AISB 2011

Learning Language Models from Multilingual Corpora

Editors:
Dimitar Kazakov &
George Tsoulas

THE UNIVERSITY of York

Foreword from the Convention Chairs

The AISB'11 call for symposium proposals particularly encouraged events drawing more strongly on the cognitive science aspect of the AISB remit. The result is a coherent programme with a very strong interdisciplinary character, which is also matched in the choice of plenary speakers. The three symposia looking at the interaction between Computing and Philosophy, the prospect of machine consciousness and the quest for a new, comprehensive intelligence test, form a coherent unit where the eternal questions of who we are and what makes us so are asked from a dual Human-Machine perspective. The Symposia on Active Vision, Computational Models of Cognitive Development and Human Memory for Artificial Agents demonstrate how better understanding of the nature and basis of cognitive processes can advance work on Artificial Intelligence and, inversely, how computational models of these processes can help better to understand them. The prominent multi-agent design and modelling paradigm links the Symposium on Social Networks and Multi-agent Systems with the one on AI and Games. Finally, the Symposium on Learning Language Models from Multilingual Corpora, which brings together some of the first attempts in this area, can also be seen through the prism of such a general notion in Philosophy and Linguistics as semiosis, and the dual role of sign and interpretant that text plays in translations.

We are delighted that after another ten successful years in its long history, the AISB convention is returning to the University of York. The 2011 convention takes place on the brand-new Heslington East campus, the result of a multi-million pound expansion that is now the new home of the Department of Computer Science, and hosts the Excellence Hub for Yorkshire and Humber, a new incubator for interdisciplinary research and interaction between academia and industry. The last few years have seen a strong involvement of the Computer Science Department in such interdisciplinary collaboration through the York Centre for Complex Systems Analysis (YCCSA), and we hope that this convention will provide a boost for more synergy between York departments, with other institutions conducting AI-related research in the region, and beyond. As the programme shows, we have also made an effort to promote cooperation with industry and use the convention to support school outreach. The convention format makes it perfect for establishing dialogue and collaboration in new areas of research, as well as across disciplines, and we hope that this year, it will play again this role to the full. We want to thank everyone who has contributed to it or otherwise made this event possible and wish all participants a fruitful and enjoyable time in York.

Dimitar Kazakov and George Tsoulas

Proceedings of the Symposium on Learning Language Models from Multilingual Corpora (LLMMC)

ISBN: 978-1-908187-05-5

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

Printed by the University of York, York, UK, April 2011

Preface

This volume contains the proceedings of the Symposium on Learning Language Models from Multilingual Corpora (LLMMC), organised on 6 April 2011 as part of the 2011 Annual Convention of the Society for the Study of Artificial Intelligence and the Simulation of Behaviour (AISB'11, 4–7 April 2011, York, UK). The symposium takes place alongside eight other symposia on various aspects of AI-related research, drawing strongly on Computer Science, Psychology and Philosophy, among other disciplines. A number of plenary speakers of international repute (Alan Baddeley, Katie Slocombe, Mark Steedman, Stephen Wolfram) will add to the excitement of this interdisciplinary international event, which takes place in the historical city of York, one of the oldest and most visited destinations in England. In this symposium, we are interested in explicit models, usable and verifiable by humans, which can be used for either translation or for modelling individual languages, e.g., as applied to morphology, where the available translations can help identify word forms of the same lexical entry in a given language, or lexical semantics, where parallel corpora can help extract instances of relations like synonymy and hypernymy, which are essential for building thesauri and ontologies. While bringing together a few first practical attempts in this area, the focus of the symposium can also be seen through the prism of such a general notion in Philosophy and Linguistics as semiosis, and the dual role of sign and interpretant that text play in translations.

April 2011
York, UK

Dimitar Kazakov
Preslav Nakov
Ahmad R. Shahid

Organising Committee

Dimitar Kazakov	University of York
Preslav Nakov	National University of Singapore
Ahmad R. Shahid	University of York

Programme Committee

Graeme Blackwood	University of Cambridge
Phil Blunsom	University of Oxford
Francis Bond	NICT - Nanyang Technological University
Yee Seng Chan	University of Illinois at Urbana-Champaign
Daniel Dahlmeier	National University of Singapore
Marc Dymetman	Xerox Research Centre Europe
Andreas Eisele	European Commission, Directorate-General for Translation
Michel Galley	Stanford University
Kuzman Ganchev	University of Pennsylvania
Corina Girju	University of Illinois at Urbana-Champaign
Philipp Koehn	University of Edinburgh
Krista Lagus	Helsinki University of Technology
Wei Lu	National University of Singapore
Elena Paskaleva	Bulgarian Academy of Science
Katerina Pastra	Institute for Language and Speech Processing
Khalil Sima'An	University of Amsterdam
Ralf Steinberger	European Commission - Joint Research Centre
Joerg Tiedemann	Uppsala University
Marco Turchi	European Commission Joint Research Centre
Jaakko Väyrynen	Aalto University School of Science

Table of Contents

Constructing a Reusable Linguistic Resource for a Polyglot Speech Synthesis

Nur-Hana Samsudin and **Mark Lee**[1]

Abstract. This paper is about constructing sharable linguistic information to be used across languages for a Text-to-Speech (TTS) system. The data is obtained from existing resources. The focus of the paper is the phonetic and linguistic aspects. A monolingual TTS architecture is introduced with descriptions on each stage of processing. A multilingual TTS architecture is also introduced. Language dependent and language independent information required in the general TTS engine is described. Finally we will propose layer of communication between the pooled language information and a TTS engine in order to make it available for information reuse where possible for different languages.

1 INTRODUCTION

Globalisation has triggered the need of multilingual and multimodal applications. As such, speech to speech translation components and devices are evolving towards this direction. That is to say, TTS is moving towards multilingualism. Multilingual TTS is not totally a novel idea. MBROLA for instance, was developed in 1996 with multilingual facilities [5]. Although MBROLA is not a complete TTS system (because it does not accept raw text as input), it is an example of a working TTS with (encrypted) multilingual speech corpora.

To create a monolingual TTS, one needs to gather its language information as well as speech recording for the target language. A TTS system development is influenced by the target language (other than the standard architecture). Even if one already familiar with developing a TTS, one will still need to do the implementation process all over again to have a complete working TTS for a different language, since the information obtained for one TTS is not necessarily fully applicable for different languages. If one creates a multilingual TTS, the variability of data to be gathered will depend on the goal of how many and which particular languages the multilingual TTS is supposed to be handle. This paper describes the work that we have carried out to minimise the pre-requisite language information gathering; i.e.: the grapheme-to-phoneme conversion, the phonological processing and the prosody assignment. It is designed in such a way that, it is possible to do rapid TTS development for resource poor languages. We also proposed to use the speech database which has already been developed for other languages. As for the proof of concept, we are using MBROLA synthesiser.

