, or during non-linguistic actions. Cognitive mechanisms are synergistically intertwined with affective and somatic components, and largely inseparable (Ziemke 2011).

3 THE COROLLARIES FOR ROBOTICS

… it is the movement which is primary, and the sensation which is secondary, the movement of the body, head, and eye muscles determine the quality of what is experienced. In other words, the real beginning is with the act of seeing; it is looking, and not a sensation of light.
—John Dewey (1896:358*f.*)

Since the official launch of AI as a new research discipline at the seminal Dartmouth conference in 1956, much of work in the field has been driven by the 'Physical Symbol Hypothesis' (Newell & Simon 1976): trying to construct systems that would possess or build internal, symbolic representations of objects and relations in the outside *world*—in other words, a "world model"—which usually had little to do with their hardware, sensorimotor *experience*, or current context[18], but were instead characterised by precisely defined states and finite lists of acceptable commands (Wang 2009:2*f.*). Under such a functionalist approach, the body is merely a platform on which cognitive operations are running. In some areas, such closed systems were able to achieve spectacular feats, for instance in defeating world chess champions.

Chess, however, is a formal game, set in a virtual world with discrete states, positions, and licit moves, a game involving complete information, and a static one: no move means no change, and the inventory of legitimate operations remains constant (Pfeifer & Scheier 1999:58*ff.*). This is quite unlike what usually happens in the real world. Hence, the last two and a half decades have witnessed recurrent appeals for situated, embodied autonomous systems actively and directly interacting with the world around (*cf. op. cit.*; Brooks 1991; Varela *et al.* 1991) and constructing knowledge via this dynamic enactment (the active learning being qualitatively different from statistical machine learning; *cf.* e.g Froese 2009; Vernon 2010). Evidently robots, even anthropomorphic ones, are far from isomorphic with humans in terms of both the 'brain' and the rest of the body, including the input and output devices (sensors and actuators). Also, as one reviewer rightly remarks, in the language technology field priority is not necessarily to make a machine as humanlike as possible, with the same architecture; rather, it is to make the machine so that it does things on a level comparable to humans (or, I would add, surpassing that) – in other words, to achieve similar—or better—functionality in terms of mode, scope, or scale. Or, going completely beyond the anthropocentric GOFAI perspective (Haugeland 1985; *cf.* Wang 2008), since passing the Turing Test is not a *sine qua non* of being intelligent, as acknowledged by the test's designer himself (Turing 1950). This, however, means that robust artificial cognitive agents can bypass the human limitations[19] inherent in most of the above points (just as they could overcome some contingencies resulting from the material properties of the human brain and bodily features such as synaptic speed and efficiency, the physical characteristics of the vocal tract, the auditory perception system, or muscular flexibility[20]). Nevertheless, they could still benefit from strengthened associative connections owing to the motor and semantic resonance in both the optimization of their processes, and reactivity and sensitivity to environmental stimuli, across a range of tasks:

(i) in grounded language understanding (*cf.* e.g. Glenberg & Kaschak 2002; Feldman & Narayanan 2004; Gallese & Lakoff 2005; Sato *et al.* 2008), where structuring the environment acts as scaffolding[21] and all inputs contribute to evidential support,

(ii) in automated articulation-based speech recognition (utilising motor information, i.e. combining spoken input with visual data—e.g. the shape of the speakers lips—and maybe even data such as strength of the incoming airstream),

(iii) while learning about context-dependent phenomena in the surrounding world (e.g. action sequences and argument structure in construction grammar; *cf.* Dominey 2007; since embodiment plays a constitutive role in the process of cognition; Vernon 2010), or in the process of language acquisition in general (because language—at least in the initial stages—is acquired by situated embodied direct engagement with the world, and not just passive perception, e.g. watching television; *cf.* e.g. Steels 2009),

(iv) to help with storage and retrieval due to the benefits of episodic memory,

(v) to support action prediction, planning and anticipation (Koelewijn *et al.* 2008; Stapel *et al.* 2010; van Elk *et al.* 2010)[22], including prediction of the next sensory feedback,

(vi) to support action execution (with linguistic input making the actor better aware of the affordances, i.e. physically feasible action possibilities), and

(vii) to reinforce feedback in 'soft robotics' and morphological computation, where there is no clear separation between the controller (or orchestrator) and the hardware (morphology), and the tasks are distributed between the brain, body, and environment (*cf.* e.g. Paul

[18] The fact that the appropriate relations to some outside world could be established by the system's designer or end-user becomes unhelpful the moment we want to deal with an autonomous agent, with the human interpreter removed from the loop, as emphasised by Steven Harnad in his seminal (1990) paper (*cf.* also Pfeifer & Scheier 1999:69*f.*).

[19] The limitations need not in themselves necessarily be a bad thing; to the contrary, they may serve a useful role in limiting the search space and focusing attention on the most vital stimuli. The restrictions imposed on the vocal apparatus in turn mean that speech is segmented and decelerated enough to facilitate comprehension. The relative absence of such constraints on computers may be the exact reason why the latter have problems tackling tasks where humans perform with ease (Tom Froese, p.c., 9 Mar 2012).

[20] Just as robots can have an advantage when equipped with e.g. infrared, or ultrasonic sensors.

[21] Sensorimotor dynamics plays a crucial part in toddlers' learning to categorise objects: it is only when the infant brings the object in front of their eyes and focuses on it that s/he learns to associate it with its name (Smith 2010).

[22] Though originally grounded in sensorimotor experience, mental imagery, or simulation of interaction with the world, may subsequently become environmentally decoupled, as in forward models (Clark & Grush 1999).

2004; Pfeifer 2011; which also has the aim of off-loading computation; Di Paolo 2009);[23]

(viii) in cognitive developmental robotics, aiming at understanding human cognitive developmental processes by synthetic or constructive approaches (Asada *et al.* 2009, Asada 2011, Ishiguro *et al.* 2011);

(ix) in common grounding and alignment, which are crucial for fruitful situated human-machine interaction, and which are another area where sensory experience must be coordinated with linguistic interaction.

Principally, if our goal were to create machines which do things on a *comparable level* to—or surpassing—humans, we could do away with attempts at embodying them in human-inspired ways (Taivo Lints, p.c., 31 May 2012) – they could function perfectly well with totally nonhuman kinds of embodiment (different 'bodies', different sensors and effectors, different internal architectures... or even with embodiment in a virtual world; Bringsjord *et al.* 2008; Goertzel *et al.* 2008). Given the role played by the morphology of the sensory apparatus and the architecture of the sensorimotor loop in shaping and structuring the information that reaches the controller, and thereby in concept formation, it would anyway be difficult for a machine to form the same concepts, categories and behaviours as us without having comparable morphology (as remarked e.g. by Barsalou 1999 or Lakoff & Johnson 1998). However, if our goal is to have machines 'thinking' and behaving in a way *compatible* with ours—which is a highly practical and desirable goal—then it is of high importance for them to develop, learn and function in a similar "experience space" (Taivo Lints, p.c.; *cf.* also Wang 2009:5).

The requirement that the behaviour, perception and conceptual apparatus of artificial intelligent agents be grounded in their experience of their *own* interaction with the outside world at once means that their concepts and categories need not necessarily rely on the same minimal constituents and grammatical categories as have been externally identified and defined in linguistics. Instead, the gradually emergent categories are more likely to be intrinsically meaningful behaviours and affordances (see also Kuniyoshi *et al.* 2004), action-oriented rather than orbocentric (Hoffmann & Pfeifer 2011). For instance, to a robot who has never kicked or observed anyone kick anything but footballs, the minimal unit of meaning may be <kick a ball> rather than <kick> alone (although this does not rule out the possibility of extrapolation and abstraction should a relevant opportunity arise).[24] Similarly, irrespective of whether the input is expressed using [$_{NP}$ kicking a ball] or [$_{VP}$ kick a ball], it should activate the same action schema.

4 TOWARDS A BROADER DEFINITION OF MULTISENSORY INTEGRATION

In order to form a meaningful experience and construct coherent, reliable and robust representations of the surrounding world, the human brain combines prior knowledge with sensory input arriving from various modalities and integrates these at multiple levels of the neuraxis. This serves to maximize the efficiency of everyday performance and learning, enhancing the salience of the events, helping increase the detection and identification of the external stimuli, disambiguate them, compensate for incomplete information, and shorten reaction times. In view of the inseparability of language and the body, the concept of multisensory integration—whether in natural or artificial cognitive agents—should be extended and cover both the linguistic input and the complementary information that the brain combines from temporally coincident sensory impressions. This does not mean that we should 'dumb down' the statistical processes where they operate successfully; instead, where the input stream in one channel is too noisy, turning on auxiliary channels[25] and interacting with the environment in an active manner may generate ancillary data and help e.g. disambiguate the signal and take the right decision (see also Pfeifer & Scheier 1997; Beer 2003).[26] An added benefit would then be significantly reduced programming costs.

CONCLUSIONS

> *A living organism* enacts *the world it lives in; its effective embodied action in the world actually constitutes its perception and thereby grounds its cognition.*
> —Stewart, Gapenne & Di Paolo (2010:vii)

I have started out with a brief depiction of the dualistic Cartesian approach that has characterised much of twentieth-century thought, including that underlying most of traditional AI. While adherence to such an outlook has in many domains led to very spectacular achievements, there are limits which purely symbolic systems cannot overcome. While the subject of the mind-body relationship is by no means new, the link, still very often ignored by cognitive science communities (logic, linguistics, computer science) may be the key element for bypassing the present limitations of AI systems.

Language, too, has for a long time been treated across scientific domains as an abstract system operating largely independently from the body (articulatory-perceptual organs notwithstanding). I have presented an inventory of heterogeneous evidence against such a view, addressing instead the issue of the link between language and body. While many of the embodied language phenomena specific to humans have little

[23] The idea of morphological computation in animals can be well illustrated on the example of cockroaches skilfully climbing over obstacles that exceed their body height, using relatively few neurons, off-loading most tasks to morphology (by reconfiguring the mesothoracic shoulder joint), exploiting mechanical change and feedback, and capitalising on the stability of the local feedback circuits; *cf.* Watson *et al.* 2002; Pfeifer *et al.* 2007; Pfeifer & Gomez 2009).

[24] See for instance the POETICON++ project (Robots need Language: A computational mechanism for generalisation & generation of new behaviours in robots; `http://www.poeticon.eu/`).

[25] These channels need not all be active at all times, especially when it might burden the cognitive load in non-essential tasks, when conflicting inputs can bring the machine to a halt, or when the benefits—e.g. in terms of speed—would be negligible (Richard Littauer, p.c., 26 May 2012). The system's available resources should be dynamically allocated to different tasks in such a way as to achieve the highest overall efficiency.

[26] One consequence for humans may be that the role of kinaesthetic modality, traditionally largely believed to dominate in children, but be negligible in adults (*cf.* e.g. Barbe & Milone 1981; Felder & Spurlin 2005), should be reassessed, as the effectiveness may be demonstrated of 'learning-by-doing' and task-based approaches to language learning and teaching where the students have to use their bodies (e.g. when acquiring novel lexis via common cookery classes).

direct translation to machines, there are others that can profitably be exploited and inspire the development of robust artificial autonomous agents that rely on semantics grounded in their past experience (both linguistic and non-verbal) as well as possible related operations on the concepts concerned. Agents which are adaptive to feedback and can, despite insufficient knowledge, time pressure and storage space constraints safely and successfully navigate, learn, and communicate in the complex and dynamic ecological niche they share with human actors.

ACKNOWLEDGMENTS

The author wishes to thank Noam Chomsky, Anna Esposito, Taivo Lints, Richard Littauer, Gary Lupyan, Vincent C. Müller, Katerina Pastra, Michael Pleyer, Yulia Sandamirskaya, Luc Steels, the Evolang IX reviewers, and the anonymous referees for this Symposium for invaluable commentary, discussion and bibliographical references. Naturally, willingness to comment does not imply endorsement; all the usual disclaimers apply. An earlier version of this paper was presented at the 4th EUCogII Members' Conference "Embodiment – Fad or Future?", Anatolia College/American College of Thessaloniki, 11-12 Apr. 2011.

REFERENCES

[1] Archibald, L.M.D., Gathercole, S.E.: Prevalence of SLI in Language Resource Units. *J Res Spec Educ Needs* 6(1), 3–10 (2006).

[2] Asada, M., Hosoda, K., Kuniyoshi, Y., Ishiguro, H., Inui, T., Yoshikawa, Y., Ogino, M., Yoshida, C.: Cognitive developmental robotics: a survey. *IEEE Transactions on Autonomous Mental Development* 1(1), 12–34 (2009).

[3] Asada, M.: Can cognitive developmental robotics cause a paradigm shift? In: Krichmar, J.L., Wagatsuma, H. (Eds) *Neuromorphic and Brain-Based Robots: Trends and Perspectives*, Cambridge: Cambridge University Press, 251–73 (2011).

[4] Bak, T.H., O'Donovan, D.G., Xuereb, J.H., Boniface, S., Hodges, J.R.: Selective impairment of verb processing associated with pathological changes in Brodmann areas 44 and 45 in the motor neurone disease–dementia–aphasia syndrome. *Brain* 124, 103–20 (2001).

[5] Bak, T.H., Yancopoulou, D., Nestor, P.J., Xuereb, J.H., Spillantini, M.G., Pulvermuller, F., Hodges, J.R.: Clinical, imaging and pathological correlates of a hereditary deficit in verb and action processing. *Brain* 129(2), 321–332 (2006).

[6] Barbe, W.B., Milone, M.N.: What we know about modality strengths, *Educational Leadership* 38(5), 378–80 (1981).

[7] Barsalou, L.W.: Perceptual symbol systems. *Behavioral and Brain Sciences* 22, 577–609 (1999).

[8] Beer, R.: The dynamics of active categorical perception in an evolved model agent. *Adaptive Behav* 11, 209–43 (2003).

[9] Bedny, M., Caramazza, A., Grossman, E., Pascual-Leone, A., Saxe, R.: Concepts are more than percepts: the case of action verbs. *J Neurosci* 28, 11347–11353 (2008).

[10] Bellugi, U., Marks, S., Bihrle, A., Sabo, H.: Dissociation between language and cognitive functions in Williams syndrome. In D. Bishop, K. Mogford (Eds) Language development in exceptional circumstances. London: Churchill Livingstone, 177–189 (1988).

[11] Bellugi, U., Wang, P.P., Jernigan, T.: Williams Syndrome: An unusual neuropsychological profile. In S.H. Broman, J. Grafman (Eds) Atypical cognitive deficit in developmental disorders: Implications for brain function. Hillsdale, NJ: Lawrence Erlbaum, 23–56 (1994).

[12] Bergen, B., Lindsay, S., Matlock, T., Narayanan, S.: Spatial and linguistic aspects of visual imagery in sentence comprehension. *Cogn Sci* 31, 733–764 (2007).

[13] Bergen, B., Wheeler, K.B.: Grammatical aspect and mental simulation. *Brain Lang* 112, 150–158 (2010).

[14] Bishop, D.V.M., Snowling, M.J.: Developmental dyslexia and specific language impairment: same or different? *Psychol Bull* 130, 858–88 (2004).

[15] Bonda, E., Petrides, M., Frey, S., Evans, A.: Frontal cortex involvement in organized sequences of hand movements: Evidence from a positron emission topography study. *Soc Neurosci Abstr* 20, 353 (1994).

[16] Bookheimer, S.: Functional MRI of language: New approaches to understanding the cortical organization of semantic processing. *Ann Rev Neurosci* 25, 151–188 (2002).

[17] Boulenger, V., Hauk, O., Pulvermüller, F.: Grasping ideas with the motor system: Semantic somatotopy in idiom comprehension. *Cereb Cortex* 19(8), 1905–1914 (2009).

[18] Boulenger, V., Mechtouff, L., Thobois, S., Broussolle, E., Jeannerod, M., Nazir, T.A.: Word processing in Parkinson's disease is impaired for action verbs but not for concrete nouns. *Neuropsychologia* 46(2), 743–756 (2008).

[19] Bringsjord, S., Shilliday, A., Taylor, J., Werner, D., Clark, M., Charpentie, E., Bringsjord, A. : Toward logic-based cognitively robust synthetic characters in digital environments. *Artificial General Intelligence 2008*, Amsterdam : IOS Press, 87–98 (2008).

[20] Broca, P. : Remarques sur le siège de la faculté de la parole articulée, suivies d'une observation d'aphémie (perte de parole). *Bull Soc Anatom (Paris)* 36, 330–357 (1861).

[21] Brooks, R.A.: Intelligence without representation. *Artif Intell* 47, 139–59 (1991).

[22] Buccino, G., Riggio, L., Melli, G., Binkofski, F., Gallese, V., Rizzolatti, G.: Listening to action-related sentences modulates the activity of the motor system: a combined TMS and behavioral study. *Cogn Brain Res* 24, 355–363 (2005).

[23] Calabretta, R., Di Ferdinando, A., Wagner, G. P., Parisi, D.: What does it take to evolve behaviorally complex organisms? *BioSystems* 69, 245–262 (2003).

[24] Capirci, O., Sabbadini, L., Volterra, V.: Language development in Williams syndrome: A case study. *Cogn Neuropsych* 13, 1017–39 (1996).

[25] Chomsky, N. *Syntactic structures*. The Hague: Mouton (1957).

[26] Chomsky, N. *Aspects of the Theory of Syntax*. Cambridge, MA: MIT Press (1965).

[27] Chomsky, N. *Reflections on language*. New York: Pantheon (1975).

[28] Chomsky, N. *Rules and Representations*. New York: Columbia University Press (1980).

[29] Chomsky, N. Three factors in language design. *Ling Inq* 36(1), 1–22 (2005).

[30] Clark, A., Grush, R.: Towards cognitive robotics. *Adaptive Behav* 7(1), 5–16 (1999).

[31] Curtiss, S. *Genie: The Case of a Modern Wild Child*. New York: Academic Press (1981).

[32] Damasio, H., Grabowski, T. J., Tranel, D., Hichwa, R. D., Damasio, A. R. A neural basis for lexical retrieval. *Nature*, 380, 499–505 (1996).

[33] Damasio, H., Tranel, D., Grabowski, T., Adolphs, R., Damasio, A. Neural systems behind word and concept retrieval. *Cognition* 92, 179–229 (2004).

[34] Dennis, M., Whitaker, H. Language acquisition following hemidecortication: Linguistic superiority of the left over the right hemisphere. *Brain Lang* 3, 404-433 (1976).

[35] Dewey, J.: The reflex arc concept in psychology. *Psychological Rev* 3, 357-70 (1896).

[36] Di Paolo, E. From sensorimotor coordination to enaction: Agency, sense-making and sociality as horizons for embodied cognition. 1st EUCogII Members' Conference "Challenges for artificial cognitive systems", University Medical Center Hamburg-Eppendorf (2009, Oct 10).

[37] Di Sciullo, M., Boeckx, C. *The Biolinguistic Enterprise. New Perspectives on the Evolution and Nature of the Human Language Faculty*. Oxford: Oxford University Press (2011).
[38] Diesch, E., Eulitz, C., Hampson, S., Ross, B. The neurotopography of vowels as mirrored by evoked magnetic field measurements. *Brain Lang* 53(2), 143–168 (1996).
[39] Dijkstra, K., Kaschak, M., Zwaan, R.: Body posture facilitates retrieval of autobiographical memories. *Cognition* 102, 139–49 (2007).
[40] Dominey, P.F.: Spoken language and vision for adaptive human-robot cooperation. In: Hackel, M. (ed.), *Humanoid robotics*. ARS International, Vienna (2007).
[41] Elsabbagh, M., Karmiloff-Smith, A.: Modularity of mind and language. In K. Brown (Ed.), *The Encyclopaedia of Language and Linguistics* [2nd ed.], Oxford: Elsevier, 218–24 (2006).
[42] Embick, D., Marantz, A., Miyashita, Y., O'Neil, W., Sakai, K.L.: A syntactic specialization for Broca's area. *Proc Natl Acad Sci USA* 97, 6150–6154 (2000).
[43] Fadiga, L., Craighero, L., Buccino, G., Rizzolatti, G.: Speech listening specifically modulates the excitability of tongue muscles: A TMS study. *Eur J Neurosci* 15(2), 399–402 (2002).
[44] Fauconnier, G., Turner, M.: *The Way We Think: Conceptual Blending and the Mind's Hidden Complexities*. New York: Basic Books (2002).
[45] Felder, R.M., Spurlin, J.: Applications, reliability and validity of the Index of Learning Styles. *Int J Engng Ed* 21(1), 103–12 (2005).
[46] Feldman, J., Narayanan, S.: Embodied meaning in a neural theory of language. *Brain Lang* 89(2), 385–392 (2004).
[47] Fischer, M.H., Zwaan, R.A.: Embodied language: A review of the role of motor system in language comprehension. *Q J Exp Psychol* 61, 825–850 (2008).
[48] Fisher, S.E., Scharff, C.: FOXP2 as a molecular window into speech and language. *Trends Genet* 25(4), 166–177 (2009).
[49] Fodor, J.A.: *The Modularity of Mind: An essay on faculty psychology*. MIT Press, Cambridge, MA (1983).
[50] Fodor, J.A.: Reply to Steven Pinker "So How Does the Mind Work?" *Mind & Language* 20, 25–32 (2005).
[51] Frak, V., Nazir, T., Goyette, M., Cohen, H., Jeannerod, M.: Grip force is part of the semantic representation of manual action verbs. *PLoS ONE* 5(3), e9728. doi:10.1371/journal.pone.0009728 (2010).
[52] Froese, T.: Hume and the enactive approach to mind. *Phenomenology and the Cognitive Sciences* 8(1), 95-133 (2009).
[53] Gallese, V., Lakoff, G.: The brain's concepts: The role of the sensory–motor system in reason and language. *Cogn Neuropsychol* 22, 455–479 (2005).
[54] Gentilucci, M., Benuzzi, F., Bertonali, L., Daprati, E., Gangitano, M.: Language and motor control. *Exp Brain Res* 133(4), 468–90 (2000).
[55] Geschwind, N.: The organization of language and the brain. *Science* 170(961), 140–4 (1970).
[56] Glenberg, A.M., Kaschak, M.P.: Grounding language in action. *Psychon Bull Rev* 9(3), 558–565 (2002).
[57] Glenberg, A., Meyer, M., Lindem, K.: Mental models contribute to foregrounding during text comprehension. *J Mem Learning* 26: 69–83 (1987).
[58] Glover, S., Rosenbaum, D.A., Graham, J., Dixon, P.: Grasping the meaning of words. *Exp Brain Res* 154 (1), 103–8 (2004).
[59] Goertzel, B., Pennachin, C., Geissweiller, N., Looks, M., Senna, A., Silva, W., Heljakka, A., Lopes, C.: An integrative methodology for teaching embodied non-linguistic agents, applied to virtual animals in Second Life. *Artificial General Intelligence 2008*, Amsterdam: IOS Press, 161–75 (2008).
[60] Goldin-Meadow, S., Nusbaum, H., Kelly, S., Wagner, S.: Explaining math: Gesturing lightens the load. *Psychological Sci* 12, 516-522 (2001).
[61] González, J., Barros-Loscertales, A., Pulvermüller, F., Meseguer, V., Sanjuán, A., Belloch, V., Ávila, C.: Reading *cinnamon* activates olfactory brain regions. *Neuroimage* 32(2), 906–12 (2006).
[62] Grodzinsky, Y., Friederici, A.D.: Neuroimaging of syntax and syntactic processing. *Curr Opin Neurobiol* 16(2), 240–6 (2006).
[63] Hagoort, P., Ramsey, N.F., Rutten, G.J.M., van Rijen, P.C.: The role of the left anterior temporal cortex in language processing. *Brain Lang* 69, 322–325 (1999).
[64] Harnad, S.: The symbol grounding problem. *Physica D* 42, 335–46 (1990).
[65] Haugeland, J.: *Artificial Intelligence: The Very Idea*. Cambridge, MA: MIT Press (1985).
[66] Hauser, M.D., Chomsky, N., Fitch, W.T.: The language faculty: What is it, who has it, and how did it evolve? *Science* 298(5598), 1569–1579 (2002).
[67] Hoffmann, M., Assaf, D., Pfeifer, R.: Cognitivism. Retrieved from `http://www.eucognition.org/index.php?page=cognitivism`
[68] Hoffmann, M., Pfeifer, R.: The implications of embodiment for behavior and cognition: Animal and robotic case studies. In: W. Tschacher, C. Bergomi (Eds) *The Implications of Embodiment: Cognition and Communication*. Exeter: Imprint Academic, 31–58 (2011).
[69] Horwitz, B., Amunts, K., Bhattacharyya, R., Patkin, D., Jeffries, J., Zilles, K., Braun, A.R.: Activation of Broca's area during the production of spoken and signed language: A combined cytoarchitectonic mapping and PET analysis. *Neuropsychologia* 41, 1868–1876 (2003).
[70] Ishiguro, H., Minato, T., Yoshikawa, Y., Asada, M.: Humanoid Platforms for Cognitive Developmental Robotics. *Intl J Humanoid Robotics* 8(3), 391–418 (2011).
[71] Iverson, J., Thelen, E.: Hand, mouth, and brain: The dynamic emergence of speech and gesture. *J Consc Stud* 6(11-12), 19–40 (1999).
[72] Jackendoff, R.: *Foundations of Language*. Oxford University Press, Oxford (2002).
[73] Jaensch, E.R : *Grundformen menschlichen Seins*. Berlin: Otto Elsner (1929).
[74] Joanisse, M.F., Seidenberg, M.S.: Specific language impairment: A deficit in grammar or processing? *Trends Cogn Sci* 2, 240–47 (1998).
[75] Johnson, M.H., Paterson, S.J., Brown, J.H., Gsödl, M.K., Karmiloff-Smith, A.: Cognitive Modularity and Genetic Disorders. *Science* 286(5448): 2355–8 (1999).
[76] Karmiloff-Smith, A.: Development itself is the key to understanding developmental disorders. *Trends Cogn Sci* 2(10), 389–98 (1998).
[77] Karmiloff-Smith, A., Brown, J.H., Grice, S., Paterson, S.: Dethroning the myth: Cognitive dissociations and innate modularity in Williams syndrome. *Developmental Neuropsychology* 23(1-2), 227–242 (2003).
[78] Katz, J. J., Fodor, J.: The structure of a semantic theory. *Lang* 39, 170–210 (1963).
[79] Kaufmann, E., Kaul, Th.: Language switch costs and dual-task costs in bimodal language production. Paper presented at the conf. 'Formal and Experimental Advances in Sign Language Theory' (FEAST 2012), Univ. Warsaw (2 Jun 2012).
[80] Kemmerer, D., Castillo, J.G., Talavage, T., Patterson, S., Wiley, C.: Neuroanatomical distribution of five semantic components of verbs: evidence from fMRI. *Brain Lang* 107, 16–43 (2008).
[81] Kemmerer, D., Gonzalez-Castillo, J.: The two-level theory of verb meaning: an approach to integrating the semantics of action with the mirror neuron system. *Brain Lang* 112, 54–76 (2010).
[82] Kintsch, W.: *Comprehension: A Paradigm for Cognition*. Cambridge University Press, Cambridge, MA (1998).
[83] Knopka G., Bomar, J.M., Winden, K., Coppola, G., Jonsson, Z.O., Gao, F., Peng, S., Preuss, T.M., Wohlschlegel, J.A., Geschwind, D.H.: Human-specific transcriptional regulation of CNS development genes by FOXP2. *Nature* 462, 213–217 (2009).

[84] Koelewijn, T., van Schie, H.T., Bekkering, H., Oostenveld, R., Jensen, O.: Motor-cortical beta oscillations are modulated by correctness of observed action. *Neuroimage* 40, 767–775 (2008).
[85] Kuniyoshi, Y., Yorozu, Y., Ohmura, Y., Terada, K., Otani, T., Nagakubo, A., Yamamoto, T.: From humanoid embodiment to theory of mind. In: F. Iida, R. Pfeifer, L. Steels, Y. Kuniyoshi (Eds) *Embodied Artificial Intelligence*. Berlin: Springer, 202–18 (2004).
[86] Lacey, S., Stilla, R., Sathian, K.: Metaphorically feeling: Comprehending textural metaphors activates somatosensory cortex. *Brain Lang* 120(3), 416–421 (2012).
[87] Lai, C.S.L., Fisher, S.E., Hurst, J.A., Vargha-Khadem, F., Monaco, A.P.: A forkhead-domain gene is mutated in a severe speech and language disorder. *Nature* 413(6855), 519–523 (2001).
[88] Lakoff, G.: *Women, Fire, and Dangerous Things. What categories reveal about the mind*. University of Chicago Press, Chicago (1987).
[89] Lakoff, G., Johnson, M.: *Metaphors we Live By*. University of Chicago Press, Chicago (1980).
[90] Lakoff, G., Johnson, M.: *Philosophy in the Flesh: The Embodied Mind and Its Challenge to Western Thought*. New York: Basic Books (1998).
[91] Lamendella, J.T.: General principles of neurofunctional organization and their manifestations in primary and non-primary language acquisition. *Language Learning* 27: 155–96 (1977).
[92] Lenneberg, E.H.: *Biological Foundations of Language*. Wiley, New York (1967).
[93] Madden, C.J., Zwaan, R.A.: How does verb aspect constrain event representations? *Mem Cognit* 31, 663–672 (2003).
[94] Marinova-Todd, S., Marshall, D. Snow, C.: Three misconceptions about age and L2 learning. *TESOL Quar* 34, 9–34 (2000).
[95] Martin, A., Wiggs, C.L., Ungerleider, L.G., Haxby, J.V.: Neural correlates of category-specific knowledge. *Nature* 379, 649–52 (1996).
[96] Meader, C.L., Muyskens, J.H.: *Handbook of Biolinguistics*. Wiley, New York (1950).
[97] Montagu, A.: *Touching: The Human Significance of the Skin*. New York: Harper & Row (1978).
[98] Nazir, T.A., Boulenger, V., Roy, A.C., Silber, B.Y., Jeannerod, M., Paulignan, Y.: Language-induced motor perturbations during the execution of a reaching movement. *Q J Exp Psychol* 61(6), 933–43 (2008).
[99] Neininger, B., Pulvermüller, F.: Word-category specific deficits after lesions in the right hemisphere. *Neuropsychologia* 41(1), 53–70 (2003).
[100] Newell, A., Simon, H.A.: Computer science as empirical inquiry: symbols and search. *Communications of the ACM* 19(3), 113–26 (1976).
[101] Nowak, M.A., Komorova, N. L., Niyogi, P.: Computational and evolutionary aspects of language. *Nature* 417, 611–617 (2002).
[102] Obleser, J., Boecker, H., Drzezga, A., Haslinger, B., Hennenlotter, A., Roettinger, M., Eulitz, C., Rauschecker, J.P.: Vowel sound extraction in anterior superior temporal cortex. *Hum Brain Mapp* 27(7), 562–571 (2006).
[103] Osterhout, L.: On the brain response to syntactic anomalies: Manipulations of word position and word class reveal individual differences. *Brain Lang* 59, 494–522 (1997).
[104] Paul, C.: Morphology and computation. *Procs Int Conf Simulation Adaptive Behavior: From animals to animats*, Cambridge, MA: MIT Press, 33–8 (2004).
[105] Penfield, W., Roberts, L.: *Speech and Brain Mechanisms*. Atheneum Press, New York (1959).
[106] Pfeifer, R.: The emergence of cognition from the interaction of brain, body, and environment. 4th EUCogII Members' Conference "Embodiment – Fad or Future?", Anatolia College, Thessaloniki (11 Apr 2011).
[107] Pfeifer, R., Gomez, G.: Morphological computation – connecting brain, body, and environment. In B. Sendhoff, O. Sporns, E. Körner, H. Ritter, K. Doya (Eds) *Creating Brain-like Intelligence: From Basic Principles to Complex Intelligent Systems*. Berlin: Springer, 66–83 (2009).
[108] Pfeifer, R., Lungarella, M., Iida, F.: Self-organization, embodiment, and biologically inspired robotics. *Science* 318, 1088–93 (2007).
[109] Pfeifer, R., Scheier, C.: Sensory-motor coordination: The metaphor and beyond. *Robotics and Autonomous Systems* 20, 157–78.
[110] Pfeifer, R., Scheier, C.: *Understanding Intelligence*. Cambridge, MA: MIT Press (1999).
[111] Piatelli-Palmarini, M.: Evolution, selection and cognition: from "learning" to parameter setting in biology and in the study of language. *Cognition* 31, 1–44 (1989).
[112] Postle, N., McMahon, K.L., Ashton, R., Meredith, M., de Zubicaray, G.I.: Action word meaning representations in cytoarchitectonically defined primary and premotor cortices. *Neuroimage* 43, 634–644 (2008).
[113] Price, C.J.: The anatomy of language: contributions from functional neuroimaging. *J Anat* 197, 335–359 (2000).
[114] Pulvermüller, F., Assadollahi, R.: Grammar or serial order? Discrete combinatorial brain mechanisms reflected by the syntactic mismatch negativity. *J Cogn Neurosci* 19(6), 971–980 (2007).
[115] Pulvermüller, F., Hauk, O.: Category-specific processing of color and form words in left fronto-temporal cortex. *Cereb Cortex* 16(8), 1193–1201 (2006).
[116] Pulvermüller, F., Hauk, O., Nikulin, V.V., Ilmoniemi, R.J.: Functional links between motor and language systems. *Eur J Neurosci* 21(3), 793–797 (2005).
[117] Pulvermüller, F., Pye, E., Dine, C., Hauk, O., Nestor, P., Patterson, K.: The word processing deficit in semantic dementia: All categories are equal but some categories are more equal than others. *J Cogn Neurosci* 22(9), 2027–2041 (2010).
[118] Raposo, A., Moss, H.E., Stamatakis, E.A., Tyler, L.K.: Modulation of motor and premotor cortices by actions, action words and action sentences. *Neuropsychologia* 47, 388–396 (2009).
[119] Rieser, J., Garing, A., Young, M.: Imagery, action and young children's spatial orientation: It's not being there that counts, it's what one has in mind. *Child Dev* 45, 1043–1056 (1994).
[120] Rubba, J., Klima, E.S.: Preposition use in a speaker with Williams syndrome: Some cognitive grammar proposals. *Center for Research on Language Newsletter*, University of California, La Jolla, CA, 5, 3–12 (1991).
[121] Rueschemeyer, S.-A., Lindemann, O., van Rooij, D., van Dam, W., Bekkering, H.: Effects of intentional motor actions on embodied language processing. *Exp Psychol* 57(4), 260–266 (2010).
[122] Shahidullah, S., Hepper, P.G. Hearing in the Fetus: Prenatal Detection of Deafness. *Intl J Prenatal and Perinatal Studies* 4(3/4), 235–40 (1992).
[123] Siakaluk, P., Pexman, P.. Aguilera, L., Owen, W., Sears, C.: Evidence for the activation of sensorimotor information during visual word recognition: The body-object interaction effect. *Cognition* 106, 433–43 (2008).
[124] Sato, M., Mengarelli, M., Riggio, L., Gallese, V., Buccino, G.: Task related modulation of the motor system during language processing. *Brain Lang* 105(2), 83–90 (2008).
[125] Simmons, W.K., Ramjee, V., Beauchamp, M.S., McRae, K., Martin, A., Barsalou, L.W.: A common neural substrate for perceiving and knowing about color. *Neuropsychologia* 45(12), 2802–10 (2007).
[126] Smith, L.B.: Grounding toddler learning in sensory-motor dynamics. Keynote lecture, EUCogII Members' Conf "Development of Cognition in Artificial Agents", ETH Zürich (29 Jan 2010).
[127] Smith, N.V.: *The Twitter Machine: Reflections on Language*. Oxford: Blackwell (1989).
[128] Smith, N.V., Tsimpli, I.-M., Ouhalla, J.: Learning the impossible: The acquisition of possible and impossible languages by a polyglot savant. *Lingua* 91, 279–347 (1993).

[129] Spivey, M.J., Tyler, M.J., Eberhard, K.M., Tanenhaus, M.K.: Linguistically mediated visual search. *Psychol Sci* 12, 282–286 (2001).
[130] Stapel, J.C., Hunnius, S., van Elk, M., Bekkering, H.: Motor activation during observation of unusual vs. ordinary actions in infancy. *Soc Neurosci* 5, 451–460 (2010).
[131] Steels, L.: The origins and evolution of languages: Darwin's unsolved mystery. International workshop "150 Years after Darwin: From Molecular Evolution to Language", Inst for Cross-Disciplinary Physics and Complex Systems, Palma de Mallorca (2009, Nov 26).
[132] Stein, B.E., London, N., Wilkinson, L.K., Price, D.D.: Enhancement of Perceived Visual Intensity by Auditory Stimuli: A Psychophysical Analysis. *J Cogn Neurosci* 8(6), 497–506 (1996).
[133] Stewart, J., Gapenne, O., Di Paolo, E.A. (Eds) *Enaction: Toward a New Paradigm for Cognitive Science.* Cambridge, MA: MIT Press (2010).
[134] Stroop, J.R.: Studies of interference in serial verbal reactions. *J Exp Psychol* 18 (6), 643–662 (1935).
[135] Tanenhaus, M.K., Spivey-Knowlton, M.J., Eberhard, K., Sedivy, J.C.: Integration of visual and linguistic information in spoken language comprehension. *Science* 268, 632–634 (1995).
[136] Tettamanti M., Manenti, R., Della Rosa, P.A., Falini, A., Perani, D., Cappa, S.F., Moro, A.: Negation in the brain: Modulating action representations. *Neuroimage* 43, 358–367 (2008).
[137] Turing, A.M.: Computing machinery and intelligence. *Mind* LIX, 433–60 (1950).
[138] Ullman, M.T.: A neurocognitive perspective on language: The declarative/procedural model. *Nature Rev Neurosci* 2(10), 717–26 (2001).
[139] Uppenkamp, S., Johnsrude, I.S., Norris, D., Marslen-Wilson, W., Patterson, R.D.: Locating the initial stages of speech-sound processing in human temporal cortex. *Neuroimage* 31(3), 1284–96 (2006).
[140] van Elk, M., van Schie, H.T., Zwaan, R.A., Bekkering, H.: The functional role of motor activation in language processing: motor cortical oscillations support lexical-semantic retrieval. *Neuroimage* 50, 665–677 (2010).
[141] Varela, F.J., Thompson, E., Rosch, E.: *The Embodied Mind: Cognitive Science and Human Experience*. Cambridge, MA: MIT Press (1991).
[142] Vernes, S.C., Nicod, J., Elahi, F.M., Coventry, J.A., Kenny, N., Coupe, A.M., Bird, L.E., Davies, K.E., Fisher, S.E.: Functional genetic analysis of mutations implicated in a human speech and language disorder. *Hum Mol Genet* 15(21), 3154–3167 (2006).
[143] Vernon, D.: Cognitive development and the iCub humanoid robot. 2nd EUCogII Members' Conference "Development of Cognition in Artificial Agents", Univ Zürich (29 Jan 2010).
[144] Volterra, V., Capirci, O., Pezzini, G., Sabbadini, L., Vicari S.: Linguistic abilities in Italian children with Williams syndrome. *Cortex* 32, 663–677 (1996).
[145] Wang, P.: What do you mean by "AI"? *Artificial General Intelligence 2008*, Amsterdam: IOS Press, 362–73 (2008).
[146] Wang, P.: Embodiment: Does a laptop have a body? In B. Goertzel, P. Hitzler, M. Hutter (Eds) *Procs 2nd Conf Artificial General Intelligence, AGI 2009*, Paris: Atlantis Press, 174–9 (2009).
[147] Watson, J., Ritzmann, R., Pollack, A.: Control of climbing behavior in the cockroach, blaberus discoidalis. ii. motor activities associated with joint movement. *J Comp Physiol A* 188, 55–69 (2002).
[148] Wernicke, C.: *Der aphasische Symptomencomplex. Eine psychologische Studie auf anatomischer Basis.* Kohn & Weigert, Breslau (1874).
[149] Willems, R.M., Hagoort, P., Casasanto, D.: Body-specific representations of action verbs: Neural evidence from right- and left-handers. *Psychol Sci* 21, 67–74 (2010).
[150] Woll, B.: What can research on atypical signing tell us about the linguistics of sign language. Inv. talk, conf. 'Formal and Experimental Advances in Sign Language Theory' (FEAST 2012), Univ. Warsaw (2 Jun 2012).
[151] Woll, B., Morgan, G.: Language impairments in the development of sign: Do they reside in a specific modality or are they modality-independent deficits? *Bilingualism: Language and Cognition* 15(1), 75–87.
[152] Yamada, J.: *Laura: A Case for the Modularity of Language*. Cambridge, MA: MIT Press (1990).
[153] Zatorre, R.J., Evans, A.C., Meyer, E., Gjedde, A.: Lateralization of phonetic and pitch discremination in speech processing. *Science* 256, 846–849 (1992).
[154] Ziemke, T.: Human embodied cognition. Scientific evidence and technological implications. 4th EUCogII Members' Conference "Embodiment – Fad or Future?", Thessaloniki: Anatolia College (2011, Apr. 12).
[155] Zwaan, R.A., Taylor, L J., de Boer, M.: Motor resonance as a function of narrative time: further tests of the linguistic-focus hypothesis. *Brain Lang* 112, 143–149 (2010).

Augmenting Interaction: Collecting Common Sense Through AR Objects

Svetoslav Dankov and **Rafal Rzepka** and **Kenji Araki**[1]

Abstract. Augmented Reality applications are becoming more popular with the continued miniaturization of technology. With the increasing use of smart phones, which often provide increased processing power, enhanced and open software platforms, Augmented Reality has become instrumental in the way we perceive our surroundings and the information that it carries. It is now possible to implement an Augmented Reality system without carrying bulky and expensive equipment. Currently, there are many systems that implement some form of Augmented Reality to provide a specialized interaction to users. However, those systems usually employ expensive, immobile components with highly specialized interfaces. In this paper we present a novel approach for building interactive interfaces using Augmented Reality. UIAR (User Interface through Augmented Reality) is an augmented reality framework that allows for the ubiquitous creation and dissemination of interactive user interfaces and augmented reality objects. Here we present the novel interaction schemes of UIAR alongside the implementation of the framework's intended purpose - collecting common sense through AR objects.

Keywords: augmented reality, human-computer interaction, common sense, ubiquitous user interfaces

1 Introduction

Augmented Reality (AR) is a fairly young area of research which is currently expanding in many of the already existing fields of Human-Computer Interaction and Computer Interfaces. In our research we implement an Augmented Reality system which will serve as an extension to existing computer interfaces, provide enhanced user-experience, and define virtual objects and their actions in an ubiquitous way.

To implement such a system, however, we must first address three major areas where we think Augmented Reality user interfaces can be improved.

Firstly, Augmented Reality objects are hard coded into applications, which makes them highly specialized and not ubiquitous. Secondly, there is no standard for defining Augmented Reality interfaces and how they react to human interaction. Finally, there is no implementation of natural interaction with such interfaces and objects.

In the field of collecting common sense knowledge data from a large number of voluntary users, one of the biggest challenges is to keep the users engaged, entertained, and focused so as to collect a sufficiently large amount of high-quality data. Most such projects start with a somewhat big user-base, which eventually dwindles in numbers and activity as the project grows older. We believe using UIAR to implement games through its interactive framework can prove beneficial in this area.

We begin by presenting a summary of the related research, connecting our approach with previous work. Next we present use cases which describe specific functionalities that our framework enables and describe the purpose of the framework. This is followed by a description of the software tools and libraries we take advantage of while implementing the framework. Next we present the proof-of-concept games we have implemented with UIAR to collect common sense. We conclude by giving short review of the steps ahead.

[1] Hokkaido University, Japan, email: {dankov, kabura, araki}@media.eng.hokudai.ac.jp

2 Related Research

The existing research into Augmented Reality and Human Computer Interaction that is relevant to this study can be divided in three areas: finger and hand-based interfaces, paper-based interfaces, and Augmented Reality applications. We will describe each one briefly, followed by a review of approaches to collecting common sense.

2.1 Hand and Finger-based Interfaces

The technologies for hand and finger-based interfaces can be roughly split in two categories - sensing-based and computer-vision based. Sensing-based systems like [21] are very robust but are often limited to detecting only "touch" behavior, not able to recognize hands or other physical object that come into view. Computer-vision based systems like [5], [6], [17] are often limited by the lighting conditions and may not respond well to sudden changes in the field of view. However, systems like [10], [11], [8] have proven to be robust and accurate enough. We are using a computer-vision based system since wearing special hardware to enable "touch" capability reduces the mobility of the system.

2.2 Paper-based Interfaces

Most of the existing paper-based interfaces fall into three categories - using paper alongside digitizing tablets like [16], using digital paper technologies like [4], and using paper tagged with markers (barcodes, fiducial markers, etc.) like [17]. In our system we will be using only 2D paper tags (AR markers) for 3D positioning of the visual objects, unlike [11] where the hand position and direction is used to determine the position of the virtual object. The interaction between the user and the interface will be entirely virtual or conveyed through AR marker motion.

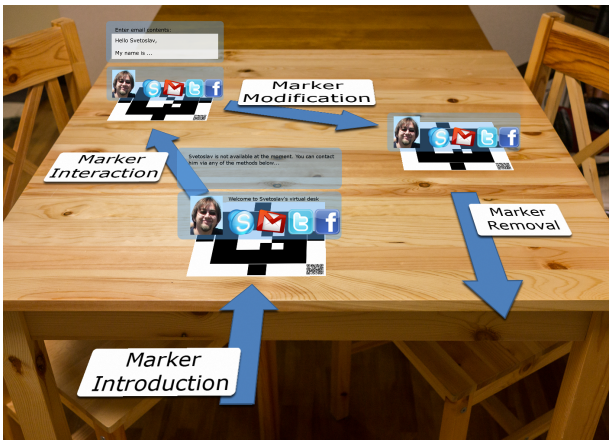

Figure 1. Example usage scenarios

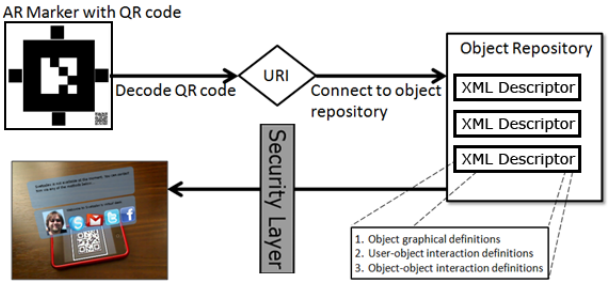

Figure 2. Basic framework model

2.3 Augmented Reality Applications

There have been many Augmented Reality applications, using either multiple-camera hand and object tracking or a single camera (like a webcam). Those applications vary in both their mobility and complexity. Our project was inspired for the most part by the Sixth Sense project developed by Pranav Mistry in the MIT Media Lab [18]. As is the purpose of [18], we strive to provide mobility, affordability and ubiquity to Augmented Reality applications.

There are two major differences between the paradigm employed by [18] and our framework, which highlight the novelty of UIAR. We let the user utilize any device to view the AR environment (mobile phone, webcam plus desktop, AR goggles, etc.) where [18] projects the AR environment over objects themselves. We also use 2D paper AR markers to determine correct 3D coordinates and scale for object placement.

The system presented in [20] showcases a collaborative AR environment, allowing multiple participants to interact with two and three-dimensional data using tangible user interfaces.

2.4 Augmented Reality Frameworks and Authoring Tools

With the popularization of software libraries like ARToolKit [7] there has been a lot of development to bring AR authoring tools in the hands of researchers and developers. However, frameworks like DWARF [2] and osgART [15] are quite complex and require an expert programmer. Our framework on the other hand gives developers with enough programming experience in ActionScript3 the ability to construct and distribute user interfaces, interactive objects, etc. with ease.

2.5 Common sense acquisition

One of the great challenges facing the artificial intelligence community is creating agents that can operate and adapt in a natural human environment. Naturally, we must be able to provide those agents with the information and the learning tools necessary for them to operate. If we want them to be able to make decisions about, relate to and have a simple understanding of the global environment in which they function, they need to be provided with basic knowledge about the world [12]. In humans, this knowledge is available to us in the form of reasoning shortcuts and factual information about each particular occasion we find ourselves in, and is generally known as common sense.

This type of knowledge is also crucial for fooling humans that they are conversing with another human - in other words passing the Turing test.

With the realization of the importance of common sense to the field of artificial intelligence, considerable research has been done towards collecting and structuring this type of knowledge. The biggest research effort by far has been the Cyc project [12] which has already collected over a million common sense assertions in little over two decades. As the project became more of a commercial venture, a much smaller set of data is available free of charge. The work required, however, has been considerable. Common sense knowledge is manually input by experts in particular areas, who first give a complete ontological structure to the data, using a specially developed knowledge representation language called CycL, and then insert domain specific data based on their expertise [12].

Another attempt to collect common sense data is the Open Mind Common Sense project. OMCS collects common sense statements from untrained volunteers over the Web in the form of natural language statements [23]. In the course of few years the project had already collected over 1.6 million statements.

Other systems, like Verbosity [1] and Common consensus [14] identified and addressed one of the major problems with such systems - user interest. In order to consistently gather quality knowledge from a large set of volunteers, they must be given enough motivation to continue to participate, especially in light of the fact that the number of volunteers drops over the life of the project.

Yen-Ling Kuo et al. [9] have successfully utilized social games to collect common sense and their findings provide useful suggestions for designing community-based games.

3 Usage Scenarios

The usage-scenarios described below serve to describe specific features of our framework that are not available in current AR systems. Via those usage scenarios we want to illustrate the particular unique functionalities that our framework provides.

3.1 Universal Marker Registry

In our system we use 2D paper markers for virtual object placement. Currently, there are multiple ways to create such a marker with the only restriction that the pattern not be too complex. This improves

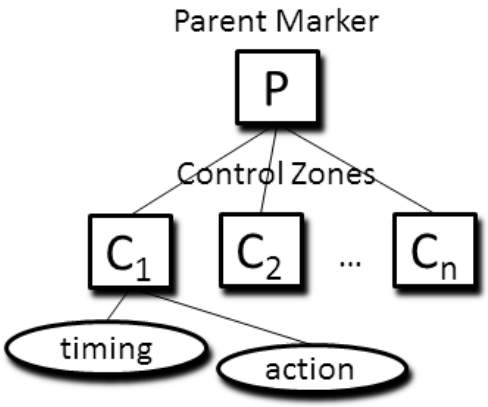

Figure 3. AR Object Control Scheme

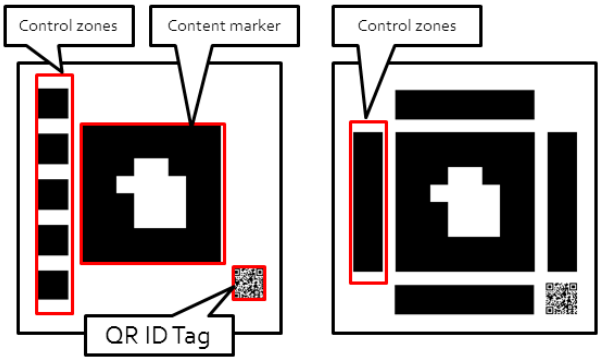

Figure 4. User-object interaction

the marker recognition which in turn allows for a scalable AR experience. The relation between the virtual object and the marker pattern is embedded in the software.

We plan to implement a universal AR marker registry so that information about the object is stored globally. Using this registry, users will be able to point their AR device to any AR marker and display its contents, regardless of whether they have seen the marker before or not. The system will recognize the URL encoded in the marker, download and display the virtual object. We are planning to include a QR code within the AR marker pattern and extract the URL from it.

This will allow for more ubiquitous AR applications. The user will no longer be restricted to using markers specifically designed for his AR system. We hope the implementation of such registry will attract interest from the Augmented Reality community and help construct a large ecology of AR objects. It will also enable developers to construct their own AR Marker ecologies, independent of the main system.

3.2 AR Object-object and Object-user Interaction Definition

The next step in our system is defining virtual object actions as part of their registry information described in the previous section. This way the AR system will know both what objects to display, as well as how those objects are supposed to interact with the user and other objects.

For the most part our virtual objects are user-interfaces. As such, their actions are defined either as user initiated or object initiated. We would like to implement both ubiquitous object-object and user-object interaction. Object-object interactions will allow us to define how virtual objects behave when in proximity of one another. A simple example would be two virtual objects positioned by 2D markers on the field of view, both objects representing a single Skype chat window with different users. If both markers are positioned close to one another the resulting action will be to open a single Skype window making a conference call to both users. User-object interactions will be described as the services the virtual object can perform upon user actions. For example, a virtual object displaying information on a person (a virtual business card) can provide information upon request, provide an email interface, a Facebook or Twitter message interface, current location, etc.

Building such a repository of objects will provide both a functional and a graphical description of the AR objects in an Augmented Reality environment which in turn will make AR applications more ubiquitous.

4 Framework purpose and novelty

- **Ubiquity of Object presence**
  Objects associated with one AR marker must remain the same independent of system used to view the AR environment. Objects are first registered via a QR code which encodes the URI of the repository from which the object graphics model and other data is to be drawn. This gives each AR marker (independent of its graphical representation) an unique identifier. This unique identifier allows us to implement the next three objectives.
- **Object Persistence**
  Objects must carry associated data and object states across different AR environment viewers. Object data and object states are stored in the database defined via the AR marker's URI.
- **Ubiquity of Object Interactivity**
  Objects must behave the same way independent of AR environment viewer. Object's interaction definitions are stored in the database defined via the AR marker's URI.
- **Definition of Object Interaction Models**
  Interaction models with AR systems have so far been system/application dependent. Each system defines for itself how users interact with the AR objects and the interaction model cannot be extended or redefined.
  Our framework allows developers to define how a specific AR Object will interact with the users and with other AR Objects introduced to the scene. They do so by assigning behaviors to extra control markers associated with the AR Object via the AR Object's URI. Figure X shows an example of AR Objects and controls.

Figure 1 shows an example usage scenario encompassing all four focus areas described above. As the AR marker is inserted into the scene, the marker's graphical and functional representation is obtained from the object repository on the server encoded in the QR code. In this case the AR marker's object is a simple business card showing a photograph, 4 buttons which when activated would present

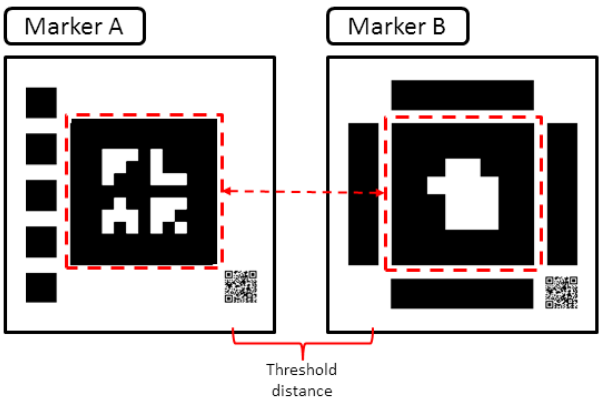

Figure 5. Object-object interaction

a different interface depending on the button, and a message board. The user can modify the content - in this case leave a new message on the message board or modify the object the marker represents - in this case by closing the message board section. Once the marker is removed from the scene it's object properties are saved and the next time the marker is introduced it will remember them. The currently existing implementations of augmented reality systems, although providing some of the above mentioned features, do not encompass all of the features UIAR intends to provide. This, in our opinion, is what distinguishes our framework from the existing ones, and gives UIAR an innovative perspective on augmented reality interfaces.

5 System design, software components and implementation

5.1 System setup

Currently, UIAR is intended to be used as a desktop application. The system has three major components - a desktop PC, a webcamera capable of minimum 640x480 resolution and printed AR markers. The test setup involves positioning the webcamera above and behind the user, who is sitting by a flat surface, with the desktop positioned in front of the user. An alternative setup is to have the webcamera positioned over and perpendicular to the flat surface. As image processing is prone to errors in different lighting conditions, we found it best to have the setup lit up by one uniform light source, away from direct sunlight. This way marker tracking and recognition is optimal. The user can then proceed to introduce markers to the scene, register them through their QR codes and interact with them.

5.2 Software components

Our system is based on several existing technologies that allow us to perform AR overlay, QR decoding, marker recognition, tracking and handling and draw our interfaces programatically. In this section we will look at each one in more detail.

- **AR overlay:** The original AR toolkit was first developed by Dr. Hirokazu Kato from the University of Washington [7] and is currently supported by the Human Interface Technology Lab at the University of Canterbury in New Zealand [3]. As we are building our framework in Adobe ActionScript programming language, we are using a language port of the ARToolKit to AS3 provided by Saqoosha [22], Nyatla [19] and Sparklib [13] named FLARToolKit.
- **QR decoding:** For decoding QR codes in ActionScript we use the QR library provided by Sparklib [13].
- **Marker handling:** To manage marker registration efficiently for multiple markers and predict marker motion we use the FLARManager 0.7 toolkit which is provided by Eric Socolofsky [24].
- **Interface Design:** To design, draw and define our interfaces we use PaperVision3D library provided by [25]. PaperVision3D is a set of libraries that give ActionScript developers a 3D engine for Flash.

All of the above mentioned libraries are distributed under licenses allowing developers to use them free of charge for non-commercial purposes. Our framework is built using Flash Builder 4 and ActionScript 3.5 SDK. Developers can produce their modules using any ActionScript compiler as long as they run the same SDK and use the same versions for PaperVision3D, FLARToolKit, and FLARManager.

5.3 System model

Figure 2 describes the basic model of our system. Here we will look at each component.

- **AR Marker with QR:** We designed our AR markers to include QR codes encoding the Unique Resource Identifiers for the object that the AR marker identifies. This allows the developer to define his own AR marker patterns and objects independent of the viewer. It also allows the AR environment viewer to recognize AR markers without the need to include the patterns in the program. The QR code can be placed either inside of the AR marker as part of the pattern or on the back of the AR marker. Note that if the QR code becomes a part of the AR marker's pattern it must do so in an asymmetrical fashion, since AR marker patterns must be asymmetrical to enable correct marker detection.
- **Database:** The database component of the system implements a simple MySQL scheme with database entries containing developer information and pointing to a local directory for specific marker id. The physical file is a precompiled Adobe SWF file that contains the AR Object's graphical and interaction definitions.

With object persistence we ensure that an AR object will retain its information and state in case it is removed from the AR environment. As we saw in Figure 1 , if an AR Marker is introduced and the user makes a modification to the state of the object it represents, the system will relate that change to the database. The next time that marker is introduced to the scene, the system will display it's previously modified state.

5.4 System interaction

In order to continue to the next section we must define the control scheme for AR Objects. Figures 3 and 4 show how we implement interaction with our objects. In the database, each parent AR Marker has associated with it a set of control patterns that define a single action. The control patterns are additional shapes printed on the side of the AR marker. Those control patterns are our equivalents of a "button". Each control pattern is defined by a "timing" parameter and an "action" parameter. The system continuously scans for the patterns

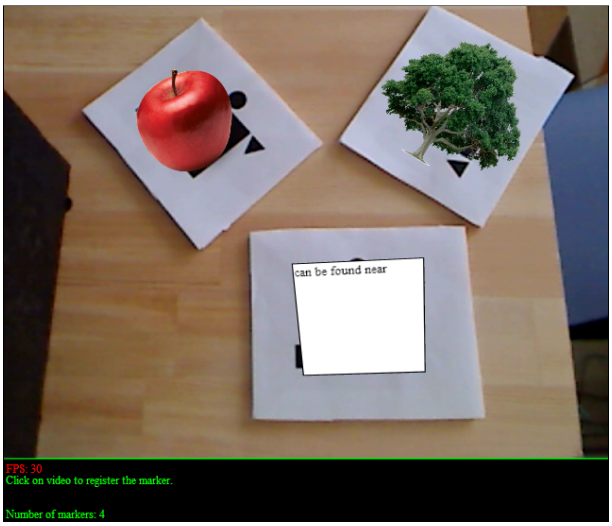

Figure 6. Visual Sentence Patterns

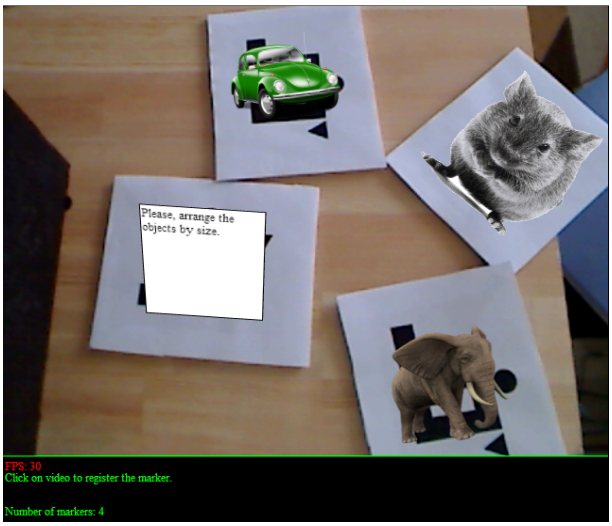

Figure 7. Arrange by Feature

and detects if the control has not been visible longer than the "timing" parameter specifies (which means that the control is activated), and performs the action based on the "action" parameter.

We can now describe how our system's user-object and object-object interaction paradigms.

5.4.1 User-object Interaction

Figure 4 describes how the users will interact with the AR Objects. For each marker ecology (defined by the database the system is connecting to) there will be multiple control patterns active. The shapes representing the control patterns do not have to be unique, since they are registered by the developer upon uploading the marker into the UIAR database. The detection technique the system uses is the same described in section **5.4**.

5.4.2 Object-Object Interaction

The last type of interaction we define in our system is object-object interaction. Figure 5 gives an example of one AR object being aware of another. In this paradigm the developers of the AR Objects are allowed to define a relative "distance threshold", which serves as a trigger for a predefined action. Since at any given time the system knows which patterns are present on the scene, along with their relative size, it can calculate relative distance between each marker. A simple example of this interaction scheme is illustrated in the next section, where aligning all markers in close distance simulates the action of a "submit" button.

6 Common sense acquisition games

We believe that using our framework can prove useful when it comes to both acquiring new volunteers and keeping the existing volunteers interested and engaged in the process. Using the UIAR framework we can implement games that are rewarding, engaging, and interactive. Moreover, those games can be targeted towards younger audiences who naturally spend more time playing games.

In this section we present a proof-of-concept implementation of UIAR for the purpose of collecting common sense.

With UIAR, developers can assign any virtual object to an AR marker (3D models, textual and media content, etc.) which will be immersive (objects will blend in with the actual environment) and interactive (objects will be aware of the surrounding objects and react to users' input). Here we present 3 implementations of games to collect specific types of common sense inspired by such games as Verbosity [1] and Common Consensus [14]. Images for those games are acquired through Google Image Search. We use Wordnet to choose related concepts when needed.

6.1 Visual sentence pattern game

Figure 6 gives an example of an instance of the visual sentence pattern game. One type of exercise commonly used to collect common sense is to simply fill in the gaps in a sentence pattern. While natural language sentences can often prove difficult to process, the use of different sentence patterns allows for collecting data that can be disambiguated, categorized and easily parsed.

Our example game uses templates such as: "X is a kind of Y", "X is used for Y", "X is typically found near/in/on Y", " X is the opposite of Y", "X is related to Y", etc. A user is provided with AR markers, which after being registered via their QR codes will correspond to X, Y, and a description of the template respectively. As the visual content of AR markers is dependent on the QR code only, the game can be setup so that every time the markers representing X or Y is re-registered, the content is changed.

For example, on first registry marker X can hold a 3D model of an apple and Y a 3D model of a tree. If the sentence pattern does not make sense, the user can re-register it until he gets the correct one (in this case, "X is typically found near/in/on Y"). To submit the entry, the user needs to arrange all three markers so that they touch each other. After submission, the game refreshes the markers with new objects and sentence patterns and the user can keep playing.

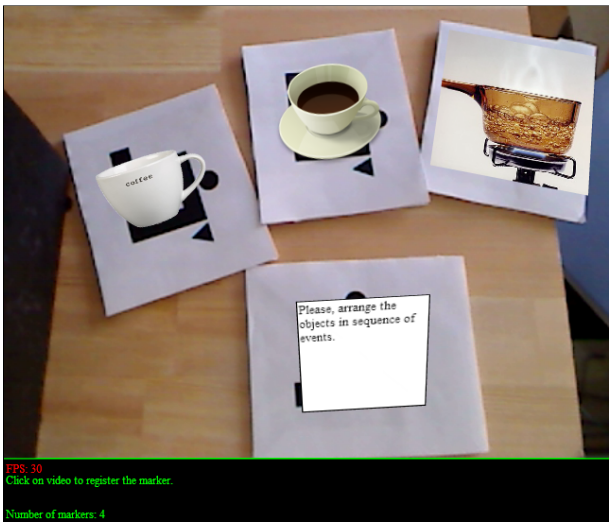

Figure 8. Arrange in Sequence

6.2 Arrange by feature game

In figure 7 we can see an example of the second type of game which is oriented toward spatially oriented knowledge collection. In this case the number of markers/objects can be as many as the screen can allow. A sample exercise of this game would be to ask the user to arrange the markers in a certain order based on a certain criteria (height, length, size, etc.). As the markers are spatially aware of each other, the user will complete the exercise by putting the markers close to each other in the order needed.

For example, the user can have 4 markers. After registering each marker, he is presented with 4 different 3D (or 2D) objects which he/she will arrange by a predefined feature and submit to the system. The same game can be used to cluster individual objects in case there is more than one feature. For example, the user can be presented with representatives of fruits and animals, in which case he can group the markers together and choose their categorization.

6.3 Arrange in sequence game

The third type of game, illustrated in Figure 8, is goal oriented. Just like the ”arrange by feature” game, in this game the user will be presented with two markers representing the beginning and the end of an activity, with the rest of the markers representing actions that must fit in a sequence.

For example, the user can start with 2 markers, one showing - the rising sun (or a person coming out of bed) and another a steaming cup of coffee. The rest of the markers could represent ”boiling water”, ”mixing water in cup”, ”opening coffee”, ”pouring sugar”. The user will complete the exercise by arranging the markers/activities in the right order.

Each object/activity can be represented either by a visual (a 2D or 3D model) or just text (the text being overlaid over the AR marker). In order to represent both simple physical objects and abstract concepts with more visual appeal it is better to use images or models. The example implementations given are specific to the realm of common sense acquisition. However, the system can be generalized to serve any number of language acquisition tasks.

7 Conclusions

In our research we are trying to address the need for enriching textual knowledge with interaction driven knowledge acquisition. We plan to implement an AR viewer for mobile devices using an HTC developer device running Android 2.3 OS. Additionally, we plan to implement additional methods for user-object interaction, improve the overall usability of the system, and implement security schemes for the AR objects.

We are currently in the process of deploying a prototype version of the UIAR framework as an open source project. We are performing evaluations on the common sense knowledge collection games, the quality of the collected data and how our system affects user retention.

We view Augmented Reality as a technological path that will keep extending and developing with constant future software and hardware improvements. We believe Augmented Reality, as opposed to Virtual Reality and Augmented Virtuality, is the medium that will be most appealing to everyday users. The medium through which we collect, interact and distribute knowledge, will diverge from the nowadays common desktop/laptop/smart-phone solution. Data interaction will become more ubiquitous as the devices through which we perform everyday tasks become themselves more ubiquitous. Even today there are multitude of projects and proof-of-concept products that look into the future of how we interact with machines. They are only one step away from becoming as mass produced and spread as the smart phone has become in the last 5 years. This proliferation of ubiquitous devices, not restricted by common keyboard/display solutions are the reason why research into interactive AR interfaces is so important to us.

REFERENCES

[1] L. V. Ahn, M. Kedia, and M. Blum, ‘Verbosity: a game for collecting common-sense facts’, *In Proceedings of ACM CHI 2006 Conference on Human Factors in Computing Systems*, **1**, 75–78, (2006).

[2] M. Bauer, B. Bruegge, G. Klinker, A. MacWilliams, T. Reicher, S. Riss, C. Sandor, and M. Wagner, ‘Design of a component-based augmented reality framework’, *Augmented Reality, International Symposium on*, **0**, 45, (2001).

[3] M. Billinghurst. Artoolkit. `http://www.hitlabnz.org`, June 2009.

[4] P. Brandl, M. Haller, J. Oberngruberand, and C. Schafleitner, ‘Bridging the gap between real printouts and digital whiteboard’, *AVI’08*, (2008). In Proceedings of the conference on Advanced Visual Interfaces.

[5] S. Do-Lenh, F. Kaplan, A. Sharma, and P. Dillenbourg, ‘Multi-finger interactions with papers on augmented tabletops’, *TEI2009*, 16–18, (2009). The 3rd International Conference on Tangible and Embedded Interaction, Cambridge, UK.

[6] D. Holman, R. Vertegaal, M. Altosaar, N. Troje, and D. Johns, ‘Paper windows: interaction techniques for digital paper’, *CHI*, 591–599, (2005). Proceedings of CHI’05.

[7] ARToolworks Inc. Artoolkit. `http://www.hitl.washington.edu/artoolkit/`, June 2009.

[8] C. Keskin, A. Erkan, and L. Akarun, ‘Real time hand tracking and 3d gesture recognition for interactive interfaces using hmm’, *ICANN/ICONIP*, (June 2003). Istanbul.

[9] Y. Kuo, K. Chiang, C. Chan, J. Lee, R. Wang, E. Shen, and J.Y. Hsu, ‘Community-based game design: Experiments on social games for commonsense data collection’, *HCOMP2009-Workshop on Human Computation*, (2009).

[10] B. Lee and J. Chun, ‘Manipulation of virtual objects in markerless ar system by fingertip tracking and hand gesture recognition’, *ICCIT’09*, (2009). Seoul.

[11] T. Lee and T. Hllerer, ‘Handy ar: Markerless inspection of augmented reality objects using fingertip tracking’, *ISWC*, (October 2007). In Proceedings for IEEE International Symposium on Wearable Computers, Boston, MA.

[12] D. Lenat, 'Cyc: Towards programs with common sense', *Communications of the ACM*, 30–49, (1990).
[13] Spark Project Actionscript Class Library. http://www.libspark.org/, June 2009.
[14] H. Lieberman, 'Common consensus: a web-based game for collecting commonsense goals', *In Proceedings of IUI*, (2007).
[15] J. Looser, R. Grasset, H. Seichter, and M. Billinghurst, 'Osgart - a pragmatic approach to mr', *ISMAR'06*, (2006).
[16] W. Mackay, G. Pothier, C. Letondal, K. Boegh, and H. Erik Sorensen, 'The missing link: augmenting biology laboratory notebooks', *UIST'02*, 41–50, (2002). In ACM Symposium on User Interface Software and Technology.
[17] C. McDonald, G. Roth, and S. Marsh, 'Red-handed: collaborative gesture interaction with a projection table', *FG2004*, 773–778, (2004). International Conference on Automatic Face and Gesture Recognition.
[18] P. Mistry, P. Maes, and L. Chang, 'Wuw - wear ur world - a wearable gestural interface', *CHI'09*, (2009). In the CHI '09 extended abstracts on Human factors in computing systems, Boston, MA.
[19] Nyatla. Nyartoolkit for as3. http://d.hatena.ne.jp/nyatla/, June 2009.
[20] H.T. Regenbrecht and M.T. Wagner, 'Interaction in a collaborative augmented reality environment', *In Proceedings of ACM CHI 2002: Changing the World, Changing Ourselves*, 504–505, (2002).
[21] J. Rekimoto, 'Smartskin: An infrastructure for freehand manipulation on interactive surfaces', *SIGCHI*, 113–120, (2002). Conference on Human Factors in Computing Systems.
[22] Saqoosha. Flartoolkit. http://saqoosha.net/, June 2009.
[23] P. Singh, 'The public acquisition of common sense knowledge', in *Proceedings of the AAAI Spring Symposium on Acquiring (and Using) Linguistic (and World) Knowledge for Information Access*, (2002).
[24] E. Socolofsky. Flarmanager. http://words.transmote.com/wp/flarmanager/, December 2009.
[25] C. Ulloa, J. Grden, R. Hauwert, T. Knip, and Andy Zupko. Papervision3d: 3d engine environment for adobe flash. http://www.papervision3d.org/, June 2009.

RhetorEthics, or – on Implementing an Aristotelian approach to Machine Ethics

Radoslaw KOMUDA[1] **, Rafal RZEPKA**[2] **and Kenji ARAKI**[2]

Abstract. We begin this paper with revisiting the differences between descriptive and normative approach to ethics and argue about the usefulness of the latter for the filed of Machine Ethics. We continue this reasoning and present our insights on previous trends in this field and highlight the need for a change in the approach. We justify that experimental approach to Machine Ethics by introducing a moral reasoning system based on Aristotelian identification of civic rhetoric. And present it as a step forward in the Machine Ethics research bypassing theoretical disputes between specialists. We finish this paper with the introduction to the CAMILLA project for adjusting our web-crawling agent and creating an Aristotelian explicit moral agent.

1 INTRODUCTION

It is said that the difference between theory and practice is reality. During our research in the Machine Ethics we have come across a number of ideas and approaches to the problem of this field of science. However, these theoretical solutions and philosophical arguments could be actually summarized in one sentence: “Socrates was right!”.

This ancient philosopher had claimed that “it is the same to know right and be righteous” [1]. His assumptions about humans' moral competence (the capacity to do what is right) were idealistic (although not in the Platonic manner) and do not cover fully human behavior, especially, when they are contrasted with human tendency to, e.g., egoistic behavior. However, machines lack this kind of tendency and that is what makes us focus on the true issue of Machine Ethics. Since machines are different from humans, the question on HOW to teach machines good from wrong has to be reformulated for the need of machine-based reasoning.

2 AGAINST NORMATIVE APPROACH TO MACHINE ETHICS

To give a better insight into this matter, lets take Asimov's First Law of Robotics into consideration. It states that “a robot may not injure a human being or, through inaction, allow a human being to come to harm”. Seemingly, it covers all situations in which an agent may cause harm to a human being: by taking an action or through inaction. It sounds perfect and fulfilling as long as we do not question agent's ability to predict or calculate potential harm caused by its (in)action.

Many current trends try to force an idea of a friendly AI [2] or present vision of the future in which robots “enjoy” working side by side with humans [3] but, e.g., lack the technical details about realizing these ideas.

Ethics is naturally divided into descriptive (saying how things are) and normative (telling how things should be). Theoretical deliberations alone rarely exceed the field of philosophy and as long as there is no engineering insight into a presented approach – its contribution to the actual research in Machine Ethics is minimal.

Another benefit from the direct implementation is the unquestionable progress in the field of ethics itself. Since a machine can only follow preprogrammed commands, it shall – until a significant progress in the field of machine consciousness is made – absolutely obey them. Thanks to that – philosophers will be able to get an unprecedented insight into the ethical system being strictly followed on a neutral ground, without any exceptions or misstatements.

This absolute obedience secondly bring us to the situation in which authors introducing the field of Machine Ethics often make references to visions known from the science-fiction scenarios. They often justify the need for the research in the field of Machine Ethics by saying that “it is clear that machines such as these (family cars that drive themselves → Author's addition) will be capable of causing harm to human beings unless this is prevented by adding an ethical component to them” [4] which is an eristic stratagem known as the a*rgumentum ad populum*. It is supposed to get listeners excited about such vision and divert their attention from the main issue, that is: Why do we do not input such essential ethical component to GPS systems in our cars?

We refer to our approach to this matter as “the Artificial Intelligence's Ockham's razor”. Following the basic rule of the original principle: “simpler explanation is better than a more complex one” – we believe in implementing the AI – not to mention Machine Ethics – solutions only if essential. Machines are task-, not – reason-oriented, e.g. an “avoid collisions” rule is enough for a self-driving car and turning it into an “avoid collisions because it may harm a human being” is a triumph of form over the content.

[1] Faculty of Theology, Nicolaus Copernicus University; Torun, Poland.
e-mail: komuda@stud.umk.pl.

[2] Graduate School of Information Science and Technology, Hokkaido University; Sapporo, Japan.
e-mail: {kabura,araki}@media.eng.hokudai.ac.jp.

3 EXPLICIT ETHICAL AGENT

Our approach is consistent with the approach by Komuda et al [5]. We are not taking an excessive part in the discussion on choosing either implicit or explicit approach to artificial moral agents. Our main focus in this paper is to highlight the need for a discussion on the essence of Machine Ethics.

3.1 WHAT "GOOD" IS?

"Good" can be classically defined after St. Thomas Aquinas [11] as "*quod omnia appetunt*" ("what everybody desires") However, can we really come up to a consensus in that matter for humans? And is it possible to find the answer when it comes to machines?

Our world is a vast place. People not only around the world but also in our countries, our cities, our neighborhoods, our communities have different values and beliefs. Are we able to reconcile these factors while pursuing our dream of a robot free in its being?

We believe that Machine Ethics is not only able to overcome these difficulties but above all – it is a great tool in search for an intercultural understanding. Though this is an argument supporting the implicit approach, we believe that we could easily extract a "do not kill" imperative from every major religious doctrine and philosophical system. The difference would be in its reasons and justification.

3.2 WHAT IS GOOD? - ARTIFICIAL MORAL INTELLIGENCE

The main problem of Machine Ethics research is the same unsolved dilemma of the ethics itself – what is good? Depending on the situation, circumstances and context – omitting our previous insights in this matter [5] – we judge the moral quality of an action differently, e.g. "stealing a car" we find wrong, especially when a thief does so to sell it or we are talking about juvenile offenders wanting to "take a ride". But the same action would not be judged that harshly if we had learned that somebody has used the car to drive a pregnant woman that was about to give birth to the hospital.

Fortunately, the way in which an agent would be supposed to collect additional information about the inquired situation does not lay in the scope of interests on Machine Ethics research. It focuses on the evaluation itself and how an artificial moral agent would be supposed to qualify an action as good or wrong on provided information. That is a kind of a moral intelligence that resembles human moral judgment, since we also do not ask additional questions about the situation.

We have decided to split the task of our research into three successive sub-tasks.

4 ARTIFICIAL MORAL:

4.1 ADVISER

Basic idea behind Artificial Moral Adviser (AMAdv.) is combining our previous experiences and research results [7, 8] and creating an agent capable of making its own conclusions based on the data extracted from the Web.

The relevant difference between an AMAdv. and an Artificial Moral Agent (AMA) itself lays in the fact that the first will not claim the right to judge the moral quality of an act in terms of good or wrong. Although it is going to possesses – essential for an explicit AMAs – the need to justify its judgment, it will use the obtained results to "suggest" a reappraisal.

4.2 CONSCIENCE

Human conscience is both pre– and post–action. It means that we are able to both determine the quality of an act before or even without taking it and feel content or remorse after it.

Since an AMAdv. could be treated as a pre–action conscience, next step in creating an AMA is making the Agent capable of judging reactions of participants on the same emotion extraction scheme and marking it as a success or a failure.

4.3 AGENT

In our assumption, creating an artificial moral agent is the ultimate goal of Machine Ethics. We believe that it may be achieved by combining pre– and post–action emotion extractions from the WWW resources.

5 ARISTOTELIAN (MORAL) ORATOR

We have decided to adapt a similar to the described in section 4.2 idea from Aristotle's "Rhetoric" [9]. This treatise on the art of persuasion distinguishes the three genres of rhetoric: a deliberative *sumbouleutikon* which considers the future and encourages to or refrains from doing something, a forensic *dikanikon* interestd in the past and prosecution or defense of the individual and the epideictic *epideiktikon*, also known as the praise-and-blame rhetoric.

The reason we have decided to use the Aristotelian approach is not only because of the usefulness of the introduced positions of the disputants but also because it provides a set of rules defined by Aristotle, e.g.: harmful things may never be advised, and useful – discouraged. We believe that this position is generally represented by humans, or – the Web-Crowd as we like to refer to the Internet contents.

6 THE CAMILLA PROJECT

Aristotle presents some important roles of premises in the deductive argument. We believe three of the premises introduced by him, namely, (gr. *tekmeria*), probability (gr. *eikota*) and signs (gr. s*emeia*) are essential not only for a proper syllogism but also – for a proper moral judgment. A thing that is impossible by its nature could not and can not happen. That is why we want our agent to be common–sense aware and introduce the Common–sense Aware Morally InteLLigent Agent, a.k.a the CAMILLA Project.

In our concept–based research, our Agent is ought to define action participants and categorize them, i.e. "John killed Jim" is going to be generalized to "A human killed a human". After the second step – ensuring that "a human" can "be killed" – our Agent will crawl the web in search for sentences corresponding to that model and perform emotion extraction. This prevents erroneous queries on one hand.

However, it might raise the risks of such since "a ball" and "a car" would be categorized as "objects" and "throwing" or "catching" it should be possible.

Common-sense dictates that:

1. An average human can throw a ball.
2. An average human can catch a ball.
3. An average human can not throw a car.
4. An average human can not catch a car.

and these are the conditions we want our Agent to be able to both find / extract and consider in its moral reasoning.

7 FUTURE WORK

Moral intelligence is the capacity to understand right from wrong. We believe that making our agent able to interpret previously extracted emotions into a decision or advise to take or withdraw an action will be a promising step ahead in achieving this goal. And since we support the Socratic approach to Machine Ethics and – the creation of a free and independent machine itself.

REFERENCES

[1] Aristotle. Eudemian Ethics. 1216b.

[2] Yudkowsky Eliezer. Creating Friendly AI. (2001)

[3] Waser Mark R. A Safe Ethical System for Intelligent Machines. Proceedings of The AAAI 2009 Fall Symposium on Biologically Inspired Cognitive Architectures (BICA-09), Washington, D.C., USA, November 5–7, 2009.

[4] Anderson, M.; Anderson, S. L. *Machine Ethics: Creating an Ethical Intelligent Agent.* [in:] AI Magazine, vol. 28, number 4, 15-26 (2007).

[5] Komuda Radoslaw, Ptaszynski Michal, Momouchi Yoshio, Rzepka Rafal and Araki Kenji. *Machine Moral Development: Moral Reasoning Agent Based on Wisdom of Web-Crowd and Emotions*" [in:] International Journal of Computational Linguistics Research, Vol. 1 , Issue 3, pp. 155-163, 2010.