Traber [11] describes the key stages in developing a multilingual TTS. These issues also influence the development of polyglot TTS systems. It is possible to create a reusable resources by producing an intermediate layer inclusive of Traber's [11] proposal which will become the medium between the TTS framework and the target language by giving attention to following issues.

- *different speech corpora are required for different languages;* Therefore separate recordings and separate recording annotations need to be conducted. This requires a huge amount of labour. And one also needs to find a good multilingual speaker to perform the recording.
- *different set of phoneme need to be introduced;* The phoneme set is different for each language. It is difficult to find a standard phoneme set to represent all sounds in all languages.
- *individual grapheme to phoneme rules need to be constructed;* A grapheme to phoneme conversion is not always orthographic for all languages. Additionally, all languages have unique phonological rules which also need to be considered before we can perform grapheme-to-phoneme conversion of any language.
- *individual prosody's value assignment need to be considered;* While some languages require stress to differentiate one word from another, other languages need them so that synthetic speech has a human-like quality. It requires linguistic information to be represented in the prosodic model.

This paper's intention is to describe a method to produce such a layer to create multilingual speech synthesisers using limited linguistic resources e.g. in cases where no corpus exists for the target language but does exist for a member of the same linguistic sub-families (but not necessarily from similar language family categories). The hypothesis is that there are useful linguistic and phonetic generalisations within selective linguistic sub-families. We however are not discussing the development of new speech corpora. In short, this paper covers a communication layer which includes the following: global phoneme and its grapheme-to-phoneme conversion, providing phonological rules for language dependent and language independent rules as well as providing generic prosody rhythm.

This paper is organised as follows. In the next section, we will describe the TTS architecture in general, and how variations of multilingual TTS are implemented in existing systems. Then we will discuss the intermediate layer that is proposed and what information is required. We will also describe the focus of phonetic and phonological processing component which is the focus of this paper. Then we will show the implementation of such intermediate layer into a complete system and conclude this paper.

2 TTS ARCHITECTURE

In a monolingual TTS architecture, the process is defined following strictly to the target language. A TTS system accepts text as input

[1] University of Birmingham, UK, email: {n.h.samsudin; m.g.lee}@cs.bham.ac.uk

and produces speech as the output. The framework of a TTS system may be represented like Figure 1. Each process will be described and developed.

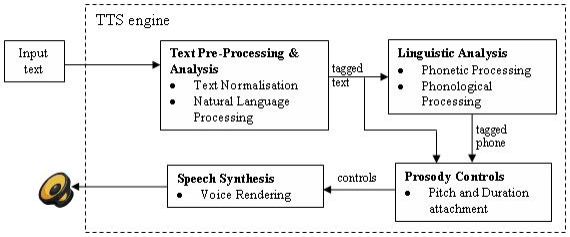

Figure 1. General TTS Architecture.

Text pre-processing and analysis process the input text of numbers, symbols and short forms into a readable orthographic form. After normalisation, the text is parsed into corresponding lexical labelling. Linguistic analysis accepts the normalised input and converts it into a phonetics transcription. Further linguistic processing in terms of pronunciation is then carried out. Prosody controls make use of the linguistic and text processing information to assign the intonation contours of the text. Finally the synthesised speech is produced. The prosodic control depends on the output from text analysis and linguistic analysis. This again depends on how the system is developed. An example of a multilingual system architecture is as shown in Figure 2.

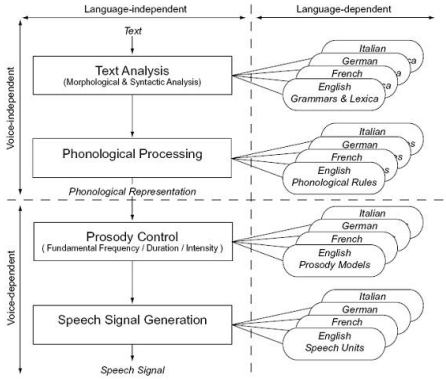

Figure 2. PolySVOX TTS Architecture [6].

In the architecture of Figure 2, Romsdorfer and Pfister [6] highlighted the relationship between language dependent and language independent components. This architecture originated from a monolingual TTS system (SVOX) and was developed into a multilingual TTS system. Pre-requisite information is gathered in isolation. The synthesiser system however is developed with enough modularity to support different language data. The linguistics and speech data have been identified as language dependent while the algorithms such as prosody control are language independent components but are speaker dependent characteristics. Text analysis and phonological processing are language dependent but not restricted to specific voice/speaker. What we propose in our work (Section 4) is how are we going to represent these language dependent and speaker dependent characteristics into our communication layer so that the implementation of a TTS for previously unknown language is efficient and rapid.

3 MULTILINGUAL VS. POLYGLOT TTS ARCHITECTURE

In this section, the difference between multilingual and polyglot speech synthesis will be described. In multilingual speech synthesis, different algorithms, rules and speech data are used for different languages [11]. In polyglot speech synthesis, there is a primary language which is identified as the main language of the synthesiser. The main aspect of polyglot speech synthesis is any systems using this framework will be able to synthesis multiple languages using the same recorded or trained voices. Code switching phenomena is a common component in both multilingual and polyglot speech synthesis [6] which may be required for the switching algorithm and switching speech databases in the cases of mixed-lingual occurrences which require the text to be processed in another language.

Both approaches have advantages and disadvantages. The goal of the development however, depends on the balance between a very native-like speech quality versus time and resources.

3.1 Multilingual Speech Synthesis

Multilingual synthesisers are suitable for use for teaching and learning of languages, and when accurate pronunciation of one language must be distinguished correctly to another or when foreign accent and dialects are not acceptable. They are also suitable when the system to be developed do not have any issues in terms of availability of linguistic resources or resource storage size. This make the multilingual speech synthesis system a very reliable framework but relatively expensive in term of cost.

Generally multilingual TTS design follows monolingual TTS architecture closely. For example, MBROLA uses one synthesiser but it has 72 diphone speech corpora from 37 languages. Each 72 diphone corpus also has its own grapheme-to-phoneme transcription. In order to make a TTS based on MBROLA framework, one needs to define the phonetic and phonological rules of the desired target language and build the language text analysis and pre-processing. The prosody modelling must also developed separately. Romsdorfer and Pfister's architecture [6] is an example of such system. In short, there are some similar aspects to monolingual in multilingual TTS architecture.

3.2 Polyglot Speech Synthesis

Contrary to multilingual speech synthesis, polyglot speech synthesis is more suitable for mixed-lingual text [6]. For example, in occurrences of xenomorphs[2] it would not be practical to switch from one corpus to another. It is also suitable for fast prototyping, like SPICE (Speech Processing - Interactive Creation and Evaluation Toolkit for New Languages) [8] which has a very short recording for training cycle for certain languages and it depends on English as the based language [3]. Although more practical, the output of a polyglot speech

[2] Words that are built from combinations of English morphemes and morphemes from the respective native language [10].

synthesis will typically have a foreign accent. This is due to the training data. The smaller the amount of the target language included in the training data, the higher likelihood that the synthesised speech quality to sounds foreign. An example of such system developed in this framework is by Latorre et al. [4]. A polyglot architecture is also suitable when obtaining resources (especially voice recording) is an issue. The architecture of mostly implemented Polyglot TTS is as shown in Figure 3. The frame with the label (1) shows the basic scheme of a HMM-based polyglot synthesis; label (2) shows the adaptation to speakers of extrinsic[3] languages and (3) shows how the synthesis of extrinsic language is implemented.

There are two phases in polyglot speech synthesis: the training phase and the synthesis phase. During the training phase, collections of speech in all the target languages are processed and the spectral features of the speech are extracted and stored. In relation to this, the most favoured method of training is Hidden Markov Model (HMM) based speech synthesis. In this technique, the information is stored in a HMM network.

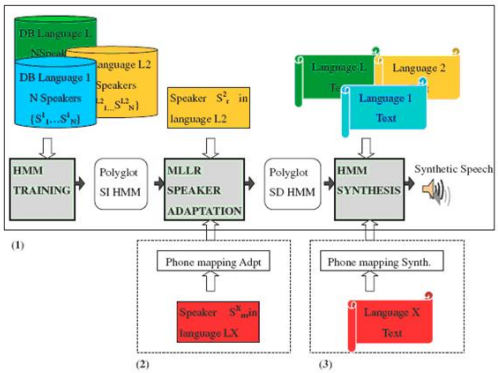

Figure 3. The distinction between the training and synthesis process in a HMM-based polyglot TTS[4].

As in the general TTS synthesis phase, the generation of speech will still go through the text analysis component. However, because the HMM-based synthesis has the ability to learn from the training information to estimate the other characteristics of the voice, the pre-processes are not as thorough as a multilingual synthesiser; although the sound may noticeably unnatural if so little natural language processing is involved.

4 LINGUISTIC MODELLING FOR POLYGLOT SPEECH SYNTHESIS

Text analysis, linguistics analysis and prosodic controls are the pre-requisite processes in a TTS framework when one either uses multilingual or polyglot approach. Each module plays a different function. In this section we will describe on how we represent the pre-requisite information to pool linguistic data in order to create a generic but applicable resources to most languages. There are two types of information at these stages: language specific and language independent information. The linguistic model represents language specific information but with flexibility for modification and improvement; according to the desired language.

[3] language not inclusive in training but needed for speech synthesis

In this section we introduce our linguistic model representation. The linguistic model will use information from the text normalisation stage as well as the original input text to construct linguistic information. The components are illustrated in Figure 4. We define two components as a sub-module to the linguistics analysis module: phonetic processing and phonological processing. In phonetic processing, a set of global phonemes is required. The transcriptions of the orthographic form mapped to the corresponding phonemes are provided. The phoneme set must be able to represent all phonetic sounds of a language at a particular time for most languages. For the phonological processing, a set of phonological rules will be implemented. We categorise the rules as follows: general phonological rules, language specific phonological rules, speaker specific rules and language pronunciation clusters. Language intonation template and pitch melody are implemented in the next module. These are all characteristics that we attend to further characterise language uniqueness. We however, are not going to implement any speaker specific rules.

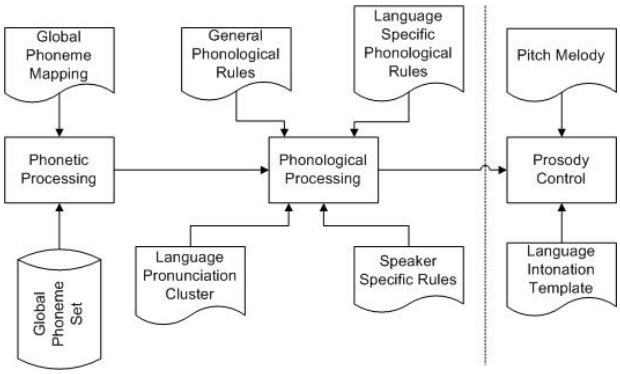

Figure 4. Components affected in our linguistic model.

General phonological rules refer to the phonological changes due to human articulation. Language specific pronunciation rules refer to the phonological rules applied in a particular language while speaker specific rules refer to the pronunciation changes because of the speaker's style. Finally, language pronunciation clusters is a categorisation of languages according to the closest similarity of speech and language features. For the time being, which cluster a language belong to are determined by the linguistic characteristics (e.g. is it a phonetically written language, is it a tonal language, is it syllabic language). These parameters are still being refined to better reflect the multilingual linguistics representation. Phonetic and phonological processing will be discussed in further detail in Section 4.1 and Section 4.2.

4.1 Phonetic Processing

A set of global phonemes is used in phonetic processing. The main idea of this set is that all orthographic representation of languages that are going to be tested in the system must be transformed into the phonetic transcription based on the declared phoneme list of the corpus. It is thus crucial for the set to represent all sounds. One way to come out with the list is to use the International Phonetic Alphabet (IPA) and provide an instance of each symbol in computer readable

format. The symbols referred here are those listed under the category of consonant (*pulmonic and non-pulmonic*), other symbols and vowels. Other categories in IPA which are diacritics, suprasegmentals and tones and word accents are not going to be considered in the phonetic processing section, but will be considered in phonological processing components. A phoneme substitution list is also provided for the synthesiser.

Global phoneme mapping rules are general mapping rules introduced for orthographic to phonemic transcription. However, to come out with this general rules, a set of irregular rules must be identified. Consider the hypothetical languages; A, B and C where the orthographic of [c] in Language A is either /k/ or /k^h/ but in Language B and Language C, [c] is always pronounced as /c^h/ or /tʃ/. Therefore, the case of [c] mapped to /c^h/ or /tʃ/ in Language A is considered as an irregular mapping. Each of them will be presented but using a different set of rules. Therefore a TTS developer who uses this information pool will need to identify which transcription the target language belong to and further refine the language transcription rules according to the target language rules.

There will be a need for a language specific mapping rules or a pronunciation dictionary in the phonetic processing component as shown in Figure 4. However, since the goal of this research is building a polyglot TTS by making use of available resources, the system can produce the output based on the available rules, even if not fully accurate. We have introduced our global phonemes in [7].

4.2 Phonological Processing

To create a global set of phonological rules for all languages, one has to identify a set of phonological rules of different languages. We are studying the phonological rules of different languages from research conducted on monolingual and multilingual TTS. Different systems (and languages) specify different ways to handle phonological processing. But there is a generic categorisation of phonological rules as listed in the Section 5.2. The main goal is not to list all the phonological rules for all languages, but how to represent different phonological rules. Based on this information, the structure of general phonological rules and language specific phonological rules can be built and further improved according to the target TTS's language. Phonological rules will process the input from the phonetic transcription based on grapheme-to-phoneme conversion.