[6] Michal Ptaszynski, Pawel Dybala, Wenhan Shi, Rafal Rzepka and Kenji Araki: "Disentangling emotions from the Web. Internet in the service of affect analysis". Proceedings of the Second International Conference on Kansei Engineering & Affective Systems (KEAS'08), pp 51-56, Nagaoka, Japan. (2008).

[7] Wenhan Shi. *Discovering Emotive Content in Utterances Using Web-mining* (in Japanese). Hokkaido University. (2008).

[8] Ptaszynski M, Dybala P., Shi W., Rzepka R., Araki K (2008). "Disentangling emotions from the Web. Internet in the service of affect analysis". In: Proc. of the Second International Conference on Kansei Engineering & Affective Systems (KEAS'08), pp 51-56, Nagaoka, Japan.

[9] Aristotle. *Rhetorics*. Book I, Chapter 3, 1358b–1359a. In: Arystoteles Dzieła wszystkie, t. 6 (in Polish), WN PWN, Warsaw 2001.

[10] Joannes Stobaeus, 2.77.

[11] St. Thomas Aquinas. De veritate 1.1.

A Domain Analytic Method in Modular-Designed Dialogue System: Application to a System for Japanese

Motoki Yatsu, Rafal Rzepka, and Kenji Araki[1]

Abstract. In this paper, we propose implicit and explicit utterance generation models and a dialogue system in which such models are implemented. Modularization of classifiers enables the agent to annotate input utterance with tags of multiple features including types of sentences and mood expressions. In the implicit model, the features extracted from the input sentences define an agent's internal state. A relativity vector to each domain is sustainably computed based on similarity in Japanese WordNet ontology and the system's internal state. The explicit answers are generated if the input text is classified as Question-Answering domain based on the tags given by classifier modules. In other cases of classification, the system generates open-domain utterances. We will discuss the result of experiments intended to show characteristics of both domain detection methods.

1 Introduction

Task-oriented interfaces for mobile devices is widely accepted as information providing agents. However, a preliminary survey showed that in several domains of utterance users have good command if an utterance is recognized not only in an 'explicit' manner but also 'implicit'. This result has motivated us to reconsider what the true design of reflexive agent means in the research area of Artificial General Intelligence (AGI).

As an introduction, this section will provide a view on the current conceptual structure of a dialogue recognition and utterance recognition methods widely used across Natural Language Processing and Human Language Technology areas.

A preliminary survey conducted on the web[2] in Japan showed that some internet users do not accept the intention of being explicit in several domains a part of daily conversations belong. This result shows that humans do not, or cannot converse making their demands or need for information only explicit in some dialogue domains.

In the first half of this paper, we review current development in utterance recognition, and propose a model which explains an important and feasible portion of present capabilities. In the latter part of the paper, we will discuss the results of a dialogue system that handles the both non-task-oriented and task-oriented utterances, based on the proposed model.

2 Current Defining of Key Concepts

Here we revise concepts developed in research on dialogic interaction, in both the linguistic observation of conventional human-human dialogues and human-computer interaction.

2.1 Models in Human-Computer Interaction: Task-Oriented vs. Non-Task-Oriented

Various research has been conducted on methods to simulate the capability to generate utterances and perform a dialogue with human, i.e. to make correct responses which satisfy intention of human user, through generated utterances. A dialogue consists of several pairs of utterances, traditionally considered as performed alternately between the two participants of a dialogue.

In such attempts, many researchers have proposed a model that distinguishes dialogues or utterances which intend to complete a specified task shared by dialogue participants. Many have utilized a task-oriented dialogue model [1]. Task-oriented domain has a relevance to knowledge based on the Kintsch and van Dijk model [14]. In the view of this model, an utterance which belongs to a task-oriented domain has also an orientation within a domain of conversational topics (topic domain).

Many systems that show some performance in resolving tasks that we can regard as relevant to a specific domain, we can also be regarded as restricting the coverage of the domain of a dialogue performed between a user and the system. This is because any utterance such a system classifies out of the specified topic domain(s) is rejected telling the user that the utterance is unrecognizable or irrelevant to the desired task.

If dialogue system design involves this classification, dialogues or utterances that the system does not classify in any task domain are marked as non-task-oriented. Non-task-oriented utterances form a domain equal to a set of utterances with the ones from the task-oriented domain excluded.

Directly following this observation, we considered a dialogue system as a state machine using non-task-oriented and task-oriented states in our previous work [16]. The system proposed in this work initiates dialogue in the non-task-oriented state where chat modules work to produce a non-task-oriented dialogue. The system is given several task domains which need to satisfy the condition of relevance, which we discuss and compare later, to generate an utterance based on the domain. The relevance condition is a threshold of similarity between the input utterance and keywords specific to the task domain.

2.2 Relevance Theory and Human-Computer Interaction

According to the Relevance Theory [15] (RT), intentions that we infer from a dialogue include the intention of the transmitter to show

[1] Graduate School of Information Science and Technology, Hokkaido University, email: {my,kabura,araki}@media.eng.hokudai.ac.jp

[2] Conducted on http://q.hatena.ne.jp/1338376413, directed to internet users older than 20. 130 people took the survey.

some information about a fact, and the intention of the same person to share the information with the hearer. The theory rejects observation in which syntactical representation of an utterance involves its meaning, but accepts the assertion that the communicator intends to elicit among the participants two intentions: that there is information to be elicited (as an informative intention; which we refer as **II**), and the fact that he/she has such an intention (as a communicative intention; **CI**). Each participant of the dialogue preforms ostensive-inferential communication holding the two kinds of intentions.

Referring to the observation of RT, the hearer makes a deduction of such two intentions, from an utterance in any domain. The present domain shared by participants limits the resulting meaning of deduction made after the ostensive-inferential communication.

To help make an II into mutualized knowledge, which can fail due to noise interference and knowledge deficiency of the hearer, the communicator may generate stimulus which consists of encoded CI. The hearer notices the mediacy of the II by the communicator by decoding the message.

2.3 Application of Both Models to Dialogue Agent

In the field of Natural Language Processing, we can estimate that presumption of information by decoding the encoded CI is more precise than deduction from the surface of the utterance. The hearer's lack of background knowledge and any ambiguity in surface features of utterance can lead to a failure of deduction.

Many ways to show informative or communicative intention in everyday conversation people use acoustic methods are usually limited to textual input to each application software. Therefore, if using a natural language as medium, text usually spoken dialogue is a rather restrictive method.

There are several works about discourse analysis [10] [12] from the point of view of computational linguistics focusing on the Relevance Theory. Stone [12] claims that intention is a mental representation with a complex structure. In this paper, we rethink his position and define intention as the main objective of the system given by the developer of the system.

One feasible example of limiting the requirement of the source intention, which comes from the mind of the participant, is limiting the intention into not completely humanlike motives that occur to sustain one's life, but an intention for the agent to only serve the user. We can consider that this limitation is successful for the following reason. The system should have an initial primitive motivator (agenda) [3] that always motivates the agent to act upon the environment. The primitive motivation in this case to filter and classify user's input to known domains grippes the system's attention, so that the system assists user's decision to select a task and let the system do the needed task. This classification is achievable in a cognition model [4], which is thus compatible with a design that separates and modularizes filtering and reactive functions to the perceptual input.

In this model, the primitive motivator perpetually motivates and fires the attention of the dialogue agent. However, the motivation must have an essential account of the agent's capabilities (tasks broadly explained) and their objectives.

In this paper spoken dialogue interface which can run on electrical devices like smartphones. The problems the users are likely to have are that they cannot make full use of functionality of the device due to lack of procedural knowledge of manipulation, which can form a strong aspiration for an interface which is more easy-to-use.

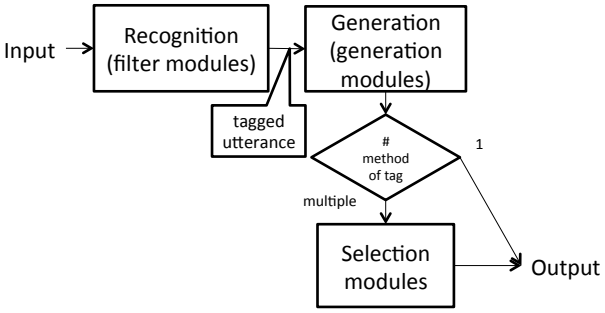

Figure 1. Schematic representation of the system.

3 Definition of Terms

In this paper, we are using a vocabulary that points to the current consensus of keywords which describe participant's behavior during a dialogue:

- Domain decision
  Attaching a domain tag to an utterance or an entire dialogue with methods discussed below.
- Open-domain (*or* non-task-oriented domain)
  A domain that does not belong to any specified domain and does not have a textual label.
- Task-oriented domain
  A domain which relates to a specific task that is expected to be achieved.

4 Methodology

In the previous sections, we have proposed an ostensive-informational model of utterance recognition and generation, which we consider to also be suitable for explaining the language division of the human cognitive system. In our view, it is possible with an all-by-simulation approach to investigate the model in question, where humans classify received utterances into non-task-domain and task-domains related to the knowledge and initial intention of the system. Fig. 1 shows a schematic representation of the system.

4.1 Implementation of Dialogue System

We have created an experimental dialogue agent. The system breaks down into two capabilities: capabilities of recognition and generation of utterance, each separated, obtaining a modular design with which the system is constructed from submodules. Functionality of each part contains a number of submodules, namely each tagger and utterance generation functionality is modularized separately.

4.2 Utterance Recognition

In this part of the system (filtering part), modules annotate its own tag to the text input. Recognizing an utterance means to annotate a tag to its text input, and tags annotated on an utterance represents a domain. In this paper, we apple a single domain tag which relates to the question answering task. Tags annotated in the filter module influence utterance generation module.

A filter module may annotate multiple tags. At the end of utterance processing, the sum of scores for all the tags is calculated and ranked.

The tag which obtains the highest score is selected and the system chooses an utterance generation method based on the selected tag.

We have also considered of another design in which a combination of tags works as data to decide the final domain of the utterance, where the machine-learning engine of the system uses the combination as learning data.

Tags, domains and functionality of those modules are listed in Table 1.

Table 1. The tags and generation (G), filter (F) modules used in the system.

Type	Name	Tag	Function/Generated Utterance
G	*Maru*	NT	Open-domain ut. based on N-gram [13]
G	*Moda*	NT	Open-domain ut. with modal expression [5]
G	*Eliza*	NT	Open-domain ut. based on scenario [8]
G	*QAC4*	QA	Answer to open-domain question [7]
G	*Recom*	RU	Domain selection inquiry
G	*Task* n	Dn	Task-oriented response of keyword a_n
Type	Score	Tag	Function
F	1.5	NT	Choose a non-task-oriented utterance
F	2.0	QA	Detect question-answering utterance
F	2.0	RU,Dn	Find a task relevance of user's utterances

Table 2. The utterance selection modules used in the system.

Type	Name	Scoring method
S	*ChatLog*	Maximum reciprocal edit distance among sentences in IRC chatlog
S	*ChatLogAbst*	*ChatLog* using POS-abstracted chatlog
S	*NGramHitnum*	Relevance of words based on frequency of coocurrence in the Web

We can name a tag with a string of ASCII characters: the name does not depend on a given structure. Though the diversity of classification of utterance types is essentially broader than the single task, in this paper we will discuss the effect of a single utterance-type classification.

5 Functionality

Here we describe the range of utterances produced by sub-modules in the generation module.

5.1 Question Answering

The task requires [7] an agent to detect a question type from utterances intended to be in other domains, and retrieve information needed to answer the question, which needs a filter's support based on WWW knowledge. Though we can classify target of this system module into a simple open-domain utterances, this method of implementation is suitable for testing the general dialogue capability (response to utterances with an explicit communication intention).

5.2 Generation of Open-Domain Utterance

The system currently uses one of three methods to generate an open-domain utterance we above listed above as type 'G' in Table 1. When creating an open-domain response, one of of the outputs of the 3 methods is selected by selection modules in Table 2.

5.3 Domain Selection Inquiry

As we showed using a survey result in **??**, the communicative or informative intention should be communicated with a form of implication. Moreover, a user's goals are included in both non-task-oriented and task-oriented domains. We vectorize averages of similarity expressed with a distance and a graphed thesaurus (Japanese WordNet [2]) between content morphemes and keywords. The filter module uses the similarity vector to select the domain keyword set most relevant to the current history of dialogue from the vector norm. Table 8 shows the sampled dialogues between the system and the user, which mainly consist of QA utterances.

5.3.1 Target Domains and Keyword Set

We chose 5 target domains, listed below. A keyword is a Japanese general noun that is bound to a domain and a task-oriented utterance generation module. Keywords are treated in a set and can express a centroid of meanings in multiple words.

5.3.2 Aim Vector $\boldsymbol{a}$

In the aim vector $\boldsymbol{a} = (a_1, a_2.a_3, ...)$ the current *aim* of the user's dialogue is calculated. Here, each element $a_1, a_2, ...$ represents average similarity between all of the content morphemes from the user's utterance and a keyword which exists as a concept in Japanese WordNet.

5.3.3 Similarity in WordNet

$\boldsymbol{a}$ is the cumulated sum of $\boldsymbol{\Delta a}$, which represents semantic similarity measured by Leacock-Chodorow [9]:

$$\Delta a_i = \frac{1}{N_U N_{K_i}} \sum_{u \in U} \sum_{k \in K_i} sim(u, k) \tag{1}$$

$$where \;\; sim(c_1, c_2) = max\left(-\log \frac{N_p}{2D}\right), \tag{2}$$

and N_U stands for the number of content morphemes in user utterance, N_{K_i} number of words in domain keyword set a_i , N_p the graph distance between c_1, c_2 in an ontology, with D as taxonomic depth in the ontology.

5.3.4 Inquiry Utterance Generation

The system acquires a norm of aim vector $\boldsymbol{a}$, $\|\boldsymbol{a}\| = \sqrt{\sum_k a_i^2}$. In a dialogue turn when $\|\boldsymbol{a}\|$ exceeds the threshold T, the system understands the user's interest is high in a task domain enough to receive a recommendation utterance. We chose a value of $T = 2.0$ in the experiment discussed later. A filter module outputs a tag 'RU' with score 3.0. The utterance generation module generates an utterance which helps the user decide the task, using k_i as the most relevant keyword which has the maximum value in $\boldsymbol{a}$.

6 Evaluation Experiments

Here we mention the experiments we performed for getting overall ratings by the users, precision of the filtering part to select an utterance generation method, and evaluation of implicit intention detection that helps user's task selection.

6.1 Questionnaire Evaluation Results

We chose Question-Answering agent [7] implemented as a dialogue system, and an ELIZA-type dialogue system [8] as baselines for this measure.

6.1.1 Questionnaire Survey

10 participants (9 male, 1 female) were requested to perform a dialogue with the system which lasts more than 20 turns. The system to evaluate was implemented as a CGI web application[3]. A questionnaire with an answering form was displayted after 20 turns elapsed, in which the participants were asked to evaluate the system using 5-point scales from 1 (I disagree) to 5 (I agree) towards these 6 statements:
(A) I would like to continue the dialogue.
(B) The dialogue is natural in grammar.
(C) The dialogue is natural in its sense.
(D) The system's vocabulary is rich.
(E) The system talked like a human.
(F) The system's recommendation had a strong relevance with my concern and interests.

Table 3. Overall ratings in 5-point scales.

Item	average of ratings
A	1.75
B	1.75
C	1.63
D	2.00
E	2.13
F	1.75
Mean	1.85
Baseline	2.54

6.1.2 Analysis of Impression Using Semantic Differential Method

We conducted an evaluation experiment intended to investigate the orientation of participants' impression toward the experimental system. The same group of participants in 6.1.1 evaluated the system using 35 pairs of adjectives in a 5-point linear scale, which are frequently used [6] in semantic differential method [11] and represent positiveness of subjective impression a participant has held to the system. The adjective pairs are listed in Table 6.

6.2 Measurement of Appropriateness of Utterance Method Selection

To measure the effectiveness of the explicit response generation, we asked another group of participants to score how precisely the system selected an utterance method. Participants choose from one of 4 grades to rank each utterance in a dialogue log (which had 150 utterances combined from the logs taken during the first experiment) the system made as response. Participants were asked to select one value from 0∼3, 3 for acceptable or appropriate utterances, 2 for grammatically correct but unacceptable, 1 for unrecognizable ones, and 0 for error outputs.

[3] http://arakilab.media.eng.hokudai.ac.jp/~my/experiment.html (in Japanese)

6.2.1 Participants

Participants were 4 male graduate students with ages of around 24. The target of this evaluation is the actual dialogue made by the participants in 6.1.1 and 6.1.2. This evaluation was also performed online.

6.2.2 Result

Table 4 shows the result of the evaluation. The mean of all scores was 2.274, with standard deviation of 0.847.

Table 4. Distribution of evaluation scores of the precision experiment.

3	2	1	0 (*errors*)
302	176	104	17
50.3%	29.3%	17.3%	2.83%

7 Discussion

The average score obtained from the experiment was by 10 and fewer participants, a number which the authors do not consider as statistically sufficient for a complete judgement of this system's potential. However, in this paper we use these results to illustrate the system's characteristics.

7.1 User's Impression and Precision of Utterance Selection

The average results of precision of utterance generation shown in Table 5 for each tag exhibit higher precision in utterance generation with tags correctly annotated.

Table 5. Utterance method selection appropriateness of each tag. (* indicates that 1 participant evaluated the item)

Tag	Average value
NT	2.32
QA	2.26
The system	2.28
Question-Answering only*	1.85
Eliza-type system only*	2.45

7.2 Evaluation of Domain Selection

The partial score of the overall rating (F) involves appropriateness of domain selection inquiries made by the module. Looking at the data, we found that when there was no selection of utterance domain in the dialogue, the value decreased.

8 Analysis of SD Evaluation Scores

We conducted factor analysis on the data obtained from the experiment (6.1.2) using GNU R environment[4]. The number of factors was 4, with their cumulated contribution rate 0.700. The factors gained from this are shown in Table 7. Further, Fig. 8 shows a dendrogram resulting from the application of cluster analysis in the furthest neighbor method to the acquired data, regarding it as a vector with 35 dimensions. When cut at score 6.0, four general clusters appear to divide the vector data.

Table 6. Adjective pairs of with contrast of positiveness. (Positive adjectives in the right column.)

1	dark	bright
2	cold	warm
3	weak	strong
4	dismal	cheerful
5	light	heavy
6	hate	love
7	hard	soft
8	passive	aggressive
9	noisy	quiet
10	inactive	active
11	bad	good
12	unkind	kind
13	violent	peaceful
14	painful	fun
15	sober	flashy
16	boring	interesting
17	dull	sharp
18	bad feeling	good feeling
19	unreliable	reliable
20	feeble	robust
21	small	large
22	flippant	serious
23	slow	fast
24	unpleasant	pleasant
25	unstable	stable
26	taciturn	talkative
27	dirty	clean
28	sloppy	neat
29	simple	complex
30	static	dynamic
31	stubborn	frank
32	irresponsible	responsible
33	sad	happy

Table 7. Result of factor analysis applied to the SD method evaluation.

Cumulated CR	Adjective pairs ID	+/-
0.231	6,9,17,19,22,24,28,31	+
0.444	2,10,13,20,26,30	-
0.580	11,12,23,32	-
0.700	7,33	+

Thus we extracted 4 factors from the factor analysis. Among these factors, the third group including "bad-good", "unkind-kind", "slow-fast", and "irresponsible-responsible" also appear in the same the dendrogram of cluster analysis. These observations represent a factor of functionality of the system. This factor explains the cause of lower

[4] http://www.r-project.org/

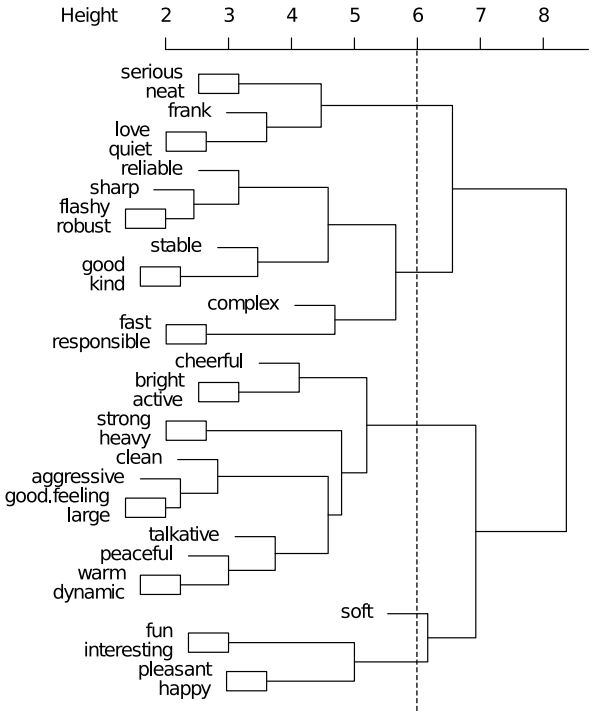

Figure 2. Dendrogram gained by cluster analysis on the SD data.

score of participant's overall ratings of impression compared to the baseline systems, as being the working speed of the system. The average of time the system took to give a response in all the dialogue made in the experiments was 33.1 seconds, against 0.163 seconds by the baseline (ELIZA-type) system.

9 Conclusion

In this paper, we discussed a method for a general-purpose modularized dialogue agent to make a suggestion based on implicit comprehension from the user's content words, as well as utterance method selection in response to explicit requests. We showed that the system that has limited primitive motivator could perform a general-purpose dialogue with minimum difference in human-rating score compared to non-modular baseline systems.

The domain keywords which were used for domain selection inquiry were limited in number and were chosen in an arbitrary manner, which resulted in high rates of incorrect domain decisions. The system should decide a domain dynamically and recommendation of task must be done with general knowledge. Combining Web-based knowledge and commonsense reasoning would suffice for the system to approach the user's need, which may need more computation resources to perform faster.

Finally, as the results of the questionnaire (see 6.1.1) suggest we can view the system's user-friendliness and grammatical, contextual acceptance as substandard (< 2.00) and yet to be improved. In order to make the utterance method selection more precise and appropriate, we are designing a method, with each filter module annotating multiple tags used as cues to form a decision process in general domains.

Table 8. Dialogue example. *User*, *QAC4*, *Recom*, *Maru*, and *Eliza* are module names shown in Table 1.

Speaker	Turn	Surface of utterance (translation by author)
User	1	Nihon no shusho ha dare desuka (Who is the prime minister of Japan?)
QAC4	1	Noda Yoshihiko datta (Was Yoshihiko Noda)
User	2	Amerika no daitoryo ha dare desuka (Who is the president of America?)
QAC4	2	Obama rashii desu kara ne (They say Obama.)
User	3	Kyo ha shokudo de yushoku o taberu yotei desu (Today I am planning to eat supper at the cafeteria)
Recom	3	Moshikashite, keno, byoki ni kyomi arimasu ka? (Perhaps, you are interested in health and diseases?)
User	1	Konnichiha (Hello)
Recom	1	Moshikashite, kenko, byoki ni kyomi arimasu ka? (Perhaps, you are interested in health and diseases?)
User	2	Iie (No)
Maru	2	Eto, gomen nasai, yoku wakaranai desu. (Sorry, but I don't understand.)
User	3	Kenko nanka kyomi arisen (I am not interested in health.)
(Error)	3	Gomen nasai, chotto era desu. (Sorry, but an error occurred.)
User	4	Moshika shite sotchi no yoga byoki nanja nai no? (Perhaps it's you who is ill?)
f QAC4	4	Kyo ja naiyo (It isn't today.)
User	5	Bokokugo ha nan desu ka? (What is your mother tongue?)
QAC4	5	Nihon desu kane (I guess it's Japan)
User	6	Nihon de umareta tte koto? (You mean you were from Japan?)
Eliza	6	Nihon ga ... tte koto kana? (You mean Japan :. ?)
User	7	Maresia itta koto aru? (Have you ever been to Malaysia?)
Eliza	7	Tashikani ... yone (Indeed so ...)
User	8	Sapporo de ichiban oishii ramen-yasan ha doko desuka (What is the best ramen shop in Sapporo?)
QAC4	8	Hokkaido desu kane (I guess it's Hokkaido)
User	9	Seibetsu ha nan desuka? (What is your gender?)
QAC4	9	.....Tokyo dana to (I thought it was Tokyo)
User	10	Ore to kekkon shite kudasai (Will you marry me?)
Eliza	10	...yo ni nareru kana? (Can I become ... ?)
User	11	Ima ikutsu desuka? (How old are you now?)
QAC4	11	.....Matsuda Seiko datta (It was Seiko Matsuda)

REFERENCES

[1] N.O. Bernsen, L. Dybkjær, and H. Dybkjær, 'A dedicated task-oriented dialogue theory in support of spoken language dialogue systems design', in *Third International Conference on Spoken Language Processing*, (1994).

[2] B. Francis, I. Hitoshi, F. Sanae, U. Kiyotaka, K. Takayuki, and K. Kyoko, 'Enhancing the japanese wordnet', in *Proceedings of the 7th Workshop on Asian Language Resources, ACL-IJCNLP*, pp. 1–8, (2009).

[3] S. Franklin and U. Ramamurthy, 'Motivations, values and emotions: 3 sides of the same coin', in *Proceedings of the Sixth International Workshop on Epigenetic Robotics, Paris, France, September 2006, Lund University Cognitive Studies*, number 128, pp. 41–48, (2006).

[4] B. Goertzel and P. Wang, 'A foundational architecture for artificial general intelligence', *Advances in Artificial General Intelligence: Concepts, Architectures and Algorithms*, 6, 36, (2007).

[5] S. Higuchi, R. Rzepka, and K. Araki, 'A casual conversation system using modality and word associations retrieved from the web', *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, (October 2008).

[6] M. Inoue and T. Kobayashi, 'The research domain and scale construction of adjective-pairs in a semantic differential method in japan', *The Japanese journal of educational psychology*, 33(3), 253–260, (1985).

[7] K. Kameyama, K. Araki, and Y. Kimura, 'Effectiveness of automatic acquisition of knowledge-source selection-rules for a question answering system', *IPSJ SIG Technical Report*, 2007(35), 85–90, (2007).

[8] Y. Kimura, K. Araki, Y. Momouchi, and K. Tochinai, 'Spoken dialogue processing method using inductive learning with genetic algorithm', *IEICE TRANSACTIONS on Information and Systems*, 84, 2079–2091, (2001).

[9] C. Leacock and M. Chodorow, 'Combining Local Context and WordNet Similarity for Word Sense Identification', *An Electronic Lexical Database*, 265–283, (1998).

[10] S. Minewaki, K. Shimada, and T. Endo, 'Interpretation of utterances based on relevance theory: Toward the formalization of implicature with the maximum relevance', *Proc. of PACLING2005*, 214–222, (2005).

[11] C.E. Osgood, G.J. Suci, and P.H. Tannenbaum, *The measurement of meaning*, volume 47, Univ of Illinois Pr, 1957.

[12] M. Stone, J.C. Trueswell, and M.K. Tanenhaus, 'Communicative intentions and conversational processes in human-human and humancomputer dialogue', *Approaches to studying world-situated language use*, 39–70, (2005).

[13] M. Takahashi, R. Rzepka, and K. Araki, 'A performance evaluation of sentence generation method using web search and word n-gram models', *Proceedings of NLP2010*, (March 2010).

[14] T.A. Van Dijk, 'On macrostructures, mental models, and other inventions: A brief personal history of the kintsch-van dijk theory', *Discourse comprehension: Essays in honor of Walter Kintsch*, 383–407, (1995).

[15] D. Wilson and D. Sperber, *Relevance theory*, Wiley Online Library, 1992.

[16] M. Yatsu, R. Rzepka, and K. Araki, 'A modules-based, task-navigational dialogue system', in *Proceedings of the Pacific Association For Computational Linguistics 2011*, (2011).

Developments in Context-sensitive Affect Detection in an Intelligent Agent

Li Zhang[1]

Abstract. Affect interpretation from multithreaded online conversations is a challenging task. Understanding context and identifying target audiences are very crucial for the appropriate interpretation of emotions implied in an individual input embedded in such online social interactions. In this paper, we discuss how context is used to interpret affect implied in conversational inputs with weak or no affect indicators embedded in multithreaded social interactions. Topic theme detection using latent semantic analysis has been applied to such inputs to identify their discussion themes and potential target audiences. Relationships between characters have also been taken into account for affect analysis. Such semantic interpretation of the dialogue context also shows great potential in the recognition of metaphorical phenomena and the development of a personalized intelligent tutor for drama improvisation.

1 INTRODUCTION

Human behaviour in social interaction has been intensively studied. Intelligent agents are used as an effective channel to validate such studies. For example, mimicry agents are built to employ mimicry social behaviour to improve human agent communication [1]. Intelligent conversational agents are also equipped to conduct personalised Turing and generate small talk behaviours to enhance users' experience. However, the Turing test [2] introduced in 1950 still poses big challenges to our intelligent agent development. Especially, the proposed question, "can machines think?", makes many of our developments shallow.

We believe it will make intelligent agents possess human-like behaviour and narrow the communicative gap between machines and human-beings if they are equipped to interpret human emotions during the interaction. Thus in our research, we equip our AI agent with emotion and social intelligence as the potential attempts to answer the above Turing question. According to Kappas [3], human emotions are psychological constructs with notoriously noisy, murky, and fuzzy boundaries that are compounded with contextual influences in experience and expression and individual differences. These natural features of emotion also make it difficult for a single modal recognition, such as via acoustic-prosodic features of speech or facial expressions. Since human being's reasoning process has taken related context into consideration, in our research, we intend to make our agent take multi-channels of subtle emotional expressions embedded in social interaction contexts into consideration to draw reliable affect interpretation. The research presented here focuses on the production of intelligent agents with the abilities of interpreting dialogue contexts semantically to support affect detection as the first step of building a 'thinking' machine.

Our research is conducted within a previously developed online multi-user role-play virtual drama framework, which allows school children aged 14 – 16 to talk about emotionally difficult issues and perform drama performance training. In this platform young people could interact online in a 3D virtual drama stage with others under the guidance of a human director. In one session, up to five virtual characters are controlled on a virtual stage by human users ("actors"), with characters' (textual) "speeches" typed by the actors operating the characters. The actors are given a loose scenario around which to improvise, but are at liberty to be creative. An intelligent agent is also involved in improvisation. It included an affect detection component, which detected affect from human characters' each individual turn-taking input (an input contributed by an individual character at one time). This previous affect detection component was able to detect 15 emotions including basic and complex emotions and value judgments, but the detection processing has not taken any context into consideration. The intelligent agent made attempts to produce appropriate responses to help stimulate the improvisation based on the detected affect. The detected emotions are also used to drive the animations of the avatars so that they react bodily in ways that is consistent with the affect that they are expressing [4]. An example of the system interface is shown in Figure 1.

Figure 1. An example user interface with three human characters (the first three characters counting from the left hand side) and one AI actor (the last character)

Moreover, the previous affect detection processing was mainly based on pattern-matching rules that looked for simple grammatical patterns or templates partially involving specific

[1]School of Computing, Engineering and Information Sciences, University of Northumbria, UK. Email: {li.zhang}@northumbria.ac.uk

words or sets of specific alternative words. A rule-based Java framework called Jess was used to implement the pattern/template-matching rules in the AI agent allowing the system to cope with more general wording and ungrammatical fragmented sentences. From the analysis of the previously collected transcripts, the original affect interpretation based on the analysis of individual turn-taking input itself without any contextual inference proved to be effective enough for those inputs containing strong clear emotional indicators such as 'yes/no', 'haha', 'thanks' etc. There are also situations that users' inputs do not have any obvious emotional indicators or contain very weak affect signals, thus contextual inference is needed to further derive the affect conveyed in such user inputs.

We have conducted context-based affect detection using emotion modelling in personal contexts using Markov chains previously with the support of contextual linguistic features. Since emotions and concepts can be expressed in multiple ways, linguistic features for contextual communication are sometimes not reliable enough. Also the previous approach was not capable enough to deal with sudden topic changes, which led to affect detection errors. Comparing with the previous approach, the new developments go beyond linguistic features and employ latent semantic analysis to reveal the underlying semantic structures embedded in the improvisational inputs to identify the discussion themes and target audiences and thus to inform affect interpretation. For example, the inspection of the collected transcripts indicates that the improvisational dialogues are often multi-threaded. This refers to the situation that social conversational responses of different discussion themes to previous several speakers are mixed up due to the nature of the online chat setting. Therefore the detection of the most related discussion theme context using semantic analysis is very crucial for the accurate interpretation of the emotions implied in those inputs with ambiguous target audiences and weak affect indicators.

A neural network implementation is then used to perform affect detection in social interaction contexts with the consideration of interpersonal relationships between speakers and the target audiences, emotions implied by target audiences and sentence types of the current inputs. The new approach proved to be robust enough to deal with emotions expressed during sudden topic changes and creative improvisation. The semantic-based analysis also shows great potential to extend the application to normal daily life situations outside of the limitations of any chosen scenarios.

2 RELATED WORK

Tremendous progress in emotion recognition has been witnessed by the last decade. Endrass, Rehm and André [5] carried out study on the culture-related differences in the domain of small talk behaviour. Their agents were equipped with the capabilities of generating culture specific dialogues. There is much other work in a similar vein. Recently textual affect sensing has also drawn researchers' attention. Neviarouskaya et al. [6] provided a sentence-level textual affect sensing system, called @AM, to recognize judgments, appreciation and different affective states. They adopted a rule-based domain-independent approach with semantic analysis of verbs. Although some linguistic contexts introduced by conjunctions such as 'but' were considered, the detection task setup was still limited to the analysis of individual input. Ptaszynski et al. [7] employed context-sensitive affect detection with the integration of a web-mining technique to detect affect from users' input and verify the contextual appropriateness of the detected emotions. The detected results made an AI agent either sympathize with the player or disapprove the user's expression by the provision of persuasion. However, their system targeted interaction only between an AI agent and one human user in non-role-playing situations, which greatly reduced the complexity of the modelling of the interaction context.

Scherer [8] explored a boarder category of affect concepts including emotion, mood, attitudes, personality traits and interpersonal stances (affective stance showed in a specific interaction). Mower et al. [9] argued that it was very unlikely that each spoken utterance during natural human robot/computer interaction contained clear emotional content. Thus, dialog modeling techniques, such as emotional interpolation, emotional profiling, and utterance-level hard labelling, have been developed in their work to interpret these emotionally ambiguous or non-prototypical utterances. Such development would benefit classification of emotions expressed within the context of a dialog. Batliner et al. [10] focused on the modeling of the frequently used seven emotional states in their study such as reprimanding, motherese, angry etc, into two dimensions: Valence and Interaction. They stated that "typical emotions are to a large extent rather private". Such emotions may not be observed very often in public settings. Their research thus focused on social interaction modeling dimension.

Moreover, naturalistic emotion expressions usually consist of a complex and continuously changed symphony of multimodal expressions, rather than rarely unimodal expressions. However, most existing systems consider these expressions in isolation. This limitation may cause inaccuracy or even lead to a contrary result in practice. For instance, currently many systems can accurately recognize smile from facial expressions, but it is inappropriate to conclude a smiling user is really happy [3]. In fact, the same expression can be interpreted completely differently depending on the context that is given [11]. It also motivates us to use semantic interpretation of social contexts to inform affect detection in our application.

3 SEMANTIC-BASED TOPIC THEME DETECTION

In our previous study, we noticed that the language used in our application domain is often complex, idiosyncratic and invariably ungrammatical. It contains abbreviations and borrows heavily from the language of chat-rooms. Compared to the language normally analysed in computational linguistics it provides significant additional challenges. We also implemented pre-processing components previously to deal with mis-spellings, abbreviations, etc.

Most importantly, the language also contains a large number of weak cues to the affect that is being expressed. These cues may be contradictory or they may work together to enable a stronger interpretation of the affective state. In order to build a reliable and robust analyser of affect it is necessary to undertake several diverse forms of analysis and to enable these to work together to build stronger interpretations. It thus guides not only our previous research but also our current developments. For example, in our previous work, we undertook several analyses of

any given utterance. These would each build representations which may be used by other components (e.g. syntactic structure) and would construct (possibly weak) hypotheses about the affective state conveyed in the input. Previously we adopted rule-based reasoning, robust parsing, pattern matching, semantic and sentimental profiles for affect detection analysis. In our current study, we also integrate contextual information to further derive the affect embedded in the interaction context and to provide affect interpretation for those without strong affect indicators.

In order to detect affect accurately from the improvisational input without strong affect indicators and clear target audiences, we employ the semantic meaning of the social interaction context to inform the affect detection processing. In this section, we discuss our approaches of using latent semantic analysis (LSA) [12] and its related packages for terms and documents comparison to recover the most related discussion themes and potential target audiences to benefit affect detection.

In our previous rule-based driven affect detection implementation, we mainly relied on keywords and partial phrases matching with simple semantic analysis using WordNet etc. However, we notice many terms, concepts and emotional expressions can be described in various ways. Especially if the inputs contain no strong affect indicators, other approaches focusing on underlying semantic structure in the data should be considered. Thus latent semantic analysis is employed to calculate semantic similarities between sentences to derive discussion themes for such inputs.

Latent semantic analysis generally identifies relationships between a set of documents and the terms they contain by producing a set of concepts related to the documents and terms. In order to compare the meanings or concepts behind the words, LSA maps both words and documents into a 'concept' space and performs comparison in this space.

In detail, LSA assumes that there is some underlying latent semantic structure in the data which is partially obscured by the randomness of the word choice. This random choice of words also introduces noise into the word-concept relationship. LSA aims to find the smallest set of concepts that spans all the documents. It uses a statistical technique, called singular value decomposition, to estimate the hidden concept space and to remove the noise. This concept space associates syntactically different but semantically similar terms and documents. We use these transformed terms and documents in the concept space for retrieval rather than the original terms and documents.

In our work, we employ the semantic vectors package [13] to perform LSA, analyze underlying relationships between documents and calculate their similarities. This package provides APIs for concept space creation. It applies concept mapping algorithms to term-document matrices using Apache Lucene, a high-performance, full-featured text search engine library implemented in Java [13]. We integrate this package with our intelligent agent's affect detection component to calculate the semantic similarities between improvisational inputs without strong affect signals and training documents with clear discussion themes. In this paper, we target the transcripts of the Crohn's disease scenario for context-based affect analysis. In this scenario, it is mainly about Peter who has had Crohn's disease since the age of 15. Crohn's disease attacks the wall of the intestines and makes it very difficult to digest food properly. Peter has the option to undergo surgery (ileostomy) which will have a major impact on his life. The task of the role-play is to discuss the pros and cons with friends and family and decide whether he should have the operation. The other characters are: Janet (Mum) who wants Peter to have the operation, Matthew (older brother) who is against the operation, Arnold (Dad) who is not able to face the situation, and David (the best friend) who mediates the discussion.

In order to compare the improvisational inputs with documents belonging to different topic categories, we have to collect some sample documents with strong topic themes. Personal articles from the Experience project (www.experienceproject.com) are used for this purpose. These articles belong to 12 discussion categories including Education, Family & Friends, Health & Wellness, Lifestyle & Style, Pets & Animals etc. Since we intend to perform discussion theme detection for the transcripts of the Crohn's disease scenario, we have extracted sample articles close enough to the above scenario including articles of Crohn's disease (five articles), school bullying (five articles), family care for sick children (five articles) and food choice (three articles). Phrase and sentence level expressions implying 'disagreement' and 'suggestion' have also been gathered from the several other articles published on the Experience website. Thus we have training documents with six discussion themes including 'Crohn's disease', 'bullying', 'family care', 'food related', 'suggestions' and 'disagreement'. The first four themes are sensitive and crucial discussion topics to this scenario, while the last two themes are intended to capture arguments expressed in multiple ways. All the gathered sample documents of the above six categories have been put under one directory for further analysis.

We have taken one example interaction of the Crohn's disease scenario produced by testing subjects during our previous user testing in the following to demonstrate how we detect the discussion themes for those inputs with weak or no affect indicators and ambiguous target audiences.

1. Matthew: *I don't think* you should have the treatment peter. [disapproval]
2. Arnold: *lets not* talk about it now. Can we just get our food and discuss this later. [disapproval]
3. Peter: *why not* hav it matt [neutral]
4. Matthew: you will *get bullied* cuz of it. [sad]
5. Janet: nobody will bully u. *shut it* matt [angry]
6. Dave: *stop arguing* every1. [disapproval]
7. Matthew: dad, *stop talking* about food. [disapproval]
8. Janet: stand up to da bullies and *do not be* afraid. [disapproval]
9. Peter: mum, I don't think u really understand what it is. Its not like a carrier bag. I will use it when I go loo [*Target audience: Janet; disapparoval*]
10. Arnold: *excuse me*, *can we not*? Im eating [disapproval]
11. Janet: I know what it is, but it will still help you [*Target audience: Peter; disapproval*]
12. Matthew: how do you feel about having a bag attached to you? [*Target audience: Peter and Janet; angry*]

Since our previous affect detection focuses on affect interpretation from inputs with strong emotion signals, it provides an affect label for such inputs in the above example. The emotion indicators are also illustrated in italics in the above examples. The inputs without an affect label followed straightaway are those with weak or no strong affect indicators

(9th, 11th & 12th inputs). Therefore further processing is needed to recover their most related discussion themes and identify their most likely audiences in order to identify implied emotions more accurately. Our general idea for the detection of discussion themes is presented in Figure 2.