5 APPLYING THE MODEL INSIDE A TTS FRAMEWORK

As stated previously, a linguistic model is going to be used in the synthesis phase of polyglot speech. The idea is that by providing richer speech information, a good synthesised polyglot speech quality is achievable as compared to the original language generation using different speech database but the same synthesiser. The model is not framework dependent. However, to prove the feasibility of the approach, we are implementing the idea into MBROLA synthesiser. Here are the components we are currently implementing.

5.1 Global Phoneme Mapping

Contrary to a phone, a grapheme is not always pronounced as the same for all languages. In global phoneme mapping, a general phone mapping rules need to be defined and additional rules required for language specific pronunciation. For the proof of concept, we have built two types of grapheme-to-phoneme rules: one which is a phonetically written language (e.g.; Spanish) and non-phonetically written language (e.g.: English and French). We then will test the grapheme-to-phoneme rules and observe the modification required to change for similar language class; phonetically written language or not; and described its applicability for based on only two classifications.

5.2 General Phonological Rules and Language Specific Phonological Rules

Phonological process can be divided into two: lexical process and post-lexical processes or also known as phrasal phonology. Phonological alternation can happen by a variety of processes as described by [9]:

- assimilations and dissimilations
- fortitions and lenitions (or strengthenings and weakenings)
- insertion, deletion and coalescence,
- lengthenings and shortenings
- metathesis and reduplication.

In the phonological processing, we will record fortitions and lenitions, insertion and deletion; and metathesis and reduplication. The lengthening and shortening is control using the phonemes and prosody control. Fortitions and lenitions are controlled in pitch melody. Processing involved in phonological processing component is processed from the original phonetic transcription from grapheme-to-phoneme conversions. In other words, applying the phonological rules is almost like running grapheme-to-phoneme transcription again using another set of rules.

5.3 Language Intonation Template

Intonation for each language is different. If one listens to a foreign speech, one can find that most language has a rhythm of speech during production. Intonation when further analysed will give both qualitative measurement and quantitative one. Qualitative data is evaluated based on human perceptions. For quantitative measurement, intonation can be modelled by the prosody of the speech - pitch, duration and intensity. By manipulating these parameters, the stress and tonal of a language can be represented. However, the implementation will have to go back to the phonetic and phonology of the language. For example, English (RP) has specific stress point at certain position of words in speech. Mandarin has 4 tones (and in some regional dialects, there are 6 tones)[1]. Estonian has three tones which two can easily defined as a shortening and lengthening of phonemes but the third one has a different type of phonological foot patterns [5] which is a combination of stress and tones. While Malay and Indonesian do not have any tone or stress standard at all which make them even more difficult to represent. Because of rich intonation characteristics and how much it varies from one to the other, representing intonation require refined categorisation. For example, tone will be dealt with at the phonological rules. Lengthening and shortening of phonemes will be controlled at the grapheme-to-phoneme transcription but stress will be handled at the intonation module.

5.4 Pitch Melody

Hirst [2] has introduced MOMEL and INTSINT algorithms to represent melody of speech. Based on the analysed pitch modulation, we are implementing the pitch changes according to the INTSINT

algorithm value to represent the contour of pitch changes according to the defined stress level.

By having a template of intonation, a set of intonation patterns will be constructed. For any language which has specific requirements for stress or accent or any intonation criteria, language specific intonation rules need to be defined. The intonation rules and template will complemented each other to produce the correct intonation for the language of the synthesised speech.

5.5 Language Pronunciation Clusters

Language pronunciation clusters plays the role of refining the process of synthesising speech. The information from this component, combined with the rules and intonation template will make it possible for the system to find a better path in constructing any utterance. It is also used to determine which speech corpus should be selected if the corpus for the language in question is not available, by finding the language with the most similar characteristics with the language in question. It is also important to emphasise that the language family (as in Germanic, Austronesian or Sino-Tibetan) is not playing a role in clustering the language. For example, the combination of the following is currently being tested: Spanish - Malay (generate Spanish using Malay speech database) and German - English. Spanish is from Indo-European language but Malay from Austronesian language family.

6 CONCLUSION

In this paper, we have described an approach for building multilingual resources from limited data for TTS systems. The purpose is to reuse this information for polyglot speech synthesis. The final goal is to come up with a linguistic model which consists of the phonetic and phonological processing components as well as prosody representation. For the phonetic processing component, a global phoneme set has been defined. The plan for the phonological processing component is to come up with the following rules: general phonological rules, language specific phonological rules and language pronunciation clusters. Language intonation template and pitch melody is going to be implemented in prosody control module. After obtaining the phonetic and phonological models as well as the prosody control module, they will be combined as a linguistic model. The model currently is implemented using MBROLA as a synthesiser to see the feasibility applying such a model in an application and to investigate the effects of different corpora.

REFERENCES

[1] David Crystal, *The Cambridge Encyclopaedia of Language*, Cambridge University Press, Cambridge, 2 edn., 2002.

[2] Daniel J. Hirst, 'Automatic Analysis of Prosody for Multilingual Speech Corpora', in *Improvements in Speech Synthesis.*, eds., E. Keller, G. Bailly, J.Terken, and M. Huckvale, Wiley Publisher, Chichester, United Kingdom, (2001).

[3] John Kominek, Tanja Schultz, and Alan W Black, 'Voice Building from Insufficient Data - Classroom Experience with web-based Development Tools', in *Proceedings of the 6th ISCA Workshop on Speech Synthesis (SSW-6)*, Bonn, Germany, (August 2007).

[4] Javier Latorre, Koji Iwano, and Sadaoki Furui, 'New Approach to the Polyglot Speech Generation by means of an HMM-based Speaker Adaptable Synthesizer', in *Speech Communication*, volume 48, pp. 1227–1242, (October 2006).

[5] MBROLA-Group. The MBROLA PROJECTS Towards a Freely Available Multilingual Speech Synthesizer. Internet, 2005. Accesed on 17th July 2009.

[6] Harald Romsdorfer and Beat Pfister, 'Text Analysis and Language Identification for Polyglot Text-to-Speech Synthesis', in *Speech Communication*, volume 49. Elsevier, (September 2007).

[7] Nur-Hana Samsudin and Mark Lee, 'Constructing a Multilingual Phoneme List for Polyglot Speech Synthesiser', in *Proceedings of 4th Language & Technology Conference*, ed., Zygmunt Ventulani, pp. 391–395, Poznań, (November 2009). Fundacja Uniwersytetu im. Adama Mickiewicza, Wydawnictwo Poznańskie Sp.

[8] Tanja Schultz, Alan W Black, Sameer Badaskar, Matthew Hornyak, and John Kominek, 'SPICE: Web-based Tools for Rapid Language Adaptation in Speech Processing Systems', in *Proceedings of Interspeech*, Antwerp, Belgium, (August 2007).

[9] Andrew Spencer, *Phonology: Theory and Description*, Blackwell Publishing, Department of Language and Linguistics, University of Essex, First edn., 1996.

[10] Jochen Steigner and Marc Schröder, 'Cross-Language Phonemisation In German Text-To-Speech Synthesis', in *Proceedings of Interspeech 2007*, pp. 1913–1916, Antwerp, Belgium, (August 2007).

[11] Christof Traber, Karl Huber, Karim Nedir, Beat Pfister, Eric Keller, and Brigitte Zellner, 'From Multilingual to Polyglot Speech Synthesis', in *Proceedings of Eurospeech 1999*, pp. 835–838, Budapest, (September 1999).

Multi-word Level Context Features: towards Context Feature Improvement

Azniah Ismail and **Suresh Manandhar** [1]

Abstract. Learning high precision bilingual lexicons is not an easy task in an unsupervised setting especially when good quality bilingual, comparable corpora are not available. In this paper, we draw attention to the *context features* that are commonly used in learning bilingual lexicons from corpora. Instead of using the conventional single-word context features, we experiment with context features at multi-word level. The method we proposed shows some potential in recuperating the precision in the unsupervised setting. Additionally, we share our experience in acquiring comparable corpora for a non-related language pair.

1 INTRODUCTION

The fact that knowledge resources such as bilingual lexicons and parallel corpora are not usually freely available means that they might become bottlenecks in many NLP problems. Although a high precision bilingual lexicon can be learned automatically, learning bilingual lexicon from non-parallel, comparable corpora is proven more challenging compared to using parallel corpora (see [7]).

As the corpora play a major role in determining the outcome, there has been effort to automatically construct such resources (see [1] and [9]). However, the major drawback is that such methods rely on other knowledge resources such as a sufficient size of initial bilingual lexicon and NLP processing tools such as word alignment tools, POS-taggers and lemmatizers. For a certain language pair, a huge initial bilingual lexicon may not be available. Furthermore, most available tools are in supervised mode and require labelled data such as the tagged corpus. Again, scarce knowledge resources can be the problem. For this reason, we are determined to learn high precision bilingual lexicon from non-parallel, comparable corpora without relying on any existing knowledge resources, which includes the initial bilingual lexicons and the parallel corpora.

A few studies have attempted to learn bilingual word pairs from this direction. [8] described five different features that can be used to extract the bilingual lexicon in such settings, namely identical spelling, similar spelling, similar context, similar word relationship and similar frequencies. [8] experimented the methods on the German-English corpora of a fairly comparable domain. The only notable results are attained using the first and second features with up to 96% and 91% of accuracy respectively. On the other hand, the first two features are very much limited to the orthographic feature between the source and the target words. It is undeniable that majority of word pairs cannot be collected this way.

[8] used just these two features to build an initial bilingual lexion and attempt to expand its size with the similar context based method. The attempt was not a huge success with only 39% accuracy is reported. Although the first and second features produce high precision word pairs, the accuracy drops briskly because the third feature relies heavily onto the size and coverage of the initial bilingual lexicon. [3] adopted similar approach to obtain initial matching pairs and used *canonical correlation analysis* to infer the most confident bilingual word pairs among them. They recorded an improvement of 20% compared to [8]'s approach for the non-related (German-English) language pair. It is important to note that both [8] and [3] used large comparable German-English corpora for their experiments.

Learning high precision bilingual lexicons from comparable corpora with identical spelling in an unsupervised setting is a proven success. However, the task of building high precision bilingual lexicons containing non-identical spelling word pairs especially in extreme settings remains an interesting question. There is a need for a bilingual lexicon extraction (BLE) method to learn such lexicons.

[6] reported that the outcomes can be improved by being selective with the context features. Hence, it might be worthy to find a method to derive a good set of context features. In this paper, we report an on-going research for a method based on the similar context (also known as the context feature approach), but extending the context features to multi-word instead of using a single word per feature. We called this method as multi-word level context features (*MWCF*) in general.

2 RELATED WORK

In the standard context feature approach, the source word and the target word are vectors with their dimensions are defined by single-word context features comprising of unigrams that co-occur within a certain window around the source or the target word, respectively. To project the source and the target word vectors onto the same space, the target dimensions are translated into the source language using initial bilingual lexicons. The similarity between the source and the target word can be computed automatically using measure such as the *cosine similarity metric*. For a further description on BLE we refer readers to [11].

The major drawback for the standard approach is the sensitivity to the corpora and the initial bilingual lexicon. According to [6], the outcomes can be affected for many reasons, which may include: (1) context features being weak or sparse, (2) the most important context features not occurring in the corpora, and (3) some context features missing due to low occurrence. These are all related to the corpora that are being used in these settings.

Based on our observation, other factors related to the initial bilingual lexicon, which includes: (1) missing important features from

[1] University of York, UK, email: {azniah,suresh}@cs.york.ac.uk

the initial bilingual lexicons, (2) the lexicon entries being too general, (3) the lexicon has low coverage, or (4) the lexicon contains many ambiguous words that might probably makes single word features to mislead, also causes sensitivity issues. Hence, both corpora and bilingual lexicon might have significant effect to the outcomes. For example, in the worst case scenario of English to Malay translation, given the source word *coach* and and a vector of $< student, bag, team, door, white >$, we do not know which Malay word it corresponds to. The situation can be a lot worse if the comparable corpora are not in similar domains. It may be equivalent to the Malay word *jurulatih* ("a trainer or instructor") or it may be equivalent to the Malay word *koc* ("a type of transportation"). As single word feature can be ambiguous, there might be other unexpected target word that also seems to share the features due to the limited size and coverage of the initial bilingual lexicon. The approach we adopt attempts to overcome this problem. We generalize the idea and introduce context features at the multi-word level. In order to construct the multi-word features, we utilize the n-grams.

Solving BLE related problems using n-grams is not new. Some properties of multi-word units to remain together in sequence that makes n-gram method favorable, for example, the specialized terminology like *Lyme Disease* and *Carpal Tunnel Syndrome*. BLE related work includes [4] that learn bilingual collocations by finding similar word chunks of n-grams and [12] that use various n-gram models to generate translation units for the BLE. Apart from the n-grams, no other similarity to our approach is found. [4] and [12] learn the n-grams from parallel corpora. They use the n-grams to extract multi-word correspondences. Context feature approach is not involved in either of the work.

In our approach, n-grams are used only as windows to capture sets of context words that co-occur together with the source or the target word. Thus the multi-word features derived are not necessarily in the same sequence as in the n-grams. We use these features to extract word-to-word correspondences for the non-related, English-Malay language pair. As far as we know, no other BLE method for the English-Malay language pair has been attempted in unsupervised settings. However, our method is not restricted to other roman character based language pairs. It is important to mention that our work follows the bag-of-word model such as [5] and [6], among others. Other methods, especially that follow certain order or pattern, usually require POS-tagged corpus (see [10] and [2]). We do not allow this in our setting.