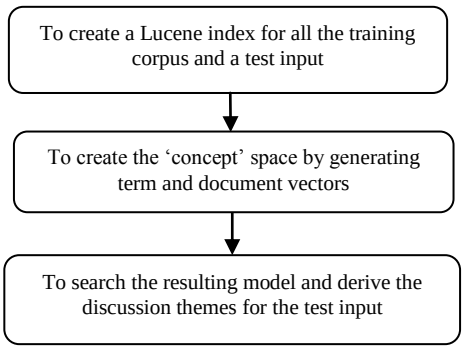

Figure 2. Discussion theme detection processing

We start with the 9th input from Peter to demonstrate the topic theme detection. First of all, this input is stored as a separate individual test file (test_corpus1.txt) under the same folder containing all the training sample documents of the six categories.

As shown in Figure 2, the corresponding semantic vector APIs are used to create a Lucene index for all the training samples and the test file. This generated index is also used to create term and document vectors, i.e. the concept space. Various search options could be used to test the generated concept model. In order to find out the most effective approach to extract the topic theme of the test inputs, we have made several experiments. First of all, we intend to provide rankings for all the training documents and the test sentence based on their semantic distances to a topic theme. We achieve this by searching for document vectors closest to the vector for a specific term (e.g. 'disease'). An example partial output is listed in Figure 3.

In the outputs shown in Figure 3, except for the test file: test_corpus1.txt, other listed files are all training corpus taken from the articles published on the Experience website. The double values shown in the first column are the semantic distance values between each document and the chosen topic theme ('disease'). We have intended to rank all the files based on their semantic closeness to the other five topic themes (such as 'bullying, 'disagreement' etc). But as we mentioned earlier, there are multiple ways to describe a topic theme such as 'disagreement' and 'suggestion'. It affects the file ranking results more or less if different terms indicating such themes are used. Experiments also show that even if we make the search respectively using the terms with the same root (such as 'bullying' and 'bullied'), the file ranking results are different for part of the files. Thus we still need to use other more effective search methods to accompany the following file ranking findings for a specific topic theme.

There is another search algorithm which can find terms most closely related to each document based on the concept space we built earlier. This algorithm has been applied to the training corpus to test its efficiency and findings. According to the first step processing to find the semantic distances to a topic term, we have the training document 'crohn4.txt' listed on the top of the ranking list. However, when this file is used to find the terms most closely related to it, we have the output presented in Figure 4. In this output, the most useful disease related theme term (such as 'disease') has not returned with a top ranking on the term list, but listed in the middle. Thus it indicates such most related term finding results to a file are not reliable enough for automatic processing. Another approach especially suitable to our application domain is to find the semantic similarity between documents. All the training sample documents are taken from articles under clear discussion themes within the 12 categories of the Experience project. The file titles used indicate the corresponding discussion themes. If the semantic distances between files, especially between training files and the test file, are calculated, then it provides another source of information for the discussion theme detection. Therefore we use the CompareTerms semantic vector API to find out semantic similarities between all the training corpus and the test document. Part of the example output is presented in Figure 5.

```
Found vector for 'disease'
Search output follows ...
0.8112944014737917:F:\test_data\test_lsa6\crohn4.txt
0.7954427523946758:F:\test_data\test_lsa6\crohn3.txt
0.6244268984149078:F:\test_data\test_lsa6\test_corpus1.txt
0.5990879215052521:F:\test_data\test_lsa6\family_care2.txt
0.5779413288260994:F:\test_data\test_lsa6\family_care5.txt
0.5200265398697822:F:\test_data\test_lsa6\crohn1.txt
0.5184941042931127:F:\test_data\test_lsa6\crohn5.txt
0.5094701704973981:F:\test_data\test_lsa6\crohn2.txt
0.4962319424068785:F:\test_data\test_lsa6\family_care3.txt
0.4606811844883898:F:\test_data\test_lsa6\family_care4.txt
0.4583252892724773:F:\test_data\test_lsa6\family_care1.txt
```

Figure 3. Example partial output for searching for document vectors closest to the vector for a topic theme, 'disease'.

```
Found vector for 'F:\test_data\test_lsa7\crohn4.txt'
Search output follows ...
0.9351587400907257:have
0.8887467992451593:else
0.8508699187263247:times
0.8069219978282258:hard
0.8027186536514817:told
0.79851066842271:think
0.7976727548113798:i
0.7947333669061222:living
0.7857429057396778:remember
0.7829770373672106:when
0.779169627545226:disease
0.777977648288496:my
0.7730323460332773:you
0.771804897514864:over
0.7682103498791507:labor
0.765358656841035:know
0.7614978983373525:rest
0.7574792496840674:understand
0.7386694834219586:own
0.73511471731328:lot
```

Figure 4. Example output for finding terms most closely related to a training document 'crohn4.txt'.

The similarity results listed in Figure 5 show there are two training files (crohn4.txt and disagree1.txt) semantically most similar to the test file (test_corpus1.txt containing the 9th input from Peter). These two training files respectively recommend the following two most related discussion themes: 'disease' and 'disagreement'. In the first step processing mentioned earlier, to find document vectors closest to that of a topic theme, the test sentence achieves the best ranking for the 'disease' topic theme and the second best ranking for the 'bullying' topic theme. With the integration of the semantic similarity results between document vectors, our processing concludes that the 9th input

from Peter relates most closely to negative topics 'disease', 'bullying' and 'disagreement' and most probably indicates 'disapproval' due to its closeness to the 'disagreement' theme.

```
similarity of "bullied1.txt" with "test_corpus1.txt": 0.5775331845539662
similarity of "bullied2.txt" with "test_corpus1.txt": 0.6457772244890972
similarity of "bullied3.txt" with "test_corpus1.txt": 0.6593339908493666
similarity of "crohn1.txt" with "test_corpus1.txt": 0.6039827766746925
similarity of "crohn2.txt" with "test_corpus1.txt": 0.6827992954523574
similarity of "crohn3.txt" with "test_corpus1.txt": 0.7054680619761999
similarity of "crohn4.txt" with "test_corpus1.txt": 0.8520039156044099
similarity of "disagree1.txt" with "test_corpus1.txt": 0.7601959945997482
similarity of "family_care1.txt" with "test_corpus1.txt": 0.574022120956489
```

Figure 5. Part of the output for the semantic similarities between training documents and the test file (the 9th input).

In a similar way, the conversation theme processing has identified the following two semantically most similar training documents (crohn2.txt and bullied3.txt) to the 11th input from Janet. These two training files respectively recommend the same two discussion themes: 'disease' and 'bullying' as those for the 9th input from Peter. The partial results are listed in Figure 6. The 11th input also achieves top four ranking for the enquiry of search for document vectors closest to the vector for 'disease'.

```
similarity of "bullied3.txt" with "test_corpus2.txt": 0.8311841420279548
similarity of "crohn1.txt" with "test_corpus2.txt":  0.6701409905059565
similarity of "crohn2.txt" with "test_corpus2.txt":  0.8875302851886464
similarity of "crohn3.txt" with "test_corpus2.txt":  0.8031293680512144
```

Figure 6. Part of the results showing the semantic similarities between training document vectors and the 11th input.

The 10th input from Arnold contains strong affect indicators (see italics). The previous affect detection algorithm labeled it with 'disapproval'. Since the 11th input did not mention clear target audiences, we have to recover the topic theme of the 10th input from Arnold. The result shows the most similar training document vectors to that of the 10th input are 'suggestion' and 'food related', which are different from the recovered topic themes of the 11th input. Therefore, as mentioned above, because of the similar discussion themes between the 9th and 11th inputs, it is assumed that Peter is the most likely target audience of the 11th input from Janet.

By searching for document vectors closest to the vector for the discussion theme 'disease', the last input (12th input) from Matthew shows high semantic closeness to this topic with a semantic distance score over 0.65 and a top four ranking. The similarity processing indicates that it is most similar to 'crohn4.txt' and 'bullied3.txt' in the semantic domain. Thus this input is a most likely further discussion aroused by the 9th and 11th inputs respectively contributed by Peter and Janet. Thus the most likely target audiences of the 12th input are Peter and Janet.

In our application domain, the conversation theme detection using semantic vectors analysis is able to help the AI agent to detect the most related discussion themes and therefore to identify the most likely target audiences. We believe these are very important aspects for the accurate interpretation of the emotion context. We also envisage the above processing would be really helpful to distinguish small talk (task un-related discussion) behaviours and task-driven talk during human agent/robot interaction. Thus it may enable the AI agent to respond more appropriately during the social interaction. In the following section, we discuss how cognitive cues such as relationships and emotion contexts of target audiences are used to inform context-based affect interpretation.

4 NEURAL NETWORK BASED AFFECT DETECTION

The cognitive emotion research of Hareli and Rafaeli [14] pointed out that "one person's emotion is a factor that can shape the behaviors, thoughts and emotions of other people". They also believed that "emotions may affect not only the person at whom the emotion was directed but also third parties who observe an agent's emotion". In our application domain, one character's manifestations of emotion can also thus influence others. Research of Wang, Lee and Marsella [15] also discussed that feedback of artificial listeners can be influenced by interpersonal relationships, personalities and culture related aspects. In our application, for example, if two characters share positive relationships and one of them experiences 'sad' emotion, then it is more likely the other character responses with an empathic response of 'sadness'. Otherwise if they have a negative relationship, then the other character is more inclined to show a gloating response of 'happiness'. Thus such interpersonal relationships (such as positive (friendly) or negative (hostile or tense) relationships) are also employed to advise the affect detection in the social context.

In the example interaction mentioned in section 3, our topic theme processing has identified the most likely audience of the 11th input from Janet is Peter. Especially in Peter's previous input (9th input), the family role 'mum' used also indicated Peter started the conversation with Janet in the first place. This 11th input is aroused by such social interaction. The above topic theme detection also noticed that in the 10th input, Arnold 'suggested' a topic change. Instead of following on the previous discussion theme, Arnold switched to the food related topic. This also shows potential less interest or indifference to the discussion of the previous 'disease' related topic suggested by Peter. Thus the 10th input may indicate a 'negative' emotion by avoiding or showing less interest in the previous discussion theme. The original version of the affect detection without any contextual analysis has also interpreted the 10th input showing 'disapproval'.

The topic theme detection also reveals that the 11th input from Janet is mainly related to negative topics such as 'disease' and 'bullying'. Therefore, the target audience of the 11th input from Janet is not Arnold but Peter. Peter showed 'disapproval' in the most recent input (the 9th input). This indicates the most related context to the 11th input is a 'negative' context. Moreover, in the original affect detection processing, we used a syntactic parser, Rasp, to obtain sentence type information for each user input at the pre-processing stage. Thus the Rasp parser outputted a 'conjunction' sentence type for the 11th input and 'but' is the conjunction word. Such a conjunction phrase normally will be used to express a contradictory opinion or another point of view in linguistics.

Moreover the mum character, Janet, wants what is best for Peter and has a positive relationship with the sick leading son character. Thus we can detect Janet's emotion purely based on the linguistic feature of the 11th input without any emotion shifters. That is Janet is more likely to provide another point of view under the above 'negative' discussion theme by showing 'disapproval' to Peter's previous point of view. Thus the 11th

input is more likely to indicate 'disapproval'. If Peter and Janet share a negative relationship and Peter showed a negative emotion in the most recent input, then Janet either may behave with a gloating response of 'happiness' or may response with an 'outrageous' emotion. A neural network implementation is used to perform such reasoning with the consideration of relationships, emotions implied by target audiences and sentence types to detect affect in the social interaction contexts.

For the 12th input from Matthew, the Rasp parser also implied that its sentence type is a question sentence. In English, the expression of question sentences is so diverse. Most of them will require confirmation or replies from other characters, while there is a small group of question sentences that do not really require any replies, i.e. rhetorical questions. Such questions (e.g. "What the hell are you thinking?", "Who do you think you are?", "How many times do I need to tell you?", "Are you crazy?") encourage the listener to think about what the (often obvious) answers to the questions must be. They tend to be used to express dissatisfaction. In our application domain, we especially detect such rhetorical questions using latent semantic analysis after Rasp's initial analysis of the sentence type information. We construct two training documents for questions sentences: one with normal question sentences and the other with rhetorical questions. We use the semantic vector API to perform semantic similarity comparison between the two training document vectors and the 12th input from Matthew. The output is in Figure 7.

```
similarity of "rhetorical_ques.txt" with "test_corpus.txt": 0.7690450440578355
similarity of "normal_ques.txt" with "test_corpus.txt": 0.7092013666709898
```

Figure 7. Semantic similarities between the training question documents and the test document (the 12th input).

The above results indicate that the 12th input is regarded as a rhetorical question, which implies 'dissatisfaction'. We previously reveal the discussion themes of the 12th input are also 'disease' and 'bullying' related and its most likely target audiences are 'Peter' and 'Janet'. Both Peter's and Janet's most recent inputs are regarded as the most related social context to the last input from Matthew. Since Peter and Janet both implied negative emotions in their most recent inputs, the 12th input is embedded in a negative interaction context. According to the scenario, Matthew also has a positive relationship with Peter and he believes that Peter will be bullied because of the side effect of the operation. Thus he is against the operation idea. On the contrary, Janet wants Peter to have the operation. Therefore, Matthew and Janet have a tense relationship. Moreover, since the 12th input is recognised as a rhetorical question reflecting dissatisfaction by itself, in such a negative interaction context and with a comparatively tense relationship with one of the target audience characters, Matthew is more likely to express 'angry' in the 12th input.

The above interpretation of emotional influence between characters with the consideration of their interpersonal relationships, recent emotions of target audiences and sentence types has been implemented by Backpropagation, a supervised neural network algorithm. Neural networks are generally well-known for classification tasks and pattern recognition. Backpropagation is also one of the most classic supervised neural network algorithms. It is chosen due to its promising performances and robustness for the modeling of the problem domain.

We intend to use this neural network implementation to accept the sentence type of the current input, most recent emotions of the current input's potential target audiences, an averaged relationship value between the target audiences and the speaking character as inputs. The number of target audiences could range from minimum one to maximum four for one social input in one drama improvisation session with altogether five characters. The output of the neural network will be the most probable emotion implied in the current input expressed by the speaking character. In this context-based affect detection application, we consider the most frequently used 10 emotions ('neutral', 'approval', 'disapproval', 'angry', 'grateful', 'regretful', 'happy', 'sad', 'worried' and 'caring') in this scenario as the output detected affective states.

Moreover, a single hidden layer can model any continuous functions and is easily trained with Backpropagation. Neural networks with two hidden layers are universal computing devices. However it is more difficult to train. The neural network with one hidden layer is also capable enough for the target problem domain. Therefore a model with one single hidden layer is chosen in our application. The three-layer topology of the neural network includes: one input, one hidden and one output layer, with six nodes in the input layer and 10 nodes respectively in the hidden and output layers. The six nodes in the input layer indicate the most recent emotional implications expressed by potential up to four target audiences, an averaged interpersonal relationship and sentence type information. We use three values to define relationships: 1 for a positive relationship, 0 for a neutral relationship and -1 for a negative relationship. An average relationship value will be calculated and used as one input to the neural network if the user input has more than one potential target audience. The 10 nodes in the output layer represent the 10 most frequently used emotions.

The 600 example inputs with agreed annotations extracted from the selected five example transcripts of the Crohn's disease scenario are also used for the training of the neural network. The training data are generated in the following way. Potential target audiences of each input have been identified by two human judges. The most recent emotions implied by the identified audiences have been collected as input values for the neural network. Scores for interpersonal relationships between characters are pre-defined. For example, Arnold has a tense relationship (-1) with Peter but a medium relationship (0) with Matthew and a positive relationship (1) with Janet, and so on and so forth for other characters. Then an average relationship score is produced. Sentence type information is obtained using Rasp for each input. The subsequent emotion experienced by the speaking character is used as the expected output. A sequence consisting of up to four emotions, a score for relationship interpretation and a sentence type is regarded as one training data. In this way, 553 training data are used to train the Backpropagation algorithm. Standard error functions of Backpropagation are used to calculate errors in the output and hidden layers. Then they are respectively used to adjust the weights from the hidden to output layer and the weights from the input to hidden layer.

In order to maintain the algorithm's generalization capabilities, we minimize the changes made to the network at each step. This can be achieved by reducing the learning rate.

Thus by reducing the changes over time, we reduce the possibility that the network will become over-learning and too focused on the training data. After the neural network has been trained to reach a reasonable average error rate (less than 0.05), it is used for testing to predict emotional influence of other participant characters towards the speaking character in the test interaction contexts.

In the above example interaction we discussed in section 3, for the emotion detection of the 11th input, we have the following sequence used as the inputs to the Backpropagation algorithm:

1. The most related emotion context: 'Disapproval (implied in the 9th input from the target audience, Peter), null, null, and null'. 'Null' is used to represent the absence of other target audiences.
2. Relationship: '1'- Peter and Janet share a positive relationship.
3. Sentence type: 'conjunction_but' – a conjunction type with a 'but 'phrase.

The neural network uses the above as inputs and outputs 'disapproval' as the implied emotion in the 11th input from Janet as described above. Similarly, for the 12th input from Matthew, the following sequence is used as the inputs to the Backpropagation algorithm:

1. The most related emotion context: 'Disapproval (implied in the 9th input from one target audience, Peter) and disapproval (implied in the 11th input from the other target audience, Janet), null and null'.
2. Relationship: '1': Peter and Matthew have a positive relationship; '-1': Janet and Matthew have a tense or negative relationship. The average value is: (1-1)/2 = 0.
3. Sentence type: 'rhetorical_que': a rhetorical question type.

Then it outputs that Matthew is most likely to be 'angry'. Another three transcripts of the Crohn's disease scenario have also been used for the testing of the neutral network-based reasoning. Two human judges are also used to provide affect annotation of the test example inputs. 230 emotional contexts with agreed affect annotation are extracted in a similar way to evaluate the performance of the Backpropagation algorithm. Each emotional context consists of the affective states expressed by target audience human characters in their most recent inputs. Character relationships and sentence types are also appended after these emotional contexts. They will be used as the inputs to the neural network to predict their influence to the emotion of the subsequent speaker. The output of the network is then used to compare with the human judges' annotation of this current speaker's input. In this way, we can provide a channel for context-based affect interpretation as emotion shifters in the social interaction context. Evaluation details are presented in section 5.

5 EVALUATION

The overall system employs a client & server architecture and is implemented in Java. Human actors, the intelligent agent and the human director work through software clients connecting with the server. Figure 8 gives an overview of the control of the expressive characters. Users' text input is analyzed by the AI agent in order to detect affect in the text. The output is an emotion label with intensity derived from the text. This is then used in two ways. Firstly it is used by the minor character (played by the AI agent) to generate a response. Secondly the label and the intensity are sent to the emotional animation system (via an XML stream) where it is used to generate animation.

The context-based affect detection will only be activated if the user inputs contain weak or no obvious strong affect indicators. It is implemented in the following way. If the user inputs do not include strong affect indicators, the APIs from the semantic vector packages are used to detect their topic themes and identify their potential target audiences. Then a neural network Java class with a standard Backpropagation algorithm is used to detect affect for such inputs with the consideration of interpersonal relationships, the inputs' sentence types and the most recent emotions implied by the target audiences. The detected affective states and the AI agent's responses to other characters have been encoded in an xml stream, which is sent to the server by the AI agent. Then the server broadcasts the xml stream to all the clients so that the detected affective states can be picked up by the animation engine to contribute to the production of 3D gestures for the user-controlled avatars. The overall affect detection component works in real-time applications with the running time of approximately 450ms on average with the following type of processor: Intel(R) Core(TM)2 Duo CPU T9500 @2.60GHz 2.60GHz.

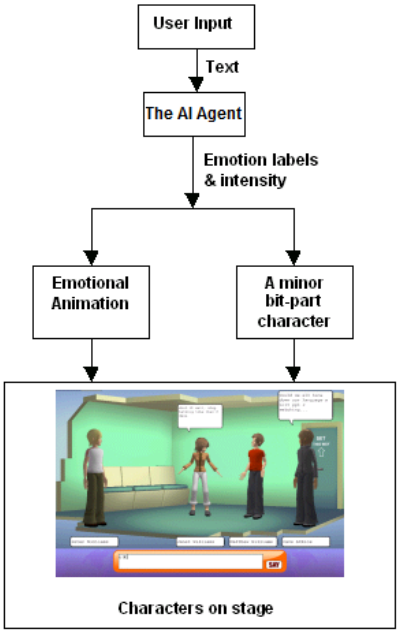

Figure 8. Affect detection and the control of characters

User testing was conducted previously with 200 British secondary school students to evaluate the affect detection component and the AI agent's performance. Subjects were 14–16 years old students at local schools. There was no control of gender for the testing. Briefly, the methodology of the testing is that we had each test subject have an experience of both scenarios, one including the AI minor character only and the other including the human-controlled minor character only. After the testing sessions, we obtained users' feedback via questionnaires and group debriefings. Improvisational transcripts were automatically recorded to allow further evaluation.

Moreover, we concealed the fact that the AI agent was involved in some sessions in order to have a fair test of the difference that is made. It surprised us that few users appeared to realise that sometimes one character was computer-controlled, although that it was not an aim of our work to ensure that human actors did not realise this.

We have taken previously recorded transcripts to evaluate the efficiency of the updated affect detection component with contextual inference. In order to evaluate the performances of the topic theme detection using latent semantic analysis and the neural network based affect detection in the social context, another three transcripts of the Crohn's disease scenario are used. Two human judges are employed to annotate the topic themes of the extracted 300 user inputs from the test three transcripts using the previously mentioned six categorizations. Moreover, Cohen's Kappa is a statistical measurement of inter-annotator agreement. It provides robust measurement by taking the agreement occurring by chance into consideration. We used it to measure the inter-agreement between human judges for the topic theme annotation and obtained 0.83. Then the 265 example inputs with agreed topic theme annotations are used as the gold standard to test the performance of the topic theme detection. A rule-based keyword pattern matching baseline system has been used to compare the performance with that of the LSA. We have obtained an averaged precision, 0.736, and an averaged recall, 0.733, using the LSA while the baseline system achieved an averaged precision of 0.603 and an averaged recall of 0.583 for the six topic theme detection. The detailed results indicated that discussion themes of 'bullying', 'disease' and 'food choices' have been very well detected by our semantic-based analysis. The discussions on 'family care' and 'suggestion' topics posed most of the challenging. For example, the following input is from Peter classified as a 'suggestion' topic by the human annotators, "This is so hard and I need your support". The semantic analysis has given the highest similarity score (0.905) to one of the 'bullying' theme training documents and the 2nd highest score (0.901) to the training document with the 'suggestion' theme. Also, "how would you all feel if you were in my situation" has been regarded as a 'bullying' related discussion with the 2nd recommendation of a 'disagreement' theme while the input was annotated to be more inclined to indicate 'disagreement'. Although the topic detection using LSA made errors like the above sometimes, the similarity scores for the ideal classifications became very close to the top score for another topic category. We also notice that sometimes without too many contexts, the test sentences themselves could also show ambiguity for topic detection tasks. Generally the semantic-based interpretation achieves reasonable and promising results.

The two human judges have also annotated these 265 example inputs with the 10 frequently used emotions. Cohen's Kappa is used again to measure the performance of the overall updated affect detection component embedded in the intelligent agent. Such a statistical result is also considered as a conservative measure of agreement and more robust than simple percent agreement calculation. It is also widely used in linguistic field to measure annotation inter-agreement. Thus it is used again as an effective channel to measure our system's performance.

In our application, since 10 emotions were used for annotation and the annotators may not experience the exact emotions as the test subjects did, it led to the low inter-agreement between human judges. The inter-agreement between human judge A/B is 0.63. While the old version of the affect detection without any contextual inference achieves 0.46 in good cases, the new version achieves inter-agreements with human judge A/B respectively 0.56 and 0.58. As mentioned earlier, due to the fact that we have considered a large category of emotions for affect annotation, the inter-agreement improvements are comparatively small. However, a detailed inspection of the annotated test transcripts by the new version of the AI agent indicates that many expressions regarded as 'neutral' by the previous version have been annotated appropriately as emotional expressions.

Moreover, in order to provide initial evaluation results for the neural network-based affect detection, the human judges' previous annotations are also converted into three emotion labels: positive, negative and neutral. Cohen's Kappa is also produced to measure the human annotators' inter-agreement using these three labels: 0.85. Then 230 inputs with agreed annotations are used as the gold standard with 45% negative inputs, 28% positive and 27% neutral expressions. The affective annotations achieved by the neural network-based affect detection are also converted into solely positive and negative. A baseline system is built using simple Bayesian networks in order to further measure the neural network-based affect detection. The Bayesian network has the following topology. Emotions implied in the last two inputs (no matter if these two previously speaking characters are the target audiences or not) are used as inputs and the output will be the predicted affect of the currently speaking character. Training was also conducted for the baseline system with 250 examples from the training transcripts. Evaluation results are presented in Table 1.

The neural network-based inference with the consideration of interpersonal relationships, most recent emotions expressed by the target audiences and sentence types has performed generally better and more stable than the baseline system with only the consideration of the last two emotional contexts. Especially our approach copes well with the sudden change of emotions in the social context due to unexpected topic change, while such situations still challenge the baseline system greatly.

We also mentioned in section 2 an affect detection system, @AM, developed by Neviarouskaya et al. [6]. The @AM system focused on affect detection from individual sentences borrowed from the Experience website. Their approach mainly used linguistic features of individual sentences to detect affect and did not take any context into consideration, while our approach mainly uses cognitive cues such as emotion contexts of target audiences and their interpersonal relationships with the speaker for affect detection. Although there are differences on technical aspects between our approach and their @AM system, we still compare the two systems to get some indication of our system's performance. Their @AM system was used to annotate 1,000 sentences using three labels (positive, negative and neutral). Their system achieved 92% precision scores for the annotation of positive inputs, 91% precision for negative inputs and 47% precision for neutral ones. Neutral sentences still challenged their system greatly. Tested on a small sample set (230 inputs), our context-based affect detection generally performs stably on the detection of each category of emotional and neutral expressions comparing with the @AM system. However, further testing is needed to prove our system's efficiency.

		Precision	Recall	F-Measure
The neural network-based affect detection	Positive	0.875	0.75	0.8
	Negative	0.867	0.813	0.839
	Neutral	0.737	0.875	0.8
The baseline system	Positive	0.587	0.711	0.643
	Negative	0.845	0.652	0.736
	Neutral	0.395	0.536	0.455

Table 1. Emotion detection results for both of the neural network-based affect detection and the baseline system

6 CONCLUSIONS & FUTURE WORK

The overall context-based affect detection integrated with the original affect sensing component is embedded in the AI agent. It has generally made the AI agent perform better for the detection of emotions embedded in the multi-threaded dialogue contexts. In future work, we aim to use more example transcripts from different scenarios (such as bullying and career training) and articles from the Experience project to further improve the performance of the semantic topic theme detection and context-based affect detection. We also intend to extend the emotion modeling with the consideration of personality and culture. We are also interested in topic extraction to support affect interpretation directly, e.g. the suggestion of a topic change indicating potential indifferent to the current discussion theme. It will also ease the interaction and make human characters comfortable if our agent is equipped with culturally related small talk behavior. We also notice that the training and testing transcripts contain more negative inputs than positive ones due to the nature of the chosen scenarios (in this case, Crohn's disease). We also intend to employ partially supervised learning to deal with imbalanced affect classification. Moreover, emotions implied by the rhetorical questions may not be obvious sometimes [16]. For example, "aren't I awesome?", could be used to indicate both an affirmation of one's deeds as well as dissatisfaction in a context when the speaker actually does something wrong.

REFERENCES

[1] X. Sun, J. Lichtenauer, M.F. Valstar, A. Nijholt, and M. Pantic: A Multimodal Database for Mimicry Analysis. In *Proceedings of ACII* 2011: 367-376. (2011).

[2] A.M. Turing, Computing machinery and intelligence. *Mind,* 59, 433-460. (1950).

[3] A. Kappas. Smile when you read this, whether you like it or not: Conceptual challenges to affect detection. *IEEE Transactions on Affective Computing*, 1 (1), 38-41, (2010).

[4] L. Zhang. Exploitation on Contextual Affect Sensing and Dynamic Relationship Interpretation. *ACM Computers in Entertainment*. Vol.8, Issue 3. (2010).

[5] B. Endrass, M. Rehm & E. André. Planning Small Talk Behavior with Cultural Influences for Multiagent Systems. *Computer Speech and Language.* 25(2):158-174. (2011).

[6] A. Neviarouskaya, H. Prendinger and M. Ishizuka. Recognition of Affect, Judgment, and Appreciation in Text. In *Proceedings of the 23rd International Conference on Computational Linguistics*, Beijing, China, pp. 806-814. (2010).

[7] M. Ptaszynski, P. Dybala, W. Shi, R. Rzepka and K. Araki. Towards Context Aware Emotional Intelligence in Machines: Computing Contextual Appropriateness of Affective States. In *Proceeding of IJCAI.* (2009).

[8] K.R. Scherer. Vocal Communication of Emotion: A Review of Research Paradigms. *Speech Communication* 40, 227-256. (2003).

[9] E. Mower, A. Metallinou, C. Lee, A. Kazemzadeh, C. Busso, S. Lee & S.S. Narayanan. Interpreting ambiguous emotional expressions. *International Conference on Affective Computing and Intelligent Interaction*. Amsterdam, The Netherlands. (2009).

[10] A. Batliner, S. Steidl, C. Hacker & E. Nöth. Private emotions vs. social interaction — a data-driven approach towards analysing emotions in speech. *User Modeling and User-Adapted Interaction*: 175-206. (2008).

[11] J.A. Russell, J.A. Bachorowski, and J.M. Fernandez-Dols. Facial and Vocal Expressions of Emotion, *Ann. Rev. Psychology*, vol. 42, pp. 329-349. (2003).

[12] T.K. Landauer and S. Dumais. Latent semantic analysis. Scholarpedia, 3(11):4356. 2008.

[13] D. Widdows and T. Cohen. The Semantic Vectors Package: New Algorithms and Public Tools for Distributional Semantics. *IEEE Int. Conference on Semantic Computing*. (2010).

[14] S. Hareli and A. Rafaeli. Emotion cycles: On the social influence of emotion in organizations. *Research in Organizational Behavior*, 28, 35-59. (2008)

[15] Z. Wang, J. Lee and S. Marsella. Towards More Comprehensive Listening Behavior: Beyond the Bobble Head. *International Conference on Intelligent Virtual Agents.* (2011).

[16] C. Potts and F. Schwarz. Exclamatives and heightened emotion: Extracting pragmatic generalizations from large corpora. Ms., *UMass Amherst*. (2008).

YACIS: A Five-Billion-Word Corpus of Japanese Blogs Fully Annotated with Syntactic and Affective Information

Michal Ptaszynski[1] **Pawel Dybala**[2] **Rafal Rzepka**[3] **Kenji Araki**[4] and **Yoshio Momouchi**[5]

Abstract. This paper presents YACIS, a new fully annotated large scale corpus of Japanese language. The corpus is based on blog entries from Ameba blog service. The original structure (blog post and comments) is preserved, thanks to which semantic relations between posts and comments are maintained. The corpus is annotated with syntactic (POS, dependency parsing, etc.) and affective (emotive expressions, emoticons, valence, etc.) information. The annotations are evaluated in a survey on over forty respondents. The corpus is also compared to other existing corpora, both large scale and emotion related.

1 INTRODUCTION

Text corpora are some of the most vital linguistic resources in natural language processing (NLP). These include newspaper corpora [1], conversation corpora or corpora of literature[6]. Unfortunately, comparing to major world languages, like English, there are few large corpora available for the Japanese language. Moreover, grand majority of them is based on newspapers, or legal documents[7]. These are usually unsuitable for the research on sentiment analysis and emotion processing, as emotions and attitudes are rarely expressed in this kind of texts. Although there exist conversation corpora with speech recordings, which could become useful in such research[8], due to the difficulties with compilation of such corpora they are relatively small. Recently blogs have been recognized as a rich source of text available for public. Blogs are open diaries in which people encapsulate their own experiences, opinions and feelings to be read and commented by other people. Because of their richness in subjective and evaluative information blogs have come into the focus in sentiment and affect analysis [2, 3, 4, 5]. Therefore creating a large blog-based emotion corpus could become a solution to overcome both problems, of the lack in quantity of corpora and their applicability in the research on sentiment analysis and emotion processing. However, there have been only a few small (several thousand sentences) Japanese emotion corpora developed so far [2]. Although there exist medium scale Web-based corpora (containing several million words), such as JpWaC [6] or jBlogs [7], access to them is usually allowed only from the Web interface, which makes additional annotations (parts-of-speech, dependency structure, deeper affective information, etc.) difficult. Furthermore, although there exist large resources, like Google N-gram Corpus [8], the textual data sets in such resources are short (up to 7-grams) and do not contain any contextual information. This makes them unsuitable for emotion processing research, since most of contextual information, so important in expressing emotions [9], is lost. Therefore we decided to create a new corpus from scratch. The corpus was compiled using procedures similar to the ones developed in the WaCky initiative [10], but optimized to mining only one blog service (Ameba blog, http://ameblo.jp/, later referred to as Ameblo). The compiled corpus was fully annotated with syntactic (POS, lemmatization, dependency parsing, etc.) and affective information (emotive expressions, emotion classes, valence, etc.).

The outline of the paper is as follows. Section 2 describes the related research in large scale corpora and blog emotion corpora. Section 3 presents the procedures used in compilation of the corpus. Section 4 describes tools used in corpus annotation. Section 5 presents detailed statistical data and evaluation of the annotations. Finally the paper is concluded and applications of the corpus are discussed.

2 RELATED RESEARCH

In this section we present some of the most relevant research related to ours. There has been no billion-word-scale corpus annotated with affective information before. Therefore we needed to divide the description of the related research into “Large Scale Corpora” and “Emotion Corpora”.

2.1 Large-Scale Web-Based Corpora

The notion of a ”large scale corpus” has appeared in linguistic and computational linguistic literature for many years. However, study of the literature shows that what was considered as ”large” ten years ago does not exceed a 5% (border of statistical error) when compared to present corpora. For example, Sasaki et al. [11] in 2001 reported a construction of a question answering (QA) system based on a large scale corpus. The corpus they used consisted of 528,000 newspaper articles. YACIS, the corpus described here consists of 12,938,606 documents (blog pages). The rough estimation indicates that the corpus of Sasaki et al. covers less than 5% of YACIS (in particular 4.08%). Therefore we mostly focused on research scaling the meaning of ”large” up to around billion-words and more.

Firstly, we need to address the question of whether billion-word and larger corpora are of any use to linguistics and in what sense it is better to use a large corpus rather than a medium sized one. This question has been answered by most of the researchers involved in the creation of large corpora, thus we will answer it briefly referring

[1] Hokkai-Gakuen University, Japan, email: ptaszynski@hgu.jp
[2] Otaru University of Commerce, Japan, email: paweldybala@res.otaru-uc.ac.jp
[3] Hokkaido University, Japan, email: kabura@media.eng.hokudai.ac.jp
[4] Hokkaido University, Japan, email: araki@media.eng.hokudai.ac.jp
[5] Hokkai-Gakuen University, Japan, email: momouchi@eli.hokkai-s-u.ac.jp
[6] http://www.aozora.gr.jp/
[7] http://www-nagao.kuee.kyoto-u.ac.jp/NLP_Portal/lr-cat-e.html
[8] http://www.ninjal.ac.jp/products-k/katsudo/seika/corpus/public/

to the relevant literature. Baayen [12] notices that language phenomena (such as probability of appearance of certain words within a corpus) are distributed in accordance with Zip's Law. The Zip's Law was originally proposed and developed by George Kingsley Zipf in late 1930's to 1940's [13, 14], who formulated a wide range of linguistic phenomena based on probability. One such phenomenon says that the number of occurrences of words within a corpus decreases in a quadratic-like manner. For example, when all unique words in a corpus are represented in a list with decreasing occurrences, the second word on the list will have a tendency to appear two times less often than the first one. This means that if a corpus is not big enough, many words will not appear in it at all. Baroni and Ueyama [7] and Pomikálek et al. [15] indicate that Zipf's Law is one of the strongest reasons to work with large-scale corpora, if we are to understand the most of the language phenomena and provide statistically reliable proofs for them. There are opponents of uncontrolled over-scaling of corpora, such as Curran (with Osborne in [16]), who show that convergence behavior of words in a large corpus does not necessarily appear for all words and thus it is not the size of the corpus that matters, but the statistical model applied in the processing. However, they do admit that the corpus scale is one of the features that should be addressed in the corpus linguistic research and eventually join the initiative of developing a 10 billion word corpus of English (see Liu and Curran [17]).

The latter, followed by Baroni and Ueyama [7], indicate at least two types of research dealing with large-scale corpora. One is using popular search engines, such as Google[9] or Yahoo![10]. In such research one gathers estimates of hit counts for certain keywords to perform statistical analysis, or wider contexts of the keywords, called "snippets" (a short, three line long set of text containing the keyword), to perform further analysis of the snippet contents. This refers to what has generally developed as the "Web mining" field. One of the examples is the research by Turney and Littman [18]. They claim to perform sentiment analysis on a hundred-billion-word corpus. By the corpus they mean roughly estimated size of the web pages indexed by AltaVista search engine[11]. However, this kind of research is inevitably constrained with limitations of the search engine's API. Pomikálek et al. [15] indicate a long list of such limitations. Some of them include: limited query language (e.g. no search by regular expressions), query-per-day limitations (e.g. Google allows only one thousand queries per day for one IP address, after which the IP address is blocked - an unacceptable limitation for linguistic research), search queries are ordered with a manner irrelevant to linguistic research, etc. Kilgariff [19] calls uncritical relying on search engine results a "Googleology" and points out a number of problems search engines will never be able to deal with (such as duplicated documents). Moreover, only Google employees have unlimited and extended access to the search engine results. Kilgariff also proposes an alternative, building large-scale corpora locally by crawling the World Wide Web, and argues that it is the optimal way of utilizing the Internet contents for research in linguistics and computational linguistics.

There have been several initiatives to build billion-word-scale corpora for different languages. Google is a company that holds presumably the largest text collection in the world. The scale makes it impossible to control, evaluate and fully annotate, which makes it a large collection not fully usable for researchers [15, 19]. However, Google has presented two large corpora. One is the "Web 1T (trillion) 5 gram" corpus [47] published in 2006. It is estimated to contain one trillion of tokens extracted from 95 billion sentences. Unfortunately, the contents available for users are only n-grams, from 1 (unigrams) to 5 (pentagrams). The corpus was not processed in any way except tokenization. Also, the original sentences are not available. This makes the corpus, although unmatchable when it comes to statistics of short word sequences, not interesting for language studies, where a word needs to be processed in its context (a sentence, a paragraph, a document). The second one is the "Google Books 155 Billion Word Corpus"[12] published recently in 2011. It contains 1.3 million books published between 1810 and 2009 and processed with OCR. This corpus has a larger functionality, such as part of speech annotation and lemmatization of words. However, it is available only as an online interface with a daily access limit per user (1000 queries). The tokenized-only version of the corpus is available, also for several other languages[13], unfortunately only in the n-gram form (no context larger than 5-gram).

Among corpora created with Web crawling methods, Liu and Curran [17] created a 10-billion-word corpus of English. Although the corpus was not annotated in any way, except tokenization, differently to Google's corpora it is sentence based, not n-gram based. Moreover, it successfully proved its usability in standard NLP tasks such as spelling correction or thesaurus extraction.

The **WaCky** (**W**eb **a**s **C**orpus **k**ool **y**nitiative) [7, 10] project started gathering and linguistically processing large scale corpora from the Web. In the years 2005-2007 the project resulted in more then five collections of around two billion word corpora for different languages, such as English (ukWaC), French (frWaC), German (deWaC) or Italian (itWaC). The tools developed for the project are available online and their general applicability is well established. Some of the corpora developed within the project are compared in table 1.

BiWeC [15], or **Bi**g **W**eb **C**orpus has been collected from the whole Web contents in 2009 and consists of about 5.5 billion words. The authors of this corpus aimed to go beyond the border of 2 billion words set by the WaCky initiative[14] as a borderline for corpus processing feasibility for modern (32-bit) software.

Billion-word scale corpora have been recently developed also for less popular languages, such as Hungarian [24], Brazilian Portuguese [46] or Polish [23].

As for large corpora in Japanese, despite the fact that Japanese is a well recognized and described world language, there have been only few corpora of a reasonable size.