3 LEARNING THE MULTI-WORD CONTEXT FEATURES

In this section, we describe the multi-word features by using sample texts shown in Figure 1. Given the source word is *coach*, the word level n-grams that contain it are extracted from the texts. For example, the first 4-grams derived from the first EN sentence is "*Argentina coach Diego Maradona*", followed by "*coach Diego Maradona has*".

Each n-gram may contribute a multi-word feature candidate of context words that co-occur with the word *coach* within the n-gram window. For example, if the n-gram is "*Argentina coach Diego Maradona*", the multi-word feature is $< Argentina, Diego, Maradona >$. Figure 2 shows the multi-word feature lists for the source word *coach* and for its potential translation equivalent in Malay, i.e. *jurulatih*. The n-gram frequencies are obtained from very small comparable corpora.

We observed that when the source word occurs highly with the multi-word feature, its corresponding word in the target language may also occur highly with the multi-word feature correspondence. We also observe that both may also share certain common multi-word features (in this case, all words between a multi-word feature pair are identical). For example, *Argentina, diego, maradona* is observed in both sides. Such multi-word features can be used to map the source word to its corresponding target word. However, word order can be different between non-related language pairs. We therefore define the multi-word context feature as a multi-word unit that contains a set of context words that co-occur together in an n-gram window but the context words are not necessarily in certain order (bag-of-word).

4 *BLE* METHOD USING *MWCF* APPROACH

Suppose that English is the source language and Malay is the target language, the method includes:

4.1 Pre-processing

The corpora are cleaned by removing all html tags (if any).

4.2 Select the source and the target words

To get the source word list, we use sentence boundary detection and tokenization on the pre-processed content before we filter out the stop words. We sort words in the text according to their frequencies. We take medium frequency words into the list (see subsection 5.1 for details).

We repeat the same procedure to obtain the target word list. Both lists may include any word type.

4.3 N-gram extraction

To learn the n-grams from the corpus, we make use of the sentence boundary. We collect the n-grams, which $3 \leq n \leq 6$ and that occur more than once. We store the count value, which is defined as the number of documents containing the n-grams divided by the total number of documents. We repeat the process for the target words.

4.4 Multi-word feature extraction

For each source word, we store their unique n-grams. We remove the source word from the n-grams and keep the remaining words (context words) in a list as a multi-word feature unit for the source word. We repeat the process for the target words.

4.5 Finding the translation pairs

To obtain the potential translation pairs, we find target word (W_T) that maximizes the similarity with the source word (W_S).

$$tr(W_S) = argmax\ sim(W_S, W_T)$$

We use cosine measure to get the similarity values:

$$cos(\theta) = \frac{A*B}{||A||*||B||}$$

where A and B are the elements in the vectors W_S and W_T respectively, and $A * B$ is the dot product of the two vectors.

The potential translation pairs are then sorted in the descending order. For each English source word, we considered high ranked Malay target words, for which the similarity values must not be below a threshold t_1.

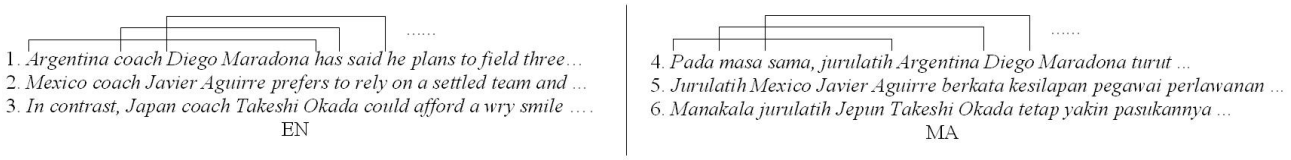

Figure 1. Sample of non-parallel English (EN) and Malay texts (MA) from comparable corpora. The EN contains the source word *coach* while the MA contains the target word that equivalent to the source word, i.e. *jurulatih*. The block of lines on the first EN sentence showing some examples of 4-grams that can be drawn from the sentence.

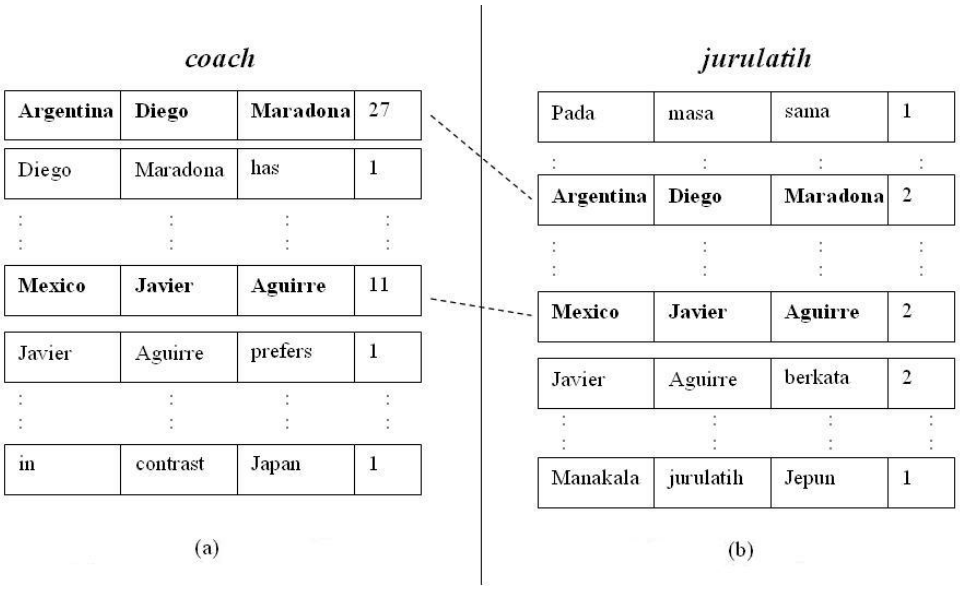

Figure 2. Examples of multi-word features for the source word *coach* and for the target word *jurulatih* (a) derived from an English corpus, and (b) derived from a Malay corpus.

5 EXPERIMENTAL SETTING

5.1 Data

Although corpora play a major role in determining the outcome, yet we have not much choice. Therefore, we built our own corpora ($Corpora_{WC}$) by compiling the *World Cup 2010* online news articles from June 11^{th}, 2010 to July 11^{th}, 2010.

This compilation was done automatically and involved several major online newspaper in Malaysia such as *Berita Harian*, *Utusan Malaysia*, *The New Straits Times* and *The Stars*. We also add news articles from the *FIFA* official website. In order to capture the content of the report, we took only texts in between $< p >$ and $< /p >$ tags using regular expression. The raw English-Malay corpora are very small, each contains 2,287 and 1,304 articles respectively.