Srdanović Erjavec et al. [20] notice this lack of freely available large corpora for Japanese. They used WAC (Web As Corpus) Toolkit[15], developed under the WaCky initiative, and Kilgariff et al.'s [21] Sketch Engine, a tool for thesauri generation from large scale corpora (applied also for English in [15]). They gathered **JpWaC**, a 400 million word corpus of Japanese. Although JpWac covers only about 7% of YACIS (400 mil. vs 5.6 bil. words), the research is worth mentioning, since it shows that freely available tools developed for European languages are to some extend applicable also for languages of completely different typography, like Japanese[16]. However, they faced several problems. Firstly, they had to normalize the character

[9] http://www.google.com
[10] http://www.yahoo.com
[11] In 2004 AltaVista (http://www.altavista.com/) has become a part of Yahoo!.
[12] http://googlebooks.byu.edu/
[13] http://books.google.com/ngrams/datasets
[14] http://wacky.sslmit.unibo.it/
[15] http://www.drni.de/wac-tk/
[16] languages like Chinese, Japanese or Korean are encoded using 2-bite characters.

Table 1. Comparison of different corpora, ordered arbitrary by size (number of words/tokens).

corpus name	scale (in words)	language	domain	annotation
Liu&Curran [17]	10 billion	English	whole Web	tokenization;
YACIS	5.6 billion	Japanese	Blogs (Ameba)	tokenization, POS, lemma, dependency parsing, NER, affect (emotive expressions, Russell-2D, emotion objects);
BiWeC [15]	5.5 billion	English	whole Web (.uk and .au domains)	POS, lemma;
ukWaC	2 billion	English	whole Web (.uk domain)	POS, lemma;
PukWaC (Parsed-ukWaC) [10]	2 billion	English	whole Web (.uk domain)	POS, lemma, dependency parsing;
itWaC [7, 10]	2 billion	Italian	whole Web (.it domain)	POS, lemma;
Gigaword [24]	2 billion	Hungarian	whole Web (.hu domain)	tokenization, sentence segmentation;
deWaC [10]	1.7 billion	German	whole Web (.de domain)	POS, lemma;
frWaC [10]	1.6 billion	French	whole Web (.fr domain)	POS, lemma;
Corpus Brasiliero [46]	1 billion	Brazilian Portuguese	multi-domain (newspapers, Web, talk transcriptions)	POS, lemma;
National Corpus of Polish [23]	1 billion	Polish	multi-domain (newspapers, literature, Web, etc.)	POS, lemma, dependency parsing, named entities, word senses;
JpWaC [20]	400 million	Japanese	whole Web (.jp domain)	tokenization, POS, lemma;
jBlogs [20]	62 million	Japanese	Blogs (Ameba, Goo, Livedoor, Yahoo!)	tokenization, POS, lemma;

Table 2. Detailed comparison of different Japanese corpora, ordered by the number of words/tokens.

corpus name	scale (in words)	number of documents (Web pages)	number of sentences	size (uncompressed in GB, text only, no annotation)	domain
YACIS	5,600,597,095	12,938,606	354,288,529	26.6	Blogs (Ameba);
JpWaC [20]	409,384,411	49,544	12,759,201	7.3	whole Web (11 different domains within .jp);
jBlogs [7]	61,885,180	28,530	[not revealed]	.25 (compressed)	Blogs (Ameba, Goo, Livedoor, Yahoo!);
KNB [2]	66,952	249	4,186	450 kB	Blogs (written by students exclusively for the purpose of the research);
Minato et al. [29]	14,195	1	1,191	[not revealed]	Dictionary examples (written by dictionary authors);

encoding for all web pages[17] (Ameba blog service, on which YACIS was based, is encoded by default in Unicode). Moreover, since they did not specify the domain, but based the corpus on the whole Web contents, they were unable to deal ideally with the web page metadata, such as the page title, author, or creation date, which differs between domains (Ameba has clear and stable meta-structure).

Baroni and Ueyama [7] developed **jBlogs**, a medium-sized corpus of Japanese blogs containing 62 million words. They selected four popular blog services (Ameba, Goo, Livedoor, Yahoo!) and extracted nearly 30 thousand blog documents. Except part of speech tagging, which was done by a Japanese POS tagger ChaSen, the whole procedure and tools they used were the same as the ones developed in WaCky. In the detailed manual analysis of jBlogs, Baroni and Ueyama noticed that blog posts contained many Japanese emoticons, namely *kaomoji*[18]. They report that ChaSen is not capable of processing them, and separates each character adding a general annotation tag "symbol". This results in an overall bias in distribution of parts of speech, putting symbols as the second most frequent (nearly 30% of the whole jBlogs corpus) tag, right after "noun" (about 35%). They considered the frequent appearance of emoticons a major problem in processing blog corpora. In our research we dealt with this problem. To process emoticons we used CAO, a system for detailed analysis of Japanese emoticons developed previously by Ptaszynski et al. [34].

Apart from the above Kawahara and Kurohashi [27] claim the creation of a large, about two-billion-word corpus. However, detailed description of this corpus is not available.

[17] Japanese can be encoded in at least four standards: JIS, Shift-JIS, EUC, and Unicode.

[18] For more detailed description of Japanese emoticons, see [34].

Finally, Okuno Yoo and Sasano Manabu from Yahoo! Japan report on developing a large scale blog corpus, similar in form to the Google "Web 1T 5 gram" with only n-grams available for processing [45]. No information on the corpus is yet available except methods of development, tools (tokenization by MeCab, a POS tagger for Japanese) and its size (1TB).

2.2 Emotion and Blog Corpora

The existing emotion corpora are mostly of limited scale and are annotated manually. Below we compare some of them. As an interesting remark, five out of six were extracted from blogs.

Quan and Ren in 2010 [5] created a Chinese emotion blog corpus **Ren-CECps1.0**. They collected 500 blog articles from various Chinese blog services, such as sina blog (http://blog.sina.com.cn/) or qq blog (http://blog.qq.com/). The articles were annotated with a variety of information, such as emotion class, emotive expressions or valence. Although the syntactic annotations were simplified to tokenization and POS tagging, the corpus is comparable to YACIS in the overall variety of annotations. The motivation for Quan and Ren is also similar - the lack of large scale corpora for sentiment analysis and emotion processing research in Chinese (in our case - Japanese). Wiebe and colleagues [38, 39] created the **MPQA** corpus of news articles. It contains 10,657 sentences in 535 documents. The annotations include emotive expressions, valence, intensity, etc. However, Wiebe et al. focused mostly on sentiment and subjectivity analysis, and they did not include annotations of emotion classes. Hashimoto et al. [2] developed the **KNB** corpus of Japanese blogs. The corpus contains about 67 thousand words in 249 blog articles. Despite its small scale, the corpus proposes a good standard for preparation

Table 3. Comparison of emotion corpora ordered by the amount of annotations.

corpus name	scale (in sentences / docs)	language	annotated affective information					syntactic annotations
			emotion classes (standard)	emotive expressions	emotive/ non-emot.	valence/ activation	emotion intensity	
YACIS	354 mil. / 13 mil.	Japanese	10 (language and culture based)	○	○	○/○	○	T,POS,L,DP,NER;
Ren-CECps1.0	12,724 / 500	Chinese	8 (Yahoo! news)	○	○	○/×	○	T,POS;
MPQA	10,657 / 535	English	none (no standard)	○	○	○/×	○	T,POS;
KNB	4,186 / 249	Japanese	none (no standard)	○	×	○/×	×	T,POS,L,DP,NER;
Minato et al.	1,191 / 1	Japanese	8 (chosen subjectively)	○	○	×/×	×	POS;
Aman&Szpak.	5205 / 173	English	6 (face recognition)	○	○	×/×	○	×

of blog corpora for sentiment and affect-related studies. It contains all relevant syntactic annotations (POS, dependency parsing, Named Entity Recognition, etc.). It also contains sentiment-related information. Words and phrases expressing emotional attitude were annotated by laypeople as either positive or negative. A disadvantage of the corpus, except its small scale, is the way it was created. Eighty students were employed to write blogs for the need of the research. It could be argued that since the students knew their blogs will be read mostly by their teachers, they selected their words more carefully than they would in private. In YACIS we included all types of information contained in KNB, and also added more, especially in the affect-related annotations. Aman and Szpakowicz [4] created a small-scale English blog corpus in 2007. They focused not on syntactic, but on affect-related annotations. They were also some of the first to recognize the task of distinguishing between emotive and non-emotive sentences. ML-Ask, a system applied in annotation of YACIS was evaluated in this matter with high accuracy. Finally, Minato et al. [29] collected a 14,195 word / 1,191 sentence corpus. The corpus is a collection of dictionary examples from "A short dictionary of feelings and emotions in English and Japanese" [25]. It is a dictionary created for Japanese language learners. The sentence examples were mostly prepared by the dictionary author. Moreover, the dictionary does not propose any coherent emotion class list, but rather the emotion concepts are chosen subjectively. All the above corpora were annotated manually or semi-automatically. In our research we performed the first attempt to annotate affect on a large scale corpus automatically. We performed this with previously developed systems, thoroughly evaluated and based on a standardized emotion class typology.

3 YACIS CORPUS COMPILATION

The corpus (named **YACIS** Corpus, or **Y**et **A**nother **C**orpus of **I**nternet **S**entences) was assembled using data obtained automatically from the pages of Ameba Blog (www.ameblo.co.jp, below referred to as Ameblo). There were two main reasons for using Ameblo. Firstly, the users are mostly Japanese so the risk that the links may lead to pages written in a language other than Japanese is small. Secondly, Ameblo has a clear structure of HTML source code, which makes it easy to extract only posts and comments omitting the irrelevant contents, such as advertisements or menu links.

All the tools used for compiling this corpus were developed especially for the purpose of this research. Although there existed several other solutions, all of them were created for crawling the whole Web and included some parts irrelevant for crawling blog service urls like Ameblo (such as the detection of "robots.txt" file, which specifies that no robots should visit any URL from the domain, used for privacy protection), or parts that can be done more easily if the crawling domain is restricted to one blog service (such as HTML code boilerplate deletion). All these parts slow down the crawling process, and sometimes influence the corpus extraction (e.g., general rules for HTML code deletion are less precise than specific rules for deletion of the HTML code that appears in Ameblo). Therefore the available tools, very useful as they are, were insufficient for our needs. All our tools were written in C# and are operating under MS Windows systems.

We developed a simple but efficient web crawler designed to crawl exclusively Ameblo Web pages. The only pages taken into account were those containing Japanese posts (pages with legal disclaimers, as well as posts written in English and other languages were omitted). Initially we fed the crawler with 1000 links taken from Google (response to a query: 'site:ameblo.jp'). All the pages were saved to disk as raw HTML files (each page in a separate file) to be processed later. All of them were downloaded within three weeks between 3rd and 24th of December 2009. Next, we extracted all the posts and comments and divided them into sentences.

Although sentence segmentation may seem to be a trivial task it is not that easy when it comes to texts written by bloggers. People often use improper punctuation, e.g., the periods at the end of sentences are often omitted. In that case we assumed that if the given parts of text are separated by two
 tags (two markers of a new line) then those parts will be two separate sentences. This does not solve the problem in all cases. Therefore we rejoined previously separated parts if the first part ended with a coma or if the quotation marks or parenthesis were opened in the first part and closed in second.

Unfortunately, these modifications were still not perfect and in several cases parts of the text remained not split while others were segmented erroneously. One of the possible improvements was to take into consideration emoticons. We observed that if an emoticon is present in the sentence it usually appears at the end of it. Even in the cases the emoticon did not appear on the very end of the sentence, it still separated two clauses of a different meaning. Moreover, the meaning of the emoticon was always bound with the clause preceding it. This suggested separating sentences after emoticons. To do that we used CAO emoticon analysis system developed by Ptaszynski, et al. [34]. Observations showed this coped with most of the remaining sentence segmentation errors. In a random 1000 sentence sample, less than 1% remained erroneously separated. Analysis of errors showed these were sentences separated by blog authors in a non-standard way and without any particular rule. However, since such cases did not exceed a 5% border of statistical error we considered them an agreeable error.

Currently the data is stored in modified-XML format. Although it looks like XML it does not comply with all XML standards due to the presence of some characters forbidden by XML specification, such as apostrophes (') or quotation marks ("). Those modifications were made to improve the communication with natural language processing tools used in further processing of the corpus, such as a text

parser, part-of-speech analyzer (e.g., MeCab [41]), affect analysis system (ML-Ask [33]) and others. Each page was transformed into an independent XML block between <doc></doc> tags. Opening tag of the <doc> block contains three parameters: URL, TIME and ID which specify the exact address from which the given page was downloaded, download time and unique page number, respectively. The <doc> block contains two other tag types: <post> and <comments>. The <post> block contains all the sentences from the given post where each sentence is included between <s></s> tags. The block <comments> contains all comments written under given post placed between <cmt></cmt> tags which are further split into single sentences placed between <s></s> tags (as described above). An example XML structure of the corpus is represented in figure 1.

The corpus is stored in 129 text files containing 100 000 <doc> units each. The corpus was encoded using UTF-8 encoding. The size of each file varies and is between 200 and 320 megabytes. The size of raw corpus (pure text corpus without any additional tags) is 27.1 gigabytes. Other primary statistics of the corpus are represented in the table 4 below.

Table 4. General Statistics of YACIS Corpus

# of web pages	12,938,606
# of unique bloggers	60,658
average # of pages/blogger	213.3
# of pages with comments	6,421,577
# of comments	50,560,024
average # of comment/page	7.873
# of characters (without spaces)	28,582,653,165
# of characters (with spaces)	34,202,720,910
# of words	5,600,597,095
# of all sentences	354,288,529
# of sentences < 500 characters	353,999,525
# of sentences after correction of sentence segmentation errors	371,734,976
# of words per sentence (average)	15
# of characters per sentence (average)	77

As mentioned in Table 4, average sentence length is 28.17 Japanese characters. Kubota et al. [44] divide sentences in Japanese according to their intelligibility into: easily intelligible short sentences (up to 100 characters) and difficult long sentences (over 100 characters long). The sentences in our corpus fit in the definition of short sentences which means they are easily understandable. After exclusion of very long sentences (consisting of over 500 characters) the number of sentences does not change significantly and is 354,169,311 (99,96%) with an average length of 27.9 characters. This means the corpus is balanced in the length of sentences.

4 YACIS CORPUS ANNOTATION TOOLS

The corpus, in the form described in section 3 was further annotated with several kinds if information, such as parts-of-speech, dependency structure or affective information. The tools we used in particular are described in detail below.

4.1 Syntactic Information Annotation Tools

MeCab [41] is a standard morphological analyzer and parts-of-speech (POS) tagger for Japanese. It is trained using a large corpus on a Conditional Random Fields (CRF) discriminative model and uses a bigram Markov model for analysis. Prior to MeCab there were several POS taggers for Japanese, such as Juman[19] or ChaSen[20]. ChaSen and MeCab have many similarities in their structures. Both share the same corpus base for training and use the same default dictionary (ipadic[21] based on a modified IPA Part of Speech Tagset developed by the Information-Technology Promotion Agency of Japan (IPA)). However, ChaSen was trained on a Hidden Markov Model (generative model), a full probabilistic model in which first all variables are generated, and thus is slower than MeCab, based on a discriminative model, which focuses only on the target variables conditional on the observed variables. Juman on the other hand was developed separately from MeCab on different resources. It uses a set of handcrafted rules and a dictionary (jumandic) created on the basis of Kyoto Corpus developed by a Kurohashi&Kawahara Laboratory[22] at Kyoto University. Both MeCab and Juman are considerably fast, which is a very important feature when processing a large-scale corpus such as YACIS. However, there were several reasons to choose the former. MeCab is considered slightly faster when processing large data and uses less memory. It is also more accurate since it allows partial analysis (a way of flexible setting of word boundaries in non-spaced languages, like Japanese). Finally, MeCab is more flexible when using other dictionaries. Therefore to annotate YACIS we were able to use MeCab with the two different types of dictionaries mentioned above (ipadic and jumandic). This allowed us to develop POS tagging for YACIS with the two most favored standards in morphological analysis of Japanese today. An example of MeCab output is represented in figure 2 (the results were translated into English according to Francis Bond's "IPA POS code in Japanese and English"[23] developed as a standard for annotation of Japanese WordNet[24]).

Cabocha [42] is a Japanese dependency parser based on Support Vector Machines. It was developed by MeCab developers and is considered to be the most accurate statistical Japanese dependency parser. Its discriminative feature is using Cascaded Chunking Model, which makes the analysis efficient for the Japanese language. Other dependency parsers for Japanese, such as KNP[25] use statistical probabilistic models, which makes them inefficient for complex sentences with many clauses. Cascaded Chunking Model parses a sentence deterministically focusing on whether a sentence segment modifies a segment on its right hand side [42]. As an option, Cabocha uses IREX[26] (Information Retrieval and Extraction Exercise) standard for Named Entity Recognition (NER). We applied this option in the annotation process as well. An example of Cabocha output is represented in figure 2. Table 5 represents all tag types included in IREX.

Table 5. Named entity annotations included in the IREX standard.

<opening tag>...</closing tag>	**explanation**
<ORGANIZATION>... ...</ORGANIZATION>	organization or company name including abbreviations (e.g., Toyota, or Nissan);
<LOCATION>...</LOCATION>	mane of a place (city, country, etc.);
<PERSON>...</PERSON>	name, nickname, or status of a person (e.g., Lady Gaga, or "me", "grandson", etc.);
<ARTIFACT>...</ARTIFACT>	name of a well recognized product or object (e.g., Van Houtens Cocoa, etc.);
<PERCENT>...</PERCENT>	percentage or ratio (90%, 0.9);
<MONEY>...</MONEY>	currencies (1000 $, 100 ¥);
<DATE>...</DATE>	dates and its paraphrased extensions (e.g., "4th July", but also "next season", etc.)
<TIME>...</TIME>	hours, minutes, seconds, etc.

[19] http://nlp.ist.i.kyoto-u.ac.jp/EN/index.php?JUMAN
[20] http://chasen.naist.jp/hiki/ChaSen/
[21] http://sourceforge.jp/projects/ipadic/
[22] http://nlp.ist.i.kyoto-u.ac.jp/index.php
[23] http://sourceforge.jp/projects/ipadic/docs/postag.txt
[24] http://nlpwww.nict.go.jp/wn-ja/index.en.html
[25] http://nlp.ist.i.kyoto-u.ac.jp/EN/index.php?KNP
[26] http://nlp.cs.nyu.edu/irex/index-e.html

```
<doc url="http://ameblo.jp/capo-del-rosso/entry-000000.html" time="2009-12-05 21:11:46" id="2000001">
        <post>
                <s>今日から十月です。</s>
                [Its October from today.]
                <s>なんか、九月はいつもよりアッという間に過ぎたような気がするなぁ。</s>
                [I have a strange feeling September passed faster than usual.]
                ...
        </post>
        <comments>
                <cmt>
                        <s>色々と忙しいですね〜！</s>
                        [Oh, you've been busy, weren't you?]
                        ...
                </cmt>
                <cmt>
                        <s>お疲れヤマです(^O^)</s>
                        [Well done! Cheers for good work (^o^)]
                        ...
                </cmt>
        </comments>
</doc>
```

Figure 1. The example XML structure of the main blog corpus.

4.2 Affective Information Annotation Tools

Emotive Expression Dictionary [30] is a collection of over two thousand expressions describing emotional states collected manually from a wide range of literature. It is not a tool *per se*, but was converted into an emotive expression database by Ptaszynski et al. [33]. Since YACIS is a Japanese language corpus, for affect annotation we needed the most appropriate lexicon for the language. The dictionary, developed for over 20 years by Akira Nakamura, is a state-of-the art example of a hand-crafted emotive expression lexicon. It also proposes a classification of emotions that reflects the Japanese culture: 喜 *ki/yorokobi* (joy), 怒 *dō/ikari* (anger), 哀 *ai/aware* (sorrow, sadness, gloom), 怖 *fu/kowagari* (fear), 恥 *chi/haji* (shame, shyness), 好 *kō/suki* (fondness), 厭 *en/iya* (dislike), 昂 *kō/takaburi* (excitement), 安 *an/yasuragi* (relief), and 驚 *kyō/odoroki* (surprise). All expressions in the dictionary are annotated with one emotion class or more if applicable. The distribution of expressions across all emotion classes is represented in Table 6.

Table 6. Distribution of separate expressions across emotion classes in Nakamura's dictionary (overall 2100 ex.).

emotion class	nunber of expressions	emotion class	nunber of expressions
dislike	532	fondness	197
excitement	269	fear	147
sadness	232	surprise	129
joy	224	relief	106
anger	199	shame	65
		sum	**2100**

ML-Ask [31, 33] is a keyword-based language-dependent system for affect annotation on utterances in Japanese. It uses a two-step procedure: **1)** specifying whether an utterance is emotive, and **2)** annotating the particular emotion classes in utterances described as emotive. The emotive sentences are detected on the basis of *emotemes*, emotive features like: interjections, mimetic expressions, vulgar language, emoticons and emotive markers. The examples in Japanese are respectively: *sugee* (great!), *wakuwaku* (heart pounding), *-yagaru* (syntactic morpheme used in verb vulgarization), (^_^) (emoticon expressing joy) and '!', '??' (markers indicating emotive engagement). Emotion class annotation is based on Nakamura's dictionary. ML-Ask is also the only present system for Japanese recognized to implement the idea of Contextual Valence Shifters (CVS) [40] (words and phrases like "not", or "never", which change the valence of an evaluative word). The last distinguishable feature of ML-Ask is implementation of Russell's two dimensional affect model [36], in which emotions are represented in two dimensions: valence (positive/negative) and activation (activated/deactivated). An example of negative-activated emotion could be "anger"; a positive-deactivated emotion is, e.g., "relief". The mapping of Nakamura's emotion classes on Russell's two dimensions was proved reliable in several research [32, 33, 34]. With these settings ML-Ask detects

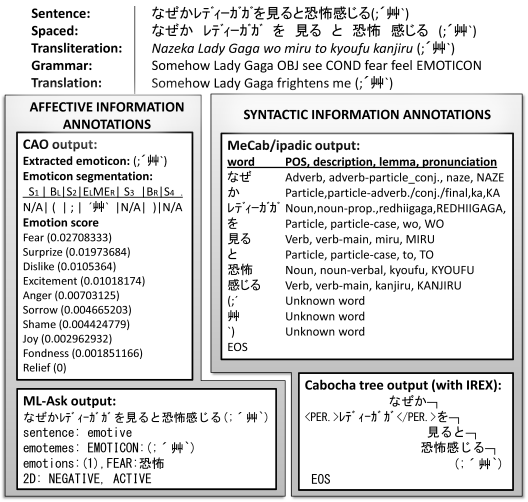

Figure 2. Output examples for all systems.

emotive sentences with a high accuracy (90%) and annotates affect on utterances with a sufficiently high Precision (85.7%), but low Recall (54.7%). Although low Recall is a disadvantage, we assumed that in a corpus as big as YACIS there should still be plenty of data.

CAO [34] is a system for affect analysis of Japanese emoticons, called *kaomoji*. Emoticons are sets of symbols used to convey emotions in text-based online communication, such as blogs. CAO extracts emoticons from input and determines specific emotions expressed by them. Firstly, it matches the input to a predetermined raw emoticon database (with over ten thousand emoticons). The emoticons, which could not be estimated with this database are divided into semantic areas (representations of "mouth" or "eyes"). The areas are automatically annotated according to their co-occurrence in the database. The performance of CAO was evaluated as close to ideal [34] (over 97%). In the YACIS annotation process CAO was used as a supporting procedure in ML-Ask to improve the overall performance and add detailed information about emoticons.

5 ANNOTATION RESULTS AND EVALUATION

5.1 Syntactic Information

In this section we present all relevant statistics concerning syntactic information annotated on YACIS corpus. Where it was possible we also compared YACIS to other corpora. All basic information concerning YACIS is represented in table 4. Information on the distribution of parts of speech is represented in table 7. We compared the two dictionaries used in the annotation (ipadic and jumandic) with other Japanese corpora (jBlogs, and JENAAD newspaper corpus) and in addition, partially to British and Italian Web corpus (ukWaC and itWaC, respectively). The results of analysis are explained below.

Ipadic vs Jumandic: There were major differences in numbers of each part-of-speech type annotations between the dictionaries. In most cases ipadic provided more specific annotations (nouns, verbs, particles, auxiliary verbs, exclamations) than jumandic. For example, in ipadic annotation there were nearly 2 billions of nouns, while in jumandic only about 1,5 billion (see table 7 and its graphical visualization in figure 3 for details). The reason for these differences is that both dictionaries are based on different approaches to part-of-speech disambiguation. Jumandic was created using a set of hand crafted syntactic rules and therefore in a corpus as large as YACIS there are situations where no rule applies. On the other hand ipadic was created on a large corpus and thus provides disambiguation rules using contextual information. This is clearly visible when the category "other" is compared, which consists of such annotations as "symbols", or "unknownw words". The number of "other" annotations with jumandic is over two times larger than with ipadic and covers nearly 40% of the whole corpus. The detailed analysis also revealed more generic differences in word coverage of the dictionaries. Especially when it comes to abbreviations and casual modifications, some words do not appear in jumandic. For example, an interjection いや *iya* ("oh") appears in both, but its casual modification いやー *iyaa* ("ooh") appears only in ipadic. In this situation jumandic splits the word in two parts: いや and a vowel prolongation mark ー, which is annotated by jumandic as "symbol".

YACIS vs jBlogs and JENAAD: It is difficult to manually evaluate annotations on a corpus as large as YACIS[27]. However, the larger the corpus is the more statistically reliable are the observable tendencies of annotated phenomena. Therefore it is possible to evaluate the accurateness of annotations by comparing tendencies between different corpora. To verify part-of-speech tagging we compared tendencies in annotations between YACIS, jBlogs mentioned in section 2.1 and JENAAD [37]. The latter is a medium-scale corpus of newspaper articles gathered from the Yomiuri daily newspaper (years 1989-2001). It contains about 4.7 million words (approximately 7% of jBlogs and 0.08% of YACIS). The comparison of those corpora provided interesting observations. jBlogs and JENAAD were annotated with ChaSen, while YACIS with MeCab. However, as mentioned in section 4.1, ChaSen and MeCab in their default settings use the same ipadic dictionary. Although there are some differences in the way each system disambiguates parts of speech, the same dictionary base makes it a good comparison of ipadic annotations on three different corpora (small JENAAD, larger jBlogs and large YACIS). The statistics of parts-of-speech distribution is more similar between the pair YACIS(ipadic)–JENAAD (ρ = 1.0 in Spearman's rank setting correlation test) and YACIS(ipadic)–jBlogs (ρ = 0.96), than between the pairs YACIS(jumandic)–jBlogs (ρ = 0.79), YACIS(jumandic)–JENAAD (ρ = 0.85) and between both version of YACIS (ρ = 0.88).

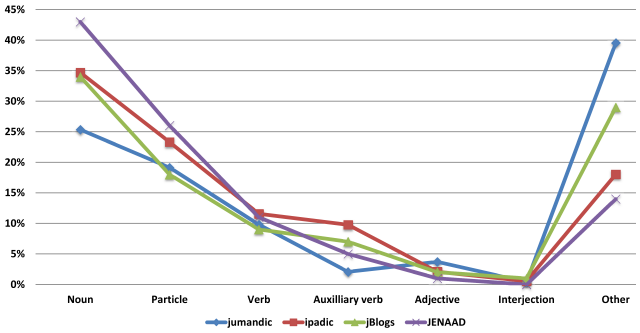

Figure 3. Graphical visualization of parts-of-speech comparison between YACIS (ipadic and jumandic annotations), Baroni&Ueyama's jBlogs and JENAAD.

Japanese vs British and Italian: As an interesting additional exercise we compared YACIS to Web corpora in different languages. In particular, we analyzed ukWaC and itWaC described in [10]. Although not all information on part-of-speech statistics is provided for those two corpora, the available information shows interesting differences between part-of-speech distribution among languages[28]. In all compared corpora the largest is the number of "nouns". However, differently to all Japanese corpora, second frequent part of speech in British English and Italian corpus was "adjective", while in Japanese it was "verb" (excluding particles). This difference is the most vivid in ukWaC. Further analysis of this phenomenon could contribute to the fields of language anthropology, and philosophy of language in general.

5.2 Affective Information

Evaluation of Affective Annotations: Firstly, we needed to confirm the performance of affect analysis systems on YACIS, since the per-

[27] Having one sec. to evaluate one sentence, one evaluator would need 11.2 years to verify the whole corpus (354 mil. sentences).

[28] We do not get into a detailed discussion on differences between POS taggers for different languages, neither the discussion on whether the same POS names (like noun, verb, or adjective) represent similar concepts among different languages (see for example [26] or [22]). These two discussions, although important, are beyond the scope of this paper.

Table 7. Comparison of parts of speech distribution across corpora (with percentage).

Part of speech	YACIS-ipadic		YACIS-jumandic		jBlogs	JENAAD	ukWaC	itWaC
	percentage	(number)	percentage	(number)	(approx.)	(approx.)		
Noun	34.69%	(1,942,930,102)	25.35%	(1,419,508,028)	34%	43%	1,528,839	941,990
Particle	23.31%	(1,305,329,099)	19.14%	(1,072,116,901)	18%	26%	[not provided]	[not provided]
Verb	11.57%	(647,981,102)	9.80%	(549,048,400)	9%	11%	182,610	679,758
Auxiliary verb	9.77%	(547,166,965)	2.07%	(115,763,099)	7%	5%	[not provided]	[not provided]
Adjective	2.07%	(116,069,592)	3.70%	(207,170,917)	2%	1%	538,664	706,330
Interjection	0.56%	(31,115,929)	0.40%	(22,096,949)	<1%	<1%	[not provided]	[not provided]
Other	18.03%	(1,010,004,306)	39.55%	(2,214,892,801)	29%	14%	[not provided]	[not provided]

formance is often related to the type of test set used in evaluation. ML-Ask was positively evaluated on separate sentences and on an online forum [33]. However, it was not yet evaluated on blogs. Moreover, the version of ML-Ask supported by CAO has not been evaluated thoroughly as well. In the evaluation we used a test set created by Ptaszynski et al. [34] for the evaluation of CAO. It consists of thousand sentences randomly extracted from YACIS and manually annotated with emotion classes by 42 layperson annotators in an anonymous survey. There are 418 emotive and 582 non-emotive sentences. We compared the results on those sentences for ML-Ask, CAO (described in detail in [34]), and both systems combined. The results showing accuracy, calculated as a ratio of success to the overall number of samples, are summarized in Table 8. The performance of discrimination between emotive and non-emotive sentences of ML-Ask baseline was a high 98.8%, which is much higher than in original evaluation of ML-Ask (around 90%). This could indicate that sentences with which the system was not able to deal with appear much less frequently on Ameblo. As for CAO, it is capable of detecting the presence of emoticons in a sentence, which is partially equivalent to detecting emotive sentences in ML-Ask. The performance of CAO was also high, 97.6%. This was due to the fact that grand majority of emotive sentences contained emoticons. Finally, ML-Ask supported with CAO achieved remarkable 100% accuracy. This was a surprisingly good result, although it must be remembered that the test sample contained only 1000 sentences (less than 0.0003% of the whole corpus). Next we verified emotion class annotations on sentences. The baseline of ML-Ask achieved slightly better results (73.4%) than in its primary evaluation [33] (67% of balanced F-score with P=85.7% and R=54.7%). CAO achieved 80.2%. Interestingly, this makes CAO a better affect analysis system than ML-Ask. However, the condition is that a sentence must contain an emoticon. The best result, close to 90%, was achieved by ML-Ask supported with CAO. We also checked the results when only the dimensions of valence and activation were taken into account. ML-Ask achieved 88.6%, CAO nearly 95%. Support of CAO to ML-Ask again resulted in the best score, 97.5%.

Table 8. Evaluation results of ML-Ask, CAO and ML-Ask supported with CAO on the test set.

	emotive/ non-emotive	emotion classes	2D (valence and activation)
ML-Ask	98.8%	73.4%	88.6%
CAO	97.6%	80.2%	94.6%
ML-Ask+CAO	100.0%	89.9%	97.5%

Table 9. Statistics of emotive sentences.

# of emotive sentences	233,591,502
# of non-emotive sentence	120,408,023
ratio (emotive/non-emotive)	1.94
# of sentences containing emoteme class:	
- interjections	171,734,464
- exclamative marks	89,626,215
- emoticons	49,095,123
- endearments	12,935,510
- vulgarities	1,686,943
ratio (emoteme classes in emotive sentence)	1.39

Table 10. Emotion class annotations with percentage.

emotion class	# of sentences	%	emotion class	# of sentences	%
joy	16,728,452	31%	excitement	2,833,388	5%
dislike	10,806,765	20%	surprize	2,398,535	5%
fondness	9,861,466	19%	gloom	2,144,492	4%
fear	3,308,288	6%	anger	1,140,865	2%
relief	3,104,774	6%	shame	952,188	2%

Statistics of Affective Annotations: At first we checked the statistics of emotive and non-emotive sentences, and its determinant features (emotemes). There were nearly twice as many emotive sentences than non-emotive (ratio 1.94). This suggests that the corpus is biased in favor of emotive contents, which could be considered as a proof for the assumption that blogs make a good base for emotion related research. When it comes to statistics of each emotive feature (emoteme), the most frequent class were interjections. This includes interjections separated by MeCab (see Table 7) and included in ML-Ask database. Second frequent was the exclamative marks class, which includes punctuation marks suggesting emotive engagement (such as "!", or "??"). Third frequent emoteme class was emoticons, followed by endearments. As an interesting remark, emoteme class that was the least frequent were vulgarities. As one possible interpretation of this result we propose the following. Blogs are social space, where people describe their experiences to be read and commented by other people (friends, colleagues). The use of vulgar language could discourage potential readers from further reading, making the blog less popular. Next, we checked the statistics of emotion classes annotated on emotive sentences. The results are represented in Table 10. The most frequent emotions were joy (31%), dislike (20%) and fondness (19%), which covered over 70% of all annotations. However, it could happen that the number of expressions included in each emotion class database influenced the number of annotations (database containing many expressions has higher probability to gather more annotations). Therefore we verified if there was a correlation between the number of annotations and the number of emotive expressions in each emotion class database. The verification was based on Spearman's rank correlation test between the two sets of numbers. The test revealed no statistically significant correlation between the two types of data, with ρ=0.38.

Comparison with Other Emotion Corpora: Firstly, we compared YACIS with KNB. The KNB corpus was annotated mostly for the need of sentiment analysis and therefore does not contain any infor-

Table 11. Comparison of positive and negative sentences between KNB and YACIS.

		positive	negative	ratio
KNB*	emotional attitude	317	208	1.52
	opinion	489	289	1.69
	merit	449	264	1.70
	acceptation or rejection	125	41	3.05
	event	43	63	0.68
	sum	1,423	865	1.65
YACIS (ML-Ask)**	only	22,381,992	12,837,728	1.74
	only+ mostly	23,753,762	13,605,514	1.75

* p<.05, ** p<.01

Table 12. Comparison of number of emotive expressions appearing in three different corpora with the results of Spearman's rank correlation test.

	Minato et al.	YACIS	Nakamura
dislike	355	14,184,697	532
joy	295	22,100,500	224
fondness	205	13,817,116	197
sorrow	205	2,881,166	232
anger	160	1,564,059	199
fear	145	4,496,250	147
surprise	25	3,108,017	129
	Minato et al. and Nakamura	**Minato et al. and YACIS**	**YACIS and Nakamura**
Spearman's ρ	0.88	0.63	0.25

mation on specific emotion classes. However, it is annotated with emotion valence for different categories valence is expressed in Japanese, such as *emotional attitude* (e.g., "to feel sad about X" [NEG], "to like X" [POS]), *opinion* (e.g., "X is wonderful" [POS]), or *positive/negative event* (e.g., "X broke down" [NEG], "X was awarded" [POS]). We compared the ratios of sentences expressing positive to negative valence. The comparison was made for all KNB valence categories separately and as a sum. In our research we do not make additional sub-categorization of valence types, but used in the comparison ratios of sentences in which the expressed emotions were of only positive/negative valence and including the sentences which were mostly (in majority) positive/negative. The comparison is presented in table 11. In KNB for all valence categories except one the ratio of positive to negative sentences was biased in favor of positive sentences. Moreover, for most cases, including the ratio taken from the sums of sentences, the ratio was similar to the one in YACIS (around 1.7). Although the numbers of compared sentences differ greatly, the fact that the ratio remains similar across the two different corpora suggests that the Japanese express in blogs more positive than negative emotions.

Next, we compared the corpus created by Minato et al. [29]. This corpus was prepared on the basis of an emotive expression dictionary. Therefore we compared its statistics not only to YACIS, but also to the emotive lexicon used in our research (see section 4.2 for details). Emotion classes used in Minato et al. differ slightly to those used in our research (YACIS and Nakamura's dictionary). For example, they use class name "hate" to describe what in YACIS is called "dislike". Moreover, they have no classes such as excitement, relief or shame. To make the comparison possible we used only the emotion classes appearing in both cases and unified all class names. The results are summarized in Table 12. There was no correlation between YACIS and Nakamura (ρ=0.25), which confirms the results calculated in previous paragraph. A medium correlation was observed between YACIS and Minato et al. (ρ=0.63). Finally, a strong correlation was observed between Minato et al. and Nakamura (ρ=0.88), which is the most interesting observation. Both Minato et al. and Nakamura are in fact dictionaries of emotive expressions. The dictionaries were collected in different times (difference of about 20 years), by people with different background (lexicographer vs. language teacher), based on different data (literature vs. conversation) assumptions and goals (creating a lexicon vs. Japanese language teaching). The only similarity is in the methodology. In both cases the dictionary authors collected expressions considered to be emotion-related. The fact that they correlate so strongly suggests that for the compared emotion classes there could be a tendency in language to create more expressions to describe some emotions rather than the others (dislike, joy and fondness are often some of the most frequent emotion classes). This phenomenon needs to be verified more thoroughly in the future.

6 CONCLUSIONS AND FUTURE WORK

In this paper we presented our research on the creation and annotation of YACIS, a large scale corpus of Japanese blogs compiled for the need of research in NLP and Emotion Processing in Text. We developed a set of tools for corpus compilation and successfully compiled the corpus from Ameblo blog service and annotated it with syntactic and affective information.

The syntactic information we annotated included tokenization, parts of speech, lemmatization, dependency structure, and named entities. The annotated corpus was compared to two other corpora in Japanese, and additionally to two corpora in different languages (British English and Italian). The comparison revealed interesting observations. The three corpora in Japanese, although different in size, showed similar POS distribution, whereas for other languages, although the corpora were comparable in size, the POS distribution differed greatly. We plan to address these differences in more detail in the future.

The affective information annotated on YACIS included emotion classes, emotive expressions, emotion valence and activation. The systems used in the annotation process include ML-Ask, a system for affect analysis of utterances and CAO, a system for affect analysis of emoticons. The evaluation on a test sample of annotations showed sufficiently high results. The comparison to other emotion corpus showed similarities in the ratio of expressions of positive to negative emotions and a high correlation between two different emotive expression dictionaries.

Although some work still needs to be done, YACIS corpus, containing over 5.6 billion words, is a valuable resource and could contribute greatly to numerous research, including research on emotions in language, sentiment and affect analysis.

YACIS corpus is meant to be used for pure scientific purposes and will not be available on sale. However, we are open to make the corpus available to other researchers after specifying applicable legal conditions and obtaining full usage agreement. In the near future we will release an additional n-gram version of the corpus to be freely accessible from the Internet without limitations and provide a demo viewable online allowing corpus querying for all types of information.

Acknowledgment

This research was supported by (JSPS) KAKENHI Grant-in-Aid for JSPS Fellows (Project Number: 22-00358).

REFERENCES

[1] Eugene Charniak, Don Blaheta, Niyu Ge, Keith Hall, John Hale and Mark Johnson. 2000. "BLLIP 1987-89 WSJ Corpus Release 1", Linguistic Data Consortium, Philadelphia, http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2000T43

[2] Chikara Hashimoto, Sadao Kurohashi, Daisuke Kawahara, Keiji Shinzato and Masaaki Nagata. 2011. "Construction of a Blog Corpus with Syntactic, Anaphoric, and Sentiment Annotations" [in Japanese], *Journal of Natural Language Processing*, Vol 18, No. 2, pp. 175-201.

[3] Kazuyuki Matsumoto, Yusuke Konishi, Hidemichi Sayama, Fuji Ren. 2011. "Analysis of Wakamono Kotoba Emotion Corpus and Its Application in Emotion Estimation", *International Journal of Advanced Intelligence*, Vol.3,No.1,pp.1-24.

[4] Saima Aman and Stan Szpakowicz. 2007. "Identifying Expressions of Emotion in Text". In *Proceedings of the 10th International Conference on Text, Speech, and Dialogue (TSD-2007)*, Lecture Notes in Computer Science (LNCS), Springer-Verlag.

[5] Changqin Quan and Fuji Ren. 2010. "A blog emotion corpus for emotional expression analysis in Chinese", Computer Speech & Language, Vol. 24, Issue 4, pp. 726-749.

[6] Irena Srdanovic Erjavec, Tomaz Erjavec and Adam Kilgarriff. 2008. "A web corpus and word sketches for Japanese", *Information and Media Technologies*, Vol. 3, No. 3, pp.529-551.

[7] Marco Baroni and Motoko Ueyama. 2006. "Building General- and Special-Purpose Corpora by Web Crawling", In *Proceedings of the 13th NIJL International Symposium on Language Corpora: Their Compilation and Application*, www.tokuteicorpus.jp/result/pdf/2006_004.pdf

[8] Taku Kudo and Hideto Kazawa. 2009. "Japanese Web N-gram Version 1", Linguistic Data Consortium, Philadelphia, http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2009T08

[9] Michal Ptaszynski, Rafal Rzepka and Kenji Araki. 2010. "On the Need for Context Processing in Affective Computing", In *Proceedings of Fuzzy System Symposium (FSS2010)*, Organized Session on Emotions, September 13-15.

[10] Marco Baroni, Silvia Bernardini, Adriano Ferraresi, Eros Zanchetta. 2008. "The WaCky Wide Web: A Collection of Very Large Linguistically Processed Web-Crawled Corpora", Kluwer Academic Publishers, Netherlands.

[11] Sasaki Y., Isozaki H., Taira H., Hirao T., Kazawa H., Suzuki J., Kokuryo K., Maeda E., "SAIQA : A Japanese QA System Based on a Large-Scale Corpus" [in Japanese], IPSJ SIG Notes 2001(86), pp. 77-82, 2001-09-10, Information Processing Society of Japan (IPSJ), http://ci.nii.ac.jp/naid/110002934347 (Retrieved in: 2011.11.11)

[12] Baayen, H. (2001) Word Frequency Distributions. Dordrecht: Kluwer.

[13] George K. Zipf (1935) The Psychobiology of Language. Houghton-Mifflin.

[14] George K. Zipf (1949) Human Behavior and the Principle of Least Effort. Addison-Wesley.

[15] Jan Pomikálek, Pavel Rychlý and Adam Kilgarriff. 2009. "Scaling to Billion-plus Word Corpora", In *Advances in Computational Linguistics*, *Research in Computing Science*, Vol. 41, pp. 3-14.

[16] James R. Curran and Miles Osborne, "A very very large corpus doesn't always yield reliable estimates", In Proceedings of the 6th Conference on Natural Language Learning (CoNLL), pages 126-131, 2002.

[17] Vinci Liu and James R. Curran. 2006. "Web Text Corpus for Natural Language Processing", In *Proceedings of the 11th Meeting of the European Chapter of the Association for Computational Linguistics (EACL)*, pp. 233-240.

[18] Peter D. Turney and Michael L. Littman. 2002. "Unsupervised Learning of Semantic Orientation from a Hundred-Billion-Word Corpus", National Research Council, Institute for Information Technology, *Technical Report ERB-1094*. (NRC #44929).

[19] Adam Kilgarriff, "Googleology is Bad Science", Last Words in: Computational Linguistics Volume 33, Number 1,

[20] Irena Srdanović Erjavec, Tomaž Erjavec, Adam Kilgarriff, "A web corpus and word sketches for Japanese", Information and Media Technologies 3(3), 529-551, 2008.

[21] Kilgarriff, A., Rychly, P., Smrž, P. and Tugwell, D., "The Sketch Engine", Proc. EURALEX. Lorient, France. 105-116, 2004.

[22] Jürgen Broschart. 1997. "Why Tongan does it differently: Categorial Distinctions in a Language without Nouns and Verbs." *Linguistic Typology*, Vol. 1, No. 2, pp. 123-165.

[23] Katarzyna Głowińska and Adam Przepiórkowski. 2010. "The Design of Syntactic Annotation Levels in the National Corpus of Polish", In *Proceedings of LREC 2010*.

[24] Peter Halacsy, Andras Kornai, Laszlo Nemeth, Andras Rung, Istvan Szakadat and Vikto Tron. 2004. "Creating open language resources for Hungarian". In *Proceedings of the LREC*, Lisbon, Portugal.

[25] Ichiro Hiejima. 1995. *A short dictionary of feelings and emotions in English and Japanese*, Tokyodo Shuppan.

[26] Paul J. Hopper and Sandra A. Thompson. 1985. "The Iconicity of the Universal Categories 'Noun' and 'Verbs'". In *Typological Studies in Language: Iconicity and Syntax*. John Haiman (ed.), Vol. 6, pp. 151-183, Amsterdam: John Benjamins Publishing Company.

[27] Daisuke Kawahara and Sadao Kurohashi. 2006. "A Fully-Lexicalized Probabilistic Model for Japanese Syntactic and Case Structure Analysis", *Proceedings of the Human Language Technology Conference of the North American Chapter of the ACL*, pp. 176-183.

[28] Radoslaw Komuda, Michal Ptaszynski, Yoshio Momouchi, Rafal Rzepka, and Kenji Araki. 2010. "Machine Moral Development: Moral Reasoning Agent Based on Wisdom of Web-Crowd and Emotions", *Int. J. of Computational Linguistics Research*, Vol. 1 , Issue 3, pp. 155-163.

[29] Junko Minato, David B. Bracewell, Fuji Ren and Shingo Kuroiwa. 2006. "Statistical Analysis of a Japanese Emotion Corpus for Natural Language Processing", *LNCS* 4114.

[30] Akira Nakamura. 1993. "Kanjo hyogen jiten" [Dictionary of Emotive Expressions] (in Japanese), Tokyodo Publishing, Tokyo, 1993.

[31] Michal Ptaszynski, Pawel Dybala, Wenhan Shi, Rafal Rzepka and Kenji Araki. 2009. "A System for Affect Analysis of Utterances in Japanese Supported with Web Mining", *Journal of Japan Society for Fuzzy Theory and Intelligent Informatics*, Vol. 21, No. 2, pp. 30-49 (194-213).

[32] Michal Ptaszynski, Pawel Dybala, Wenhan Shi, Rafal Rzepka and Kenji Araki. 2009. "Towards Context Aware Emotional Intelligence in Machines: Computing Contextual Appropriateness of Affective States". In *Proceedings of Twenty-first International Joint Conference on Artificial Intelligence (IJCAI-09)*, Pasadena, California, USA, pp. 1469-1474.

[33] Michal Ptaszynski, Pawel Dybala, Rafal Rzepka and Kenji Araki. 2009. "Affecting Corpora: Experiments with Automatic Affect Annotation System - A Case Study of the 2channel Forum -", In *Proceedings of the Conference of the Pacific Association for Computational Linguistics (PACLING-09)*, pp. 223-228.

[34] Michal Ptaszynski, Jacek Maciejewski, Pawel Dybala, Rafal Rzepka and Kenji Araki. 2010. "CAO: Fully Automatic Emoticon Analysis System", In *Proc. of the 24th AAAI Conference on Artificial Intelligence (AAAI-10)*, pp. 1026-1032.

[35] Michal Ptaszynski, Rafal Rzepka, Kenji Araki and Yoshio Momouchi. 2012. "A Robust Ontology of Emotion Objects", In *Proceedings of The Eighteenth Annual Meeting of The Association for Natural Language Processing (NLP-2012)*, pp. 719-722.

[36] James A. Russell. 1980. "A circumplex model of affect". *J. of Personality and Social Psychology*, Vol. 39, No. 6, pp. 1161-1178.

[37] Masao Utiyama and Hitoshi Isahara. 2003. "Reliable Measures for Aligning Japanese-English News Articles and Sentences". *ACL-2003*, pp. 72-79.

[38] Janyce Wiebe, Theresa Wilson and Claire Cardie. 2005. "Annotating expressions of opinions and emotions in language". *Language Resources and Evaluation*, Vol. 39, Issue 2-3, pp. 165-210.

[39] Theresa Wilson and Janyce Wiebe. 2005. "Annotating Attributions and Private States", In *Proceedings of the ACL Workshop on Frontiers in Corpus Annotation II*, pp. 53-60.

[40] Annie Zaenen and Livia Polanyi. 2006. "Contextual Valence Shifters". In *Computing Attitude and Affect in Text*, J. G. Shanahan, Y. Qu, J. Wiebe (eds.), Springer Verlag, Dordrecht, The Netherlands, pp. 1-10.

[41] T. Kudo "MeCab: Yet Another Part of Speech and Morphological analyzer", http://mecab.sourceforge.net/

[42] http://code.google.com/p/cabocha/

[43] Information Retrieval and Extraction Exercise, http://nlp.cs.nyu.edu/irex/index-e.html

[44] H. Kubota, K. Yamashita, T. Fukuhara, T. Nishida, "POC caster: Broadcasting Agent Using Conversational Representation for Internet Community" [in Japanese], Transactions of the Japanese Society for Artificial Intelligence, AI-17, pp. 313-321, 2002

[45] Okuno Yoo and Sasano Manabu, "Language Model Building and Evaluation using A Large-Scale Japanese Blog Corpus" (in Japanese), The 17th Annual Meeting of The Association for Natural Language Processing, 2011.

[46] Tony Berber Sardinha, José Lopes Moreira Filho, Eliane Alambert, "The Brazilian Corpus", American Association for Corpus Linguistics 2009, Edmonton, Canada, October 2009. http://corpusbrasileiro.pucsp.br

[47] Thorsten Brants, Alex Franz, "Web 1T 5-gram Version 1", Linguistic Data Consortium, Philadelphia, 2006, http://googleresearch.blogspot.com/2006/08/all-our-n-gram-are-belong-to-you.html

Emotion Valence Shifts in Humorous Metaphor Misunderstandings Generation

Pawel Dybala[1], Michal Ptaszynski[2], Rafal Rzepka[3], Kenji Araki[3] and Kohichi Sayama[4]

Abstract. In our previous work we proposed an idea of a system able to generate humorous metaphor misunderstanding during conversations with users, employing the mechanism of salience imbalance. However, according to existing research in the field of cognitive science, lexical salience imbalance might not be enough to constitute humorous metaphors. Another important factor in this process can be emotive salience imbalance, i.e. emotional shifts, which occur within metaphorical expressions. In this paper we propose how to employ this mechanism in our system, by implementing an emotion from text detector.

1 INTRODUCTION

Despite the fact that Turing Test [1] is often criticized for its inappropriateness, its impact on today science is undeniable. Alan Turing did not invent a golden mean to measure computer systems' human-likeness in terms of linguistic proficiency. He did, however, trigger scientists all over the world to investigate, what can and what should be done to create machines able to talk naturally.

This naturalness can be seen as the key not only to pass the Turing Test, but in general to create computer systems humans would like to interact with. Needless to say, pure grammatical correctness is by all means not enough to constitute natural interaction. Thus, we need to take into consideration also other aspects, which greatly influence this naturalness. In our research so far we focused on two such factors: emotions and humor. A summary of some of our work can be found at [2] i [3]. Currently we are working on a project in which we also plan to incorporate metaphor processing in human-computer interaction. To our best knowledge, no such system has been developed so far.

In this paper we first briefly summarize an idea of a humorous metaphor misunderstanding system HumMeR, which we proposed in our earlier work [4]. Next we mention a work of Shen and Engelmayer [5], which shows that humorous metaphors often include a sort of "emotional shift", which influences their funniness. Then we describe ML-Ask emotiveness analysis system, developed in our previous research [3, 6], which detects users emotions from text, and propose how it can be implemented into the HumMer system.

The research described in this paper is being conducted in Japanese, although the authors believe that most of the components that will be developed should be easily transferable to other languages. The research is text based, i.e. we focus on textual (and not visual or audial) aspects of conversation.

2 HUMMER SYSTEM

Proposed in our previous work [4], HumMer system is currently under development. It is designed to generate humorous metaphor misunderstandings during conversations with users. This algorithm was based on a conception of salience imbalance, commonly used to explain mechanisms working in metaphor understanding. Proposed by Ortony [7], it states that in metaphorical expressions certain highly salient properties of the metaphor source are matched with much less salient properties of metaphor target. In other words, certain properties of the target, which are normally perceived as not very salient, become more salient by comparing the common ground between the target and the source [7].

The salience imbalance theory was also showed by Shen and Engelmayer [5] to be applicable to humorous metaphors. Basing on results of experiments on humans, they showed that the degree of salience imbalance (the difference between salience of target properties and salience of source properties) should be higher in humorous than in non-humorous metaphors. In other words, salience imbalance is higher in these metaphors which include humor and are perceived as funny by humans.

In HumMeR system development we based on these findings. The input of the system is user's utterance, which is first analyzed to check if it includes any known metaphor, and, if not, if it fulfils the conditions allowing to assume that it can be a metaphorical expression. Then the system checks the salience imbalance between the concepts constituting the metaphor. This is done by using database of salience of concepts in existing metaphors as well as by querying the Internet to check co-occurrence of concepts and their descriptions (which can be seen as equivalent of salience). Next, the system recalculates the salience imbalance of the two concepts, i.e. it chooses another pair of concept properties, in which the difference in salience (salience imbalance) is higher than in the inputted expression. To do that, it uses a database of salience imbalance thresholds in humorous and non-humorous metaphors. Finally, the system uses the selected pair of properties to generate humorous metaphor misunderstanding including response to user utterance, using a database of templates commonly seen in such expressions.

The HumMeR system's algorithm outline is presented in Figure 3. The figure shows the flow of the novel metaphor

[1] JSPS Research Fellow / Otaru University of Commerce, Midori 3-5-21, 047-8501 Otaru, Japan. Email: paweldybala@res.otaru-uc.ac.jp
[2] JSPS Research Fellow / Hokkai-Gakuen University, Toyohira, Asahimachi 4-1-40, 062-0911 Sapporo, Japan.
Email: ptaszynski@media.eng.hokudai.ac.jp
[3] Graduate School of Information Science and Technology, Hokkaido University, Kita 14 Nishi 9, Kita-ku, 060-0814 Sapporo, Japan. Email: kabura@media.eng.hokudai.ac.jp
[4] Otaru University of Commerce, Department of Information and Management Science, Midori 3-5-21, 047-8501 Otaru, Japan.
Email: sayama@res.otaru-uc.ac.jp

processing procedure. If the metaphor in user's utterance is found to be an existing metaphor (i.e. it is found in our metaphor database), the system uses existing resources (like the salience database) to generate humorous misunderstandings.

Figure 3 also shows how emotiveness analysis can be implemented to facilitate HumMeR system's performance.

3 EMOTIONAL SHIFT IN HUMOROUS METAPHORS

In Section 2 we briefly summarized the salience imbalance theory and its applicability to humorous metaphors processing, which was experimentally showed by Shen and Engelmayer [5]. In the same work, however, the authors show also that extended degree of salience imbalance between the concept properties is not the only difference that occurs between humorous and non-humorous metaphors. Another important feature of the former is that they often include what Shen and Engelmayer call "a shift in emotional load of the two concepts" that constitute the metaphorical expression [5]. By this the authors understand that humoristic effect in metaphors (and in humorous contents in general) can be enhanced or even co-produced by a discrepancy between emotional valence (positive or negative) of two concepts that constitute the metaphor. For example, in the humorous metaphor:

"A friend is like an anchor – sometimes you want to throw them out of the boat." [5]

we can see that it joins two emotionally opposite properties, that are common for friends and anchors. An anchor-like friend, being a reliable and steady ally, is emotionally positive, while an idea of throwing a friend out of the boat is commonly associated as negative.

Shen and Engelmayer conducted an experiment, which results back up this claim. They investigated the degree of congruency between the emotional connotations of the two parts of humorous and non-humorous metaphors. The participants evaluated the sentence parts for their valence: positive, negative or neutral. The results showed that in most humorous metaphors a shift between positive and negative emotions occurred, while non-humorous metaphors rather tend to join emotionally similar concepts.

Thus, it can be stated that in order to generate humorous metaphors (or humorous metaphorical misunderstandings, as in our project), we should take into consideration also emotive valence of concepts and their properties. In order to do that, we need a tool that will allow us to assess sentences (or their parts) emotiveness. In HumMeR system, this role will be performed by Ptaszynski et al.'s ML-Ask Emotiveness Analysis System [3, 6].

4 ML-ASK EMOTIVENESS ANALYSIS SYSTEM

ML-Ask Emotiveness Analysis System was developed by Ptaszynski et al. [3, 6]. It which detects emotions from the textual layer of speech. Its algorithm is presented on Figure 1.

The system first analyses the inputted sentence to check its emotiveness. This is done by checking if it contains so-called "emotive elements". For example, the sentence:

"Kono hon saa, sugee kowakatta yo. Maji kowasugi!"

(That book, ya know, 'twas a total killer. It was just too scary.),

is recognized as emotive, as it contains emotive elements: *saa* (emphasis), *sugee* (totally), *yo* (emphasis), *maji* (really), *-sugi* (too much) and an exclamation mark. If the sentence was recognized as emotive, the system next detects emotion types it contains. This is done by checking if the sentence contains any "emotive expressions", i.e. expressions that convey particular emotions. For example, in the sentence above, the system found the emotive expression *kowai* (scary), which belongs to the group called *kyoufu* (fear).

If no such expression is recognized, the system uses Shi et al.'s web-mining technique [8] to extract emotive associations from the Internet. It first extract a phrase to be queried in the Internet, and transforms it to widen the search spec. If the phrase is, for instance, "it is hot today", the system would transform it into phrases like "it is hot today and...", "it is hot today, so..." etc. This procedure is called phrase modification. Next, the phrase and all its modified versions are queried in Yahoo to check its emotive associations by counting which emotive expressions follow it most often. This procedure is showed on Figure 2.

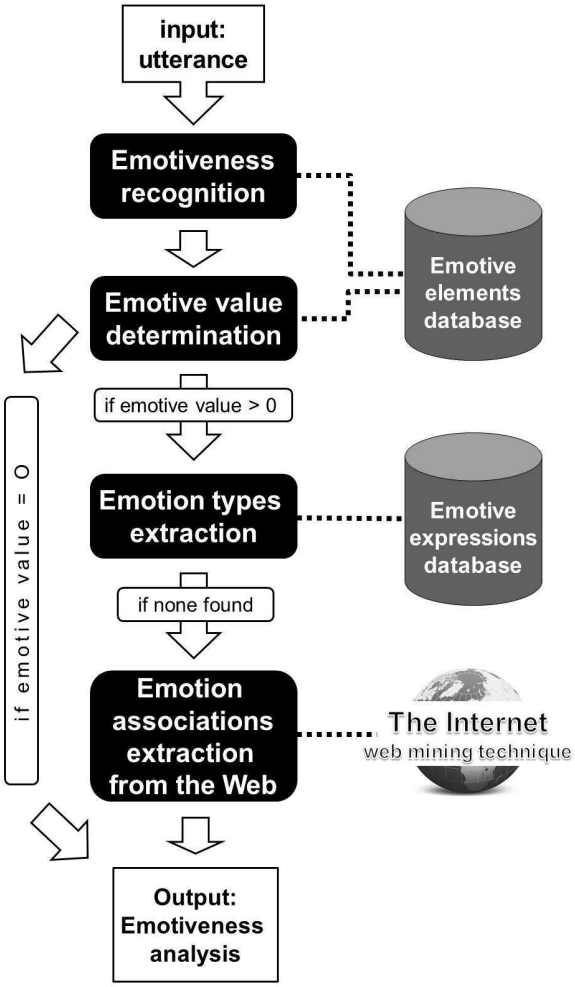

Figure 1: ML-Ask system algorithm outline

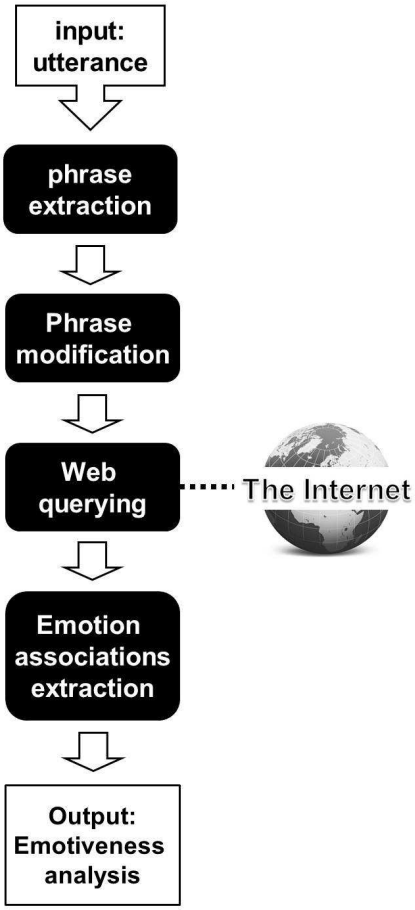

Figure 2: ML-Ask system – web mining procedure algorithm outline

As the result, we obtain an emotiveness analysis summary, such as below:

Sentence: *Kono hon saa, sugee kowakatta yo. Maji kowasugi!"* (That book, ya know, 'twas a total killer. It was just too scary.)

Emotive elements: *saa* (emphasis), *sugee* (totally), *yo* (emphasis), *maji* (really), *-sugi* (too much), exclamation mark
Emotive value: 6 (above zero -> specify types of emotions)
Emotive expressions: *kowai* (frightening)
Emotions found: fear
Valence: negative

Sentence: *Kyou wa atatakai desu ne.* (It's warm today, isn't it?)

Emotive elements: *-ne* (-isn't it)
Emotive value: 1 (above zero -> specify types of emotions)
Emotive expressions: none (-> use web mining procedure)
Emotions found on the Web: joy
Valence: positive

Performance of the ML-Ask system was tested in numerous evaluation experiments, which showed that it can successfully detect emotions from users utterances [3]. We also used it in our previous research on humor-equipped conversational systems [2], in which it also proved useful and usable. Thus, the ML-Ask system should be a proper tool to incorporate emotional shift in the HumMeR system's metaphor misunderstanding generation.

5 EMOTIONAL SHIFT IN HUMOROUS METAPHOR MISUNDERSTANDING GENERATOR

The role of ML-Ask system in the process of humorous metaphor misunderstanding generation will be to detect emotions present and associated with the candidates generated to create the misunderstandings. Next the system will assess each phrase's valence, which will allow to choose the pair in which emotional shift occurs. The outline of the system is shown on Figure 3.

The system's algorithm was explained in section 2. If, for example, user's utterance would be "a good friend is like an anchor", the system would presumably detect it as an existing metaphor and then it would extract salience of its components (descriptions of anchor and friend) from the database. Then, the system will query the Internet and offline corpora to extract common descriptions of these two concepts (anchor and friend) for which salience imbalance degree would be higher than in the inputted metaphor. These descriptions along with the concepts they belong to will then be analyzed by the ML-Ask system to check their emotional valence. In the above example, the system would check the valence of "good friend" and, if such description is generated, "throwing someone out of the boat". This will be done by querying the Internet for emotive associations, as described in section 4. Next, the system will check the valence of extracted emotive associations in order to choose the description with the opposite valence than the concept ("a good friend" is commonly associated with positive valence, while "throwing someone out of the boat" should be seen as rather negative). In the next step, the system would use metaphor misunderstanding templates database in order to generate a humorous response to user's utterance. In the above example, the response could be "Like an anchor? You mean, sometimes you want to throw him out of the boat?".

In the final stage of the HumMeR system development, we are planning to implement it into a chatterbot (see [4] for details). This will allow the system to place metaphor misunderstanding generation in daily conversations with users.

6 CONCLUSION AND FUTURE WORK

The HumMeR system project is currently under development. That said, we realize that achieving our goal may not be sufficient to create a system able to generate humorous metaphor misunderstanding in perfectly natural and human-like manner. There are numerous factors that will have to be taken into consideration in the future, such as proper timing of misunderstandings (i.e. deciding whether a metaphor should be answered by misunderstanding or not) or individual approach to every user. Some ideas on these aspects are given in [3] and [9].

Another important issue we will need to deal with in our research project is the evaluation of our system. To do that, we will use methodology proposed and tested in our earlier works (see [10] for summary).

ACKNOWLEDGEMENTS

This work was supported by KAKENHI (Project Number: 23-01348)

REFERENCES

[1] A. Turing. Computing Machinery and Intelligence. In: *Mind LIX* (236): 433–460 (1950).

[2] P. Dybala. *Humor to Facilitate HCI: Implementing a Japanese Pun Generator into a Non-task Oriented Conversational System*, Lambert Academic Publishing (2011).

[3] M. Ptaszynski. *Emotion Awareness in Dialog Agents: Affect Analysis of Textual Input Utterance and its Application in Human-Computer Interaction*. Lambert Academic Publishing (2011).

[4] P. Dybala, M. Ptaszynski, R. Rzepka, K. Araki, and K. Sayama. Beyond Conventional Recognition: Concept of a Conversational System Utilizing Metaphor Misunderstanding as a Source of Humor. In: *Proceedings of The 26th Annual Conference of The Japanese Society for Artificial Intelligence (JSAI 2012), Alan Turing Year Special Session on AI Research That Can Change The World.* Yamaguchi, Japan (to appear in June 2012).

[5] Y. Shen, and G. Engelmayer. A friend is like an anchor - sometimes you want to throw them out of the boat: What makes a metaphorical metaphorical comparison humorous?. Submitted for publication in Metaphor and Symbol (2012) (available at: http://www.tau.ac.il/~yshen/publications/%20friend%20is%20like%20an%20anchor.pdf)

[6] M. Ptaszynski, P. Dybala, R. Rzepka, and K. Araki. An Automatic Evaluation Method for Conversational Agents Based on Affect-as-Information Theory. In: *Journal of Japan Society for Fuzzy Theory and Intelligent Informatics, Special Issue on Emotions, Vol. 22*, No. 1: 73-89 (2010).

[7] A. Ortony. Beyond literal similarity. In: *Psychological Review, 86*(1): 161-180 (1979).

[8] W. Shi, R. Rzepka, and K. Araki. User Textual Input Using Causal Associations from the Internet. In: *Proceedings of FIT2008*, Fukusawa, Japan, 2008, pp. 267-268 (2008)

[9] P. Dybala, M. Ptaszynski, R. Rzepka, and K. Araki. Extending the Chain: Humor and Emotions in Human Computer Interaction. In: *International Journal of Computational Linguistics Research, Vol. 1*, No. 3: 116-128, Digital Information Research Foundation (2010).

[10] P. Dybala, M. Ptaszynski, R. Rzepka, and K. Araki. Evaluating subjective aspects of HCI on an example of a non-task oriented conversational system. In: *International Journal of Artificial Intelligence Tool., Special Issue on AI Tools for HCI Modeling 19*(6): 819-856 (2010).

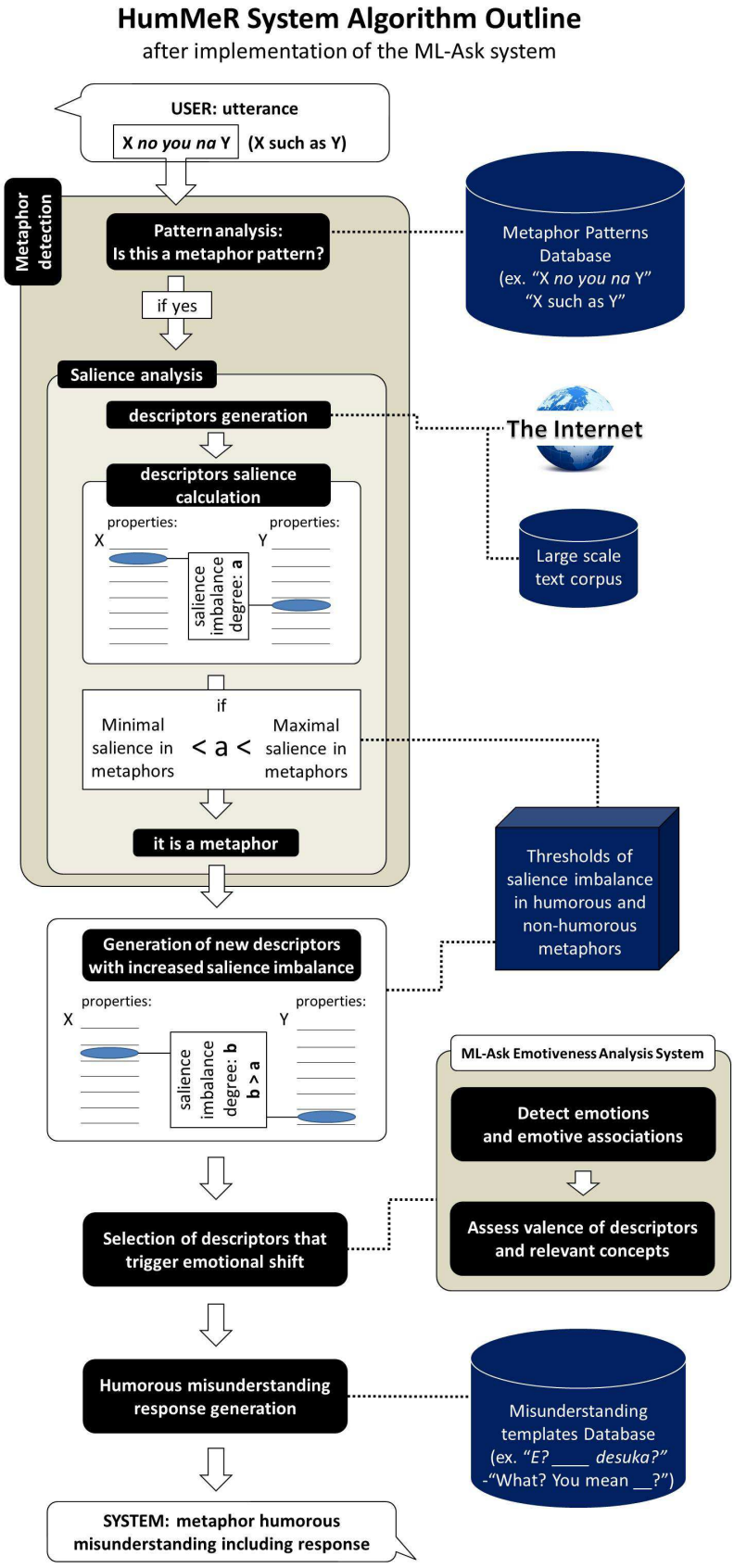

Figure 3: HumMeR humorous metaphor misunderstanding generator algorithm outline after implementation of the ML-Ask emotiveness analysis system

Affect Listeners - From dyads to group interactions with affective dialog systems

Marcin Skowron and **Stefan Rank** [1]

Abstract. Affect Listeners are applied as tools for studying the role of emotions in online communication. They need to interact both in dyads as well as in group settings with multiple users. In this paper, we present the evolution of such affective dialog systems from a focus on dyadic interaction to multi-party interaction on chat networks. Starting from experiments on the use of these dialog systems in virtual dyadic settings, we outline the requirements, design and implementation decisions necessary to apply the systems to affective interactions with multiple users. Finally, we introduce two realisations of Interactive Affective Bots designed for such interaction scenarios that integrate modelling of individuals and groups as part of their decision mechanism.

1 Introduction

The project CyberEmotions[2] deals with modelling and understanding of the role of *collective emotions* in creating, forming and breaking-up of online-communities. As part of the Cyberemotions project, the development and experimental evaluation of affective dialog systems that interact with users of network communication channels is undertaken [22, 19]. These systems serve two purposes: i.) a study tool for investigating the role of emotions in online communication and affective human-computer interaction, ii.) a support tool for e-communities providing online analysis, simulations and predictions for group dynamics, in particular addressing their affective dimension.

To date, Affect Listeners were applied in a range of experiments in dyadic settings which served to evaluate the systems' ability to participate in a realistic and coherent dialog and to establish and maintain an emotional connection with a user; and extended the understanding of the impact of affective system profiles and fine-grained communication scenarios on the self-reported emotional changes of users, their communication style and textual expressions of affective states. The next step, the application of such systems to group interactions created a new set of challenges related with, e.g., simultaneous communication with multiple users, capacities to interact in a way which intentionally follows or violates the typical communication patterns of members of a particular e-community, including the affective dimension of such interactions, or the ability to observe such a behaviour in other participants. These functionalities impact the system's ability to generate consistent or intentionally inconsistent interactive behaviour, the required affective coherence and the event-dependent adaptation of its communication patterns to other members in a group. In parallel, the system needs to represent and model discussions and emotional exchanges, at the individual and group levels, to provide the foundation for predicting the possible outcomes of the observed group dynamics, for simulating the effects of system's interactions with individuals or a group, and finally to assess the real effect of its interventions and to correspondingly update the used models.

To address the requirements related with transferring Affect Listeners from dyadic to multi-user interaction settings, the proposed approach integrated experience gathered in experiments in dyadic settings with insights acquired from a wide range of studies on the role of emotions in online communication: psychological studies and experiments on perception and generation of emotionally charged online content [11] [12] [14], agent based models of emotions [21] [5], valence trends [2], agent based model on bipartite networks [15], and event-based network discourse analysis [8]. Potential applications of such systems include support tools for online communities, e.g., providing information on the current affective state of groups, or forecasting the changes in groups' affective states or interaction dynamics.

In the remainder of the paper, we present the concept of Interactive Affective Bots (IAB) and an overview of experiments with one of the system's realization in dyadic settings. Next, based on the experiences with Affect Listener systems obtained in dyadic interaction settings and modelling of affective interactions in e-communities, we outline the requirements, design and implementation decisions necessary to apply the systems to affective interactions in multiple users environments. Finally, we introduce two realisations of IABs designed for such interaction scenarios that integrate modelling of individuals and groups as part of their decision mechanism. We conclude by discussing the relevance of the presented approach to the goals of the Turing Test, and discuss new challenges and opportunities related with the application of artificial systems for interactions and cooperation in online environments that include large number of users.

2 Interactive Affective Bots

The specifics of online, real-time and unrestricted interactions with a wide range of users influenced the selection of methods and design decisions in IAB. In particular, we aimed at: (i) robustness regarding erroneous natural language input, (ii) extensibility regarding system components and application scenarios. (iii) responsiveness; both for the generation of system responses and for simulations of individ-

[1] Austrian Research Institute for Artificial Intelligence (OFAI), Austria, email: {marcin.skowron, stefan.rank}@ofai.at

[2] http://www.cyberemotions.eu/ (all URLs last accessed 2012-04-05)

ual users' and collective emotions of the e-communities. Below we provide an overview of the main system components. For a detailed description of the system architecture and components used refer to [22, 24, 25]. At the top level layer, each realisation of IAB share the same structure, which includes Communication, Perception and Control layers presented in Fig 1.

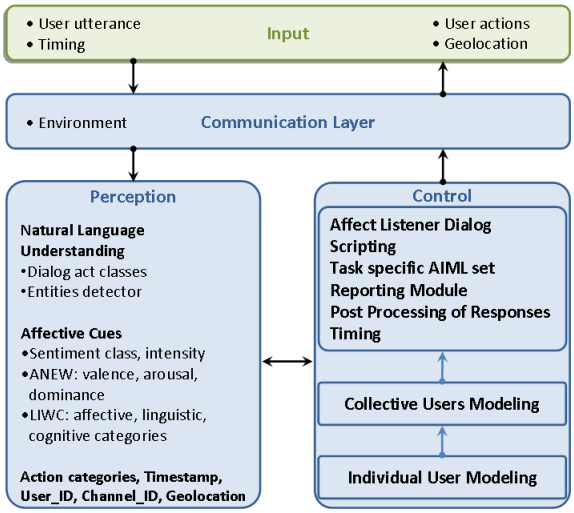

Figure 1. Interaction loop and generic architecture of Interactive Affective Bots

The Perception Layer, cf. [23], annotates both user utterances and system response candidates. This includes sentiment class and negative/positive sentiment strength [17]; valence, arousal and dominance [1]; various linguistic, cognitive and affective categories from the LIWC dictionary [18]; dialog act classes [23].

The Control Layer manages the dialog progression by relating observed dialog states to intended ones (e.g., querying and follow-up questions on the user's affective states, realizing a particular communication scenario) using the cues provided by the Perception Layer. This layer selects the system response from a number of generated response candidates, integrating rule-based action selection - Affect Listener Dialog Scripting (ALDS) - with the command interpreter for the task specific Affect Listener AIML-set[3]. As detailled in this paper, the control layer also integrates modelling of users, individuals as well as groups such as chat rooms, as part of its decision mechanism.

The Communication Layer handles the reception and dispatching of user/system utterances and provides the system with an interface to a range of interaction environments such as: Web Chat, 3D event engine[7], ICQ, XMPP (Jabber, Google or Facebook Chat).

The specification of an IAB includes the following layers of perception and interaction analysis:

1. Single utterance: annotation based on the Input Perception tool, a set of rules for generating system responses (Input Processing, AL-AIML [23])
2. Ongoing conversation: perception and analysis of the conversation context, tracking of effects of previous utterances (Input Processing, ALDS [22])
3. Individual and collective user modelling: long-term communication patterns or "personalities" of users regarding their textual expressions of affective states, characteristics of interaction patterns and sentiment expressions of groups of users and user populations (CybABMod - see section 4.4)

[3] Artificial Intelligence Markup Language (AIML)

2.1 Dyadic Interactions - Dialog Participant

The Dialog Participant realization of IAB is primarily applied for managing text-based communication with a user in an online, 1-on-1 interactive environment. It analyses and responds to the changes of the user's emotional state, i.e., textual expressions of affective states detected in the user's utterances. The typical objectives for the system in this interaction scenario include:

1. Realistic and coherent dialogs,
2. Conducive setting for communication (i.e. acquisition of large data sets),
3. Task-oriented dialogs related to "hot topics" in order to acquire users' affective states and stances towards the issues,
4. Studying the role of emotions in 1-on-1 HCI, e.g., the ability to consistently generate a particular affective profile and analysing its impact on users' self-reported emotional changes or textual expressions of affective states; or to convincingly realize a specific communication scenario, e.g., "getting acquainted with someone", "social sharing of emotions", throughout the whole time-span of communication with users and measuring their effects on users' communication patterns or their influence on system evaluation results.

3 Experimental results - Dialog Participant

System evaluation in a Wizard-of-Oz setting The first round of experiments was conducted in a Virtual Reality environment (see [7] and [24] for more details on the evaluation), where the dialog system was compared with a Wizard-of-Oz setting (WOZ)[4], in terms of its ability to: establish an emotional connection, dialog realism, and providing an enjoyable chatting experience. After each of the experimental interactions, participants were asked the following questions for assessing the dialog system, represented as a Virtual Human (VH):

1. Did you find the dialog with the VH to be realistic?
2. How did you enjoy chatting with the VH?
3. Did you find a kind of emotional connection between you and the VH?

The results achieved by the dialog system matched those obtained for the WOZ condition, i.e., there was no significant difference between the two settings.

Impact of system's affective profile We define *artificial affective profiles* as a coarse-grained simulation of a personality, corresponding to dominant, extroverted character traits, that can be consistently demonstrated by a system during the course of its interactions with users [25]. In a second round of experiments, three distinct affective profiles were implemented in the dialog system: positive, negative and neutral. Each affective profile aimed at a consistent demonstration of character traits of the system that could be described as:

- cooperative, emphatic, supporting, positively enhancing, focusing on similarities with a user - (positive),

[4] Participants believe that they communicate with a dialog system, while responses are actually provided by a human operator. In the presented experiments, the operator was asked to conduct a realistic and coherent dialog and provided free text input to user utterances.

- conflicting, confronting, focusing on differences - (negative),
- professional, focused on job, not responding to affective expressions - (neutral).

In these experiments, a browser-based communication interface, resembling a typical web chat-room environment was developed: a user input field at the bottom of the screen and a log of communication above. Participants interacted with the IAB in an unsupervised manner and were aware that they talk with an artificial system.

The results, presented in more details in [25], demonstrate that the implemented affective profiles to a large extent determined the assessment of the users' emotional connection and enjoyment from the interaction with the dialog systems, while the perception of core capabilities of the system, i.e. dialog coherence and dialog realism, were only influenced to a limited extent. Further, the self-reported emotional changes experienced by the experiment participants during the online interactions were strongly correlated with the type of applied profile. The affective profile also induced changes to various aspects of the conducted dialogs, e.g., communication style and the users' expressions of affective states. These results suggests that the participants, under this condition, assumed a more open, positive, sharing oriented attitude [26], which is also in line with the theory on Interpersonal Complementarity [13], which suggests that people in dyadic interactions negotiate their relationship through verbal and nonverbal cues, where dominant-friendliness invites submissive-friendliness whereas dominant-hostility invites submissive-hostility, and vice versa.

4 Scaling up to multiple users environments

4.1 Goals of IABs in multiple users environments

The experiments described above demonstrated that in the presented evaluation settings (VR or online web chat, relatively short interaction time) the system is able to conduct a realistic, enjoyable interaction and to establish an emotional connection with a user matching the results obtained in the WOZ setting. Further, the application of affective profiles showed an effect on self-reported emotional changes experienced by participants during the interaction with IABs, influencing also the textual expression of affective states and, to a smaller degree, the perception of core functionality of the artificial interlocutor, i.e., dialog realism and coherence.

The systems targeted at multi-user environments are primarily focused on supporting these e-communities by providing information - on demand or time-based - about a group's interaction dynamic, affective state, outcomes of the simulations regarding those parameters and exhibit relatively low activity levels in terms of direct communication with users. For groups, an IAB can track both content and affective dimensions of the communication between multiple users.