In order to get the source word list, we use sentence boundary detection and tokenization before we sort the corpus words. The raw English corpus contains 27,642 unique English words. However, according to [5], high frequency words tend to be noisy. Thus, we remove the first 25 words from the frequency list. We take only words that occur more than 15 times in the corpus (i.e. about 1500 words). From the list, we filter the stop words, remove words that occur in the initial bilingual lexicon entries and also words of length less than 4. We repeat the same procedure to obtain the target word list from the raw Malay corpus. The Malay corpus contains 9,810 unique Malay words.

5.2 Lexicons

- *Initial billingual lexicon*
  Since the $Corpora_{WC}$ contains many identical names and places for both languages such as *Diego*, *Maradona* and *Argentina*, we take advantage of these bilingual word pairs. We follow the methods in [7] to obtain the initial bilingual lexicon. However, we only consider word pairs with identical spelling. For that, we compile bilingual word pairs using *string edit distance* with the zero distance. Each word must be longer than 4 characters.

- *Evaluation lexicon*
  For the evaluation lexicon, we extract the English-Malay word pairs from the *Websters dictionary* [2]. If the English word has multiple Malay translations, only Malay translation correspondence that occurs in the Malay corpus is considered.

[2] http://www.websters-online-dictionary.org

Method	$P_{0.10}$	$P_{0.25}$	$P_{0.33}$	$P_{0.50}$	$F1score$
$SWCF_{WC2}$	0.00	9.75	14.78	14.63	18.00
$MWCF_{WC2}$	-	-	-	-	-
$SWCF_{Web}$	4.74	3.79	2.87	1.89	1.87
$SWCF_{Web,n-word}$	9.48	7.56	6.55	4.86	7.79
$MWCF_{Web,n-gram}$	33.00	28.57	20.00	20.00	28.57

Table 1. Performance of different methods.

5.3 Evaluation

In the experiments we considered the task of building Malay-English lexicon of word-to-word correspondences. Since the data is small, we are not being too ambitious. We evaluate the proposed lexicon against the evaluation lexicon. The precision is given by the number of correct translation pairs at a certain proportion of the proposed lexicon. We follow [3] that defines the recall as the proportion of the proposed lexicon divided by its full size. We also used the $F1$ measure to compare the performance among the methods:

$$F1 = \frac{2*P*R}{P+R}$$

Here, P and R are the precision and recall values, respectively.

5.4 Baseline method

For comparison purposes, we implement the standard approach (which, in general, we name it as the single-word level context features approach or $SWCF$) in a similar setting. The single-word features are collected from context words that co-occur with the source word in the source corpus (or target word in the target corpus) within a window of a sentence.

6 EXPERIMENT RESULTS

In this section, we report the experiment results. Table 1 shows the $F1$ scores and various precisions p_x at recall values x for the methods that we used.

6.1 Single vs multi-word feature approach

We attempt to learn bilingual lexicons from the $Corpora_{WC}$ using the standard approach. We labelled the approach as $SWCF_{WC}$. Unfortunately, this method has poor performance. It seems that $Corpora_{WC}$ have low reliability between the corpus. Thus, we deliberately filter the corpora, into $Corpora_{WC2}$, and re-implement the standard approach for the second attempt (labeled as $SWCF_{WC2}$). We use date and document length similarity features for the filteration.

With $Corpora_{WC2}$, the $SWCF$ performance improves to about 18% of $F1$ score. We use SWF_{WC2} as our baseline method. The $Corpora_{WC2}$ replace the $Corpora_{WC}$ in the experiments for multi-word feature. However, poor results has been recorded with the n-gram based multi-word feature approach ($MWCF_{WC2,n-gram}$).

6.2 Incorporating data from the web

Since the data is too small and the $MWCF_{WC2,n-gram}$ approach aggravates the data sparse problem, we incorporate data from the web. Each existing feature unit in the source language accompanied by its source word may contribute as a query to the search engine. In the $SWCF$ approach we look for sentences that contain the source word from the returned documents. Again, we collect context words occurring around the source word within a sentence adding to the existing context features. We repeat the process for the target. With this new added data, the performance of the $SWCF$ approach (now labeled as $SWCF_{Web}$) has deteriorated with less than 2% $F1$ is recorded.

Meanwhile, for the $MWCF$ approach using the web data ($MWCF_{Web,n-gram}$), we use each multi-word feature unit accompanied by its source word as a query. We derive n-grams containing any source word from the matched document. From these n-grams we extract and add the multi-word features to the existing features for that particular source word. Note that each of context words in the features must be part of the initial bilingual lexicon entries. We use similar queries to obtain documents in the target language, except that the source word is replaced with each target word from the target word list, one at a time. We repeat the multi-word feature extraction but this time we also obtain Malay words that have potential to become the correspondence of the source word. The $MWCF_{Web,n-gram}$ achieves almost 30% of $F1$.

6.3 *N-word* feature

For fair comparison to the single-word context feature approach, we also implement a method of *n-word* feature approach. We use similar single-word feature setting, except that more than one (bag-of-word) context words are extracted to build every single feature unit. Thus, the approach ($SWCF_{Web,n-word}$) is at multi-word level with larger, relaxed context window. The results improve slightly by almost 6% of $F1$ score when $SWCF_{Web,n-word}$, with a combination of 2 and 3 word, is used instead of just a single word ($SWCF_{Web}$).

6.4 *Top 1* based evaluation

Table 2 presents some examples of translation pairs proposed by $MWCF_{Web,n-gram}$. We observed that the proposed lexicon suggests more than one target word for each source word. Precision score could actually be improved further if only one candidate (the Top 1) is proposed for the source word. With Top 1, the final precision for $MWCF_{Web,n-gram}$ at 50% recall is significantly improved from 20.00% to 43.56%. However, the approach does not work well with $SWCF_{WC2}$, where the precision at 50% recall drops from 14.63% to 7.14%.

English	Malay	Sim score	Correct?
former	*presiden*	0.1294	No
president	*presiden*	0.1216	Yes
playmaker	*presiden*	0.0958	No
believes	*pengurus*	0.0540	No
coach	*jurulatih*	0.0318	Yes
league	*pemain*	0.0250	No
former	*pemain*	0.0250	No
coach	*pengurus*	0.0242	No
president	*liga*	0.0236	No
striker	*penyerang*	0.0214	Yes

Table 2. Some examples of translation pairs proposed by $MWCF_{Web,n-gram}$ ranked by similarity scores.

7 DISCUSSION

7.1 The potential of multi-word context features

Our experiments clearly demonstrate that multi-word context feature method achieves higher $F1$. We also experimented with multi-word in SWCF setting ($MWCF_{Web,n-word}$) to make a fair comparison in very similar setting. $MWCF_{Web,n-word}$ achieves higher $F1$ score compared to $SWCF_{Web}$.