4.2 Role of simulations in IAB

The role of agent-based modelling and simulation in this kind of interactive system is two-fold: to provide part of the information provided to participants and to serve as decision support. Based on a request from a particular online community the tools can support it with analysis on affective dimension of their interactions and provide suggestions on ways for counter-acting negative tendencies observed in a group, e.g., a decrease of cooperation or growing hostility between members. Further, simulation results can indicate targets for interventions. This entails several requirements that concern both the results of simulation runs as well as runtime characteristics and the adaptability of the simulation based on data collected during previous interactions. At this level, several questions relevant as a potential input for the systems' decision making mechanisms were identified [19]:

- Which individual in a group will be most likely to provide an accurate response to probing about the group's emotional state, and which one will be most reliable?
- What influence can individuals have on the evolution of the collective emotions in an e-community, and which of the specific participants is likely to have the biggest influence?
- Can potential escalations, both in the negative and in the positive direction, be detected early on?
- What influence will a specific intervention of the system have at the current moment, and which style of intervention is most effective?

Running a simulation on demand to query about the above questions adds the requirement of timely, or possibly anytime, responses but also the need to parameterise the simulation to quickly adapt to the current state of an e-community, ideally using the recorded history as input.

An important part of the decision-making structures of IABs is the modelling of conversation participants. This component of the agent control structure is analogous to adaptive user modelling in standard Human-Computer Interaction: the system initially has a default model of the interaction partner, adapts it over time, and complements missing information based on the knowledge derived from interaction events. In the case of multi-user environments, this includes modelling several participants, simplifying the employed models and abstracting from specific individuals.

The modelling eventually serves the purpose of deciding on utterance selection, utterance modification, timing of utterances, and the selection of conversational partners in multi-user environments. As such, the main questions that modelling efforts helps to answer for the purposes of affective interactive systems are, from general to specific:

1. What potential influence will certain interventions have on the collective state of an online community?
2. What is the influence of particular interventions on the future development of a specific group?
3. What type of intervention (affective charge, topic, timing) will have which effect?
4. What relation does a particular individual have to the state of a specific group?
5. Which intervention is most appropriate when addressing a particular individual of a specific group?

4.3 Input from theoretical modelling and analysis

The general framework for modelling the emergence of collective emotions in IABs is based on the concept of Brownian agents [21]. This framework also serves as the reference point for modelling individual emotions. A promising effort is the study of emotional trajectories in online communities. Based on the analysis of Internet Relay Chat (IRC) data [5] [20] [5], candidate methods have been identified

[5] Analysed data-set included 2.5 million posts acquired from EFNET IRC chats: http://www.efnet.org, covering a range of topics including music, casual chats, business, sports, politics, computers, operating systems and specific computer programs.

for inclusion in IABs. Similar to the idea of following the norm behaviour of a channel, further channel characteristics regarding fluctuations of behaviour over time allow to use the concepts of "synchronisation" and "desynchronisation" with a dominant channel specific interaction pattern as decision factors. Synchronisation here refers to the relation of the affective content of system utterances to the affective content of users' communication behaviours, i.e., positive or negative valence, high or low arousal. The application of Detrended Fluctuation Analysis[10] and the use of the Hurst exponent[9] to classify user behaviour in the long-term provide the basis for a long-term variant of synchronisation between user behaviour and system behaviour. Applied to a channel as a whole rather than per user, it provides an additional input for decision making that aims for influencing the overall channel state into a particular direction.

As demonstrated in [2], the probabilities of submitting consecutive messages of the same emotional valence increase as power-law with the number of already inserted messages [6]. For the purpose of decision-making in IABs, this can be directly translated to utterance modification, using a different target valence depending on the number and dominant valence of previously seen utterances. With this approach the IABs can conform to or violate a typical behaviour on a specific online communication channel or observe such interaction patterns in other participants. Further, based on the analysis of dialogue regarding ending and entropy, it was shown that the average emotional value increases during the dialog and that the probabilities of positive and objective valences in the comments equalize in the vicinity of the discussion end.

The network mapping approch [6] analysis [7] account for the properties of activity patterns and underlying network topologies characteristic for various types of users, including those identified as important/influential in a given online interaction environment. In IABs, the spanning trees analysis can be specifically applied to analyze: i) user's activity, i.e.: by the creation and analysis of evolution of the network links, e.g., positive, negative, and ii.) users collective behaviour patterns. As demonstrated in high-resolution analysis of user-to-user communication in IRC channels, only certain links survive over one day period and support a particular type of network structure. Based on this observation, the presented system realizations and application scenarios presented in sections 5.1, 5.2 are primarily tageted at serving on-demand information requests of e-communities or individuals, i.e., establishing a relatively short-time direct communication links. As experimental evidence demonstrates (see section 3), such direct, limited in time interactions with users, also contribute to an overall higher level of the perceived realism, dialog coherence, the feeling of an emotional connection and chatting enjoyment. In practical application and deployment scenarios, this often translates to a higher acceptance rate of interactive systems in online communities.

Overall, in the case of online communication channels such as IRC, an IAB can use observed increase of entropy in dialogues as a guiding factor for decision making. The insertion of objective comments or of equalising comments can potentially be used to further different goals regarding the wanted discussion length, i.e. either extending it or ending it earlier. Finally, another helpful global characterisation is provided by the analysis of the length of emotional clusters, indicating presence or absence of collective emotions in the discussion.

[6] Analysed data-set included 4 million post acquired from Blogs, Digg and BBC forums.

[7] Analysis conducted on extensive data-sets from Ubuntu IRC channels: http://irclogs.ubuntu.com/

4.4 IAB - Modelling Individuals and Groups

Based on data on previous or the recent part of current interactions with a certain group (e.g., a chat room or a discussion channel), the agent control architecture of the IABs uses an online simulation model, *CybABMod* - Cyberemotions Agent-Based Modelling module, parameterised by the population and the history of the current channel, to derive particular models for the individuals it interacts with as well for the group as a whole. The output of such a simulation is a very short-term prediction, with a necessarily modest precision, of suitable candidates for interaction.

On the individual level, this model corresponds to the inference of specific "personalities" or personality types. These personalities are characterised by a collection of decision rules that abstract from previous interactions for that individual. A default personality is assigned to newcomers to a channel based on average behaviour of the channel so far. The simulation is updated online based on the tracked history and provides both short-term predictions as well as global attributes based on the theory input described above.

Representation of Networks and Interactions A prerequisite to use modelling and simulation as part of the online decision structure, i.e. while interacting, is a suitable and flexible representation of interactions with the specific group that the system faces. In order to provide for that, we developed online data structures using the common terminology introduced by [8] layered on top of the HDF5[8] disk and memory data format [27]. The latter provides a suitable framework for both logging and analysis of chat-specific data as well as for configuration and initialization of online simulation at runtime, both for several networks concurrently.

Varying Degrees of Affective Capabilities The baseline for the simulation of communicating agents or nodes uses stochastic modelling. One of the goals of this approach is to iteratively enhance the operationalisation of appraisal processes as described in theories of emotion [3, 4]. The minimum requirements for the modelling of appraisal processes in our nodes are representations of an agent's concerns or desires, including standards about praise- or blameworthy behaviour, as well as preferences for certain types of objects or situations. Further, a method for evaluating changes in an agent's environment based on these conditions is needed. For the purpose of e-communities, the changes to be evaluated encompass the posted messages and their content as far as it is modelled, but also the perceived entrance or exit of a participant in a discussion thread.

The modelling of a complete affective architecture is not the goal for these simulations. However, the introduction of specific *surface concerns*, i.e., concerns related to actual message exchanges as far as they can be observed, and a suitable approximation of an agent's evaluation processes can be used to account for observed behaviour in e-communities.

5 Realization of Interactive Affective Bots for Multiple User Environments

Below, we introduce two realisations of IABs that integrate modelling of individuals and groups as part of their decision mechanism.

[8] http://www.hdfgroup.org/HDF5

5.1 Affective Interaction Analyser

In the default settings for mult-user environments, such as e.g. IRC or Reddit[9], the Affective Interaction Analyser (AIA) focuses on the analysis of the interaction patterns, affective content of the exchanged textual messages between the discussion participants and the tracking of group-level attributes characteristic of the affective group dynamic (see section 4.3). The content of the collected messages feeds into the tracking of the current observed affective state of a group, is part of the input to the CybABMod simulation, and thus influences the predictions of the possible outcomes of ongoing interactions. Similarly, like in previous realizations of the system, the AIA's architecture includes three layers: perception, control and communication. The functionality of the layers was however extended to allow for simultaneous perception of users' actions in a multiple users environment. This functionality provides a base, both for the analysis of individual users activities and for modelling of the whole group. Based on the input from the environment (e.g. system messages for an IRC channel as well as the formatting of messages), the perception layer identifies users' IDs and the range of actions typically performed in an interaction environment, e.g., joining and leaving a channel, changing a nick-name, posting a link or utterance. This realization of the system, is the least active in terms of interactions with casual users. In these settings, the bot's interaction capabilities are typically limited to infrequent messages that can be provided to selected participants, in particular to the channel operators in the case of IRC channels or the subreddit moderators in the case of Reddit, e.g., on demand or based on a set interval or threshold set for the observed affective and interactive states of a group.

5.2 Affective Supporter and Content Contributor

Depending on the foreseen experimental settings and tasks for IAB, the activity level settings for the bots in an environment such as e.g. IRC or Reddit, can be set between the two above presented conditions "Dialog Participant" (highest activity level) and "Affective Interaction Analyser" (lowest activity level), enabling the Affective Supporter and Content Contributor (ASCC) to participate to a moderate extent in an ongoing discussion by providing both new content, related to the discussed topic (e.g., link to a relevant website) or the results of affective group dynamic analysis and real-time simulations. This realization of the system, relies on the architecture presented above, i.e., "Affective Interaction Analyser", extended to provide additional interactive capabilities targeted to the whole group such as posting comments (i.e., affect analysis based, relevant to the observed affective states of the group) or website links relevant to the ongoing discussion (i.e., content related [22]). In the Affective Supporter and Content Contributor scenario, the IAB combines the ability to directly respond to a range of events, such as:

- changes in the environment, e.g., a user joins/leaves the channel, posts a link or comment, (updating the interaction and group status, based on the set interval or threshold - sending messages to channel operator),
- changes in the affective state of the group, e.g., sudden decrease of the valence, increase of sentiment polarity in the posts exchanged between users,
- changes in the activity of the group. For example, the detection of a decrease of the participants activity might lead to emitting a message or posting a new link, or comment respectively, to a single user as selected by the simulation model. Further, in this scenario the interactive bot can also provide information about an event from the "offline" world related to the discussed topics, or emit questions aiming to stimulate the interactions,
- responding to utterances or comments emitted directly to the IAB by users.

[9] http://www.reddit.com

6 Conclusions and Discussion

To summarise, in the presented system applications, the system-user communication is text-based, real-time and oriented at the detection and acquisition of users affective states. Communication with the system is not limited to a specific domain, topic, or one particular ICT-mediated community. Naturally, interactive systems like these, are strongly limited in the sense that they cannot match the conversational abilities of a human, in particular in interaction scenarios that include long-term communication, and further which need to combine open- and closed-domain dialog and discourse processing. However, as the experimental evidence presented in section 3 demonstrates, in 1-on-1 communication settings and relatively short communication scenarios, i.e., chat sessions that are a few minutes long, the systems could match the WOZ results in terms of dialog realism, chatting enjoyment and the ability to establish an emotional connection with users. Further, the analysis of activity patterns and affective dimensions of users' communication in multi-user environments presented in section 4.3 showed that the majority of links are established only temporarily and primarly used to exchange relevant information, to share or respond to a sentiment expressed. These results support the proposed application scenarios where the systems establish communication links in a way similar to their human counterparts: on demand basis, and for a limited time-span. Consequently, the IABs communicate directly with users in situations where high confidence scores for a potential contribution's relevance, i.e. added informational or affective value - *contribution value*, can be foreseen. These estimates are based on the outcomes of the simulation of the reception of a specific content by a particular individual or a group. Additional *action costs* are associated with interaction scenarios where posts need to be emitted to a large number of participants or to the whole e-community.

Related to the classical Turing Test setup [28], the focus on other aspects than discussion content is sometimes used to deal with system confusion, e.g., system inability to respond based on analysis of semantic content, expected states in a dialog, pragmatic context or a simple detection of keywords. In the case of the presented systems, the focus on affective content is deliberate. This approach, i.e., generation of the selected system responses based on the detected affective states of individuals or groups can be seen as complementary to continuous extending and updating of knowledge bases necessary to respond to open-domain inputs. In a range of application scenarios, a pre-requisite for a successful application of such interactive systems could be the ability to adjust (or at least foresee the outcomes of an intentional violation of) one's communication behavior or affective stance according to: the overall mood detected in a group; individuals' preferences to various entities or fellow participants; the established or evolving "social norms"; or dynamic changes in a hierarchy of interaction patterns of users. Social intelligence also plays a role in a 1-on-1 interaction scenario, and as such is relevant for the Turing Test. However, the classical Turing Test was neither the primary goal of the Affect Listener systems nor required for them to fulfill their purpose. While the interaction is, as mentioned above, unrestricted, the domain is constrained in so far as the system concerns

itself mainly with the emotional states of individual participants as well as the dynamics of collective emotions and employs suitable strategies to keep the conversation or interaction between members going.

The setups in which an agent or group of agents interacts and cooperates with a large number of users provides new challenges but also offers new opportunities for the evolving intelligence and adaptability of the artificial systems. As demonstrated in nature, when the communities begin to evolve from a scenario of low cooperation, towards a more cooperative scenario, the more advanced solutions for intelligence are obtained. This is particularly relevant for the evolution of social intelligence: interactions that require indirect reciprocity, are cognitively demanding, or where individuals need to constantly monitor the social constellation of a group. Clearly, all of these factors - to a different degree - are present in different online communities, and need to be addressed to a possibly large extent, when envisaging new supportive roles for artificial agents in such multi-user settings. Such interaction settings also influenced the evolution of human language [16].

In this paper, we presented the design choices for using agent-based simulation as part of the decision mechanism of Affect Listeners. The use of the simulation as decision support for interactive affective systems adds different requirements including real-time and online use. Both of these connections contribute to the design of the simulation. The ALs are interactive affective systems that are also used for acquiring data, and for studying online interaction. The initial realisations of the systems were applied in dyadic experimental settings, demonstrating the ability to generate an enjoyable and realistic dialog on par with WOZ settings. Further, using the AL, we studied the role of affective profiles in dyadic settings. For future work, we are interested in measuring the effect of the interactions with the systems on participants' emotional (physiological) responses and to relate and align those with the textual expressions of users' affective states observable during interaction.

ACKNOWLEDGEMENTS

The work reported in this paper is partially supported by the European Commission under grant agreement CyberEmotions (FP7-ICT-231323). The Austrian Research Institute for AI is supported by the Austrian Federal Ministry for Transport, Innovation, and Technology.

REFERENCES

[1] M.M. Bradley and P.J. Lang, 'Affective norms for english words (anew): Stimuli, instruction manual and affective ratings', Technical Report C-1, The Center for Research in Psychophysiology, Univ. of Florida, (1999).

[2] A. Chmiel, J. Sienkiewicz, M. Thelwall, G. Paltoglou, K. Buckley, A. Kappas, and J.A. Holyst, 'Collective emotions online and their influence on community life', *PLoS ONE*, **6**(7), e22207, (2011).

[3] Phoebe C Ellsworth and Klaus R Scherer, 'Appraisal Processes in Emotion', in *Handbook Of Affective Sciences*, eds., Richard J Davidson, Klaus R Scherer, and H Hill Goldsmith, chapter 29, 572–595, Oxford University Press, Oxford New York, (2003).

[4] Nico H Frijda, *The Laws of Emotion*, Lawrence Erlbaum Associates Publishers, Mahwah NJ USA London UK EU, 2007.

[5] A. Garas, D. Garcia, M. Skowron, and F. Schweitzer, 'Emotional persistence in online chatting communities', *Scientific Reports*, **2**(402), (2012).

[6] V. Gligorijevic, M. Skowron, and B. Tadic, 'Evolving topology on the network of online chats', in *submitted*, (2012).

[7] S. Gobron, J. Ahn, S. Quentin, D. Thalmann, M. Skowron, S. Rank, G. Paltoglou, M. Thelwall, and A. Kappas, 'An interdisciplinary vr-architecture for 3d chatting with non-verbal communication', in *Proc. of EGVE 2011*, (2011).

[8] R. Hillmann and M. Trier, 'Sentiment polarization and balance among users in online social networks', in *Proc. ACIS 2012*, LNCS, (accepted, 2012).

[9] H. Hurst, 'Long-term storage capacity of reservoirs', *Transactions of the American Society of Civil Engineers*, **116**, 770–799, (1951).

[10] J. Kantelhardt, E. Koscielny-Bunde, H. Rego, S. Havlin, and A. Bunde, 'Detecting long-range correlations with detrended fluctuation analysis', *Physica A*, **295**, 441–454, (2001).

[11] A. Kappas, D. Kuester, M. Theunis, and E. Tsankova, 'Cyberemotions: Subjective and physiological responses to reading online discussion forums', in *50th Annual Meeting of the Society for Psychophysiological Research*, (2010).

[12] A. Kappas, E. Tsankova, M. Theunis, and D. Kuester, 'Cyberemotions: Subjective and physiological responses elicited by contributing to online discussion forums', in *51th Annual Meeting of the Society for Psychophysiological Research*, (2010).

[13] D. Kiesler, 'The 1982 interpersonal circle: A taxonomy for complementarity in human transactions', *Psychological Review*, **90**, 185–214, (1983).

[14] D. Kuester and A. Kappas, 'Measuring emotions in individuals and internet communities', in *Internet and Emotions*, eds., T. Benski and E. Fisher, Research Network of the European Sociological Association, (accepted, 2012).

[15] M. Mitrovic and B. Tadic, 'Patterns of emotional blogging and emergence of communities: Agent-based model on bipartite networks', *arXiv*, (2011).

[16] M. Nowak and K. Sigmund, 'Evolution of indirect reciprocity', *Nature*, **437**, 1291–1298, (2005).

[17] G. Paltoglou, S. Gobron, M. Skowron, M. Thelwall, and D. Thalmann, 'Sentiment analysis of informal textual communication in cyberspace', in *In Proc. Engage 2010, Springer LNCS State-of-the-Art Survey*, pp. 13–25, (2010).

[18] J. W. Pennebaker, M. E. Francis, and R. K. Booth, *Linguistic Inquiry and Word Count: LIWC 2001*, Erlbaum Publishers, 2001.

[19] S. Rank, 'Docking Agent-based Simulation of Collective Emotions to Equation-based Model and interactive agents', in *ADS 2010 Proceedings*, pp. 82–89. The Society for Modelling and Simulation International, (2010).

[20] F. Schweitzer, D. Garcia, A. Garas, R. Pfizner, and D. Tanasse. Deliverable 5.3. CyberEmotions project deliverables 2012, 2012.

[21] Frank Schweitzer and David Garcia, 'An agent-based model of collective emotions in online communities', *European Physical Journal B*, **77**(4), 533–545, (2010).

[22] M. Skowron, 'Affect listeners. acquisition of affective states by means of conversational systems', in *Development of Multimodal Interfaces - Active Listening and Synchrony*, Lecture Notes in Computer Science, pp. 169–181. Springer, (2010).

[23] M. Skowron and G. Paltoglou, 'Affect bartender - affective cues and their application in a conversational agent', in *IEEE Symposium Series on Computational Intelligence 2011, Workshop on Affective Computational Intelligence*. IEEE, (2011).

[24] M. Skowron, H. Pirker, S. Rank, G. Paltoglou, J. Ahn, and S. Gobron, 'No peanuts! affective cues for the virtual bartender', in *Proc. of the Florida Artificial Intelligence Research Society Conf.* AAAI Press, (2011).

[25] M. Skowron, S. Rank, M. Theunis, and J. Sienkiewicz, 'The good, the bad and the neutral: affective profile in dialog system-user communication', in *Proc. of ACII 2011*, LNCS, pp. 337–346, (2011).

[26] M. Skowron, M. Theunis, S. Rank, and A. Borowiec, 'Effect of affective profile on communication patterns and affective expressions in interactions with a dialog system', in *Proc. of ACII 2011*, LNCS, pp. 346–356, (2011).

[27] M. Trier, 'Towards dynamic visualization for understanding evolution of digital communication networks', *Information Systems Research*, **19**(3), 335–350, (2008).

[28] A. Turing, 'Computing machinery and intelligence', *Mind*, **59**(236), 433–460, (1950).

Chatterbots with Occupation – Between Non Task and Task Oriented Conversational Agents

Michal Mazur[1], **Rafal Rzepka**[1] **and Kenji Araki**[1]

Abstract. Chatterbots are computer systems specialized in simulating intelligent conversation. Dialog systems are often classified as task and non-task oriented: one focusing on the completion of particular tasks, and the other representing the less formal aspects of interaction with an system, such as free talking. In our research we aim to create a performing conversational system that would act as a language tutor. Performing such function means not only passing knowledge, but also assisting students in developing a variety of other attributes, as well as creating a necessary teacher – student bond. In this paper we would like to present an idea of a conversational system that would perform both tasks, reaching beyond the boundaries of existing dialog systems classification. First, we discuss the aspects of currently used chatterbot systems with a certain purpose. Then we address the existing problems of maintaining conversation with users and reacting to their emotional states in tutoring systems. We also present the current work on our tutoring system as an example of chatterbot with occupation. Finally we discuss the potential applications of such chatterbots for education and other fields.

1. INTRODUCTION

In 1950 Alan Turing started his paper with famous question: "Can machines think?" [28]. Soon after, the possibility for machines to think started to be explored by a newly founded scientific field of Artificial Intelligence (AI). A number of technologies that tried to attempt answering Turing's question have emerged in Computer Science over the following years. Diaogue systems are the example of technology that took an effort to make interaction between human and machine possible. Weizenbaum's ELIZA (1966) became the inspiration for linguistic researchers all over the world in creating a computer program that can understand and respond to the natural language [27]. Nowadays, conversational systems integrate computational linguistics techniques with the widely spread communication over the Internet to interpret and respond to statements made by users in natural language [1]. Brennan et al. describe chatterbots as "artificial constructs that are designed to converse with human beings using natural language as input and output" [2]. The typical usage involves receiving the user utterance in a given human language and providing the possibly reasonable or intelligent response to the given sentence. Then, the sequence is repeated as long as user keeps the conversation going. Since their early days, chatterbots have become more sophisticated along with maturing technology and able to respond to user utterance in both text and synthesized voice speech. Moreover, they evolved from simple text environment into modern visual interfaces that incorporate human-like avatars or virtual presence through augmented reality interfaces [3].

2. CHATTERBOTS: PAST AND PRESENT

Since their origin chatterbots were designed to perform certain functions or, as it could be said, perform certain roles or occupations. ELIZA, simple but revolutionary chatterbot that still influences the research in that field, used simple pattern matching algorithm to find a corresponding response to a given pattern. Despite of its simplicity, this program also performed a role, being a Rogerian psychotherapist able to answer users' questions without any traces of its own personality. ELIZA was later criticized for its lack of an internal world model that could influence and track conversation. Then, a chatterbot named Parry developed by Colby in 1975 simulated paranoid behaviour [4]. By using different tricks like admitting ignorance, changing the level of the conversation, rigidly and continuing previous topics it successfully fooled its human judges into believing that he is one of the patients showing typical paranoid behaviour. That sufficiently explained its unusual responses that would be attributed to illness and make its appearance real. But a real progress in chatterbot technology came in 1995 when Wallace introduced Artificial Linguistic Internet Computer Entity (ALICE)[3]. This chatterbot, often described as modern ELIZA, is an open source project that became the prototype of many current chatterbots. When ELIZA held no memory of the conversation and its knowledge base was embedded right into the code, ALICE introduced Artificial Intelligence Markup Language (AIML) used to store the knowledge-based data. This solution responded to the existing necessity of a bigger knowledge base. The chatterbot created by Dr. Wallace has won Loebner's annual competitions that awards prizes to the most human-like chatterbots, which is the realization of Turing Test.

Exploiting natural language techniques to build conversational systems has been used for a broad range of applications. Nowadays, dialogue systems are deployed on commercial websites to respond to customers' inquiries about provided services. They also answer questions about financial services or verbally demonstrate the company portfolios. Many companies, such as VirtuOz [2], use conversational systems for customer support. They also have a potential to assist the patients as virtual physicians/doctors [5], serve as bully and harassment advisors [6] or act as storytellers [7]. Some chatterbots are created purely for entertainment purposes, deployed in video games to inform players about in-game events [8] or as companions assisting players in a virtual environment [9].

[1] Graduate School of Information Science and Technology Hokkaido University, Sapporo, Japan; Email: {mazzi, kabura, araki}@media.eng.hokudai.ac.jp

[2] http://www.virtuoz.com/

[3] http://www.alicebot.org/

Finally, educational conversational systems assist students by providing guidance as they learn [10].

In recent times, chatterbots research has been drawn towards embodied conversational systems where body language and gestures are almost of the same importance as the natural language dialogue [3]. This new approach is motivated by necessity of making conversational systems appear more human like, thus become more engaging with the user. However, when a significant proportion of chatterbots research is concentrating on evaluating such interfaces, evaluating the quality of actual conversation seems to be a research challenge. Dybala et al. [24] mention that in case of task oriented systems it is much easier to achieve user satisfaction because of the very existence of task that requires an interaction between human and system to be completed. In case of non-task oriented dialogue, the interaction depends more on the content of the conversation.

3. TASK AND NON-TASK ORIENTED CLASSIFICATION

Chatterbots are often classified into two conventional groups. The first one often referred to as non-task oriented or free talking, considers the ability of program to chat freely on any given topic, without any obligations, mostly for entertainment purposes. In opposition to non-task oriented systems are task-oriented, specialized systems which put some limitations to system functionalities and usually aim to achieve specific goals. Most previously built conversational systems can engage in either task oriented dialogues to better understand human utterances or non-task oriented dialogues to allow users to enjoy the conversation. To our knowledge, even though there have been some attempts to provide such conversational systems, there are not many existing ones that would successfully employ both functions. A system presented by Nakano et al. [11] can dynamically change dialogue strategy based on speech recognition results. This system features different control modules called experts and each of it is dedicated to perform different kind of task. Non-task oriented functionality is incorporated into this model as one of the modules, the chat expert, which employs standard chatterbot technology (ALICE). This integrated system for both kinds of dialogues can adaptively change the strategies and more accurately respond to human utterances. However, the authors mentioned that it is important to improve chat expert and operate on a larger base of vocabulary – currently limited due to the problems with voice recognition accuracy.

Crocket et al. [12] presented an overview of methodology for constructing goal oriented conversational systems (GO-CA) with a mixed-initiative strategy (from time to time either human or chatterbot may take control of the conversation). By goal-oriented they understand a deep strategic purpose of conversation and directing it to achieve a certain goal, e.g. task oriented dialogues for services and non-task oriented dialogues for chat and entertainment. Authors express the concern that traditional chatterbots prolong the conversation with humans using pointless chat. Therefore, GO-CA attempts to get the conversation back on track using its rule base and knowledge obtained during the conversation. The authors presented two different chatterbots with different goals. The first one called Adam served as University Student Debt Advisor and was focused on providing advice to students. The main requirement of the system was to cope with upset students and deal with abusive language. During the usability evaluation with 200 undergraduate students the subjects were asked to provide their opinion about Adam. 75.0% of students were satisfied by the advice and 47.0% stated that they would use this system instead of going to a real advisor. The majority of student comments were positive, e.g. "(…) I felt like I was seeing a real advisor". Among the areas where Adam could be improved, subject mentioned better understanding of slang and mobile text talk. The current version of Adam has been run at the University and system continues to learn how to relate to life as a student. The latter one was used as an advisor on "bullying and harassment in the workplace" [11]. A complex policy was automated using a conversational system for advice to be available 24 hours a day. The chatterbot allows using natural language to discuss a person's individual cases and offer valuable advice. During the evaluation process 94.0% of responders indicated that they found the advice without difficulty and they didn't feel like consulting the human advisor.

We found these results to be consistent with the research on chatterbots as language learning tools by Fryer et al. [29]. The researchers conducted an experiment with 211 students participating in a sessions with two popular chatterbots ALICE and Jabberwacky[4]. According to the results, most of participants enjoyed using the chatterbots and generally felt more comfortable conversing with the chatterbots than a human language partner or teacher. Such positive foreign language communication experience could enhance students' interest in language learning and improve motivation. 74.0% of participants were asked to write about their experience, and they defined the communication with a system as "funny and entertaining". We can interpret those results with a claim that conversational systems offering both pedagogical functions and a chance of engaging, non-stressful conversation have potential to become artificial teachers. However, the authors did not provide conclusive data and there are still questions whether their results can be treated as reliable. Therefore, more conclusive results should be provided to confirm their claims. Of course a mere chatterbot system that correctly answers the students' questions is not enough to create the artificial representation of a teacher. Among the functionalities that should be considered is to provide conversational agent with a sense of humor and the ability to react to changing emotional states of students, as well as offering some representation of emotions from the side of conversational system in order to make it more human-like. By fulfilling this condition this technology has a potential to offer much more than just mere engaging dialogue.

4. CHATTERBOTS WITH OCCUPATION

From the beginning chatterbots performed different roles supplying counsel and service to others. We claim that just like their human counterparts, conversational systems tend to be

[4] http://www.jabberwacky.com/

specialized in their fields and aim to bring the high standards of professional and intellectual excellence. By the term "occupation" we describe the activity to which one a certain individual devotes oneself that also requires special knowledge or training in a given field. However, performing a specified function does not mean that chatterbot should be strictly task-oriented. In order to develop a system that could meet users' expectations the non-task oriented aspects of chatterbot should also be taken into consideration. We are not proposing a new type of chatterbot systems, but a new way they should be classified. After all, the ability to interact with others by asking a question, commenting and receiving an answer is the foundation of human communication. A mere task oriented conversation seems to be in opposition to that fundamental truth. According to Bogdanovych et al. [13] the purpose of increasing amount of chatterbots in e-commerce is to provide consumers with answers for their questions from an entity that resembles human.

Chatterbots have been widely used in the human-computer interaction and in learning they can potentially serve not only as task-oriented tools, but also as companions that would build students' self-confidence in foreign language communication. Non-task oriented functionality seems to be a great way to allow students a chance to practice their language skills after finishing the main assignment. They would have a chance to review a newly acquired material and have an opportunity to actually use it in practice. Students could also do the self-analysis by seeing the transcript of their conversation and finding/correcting the problematic sentences. A more advanced system could also do it automatically by preparing the summaries of most occurring mistakes and offer guidance how to correct them or dynamically change the teaching strategy to concentrate on filling the gaps in the knowledge of currently studied second language. Also for the teacher it is a chance to track progress of the students and see what language they use and spot the most common mistakes. Finally, by turning text-based conversation into audio they could read and listen at the same time. Creating a chatterbot with task and non-task oriented functions has one more potential advantage – it could be a good language tool for beginners to deal with specific language tasks as well as casual conversation partner for higher level students. As teachers, chatterbots should appear as human as possible to become believable teachers.

So what actually makes the ideal teacher? One of the possible answers has been brought in a detailed report that sets out the ideal educator [14]. According to the simplified definitions provided by surveyed 12-13 years old pupils a good teacher should: listen to students, encourage them, like teaching his/her subject, take time to explain things, help students when they are stuck, allow them to have a chance to speak, not give up on them, care on their opinion, take time to explain thing, be friendly and supportive ("tells you how you are doing"). There are still many possible drawbacks in bringing a full-fledged chatterbot able to serve as language instructor. Coniam [15] in an evaluation of language resources of a few popular chatterbots (including Jabberwacky, ALICE and Jenny[5]) for their potential in teaching English describes a few commonly occurring problems. According to this research, current chatterbots work best when they process one-clause sentences and the topic of conversation is common. Some of them can deal with misspellings but colloquial forms of English cause interaction problems, especially for younger learners accustomed in using this kind of language in everyday conversation. In general, if user inputs incorrectly spelt words, the chatterbot experiences problems or does not understand.

Therefore, a chatterbot that would suggest corrections to certain ungrammatical utterances is necessary for better communication with artificial language tutor. Teenagers often use colloquial and simplified forms of language in their everyday communication, so a successful teacher should be able to recognize such language and, if possible, suggest a correctly spelt form or more appropriate word. Second language students often use different audio-visual materials, i.e. movies, music, games, to improve their language skills. However, they may encounter inappropriate expressions and try to use them during the chatting sessions. Teaching chatterbot should be able to respond this issue in an appropriate way. Recently, different conversational systems have been developed for second language learning [16-17]. The one who has received the most attention is CSIEC (Computer Simulation in Educational Communication) is an Artificial Intelligence (AI) framework developed at Peking University [10]. This system is worth attention because it has been developed over the years and has been successfully integrated into selected Chinese classrooms where it has been used as a language tool to assist teachers during English classes. Even though CSIEC provides learners with a chatting partner, the creators of the system admit that there are still many unsolved problems regarding the ability of understanding and generating natural language or dealing with textual ambiguity.

5. OUR RESEARCH

In our research we aim to create a system that would perform both task and non-task oriented functions to give students not only a chance to expand their language knowledge, but also practice with a virtual language partner who could support their endeavors.

5.1. Baseline chatterbots

Currently, we are using two different baseline chatterbots. First, we have been adapting casual conversation system using modality and word association retrieved from the Web (Modalin), developed by Higuchi et al. [18] to create the baseline of our project. Modalin uses those words association to create propositions and automatically extracts sets of words related to conversation topic selected by user. It also adds the modality to generated utterance and verifies the semantic reliability. It extracts word associations in real time using Goo search engine [6] snippets and does not rely on any off-line database. After applying the extracted word associations into proposition templates, it checks its naturalness on the Internet. If the proposition has low result frequency, it is considered unnatural and system simply omits it to generate a new one in the analogic way.

[5] Jenny, http://juan.vhost.pandorabots.com/botmaster/en/home

[6] Goo search engine, http://www.goo.ne.jp

As a consequence of occurring problems with Modalin system, highly dependent on search engines to answer the user's utterance, we decided to also incorporate the other available chatterbot in our research. For each turn of the dialogue multiple queries has to be made and it leads to overuse of search engine and, in consequence, treating the system as potential spam sender. Therefore, we decided to use the second conversational agent that, unlike Modalin, uses limited amount of queries. A conversational system called Maru-chan [19] extracts keywords from the user's utterance (nouns and adjectives) and uses them to perform fewer queries in the Google search engine. Maru-chan does not only extract world associations from the Internet, but also n-grams, the strings of words containing sets of 3- and 4-grams for every word from the list. Next, these extracted n-grams and candidates are used in order to generate chatterbot response candidates. This system also checks the length of the candidate and automatically deletes the longest and the shortest candidate to deal with sentences that convey too little or too much information. The priority goes to the medium candidates that are scored according to manually set threshold (decided after a preliminary experiment) and the top one are selected as the agent's response.

Meanwhile, we are working on a new version of Modalin that would overcome the occurring problems and we aim to possibly integrate functionalities of both chatterbots to utilize their best features and enrich chatterbot system's conversational strategies.

5.2. System components

The architecture of our system consists of a few independent modules that cooperate with each other to perform the main task (teaching English) as well as the other, supportive functions, e.g. sentence correction, language normalization and responding to users' emotional states in an appropriate way (Fig. 1). The system is built upon the baseline chatterbot (Modalin) with other modules performing a supportive role during the input utterance analysis and generation of output utterance. The functions of each module are briefly described above.

We currently work on implementation of a context learning method based on the phenomenon of code mixing into chatterbot in order to generate code-mixed phrases. This is an innovative way to use this occurrence with conversational agents and to our best knowledge it has not been yet done in the field of computer science. Code mixing is the transition between linguistic units (words or phrases) from one language into another, within one sentence, where original grammar of the native language is usually preserved unchanged [20]. We aim to create the environment where students will be able to learn new language units by context in a natural way, just like children who learn language by connecting words to the people, things, and activities around them. Namely, during conversations with a chatterbot using the base language (Japanese) students will encounter words in second language. By establishing a connection between the meaning of a certain difficult word in the first (L1) and second (L2) language, students may understand the meaning without other references, e.g. dictionaries. For instance: "Let's have a lunch together in the *shokudou* (cafeteria)" (English-Japanese code-mixed phrase; a real life example).

The results of preliminary experiments described by Mazur et al. [21] indicate that code-mixed sentences may be an effective way of expanding one's L2 vocabulary. Our recent experiment concluded with high percentage of correct answers in a given language understanding test (82.5% and 79.1% for each test phase) also brought and insight that code mixing method may be an effective way of presenting new vocabulary units to the students [22]. Code mixing module will let users engage in both task and non-task oriented conversation. In a free conversation mode this system will automatically generate Japanese sentences with English words and allow users to elaborate on their meaning while conversing with a system.. The non-task oriented functionality will allow users to select vocabulary from a given list (currently nouns) and system will use them to generate sentences using Conceptnet[7], a free common sense knowledge base. Users will encounter both real commonsense sentences and fake sentences that do not convey the meaning of a studied word. In a sentence generation we also would like to use Wordnet[8] to check the semantic distance between response candidates and eliminate the ones which are too close and may interfere with our context learning approach.

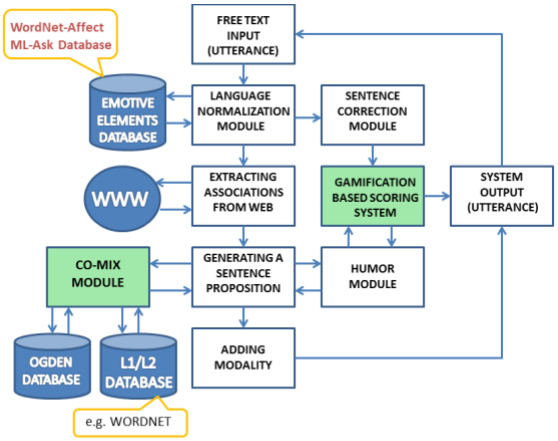

Fig. 1 System overview

Next, we plan to make use of an emotion recognition module that will recognize emotive sentences with ML-Ask Emotive Elements/Emotive Expressions Analysis System for Japanese, developed by Ptaszynski et al. [23]. This system determines utterances' emotiveness and detects types of emotions. The two main features are determining general emotiveness of a sentence (emotive/non emotive) and specifying the types of emotions. For example, the sentence "Kyo wan nante kimochi ii hi nanda! (Today is such a nice day!)!" is recognized as emotive, and contains emotive expression ("kimochi ii (nice)") as well as emotive elements ("nanda (emphasis)") and exclamation mark that make the utterance more emotive. In the second step, system checks the emotive values of the utterance and conduct the analysis of specific types of emotions. For example, the above example sentence contains an emotive expression "kimochi ii

[7] Conceptnet, http://web.media.mit.edu/~hugo/conceptnet/
[8] Wordnet, http://nlpwww.nict.go.jp/wn-ja/

(nice)" which belongs to the group called "yorokobi (joy)". This group is considered as positive, in opposition to negative emotion group, such as "iya (dislike)". Emotion recognition system will help us deal with shifting attitude of users towards the system, as well as the problem of keeping user engaged and eager to continue the discussion with conversational system.

The other aspect we take into consideration is humor that can be used to enhance positive and reduce negative engagement. We intend on using pun generation system made by Dybala et al. [24]. This system extracts a base word from a user utterance (mostly nouns) and transforms it using Japanese pun phonetic generation patterns to create candidate list. After checking the candidates in the Goo search engine and selecting the one with the highest hit rate it looks for a sentence with a chosen word and extracts its part to form a response or randomly selects a pun from its pun database. Dybala et al. also conducted a study on the role of engagement in non-task oriented conversation and the influence of humor as a measure to improve the engagement. The authors underline the fact that there is no comparative study concerning the role of engagement in conversation with task and non-task oriented systems. Finally, we would like to aid non-native speakers' reading comprehension of informal English with language normalization module based on a CEGS, a system for generating casual English short sentences from regular English input using a phonetic rule-based approach, developed by Clark et al. [25].

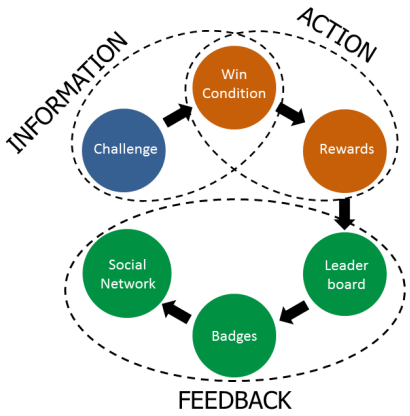

Fig. 2 Gamified point system

The other way we would like to deal with user willingness to enforce our system with a gamification module [26], dedicated to encourage user engagement and increase willingness to interact with a chatterbot. Gamification is an informal term for the use of elements of video game design in non-gaming systems. It is an innovative way of designing modern applications highly inspired by electronic entertainment. The chatterbot-based nature of our system requires other ways to encourage user engagement. Some of existing works on dialogue systems, especially the one by Jia at al. [10] show that the quality of conversation with chatterbot can be measured by its duration and general willingness of the user to interact with a system. According to this research, without a specific goal of the conversation user engagement slowly decreases. Since our initial target group consists of high school students, we decided to provide a scoring system inspired by gamification ideas, with a sole purpose to motivate users to learn foreign language and reward their efforts. Most of the schools use grades to motivate students and control their progress – high grades for excellent results, and low grades as a consequence of the lack of studying. There has been some controversy about whether grading students is an appropriate practice, because it may actively discourage them. We, on the other hand, believe that modern teaching system should offer more than just mere grading system. Using gamification ideas may bring the motivation and positive attitude of users towards the system. Therefore, such mechanism may encourage chatting with an agent and serve as a mean to keep user engaged and eager to study at longer intervals of time.

Currently used scoring systems, like the one presented by Jia et al., seem to be limited with their progress assessment methods and founded on simple grading rules, thereby allowing users to review their performance and scores only in a very basic way. Nevertheless, this work proves that such function is important to self-learning and evaluation. Therefore, we would like to introduce a more recent approach to this matter with a gamified point system that not only assess student progress, but also rewards them for their achievements and encourage doing more effort in studying. This point system is an iteration starting with a presenting a challenge for students (Fig. 2). Firstly, we explain rules of the game: present the challenge (e.g. acquiring and actively using newly learnt vocabulary units), specify winning conditions and rewards for reaching certain goals. Among the possible gratifications of user efforts the elements such as leaderboard and badges - commonly used game mechanics - will provide motivating competitiveness to users. Leaderboards show the top rank students and their best scores. Just like in video games, the urge to beat the best score and place one's own name among the top rank users has a positive motivation quality. Badges are commonly used in games and also effectively serve its purpose many popular social network applications, such as Foursquare[9]

Finally, the popularity of social media networks will give the opportunity to reward users with recognition of their achievement in the web community among their friends and relatives. A reward, such as dedicated badge, can be obtained automatically, after completing the necessary set of tasks.

6. CONCLUSIONS

This paper presented the insight on the roles the conversational systems play to benefit the modern society. We underlined the fact that existing informal classification of chatterbots is not sufficient enough to express the growing expectations of potential consumers. In a discussion about chatterbot occupation we emphasize the possible adaptation of conversational agents in learning. As an example, we described our work on an English tutoring system that presents a new approach to the subject by taking into consideration both aspects of task and non-task oriented conversation to face the student expectations towards a

[9] https://foursquare.com/

successful artificial teacher. We presented some examples showing that such system should perform both task and non-task oriented functions to become a useful language-learning tool and properly perform its designated role. We also discussed some existing problems in currently used chatterbots and presented some possible solutions to solve these issues. The originality of our project lays comes from the idea of using different modules performing separate functions, e.g. a module that use code mixing-based context method to generate sentences used in conversation with a chatterbot; gamification-inspired scoring system to encourage user engagement; a humor module that generates simple language jokes to deal with changing user attitude towards system and motivate; a language normalization module to improver comprehension of informal English; finally a module for affect analysis to generate and analyze emotive sentences.

A successful chatterbot must offer more than just performing its designated tasks, because potential users are expecting more human-like interlocutors that will not only answer their questions but also allow them to have a free conversation. Therefore, designing a new chatterbot should take into consideration both aspects of conversational systems and create a bridge between them in order to fulfil the expectations of potential users.

REFERENCES

[1] J. Lester, K. Bratning, B. Mott. Conversational Agents. CRC Press LLC. (2004).

[2] K. Brennan. The managed teacher, emotional labour, education and technology. Educational Insights 10(2), p. 55-56 (2006).

[3] M. Rehm, E. Andre. From chatterbots to natural interaction - Face to face communication with Embodied Conversational Agents. IEICE Transactions on Information and Systems, Special Issue on Life-Like Agents and Communication (2005).

[4] K. Colby. Artificial Paranoia: A Computer Simulation of Paranoid Process. Pergamon Press, New York, 1975.

[5] A. S. Lokman, J. M. Zain. Designing a Chatbot for diabetic patients. In: International Conference on Software Engineering & Computer Systems (ICSECS'09).

[6] A. Latham, K. Crockett, Z. Bandar. A Conversational Expert System Supporting Bullying And Harassment Policies. In: Proceedings of the 2nd International Conference on Agents and Artificial Intelligence, pp. 163–168 (2010).

[7] W. Chamberlain, The Policeman's Beard is Half Constructed. Warner Books, (1984).

[8] P. Hingston. A Turing Test for Computer Game Bots. IEEE TRANSACTIONS ON COMPUTATIONAL INTELLIGENCE AND AI IN GAMES, VOL. 1, NO. 3. (2009).

[9] V. Camilleri, M. Montebello. SLAVE – Second Life Assistant in a Virtual Learning Environment. RELIVE08 – Researching Learning in Virtual Environments. Milton-Keyes: The Open University. (2008).

[10] J. Jia. CSIEC: A Computer Assisted English Learning Chatbot Based On Textual knowledge and Reasoning. Elsevier B.V. (2009).

[11] M. Nakano, A. Hoshino, J.Takeuchi, Y. Hasegawa, T. Torii, K. Nakadai, K. Kato, H. Tsujino. A Robot That Can Engage in Both Task-Oriented and Non-Task-Oriented Dialogues. IEEE P. 404-411. (2006).

[12] K. Crockett, J. O'Shea, Z. Bandar. Goal Orientated Conversational Agents: Applications to Benefit Society. J. O'Shea et al. (Eds.): KES-AMSTA 2011, LNAI 6682, pp.16–25, (2011).

[13] A. Bogdanovych, S. Simoff, C. Sierra, H. Berger. Implicit training of virtual shopping assistants in 3D electronic institutions. In: Proceedings of the IADIS International e-Commerce 2005 Conference, Porto, Portugal, December 15-17. Lisbon: IADIS Press, 50–57. (2005).

[14] "What makes the ideal teacher". BBC NEWS. (2000/6/16) http://news.bbc.co.uk/2/hi/uk_news/education/793594.stm (accessed 2012/05/30).

[15] D. Coniam. Evaluating the language resources of chatbots for their potential in English as a second language. ReCALL 20(1): 98-116. 2008.

[16] G. Tatai, A. Csordas, A. Kiss, A. Szalo, L. Laufer. Happy chatbot, Happy User. Intelligent Virtual Agents, vol. 2792, 5-12.

[17] I. Stewart, P. File. Let's chat: A conversational dialogue system for second language practice. Computer Assisted Language Learning, 20, 97.116. (2007).

[18] Higuchi S., Rzepka R., Araki K. A casual conversation system using modality and Word associations retrieved from the Web. Proc.EMNLP `08. Pp.382-390, Honolulu, USA. (2008).

[19] M. Takahashi. Utterance Generation Method Using Web Search Results and Word n-Grams. Bachelor Dissertation, Hokkaido University, Japan, 2009.

[20] S. N. Sridhar, K. Kamal. The syntax and psycholinguistics of bilingual code-mixing. Canadian Journal of Psychology 34(4): 407-416 (1980).

[21] M. Mazur, R. Rzepka, K. Araki. Co-Mix Project: Towards Artificial Tutors Using Code Mixing as Foreign Language Teaching Method. IWMST 2010, pp.196-201 (2010).

[22] M. Mazur, R. Rzepka, K. Araki. Mixing Words and Emotions – New Natural Methods for Artificial Language Tutors. The 26th Annual Conference of the Japanese Society for Artificial Intelligence (2012). *To appear in the proceedings of The 26th Annual Conference of the Japanese Society for Artificial Intelligence (2012).*

[23] M. Ptaszynski, J. Maciejewski, P. Dybala, R. Rzepka, K. Araki. (2009). A System for Affect Analysis of Utterances in Japanese Supported with Web Mining. Journal of Japan Society for Fuzzy Theory and Intelligent Informatics, Vol. 21, No. 2 (April), pp. 30-49 (194-213).

[24] P. Dybala, M. Ptaszynski, R.Rzepka, Kenji Araki. Activating Humans with Humor – A Dialogue System That Users Want to Interact With. IEICE TRANS. INF & SYST. VOL.E92-D, NO.12 (2009)

[25] E. Clark, K. Araki. Text Normalization in Social Media: Progress, Problems and Applications for a Pre-Processing System of Casual English. Elsevier B.V (2011).

[26] M. Mazur, R. Rzepka, K. Araki. Proposal for a Conversational English Tutoring System that Encourages User Engagement. In: T. Hirashima et al. (Eds.) Proceedings of the 19th International Conference on Computers in Education. (ICCE2011), pp.10-12. (2011).

[27] J. Weizenbaum. ELIZA — A Computer Program For the Study of Natural Language Communication Between Man And Machine. Communications of the ACM 9 (1): 36–45 (1966).

[28] A. Turing. Computing Machinery and Intelligence. Mind LIX (236): 433–460 (1950).

[29] L. Fryer, R. Carpenter. Emerging technologies: Bots as language learning tools. Language Learning and Technology, 10(3). Pp. 8-14 (2006).

Multi-modal Belief Updates in Multi-Robot Human-Robot Dialogue Interactions

Gordon Briggs[1] and **Matthias Scheutz**[2]

Abstract. Humans working in teams typically use task-based natural language dialogues to coordinate activities. And they use mental models of team mates which they update automatically based on perceived and communicated information to predict the actions of their team mates. It is thus reasonable to assume that humans will expect future robots interacting with humans in natural language as part of mixed-initiative teams to exhibit the same kinds of belief modeling exhibited by humans.

In this paper, we propose principles that robots can use to represent beliefs and goals of other agents based on task-based natural language dialogues and use automatic inference based on communicated information to update their mental models of other agents. We demonstrate the proposed principles in a simple case study involving two robots and a human operator performing simple tasks in a laboratory environment.

1 Introduction

Mixed human-robot initiatives – teams that consist of both human and robotic team members – are widely seen as an important application domain for future autonomous robots. The goal in such teams is to utilize unique strengths of both humans and robots in order to accomplish joint goals. For example, NASA envisions space robots to help astronauts with the construction of planetary space stations. Or rescue robots in disaster areas are envisioned to aid human rescue workers in finding and retrieving wounded people. From a robotics perspective, the research challenge here is twofold: to provide the robotic capabilities necessary for a given task and to provide appropriate mechanisms for human-robot interactions that are effective and natural for humans.

While human teams typically use natural language to coordinate activities (such as discussing goals, developing plans, adjusting behaviors, etc.), mixed initiative teams are severely limited by current robots' cognitive limitations. Current robotic systems do not have the necessary modeling and inferencing capabilities for extensively emulating human mental models, nor do they have the natural language capabilities to engage in natural task-based dialogues, though progress is being made on these fronts [1]. Specifically, humans in teams are capable of (1) following multi-agent dialogues, (2) automatically updating their mental models of the involved agents based on the information communicated in natural language, and (3) automatically drawing inferences from the obtained information which may prompt them to confirm, augment or correct information and communicate those updates to their interlocutors effectively [6, 8]. If we want mixed initiative teams to interact in natural human-like ways, then robots will need mechanisms (built-in or learned) for performing mental and belief modeling and updating very much like humans.

In this paper, we propose simple belief update schemes for multi-agent dialogues that can be integrated into a cognitive architecture, thus allowing artificial agents to engage in more natural dialogues for activity coordination in mixed initiative teams. Specifically, we show how robots can use information gained from listening to dialogues among other agents to make inferences about those agents' belief states and goals, and how agents can use automatic inferences applied to mental models of other agents to better understand natural language directives and arrive at explicit goal representation (of their own and other agents' goals).

The paper is organized as follows. We begin with a few motivating mixed initiative scenarios where a human commander instructs autonomous robots to perform various tasks. These scenarios are intended to isolate several of the principles humans automatically employ in the context of teams and underscore the importance of belief-modeling mechanisms and perceptual integration. Then, we formalize the principles and describe our framework for belief modeling and updating, which also includes principles for inferring belief state based on particular natural language expressions. Next, we introduce the evaluation scenario and present the implementation details of the previously introduced principles in a distributed robotic architecture run on two robots. A particular dialogue interaction between the operator and the two robots then demonstrates the operation and utility of the proposed principles and framework. The subsequent discussion section addresses some of the challenges for larger, more capable systems while the conclusion summarizes our accomplishments and briefly touches on future work.

2 Motivation

Robotic systems are often well-suited for operating along-side or in place of humans, for example, in hazardous environments such as nuclear power plants or outer space. However, for humans to work with robots effectively, the interaction and cognitive capabilities of the robot become a critical factor, in addition to its physical characteristics and behavioral repertoire. While it is possible to devise special purpose interfaces that allow for teleoperation of robots or interactions with robots capable of limited autonomy, the much more natural case – from a human perspective – is where humans interact with *autonomous* robotic teams members as they would with other human team members: *using natural language*. This is particularly important in cases where typical human-machine communication

[1] Human-Robot Interaction Lab, Tufts University, Medford, MA, 02155, USA. gbriggs@cs.tufts.edu
[2] Human-Robot Interaction Lab, Tufts University, Medford, MA, 02155, USA. mscheutz@cs.tufts.edu

modalities are impractical [7]. For example, human searchers during rescue operations in disaster zones typically coordinate their activities through spoken natural language interactions via wireless audio links, while simultaneously occupying their eyes and hands with time-critical work. To efficiently interact with human team members, robotic team members would then also have to be capable of using spoken natural language.

However, being capable of using (rudimentary) spoken language alone is not sufficient because so much other cognitive activity is triggered in humans when humans engage in even simple dialogue exchanges. For example, human team members automatically form *mental models* of the beliefs and goals other team members have based on dialogue context and use those mental models to make inferences about other states that obtain (e.g., if a searcher says that she is done searching area X and has a goal to search another area, then she will likely leave X) and to adapt their natural language interactions (e.g., A telling B that A searched X will lead allow C to assume that B knows that A left X in questions like "Do you know where A is going next?"). Moreover, if A knows that B wants to be informed if a subtask is completed (e.g., area X search), A will automatically update B ("I finished searching X"). And B can then update C on A's activity if A is temporarily not reachable when C inquires about the status of area X; or B can correct C if C makes a statement that indicates a false belief (e.g., "Area X still needs to be searched" when A previously informed B that the search of area X was completed). Working in teams thus requires agents to monitor the dialogues and employ similar mental modeling and automatic model updates as in the human case. And it also requires similar automatic application of inference principles to communicated information.

To make these types of example more concrete in a simple robotic domain (as we will later use for evaluating our proposal mechanisms), consider a simple environment consisting of three navigation-points. A human-operator (O) is charged with coordinating in natural language the goals and behaviors of two autonomous robots, a quadrotor (Q1) and a ground-transport (T1) via radio. Let us use $at(\alpha, \lambda)$ to denote that agent α is located at nav-point λ and $B(\alpha, \phi)$ to denote that agent α believes that ϕ is true. Consider the following scenario:

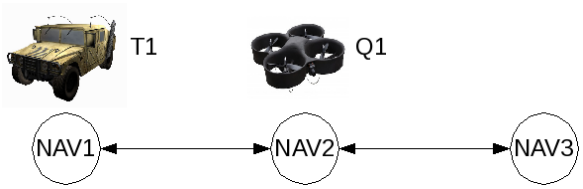

Figure 1. Sample environment for joint human-robot tasks.

```
O:  Transport 1, travel to Nav-point 3.
T1: Okay.
O:  Quadrotor 1, follow Transport 1.
Q1: Okay.
```

The scenario is illustrated in Figure 1. In order for the transport to get to Nav-point 3 in the above scenario, it must pass through Nav-point 2. Q1 can thus simply wait for T1 to show up at Nav-point 2 at which point it can follow it to Nav-point 3.

Note that autonomous agents must not only be able to update their beliefs to reflect the propositions communicated in utterances, but they must be able to compare these propositions to their own perceptions. If contradictions occur, agents must be able to ask for clarification or offer corrective statements as in the following example:

```
O:  Transport 1, are you at Nav-point 2?
T1: Yes, I am at Nav-point 2.
Q1: I do not see Transport 1.
```

In addition to the need for integrating the agent's perceptions into the dialogue system, a belief-modeling competency is necessary to generate the proper clarification or correction statement. Quadrotor 1's response would only make sense if $B(O, at(Q1, N2))$ was true. If $B(O, \neg at(Q1, N2))$, the quadrotor would have to produce a corrective utterance that implies $at(Q1, N2) \wedge \neg Sees(Q1, T1)$, such as "I am at Nav-point 2 and do not see Transport 1, " or "I do not see Transport 1 at Nav-point 2."

In sum, two important competencies must be present in a robotic agent for it to be able to communicate efficiently with human operators and team members: (1) the ability to build and maintain mental models or belief-model of the other agents (based on the synthesis of perceived, communicated, and inferred information) in order to maintain situational awareness; and (2) the ability to support the maintenance of other's mental models of oneself through communicating new information. For the rest of the paper we will use the term "mental model" in a technical sense to refer to the set of beliefs $B(\beta, \phi)$ agent α has about other agents β.

3 Belief Modeling and Dialogue

For robots to be able to engage in simple but natural sounding dialogues and automatically perform the types of belief modeling presented above, we need to add explicit rules that represent relationships among linguistic expressions as well as past and future beliefs. In particular, for task-based dialogue interactions we need to add rules that allow agents to reason about (1) the effects of perceptions, actions, and past beliefs on new/updated beliefs and (2) the effects of different utterance types (i.e., statement, questions, commands, and acknowledgments) on beliefs. The former includes all relevant properties of the agent in the world for it to be able to understand task-based dialogues; the latter includes all kinds of pragmatic implications of the employed utterance types both general (e.g., adverbial modifiers) as well as specific to the communicated context (e.g., the location predicate).

Indeed, these rules that enable belief-modeling of interlocutors are necessary to enable plan-based dialogue agents, which were first explored by Cohen and Perrault (1979) and Perrault and Allen (1980). Traum (1999) provides a review of dialogue agents enabled by the modeling of beliefs, desires, and intentions (BDI) and articulates the advantages of this approach, stating that modeling the changes communicative acts have on the mental models of agents "...allows an agent theorist or designer to place agent communication within the same general framework as agent action."

3.1 Agent Behavioral Rules

For the purposes of the employed example and the given space limitations, we make several simplifying assumptions: (1) we will not worry about employing generalizable and robust mechanisms for the translation from natural language expression to logical formulas for

the simple domains employed here (e.g., which we have done elsewhere [5]), we simply use pattern-matching to convert from natural language to logical forms; (2) we assume that all agents are truth tellers and never lie; (3) we assume that all agents immediately execute the most recent order and only that order (e.g., we have previously dealt with the more complex case of giving agents multiple possibly contradictory orders in natural language [12]); and (4) we make each agent first utter its name and then the name of the addressee so that it is easy for each agent to determine the speaker and the intended listener based on the linguistic information alone.

These rules include facts about agent behaviors that other agents can use to predict the other agents' behaviors. The first rule is concerned with an agent's perceptual system which is taken to automatically generate beliefs about what it perceives. In particular, if an agent α perceives the presence of another agent β at location λ, then it generates the belief that $B(\beta, \lambda)$.

$$Perceives(\alpha, at(\beta, \lambda)) \Rightarrow B(\alpha, at(\beta, \lambda)) \quad (1)$$

The next three rules are about agent actions: If agent α has a goal to be at location λ, then α is heading there:

$$goal(\alpha, at(\alpha, \lambda)) := goingTo(\alpha, \lambda) \quad (2)$$

If agent α is supposed to follow agent β, and β is heading to location λ, α is also going to λ:

$$follow(\alpha, \beta) \wedge goingTo(\beta, \lambda) := goingTo(\alpha, \lambda) \quad (3)$$

The next rule pertains to triggering a notification event. If you are supposed to inform agent β when a condition ϕ is achieved, then when ϕ is achieved, generate an intention-to-know ϕ by β, which will leverage the dialogue generation capabilities of the agent:

$$Inform(\beta, \phi) \wedge \phi := IK(\beta, \phi) \quad (4)$$

3.2 Belief Update Rules for Utterances

We need to add rules for handling utterances both from a speaker's and a listener's perspective. Here we will build on our recently introduced formal framework [4] where we use $[[u]]_c$ to denote the "pragmatic meaning" of an utterance u in context c (which includes task, goal, belief and discourse aspects).

The first general rule (based on the above discussed simplifications) is that an agent always believes all propositions it is able to infer from the utterance of another agent:

$$([[u]]_c \Rightarrow_\alpha^b \phi) \wedge Heard(\alpha, u) \Rightarrow B(\alpha, \phi) \quad (5)$$

Note that the inferences here are bounded by the agent's computational and algorithmic inference limitations (indicated by the agent's inference $\Rightarrow_\alpha^b$ mechanism bounded by b). While this rule is reasonable for simple agents in limited task-domains and might allow an agent to generate all implications given by an utterance in context C, it is likely that more sophisticated agents in more complex domains will not be able to generate all implications.

The second rule is that an agent believes everything it said itself:

$$([[u]]_c \Rightarrow_\alpha^b \phi) \wedge Said(\alpha, u) \Rightarrow B(\alpha, \phi) \quad (6)$$

This rule comports with Gricean conversation maxims, specifically the maxim of quality, which requires one to not say what one believes is false [9]. Though it would fail in cases of intentional deception, it is assumed that in the domain of collaborative HRI, such cases are not to be expected. Such a rule would rely on feedback utterances, such as acknowledgments (e.g. "OK") and/or reiterations (e.g. "Yes, I am going to nav-point 3.") to maintain correct mental-models.

In addition to adding general rules for utterances based on speaker and listener roles, we need to add more specific rules for capturing the pragmatic implications of different utterance types such as statements, questions, commands and acknowledgments based on prior dialogue history and sentential modifiers. We present several pragmatic rules below in the form of $UtteranceType(\alpha, \beta, X, M)$, where α denotes the speaker, β denotes the audience, X denotes the surface semantics, and M denotes the set of sentential modifiers present in the utterance (which may be the empty set, denoted here as $\{\}$). Pragmatic rules for various adverbial modifiers in this domain (such as "still" and "now") were presented in [4].

3.2.1 Statements

If α informs β that it is at λ, then we can assume that α is indeed at that location:

$$[[Stmt(\alpha, \beta, at(\alpha, \lambda), \{\})]]_c := at(\alpha, \lambda) \quad (7)$$

If α informs β that it is going to λ, then we can assume that α is indeed going to location λ:

$$[[Stmt(\alpha, \beta, goingTo(\alpha, \lambda), \{\})]]_c := goingTo(\alpha, \lambda) \quad (8)$$

If α informs β that it is going to λ with γ, then we can assume that α is indeed going to location λ and believes γ is doing the same:

$$[[Stmt(\alpha, \beta, at(\beta, \lambda), \{with(\gamma)\})]]_c := goingTo(\alpha, \lambda) \wedge B(\alpha, goingTo(\gamma, \lambda)) \quad (9)$$

If α informs β that it is engaged in action θ, then we can assume that α is indeed doing that action:

$$[[Stmt(\alpha, \beta, doing(\alpha, \theta), \{\})]]_c := doing(\alpha, \theta) \quad (10)$$

3.2.2 Questions

If α asks β about its location in the general sense ("where are you?"), then one can infer that α has an intention to know (expressed via the "IK" operator) where β is located:

$$[[Ask_{loc}(\alpha, \beta, \{\})]]_c := IK(\alpha, at(\beta, \lambda)) \quad (11)$$

for some λ.

If α asks β about its heading in the general sense ("where are you going?"), then one can infer that α has an intention to know the location where β is traveling to:

$$[[Ask_{goto}(\alpha, \beta, \{\})]]_c := IK(\alpha, goingTo(\beta, \lambda)) \quad (12)$$

for some λ.

If α asks β about its current action in the general sense ("what are you doing?"), then one can infer that α has an intention to know the current action that β is engaged in, which is specified by the $doing()$ predicate:

$$[[Ask_{doing}(\alpha, \beta, \{\})]]_c := IK(\alpha, doing(\beta, \theta)) \quad (13)$$

for some action θ.

3.2.3 Commands

If α orders β to travel to λ, then one can infer that β has a goal to be at λ, α wishes to be informed when β reaches λ, and that α wants to know whether β heard the command:

$$[[Cmd(\alpha, \beta, at(\beta, \lambda), \{\})]]_c := \quad (14)$$
$$G(\beta, at(\beta, \lambda)) \wedge Inform(\alpha, at(\beta, \lambda))$$
$$\wedge IK(\alpha, Heard(\beta, G(\beta, at(\beta, \lambda))))$$

If α orders β to follow γ, then one can infer that β has a goal to follow γ and α wants to know whether β heard the command:

$$[[Cmd(\alpha, \beta, at(\beta, \lambda), \{\})]]_c := \quad (15)$$
$$follow(\beta, \gamma) \wedge IK(\alpha, Heard(\beta, follow(\beta, \gamma)))$$

If α orders β to travel to λ, then one can infer that β has a goal to be at λ, α wishes to be informed when β reaches λ, and that α wants to know whether β heard the command:

$$[[Cmd(\alpha, \beta, at(\beta, \lambda), \{\})]]_c := \quad (16)$$
$$G(\beta, at(\beta, \lambda)) \wedge Inform(\alpha, at(\beta, \lambda))$$
$$\wedge IK(\alpha, Heard(\beta, G(\beta, at(\beta, \lambda))))$$

3.2.4 Acknowledgments

If α utters an acknowledgment (e.g., "OK.") when the previous utterance was a positive statement of location by β, then one can infer α no longer has the intention to know β's location:

$$[[Ack(\alpha, \beta, \{\})]]_c := \neg IK(\alpha, at(\beta, \lambda)) \quad (17)$$

for some λ where for any M $Prior(Stmt(\beta, \alpha, at(\beta, \lambda), \{M\})) \in c$. If α utters an acknowledgment (e.g., "OK.") when the previous utterance was a command by β to be at λ, then one can infer that

$$[[Ack(\alpha, \beta, \{\})]]_c := \quad (18)$$
$$G(\alpha, at(\alpha, \lambda)) \wedge heard(\alpha, G(\alpha, \lambda))$$

where $Prior(Cmd(\beta, \alpha, at(\alpha, \lambda), \{M\})) \in c$ If α utters an acknowledgment (e.g., "OK.") when the previous utterance was a command by β to follow γ, then one can infer that

$$[[Ack(\alpha, \beta, \{\})]]_c := \quad (19)$$
$$follow(\alpha, \gamma) \wedge heard(\alpha, meet(\alpha, \gamma))$$

where $Prior(Cmd(\beta, \alpha, at(\alpha, \lambda), \{M\})) \in c$

3.3 Belief Updates

Each agent γ updates its beliefs whenever it hears an utterance u from speaker α addressing another agent β (which may or may not be the same agent as γ) or whenever it receives a set of perceptual updates Ψ_γ. It uses the above specified principles to determine all pragmatic implications of the utterance and also to detect any beliefs inconsistent with existing beliefs (both pragmatic implications and inconsistency detection are determined by γ's inference algorithm $\Rightarrow^b_\gamma$ and are thus subject to b – for low values of b the agent might fail to compute all implications or to derive a contradiction); the set of conflicting beliefs P_γ are then removed from the agent γ's sets of beliefs.

4 CASE STUDY

All principles and belief updates described above were implemented as a special dialogue component in the Java-based *Agent Development Environment* (ADE) (see http://ade.sourceforge.net/) which is a framework for implementing distributed architectural components for robotic architectures. A simple resolution-style inference mechanism with a shallow one-step look-ahead search limit was used. The new dialogue component (in conjunction with previous algorithms for utterance generation and response selection as detailed in [4]) was used integrated into the existing robotic DIARC architecture which comprises components for perceptual processing (using camera-based vision) and navigation (for ground-based and air-based vehicles), action planning and natural language processing and has been used extensively for human-robot interactions in natural language [13]. For the case study, we used a Videre Erratic mobile robot and Parrot AR Drone Quadricopter from ExPansys. A picture of the platforms used can be found in Figure 2, while video of the interaction can be found online.

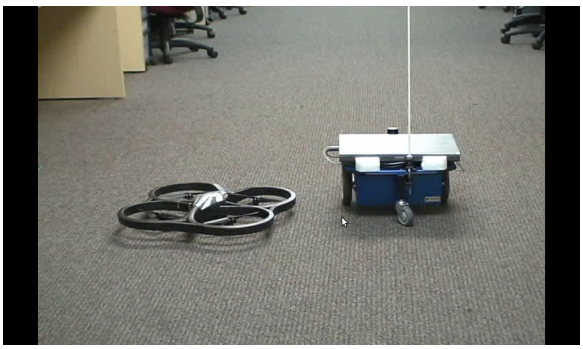

Figure 2. The quadricopter (left) and Videre (right) robotic platforms utilized for the study[4].

As illustrated in Figure 2, both the quadrotor (Q1) and the Videre transport (T1) start in the same location, which we designate S. The belief-spaces of both agents are initialized to be empty, though both are able to perceive that they are at the starting location:

$$\Psi_{Q1} := \{at(Q1, S)\} \rightarrow B_{Q1} := \{at(Q1, S)\}$$
$$\Psi_{T1} := \{at(T1, S)\} \rightarrow B_{T1} := \{at(T1, S)\}$$

The human operator (O) then queries the quadrotor:

```
O: Drone, what are you doing?
```

Since the quadrotor is idle, and has no $doing(Q1, \theta)$ terms in its belief-space, the quadrotor replies accordingly and the operator acknowledges:

```
Q1: Commander, I am not doing anything.
O: Okay.
```

The operator then gives the transport an order to travel to location alpha:

[4] http://www.youtube.com/watch?v=40_Ee2g5ztg

```
O: Transport, go to alpha.
```

Both the transport and quadrotor hear this utterance, and they update their beliefs accordingly:

$$
\begin{array}{l}
u := parse(\text{"O: T1, go to alpha."}) \\
\rightarrow u := Cmd(T1, O, at(T1, \alpha), \{\}) \\
[[u]]_c := \{G(T1, at(T1, \alpha)), Inform(O, at(T1, \alpha)), \\
...IK(O, heard(T1, G(T1, at(T1, \alpha))))\} \\
P_{Q1}^{-} := contradictedTerms([[u]]_c, B_{Q1}) \\
P_{T1}^{-} := contradictedTerms([[u]]_c, B_{T1}) \\
B_{Q1} := (B_{Q1} - P_{Q1}^{-}) + [[u]]_c \\
B_{T1} := (B_{T1} - P_{T1}^{-}) + [[u]]_c
\end{array}
$$

The belief-spaces of both agents are consequently:

$$
\begin{array}{l}
B_{Q1} := \{at(Q1, S), G(T1, at(T1, \alpha)), \\
...Inform(O, at(T1, \alpha)), \\
...IK(O, heard(T1, G(T1, at(T1, \alpha))))\} \\
B_{T1} := \{at(T1, S), G(T1, at(T1, \alpha)), \\
...Inform(O, at(T1, \alpha)), \\
...IK(O, heard(T1, G(T1, at(T1, \alpha))))\}
\end{array}
$$

As the previous command utterance was directed at T1, the agents assume the dialogue's turn is passed to T1. The transport subsequently satisfies the operator's intention to know the command was heard by generating an acknowledgment utterance:

```
T1: Okay.
```

The transport then begins to travel to location α, adding the $doing(T1, goingTo(T1, \alpha))$ term to its belief-space. The operator then gives the quadrotor an order to follow the transport, which is followed by the transport's acknowledgment:

```
O: Drone, follow transport.
Q1: Okay.
```

Having no other goals, the quadrotor then begins to follow the transport, adding the $doing(Q1, follow(Q1, T1))$ term to its belief-space. At this point, the belief-spaces of each agent are:

$$
\begin{array}{l}
B_{Q1} := \{at(Q1, S), G(T1, at(T1, \alpha)), \\
...Inform(O, at(T1, \alpha)), follow(Q1, T1), \\
...doing(Q1, follow(Q1, T1))\} \\
B_{T1} := \{at(T1, S), G(T1, at(T1, \alpha)), \\
...Inform(O, at(T1, \alpha)), follow(Q1, T1), \\
...doing(T1, goingTo(T1, \alpha))\}
\end{array}
$$

Later, the operator queries the quadrotor:

```
O: Drone, what are you doing?
```

Retrieving the appropriate term from its beliefs, the quadrotor responds in accordance with Rule 10:

```
Q1: Commander, I am following transport.
O: Okay.
```

The operator subsequently asks the quadrotor about its destination:

```
O: Drone, where are you going?
```

Because the quadrotor has previously heard that the transport was commanded to go to α, Q1 is able to infer:

$$
\begin{array}{l}
G(T1, at(T1, \alpha)) := goingTo(T1, \alpha) \\
goingTo(T1, \alpha) \wedge follow(Q1, T1) := \\
goingTo(Q1, \alpha) \\
\rightarrow goingTo(Q1, \alpha)
\end{array}
$$

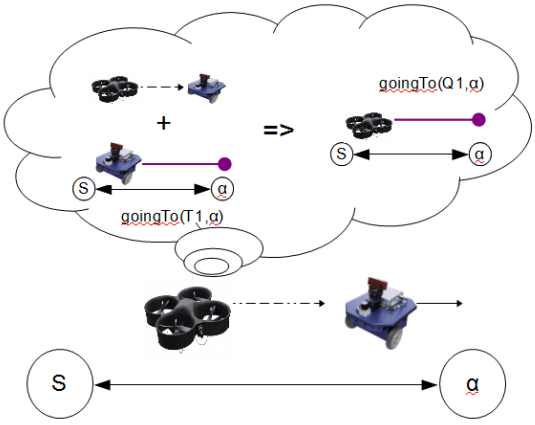

Figure 3. The quadrotor infers its destination.

The above inference is depicted in Figure 3. Based on this inference, Q1 responds accordingly:

```
Q1: Commander, I am going to alpha with
transport.
O: Okay.
```

At this point, we should clarify that there is a distinction between the robot having a explicit goal to undertake and action and having the knowledge that it is undertaking an action. Having a goal to perform an action (or achieve some state) will modulate the behavior of the robot, while just having knowledge of the current action will not necessarily affect the system's behavior. In this case the inference made in Figure 3 is only resulting in knowledge that the robot is going to the final location alpha, but not a *goal* to go to location alpha. Thus, the quadrotor will not be in danger of incorrectly following the transport by predicting the final location and traveling there (instead of following).

We should also contrast the two questions we have just examined. "What are you doing?" we have interpreted as a query to ascertain the robot's current goal, whereas "Where are you going?" seeks more specific knowledge from the robot that can only be answered by making the inference in Figure 3. We believe that, given the same level of knowledge and dialogue history, it would be plausible for a human to answer the question in a similar manner, and as such, the dialogue interaction is made more natural by enabling the robot to do the same.

Later, when the transport finally reaches location alpha, the transport is able to perceive it's new location. The transport updates its belief-space accordingly:

$$
\begin{array}{l}
\Psi_{T1} := \{at(T1, \alpha)\} \\
\rightarrow B_{T1} := \{at(T1, \alpha), Inform(O, at(T1, \alpha)), ...\}
\end{array}
$$

Inferring via Rule 4 the operator's intention to be notified of its arrival at the intended destination, the transport states:

```
T1: Commander, I am now at alpha.
```

5 DISCUSSION

The dynamics of human teams are complex and multifarious, deeply integrating and intertwining natural language exchanges and actions. Humans are extremely good at building mental models of their team mates that include general team mate characteristics as well as particular team mate beliefs, goals and intentions. And humans can effortlessly use all this knowledge to make quick inferences about the mental states of their team mates based on information gleaned from natural language interactions and the details of how that information was linguistically expressed. Most importantly, humans will expect future robots, in particular if they are otherwise very capable, to be able to perform the same kind of mental modeling and to make the same kinds of quick automatic inferences as part of task-based natural language dialogues.

We have introduced a set of principles that can form the basis of a mental modeling mechanism that is deeply integrated with the natural language dialogue mechanisms. The formalism captures perceptual and behavioral aspects of agents as well as their beliefs and intentions/goals. It also allows for different models and model updates for different agents (e.g., how an agent reacts to a particular command given by the operator) by allowing for the definition of agent-specific update rules. And it provides a natural level of abstraction where agents can introspect on their own behaviors and behavioral dispositions in an effort to model themselves and other agents.

Similar challenges involving utilizing natural language communication and maintaining situation-awareness have been investigated in [2] and in the multi-robot domain in [3]. In contrast with these approaches, our approach so far involves simple reactive agents, rather than agents with planning capabilities.

Beyond the sophistication of our agents, our current approach has additional shortcomings. First, it is unclear how far the search depth of the inference algorithm can be reasonable extended if more dialogue principles are added without losing real-time processing. Clearly, there will be limits to the set of propositions an agent can derive automatically given the number of pragmatic and agent-based rules. To curb the complexity and avoid generating thousands of irrelevant beliefs, it will become necessary to incorporate a notion of relevance that allows for targeted inference (also to derive contradictions as part of belief updates). Finally, the current version makes several simplifying assumptions (e.g., about perceptions and behavioral decisions) that will clearly be too simple for more complex tasks and agents. For instance, our communication is currently accomplished individually between single agents. Belief update rules need to be extended to account for group communication [14, 11]. The problem of collaborative planning, in which agents must work together to develop a joint plan, poses further challenges in that agents must have the ability to communicate and reason about partial candidate plans [10].

6 CONCLUSIONS

In this paper, we introduced new principles for belief modeling and updating for autonomous agents (such as robots or virtual characters) interacting with humans and other autonomous agents in mixed initiative teams through spoken natural language dialogues. We showed how we can represent beliefs and intentions of other agents to generate mental models that are rich enough to capture task-based aspects of other agents and their beliefs. We also showed how a robot can update its mental model of another robot based on task-based utterances it heard and how it can automatically apply inference-rules to the information obtained from the utterance to model and predict other agents beliefs and behaviors.

Future work will address the issues of scalability, relevance, an scope mentioned in the discussion section above. And we will conduct simple HRI evaluation experiments that will allow a human operator to command a mixed initiative team with one ground and one aerial robot as described in the case study, with and without belief modeling. This will allow us to determine whether and to what extent belief modeling as proposed in this paper can lead to objectively better task performance and subjectively better acceptance by human team mates.

ACKNOWLEDGEMENTS

This work was in part funded by ONR grants #N00014-07-1-1049 and #N00014-11-1-0493. We would also like to thank the referees for their userful comments which helped improve this paper.

REFERENCES

[1] Cynthia Breazeal, Andrew Brooks, Jesse Gray, Guy Hoffman, Cory Kidd, Hans Lee, Jeff Lieberman, Andrea Lockerd, and David Mulanda, 'Humanoid robots as cooperative partners for people.', *Journal of Humanoid Robots*, **1**, (2004).

[2] M Brenner, 'Situation-aware interpretation, planning and execution of user commands by autonomous robots', in *Proceedings of the 16th International Symposium on Robot and Human Interactive Communication (RO-MAN)*. IEEE, (2007).

[3] M Brenner and B Nebel, 'Continual planning and acting in dynamic multiagent environments', *Autonomous Agents and Multi-Agent Systems*, **19**(3), 239–331, (2009).

[4] Gordon Briggs and Matthias Scheutz, 'Facilitating mental modeling in collaborative human-robot interaction through adverbial cues', in *Proceedings of the 12th Annual SIGdial Meeting on Discourse and Dialogue*. ACL, (2011).

[5] Rehj Cantrell, Matthias Scheutz, Paul Schermerhorn, and X Wu, 'Robust spoken instruction understanding for hri', in *Proceedings of the 6th International Conference on Human-Robot Interaction*. ACM/IEEE, (2010).

[6] P. R. Cohen and Hector Levesque, 'Teamwork', Technote 504, SRI International, Menlo Park, CA, (1991).

[7] P. R. Cohen and S. L. Oviatt, 'The role of voice input for human-machine communication', *Proceedings of the National Academy of Science, USA*, **92**, 9921–9927, (1995).

[8] Philip R. Cohen, Hector J. Levesque, Jose Nunes, and Sharon L. Oviatt, 'Task-oriented dialogue as a consequence of joint activity', in *Pacific Rim International Conference on Artificial Intelligence*, (1990).

[9] H. P. Grice, 'Logic and conversation', *Syntax and Semantics*, **3**(1), 43–58, (1975).

[10] Barbara J. Grosz and Sarit Kraus, 'Collaborative plans for complex group action', *ARTIFICIAL INTELLIGENCE*, **86**(2), 269–357, (1996).

[11] Sanjeev Kumar, Marcus J. Huber, David R. Mcgee, Philip R. Cohen, and Hector J. Levesque, 'Semantics of agent communication languages for group interaction', in *In Proceedings of the 17 th Int. Conf. on Artificial Intelligence*, pp. 42–47, (2000).

[12] Paul Schermerhorn and Matthias Scheutz, 'Using logic to handle conflicts between system, component, and infrastructure goals in complex robotic architectures', in *Proceedings of the International Conference on Robotics and Automation*, (2010).

[13] Matthias Scheutz, Paul Schermerhorn, J Kramer, and D Anderson, 'First steps toward natural human-like hri', *Autonomous Robots*, **22**(4), 411–423, (2007).

[14] Ira A. Smith and Philip R. Cohen. Toward a semantics for an agent communications language based on speech-acts, 1995.