7.2 Determining the window of the n-grams

Effect of window size of is listed in Table 3. It seems that the precision is improved with larger n-gram size. However, the number of such features is very limited. Hence, larger windows of the n-grams are good discriminators but sparse. In addition, we observed that major part of the few articles found in the web for n-grams with larger window are actually the articles that we downloaded to the $Corpora_{WC}$ previously. Therefore, the reliability of the documents is very important.

Method	$P_{0.50}$
$MWCF_{Web,3-gram}$	33.33
$MWCF_{Web,4-gram}$	100.00
$MWCF_{Web,5-gram}$	100.00

Table 3. Effects on precision score at 50% recall for $MWCF_{Web,n-gram}$ (with Top 1 evaluation) in different n-gram windows.

7.3 Web data alleviates data sparsity problem

By incorporating data from the web into our $MWCF$ methods, the data sparsity has been relieved. Whilst, the single-word context feature based approach seems to add more noise to the existing data with this approach.

The multi-word level context features have more potential than single-word level context features. Taking more than one word for consideration might alleviate the ambiguity of each feature. The approach requires massive data but incorporating data from the web has in fact relieved the data sparsity and making this approach possible.

The overall outcome is still not so impressive but it could be improved further. From our observation, not the entire multi-word level context features are highly relevant to the source word (or target word, respectively). We could consider a feature verification model which allow only highly recognized features to be considered as the multi-word level features. However, the major drawback of the multi-word context feature approach is the fact that a majority of the features does not occur with the target words in the web although the web offers a huge Malay corpus.

8 CONCLUSION

We have developed an unsupervised bilingual lexicon extraction method to improve the precision score in extreme setting. Resolving some issues like context features is useful to improve the precision of the extracted bilingual lexicons. In this paper, we demonstrate that the use of context features at multi-word level achieves higher precision compared to single-word context feature approach. Moreover, incorporating data from the web has helped to make this technique possible. Nonetheless, using n-grams keeps the technique simple.

REFERENCES

[1] Pascale Fung and Percy Cheung, 'Mining very non-parallel corpora: Parallel sentence and lexicon extraction via bootstrapping and em', in *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Barcelona, Spain, (2004).

[2] N. Garera, C. Callison-Burch, and D. Yarowsky, 'Improving translation lexicon induction frommonolingual corpora via dependency contexts and part-of-speech equivalences', in *Proceedings of the Conference on Computational Natural Language Learning 2009*, (2008).

[3] Aria Haghighi, Percy Liang, Taylor Berg-Krikpatrick, and Dan Klein, 'Learning bilingual lexicons from monolingual corpora', in *Proceedings of the ACL 2008*, Columbus, Ohio, USA, (June 2008).

[4] Masahiko Haruno, Saturo Ikehara, and Takefumi Yamazaki, 'Learning bilingual collocations by word-level sorting', in *Proceedings of the COLING 16*, (1996).

[5] Azniah Ismail and Suresh Manandhar, 'Utilizing contextually relevant terms in bilingual lexicon extraction', in *Proceedings of the Workshop on Unsupervised and Minimally Supervised Learning of Lexical Semantics*, Boulder, Colorado, USA, (June 2009). Association for Computational Linguistics (ACL).

[6] Azniah Ismail and Suresh Manandhar, 'Bilingual lexicon extraction from comparable corpora using in-domain terms', in *COLING 2010: Posters*, Beijing, China, (August 2010).

[7] Philipp Koehn and Kevin Knight, 'Knowledge sources for word-level translation models', in *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, (2001).

[8] Philipp Koehn and Kevin Knight, 'Learning a translation lexicon from monolingual corpora', in *Proceedings of the ACL*, pp. 9–16, Philadelphia, USA, (2002).

[9] Dragos Stefan Munteanu and Daniel Marcu, 'Extracting parallel subsentential fragments from non-parallel corpora', in *Proceedings of the ACL 2006*, Sydney, Australia, (2006).

[10] P. Otero and J. Campos, 'Learning spanish-galician translation equivalents using a comparable corpus and a bilingual dictionary', in *LNCS 2008*, (2008).

[11] Reinhard Rapp, 'Identifying word translations in non-parallel texts', in *Proceedings of ACL 33*, pp. 320–322, (1995).

[12] Kaoru Yamamoto, Yuji Matsumoto, and Mihoko Kitamura, 'A comparative study on translation units for bilingual lexicon extraction', in *Proceedings of the Workshop on Data-Driven Machine Translation, ACL 39*, pp. 87–94, Toulouse, France, (2001).

Using Multilingual Corpora to Extract Semantic Information

Ahmad R. Shahid and **Dimitar Kazakov** [1]

Abstract. This paper presents a technique to build a lexical resource used for annotation of parallel corpora where the tags can be seen as multilingual 'synsets'. The approach can be extended to add relationships between these synsets that are akin to WordNet relationships of synonymy and hypernymy. The paper also discusses how the results can be evaluated. The reported results are for English, German, French, and Greek using the Europarl parallel corpus.

1 INTRODUCTION

The aim of this work is to build a WordNet-like resource which can be used for Word Sense Disambiguation (WSD) and other such tasks where semantics of words and phrases is the main objective. The multilingual aspect of the approach helps in reducing the ambiguity inherent in any words/phrases in the pivotal language, which is English in the case shown here.

In order to create such a resource we used proceedings from the European Parliament (Europarl)[2]. Four languages were selected with English as the pivotal language in addition to German, French and Greek.

The paragraph-aligned bilingual corpora were fed into a word-alignment tool, GIZA++, to obtain the pair-wise alignments of words in each language with English. These pair-wise aligned words were later merged into phrases where one word in one language was aligned with one or more than one word in the other language, to create minimal closures spanning words from all four languages (see Figure 1) according to the following principle: If each pair of word-aligned corpora is seen as a bipartite graph, consider the graph formed by the union of all (four) such bipartite graphs. Then, for each word in the pivotal language, find the largest connected graph that includes that word and spans all four languages. Using English as the pivotal language, these were combined into 4-tuples, effectively resulting in a database of multilingual synsets, in total more than 500,000 unique synsets. The index numbers were later assigned as sense tags to the original English corpus, essentially forming the disambiguation task. The sense tagged corpus needs to be avaluated to ascertain the veracity of the hypothesis that other languages help in disambiguating the senses of a word in the pivotal language by narrowing down its senses. The results of evaluation could determine if the proto-synsets need further refinement by merging the ones that are syntactically and/or semantically related.

Refining the proto-synsets is not trivial and requires auxiliary information. For the said purpose edit distances have been measured between every two synsets that share the same English word/phrase. The edit distances contain syntactic and semantic information that needs to be extracted if it is to be used for refinement. Synonymy relationships between different English phrases, based on their context, could also help in merging the proto-synsets. The English corpus has also been POS-tagged using the Brill Tagger. All these are tools that could be used to further the goal of refinement, and are goals for future development and research.

[1] University of York, UK, email: {ahmad,kazakov}@cs.york.ac.uk

[2] http://www.statmt.org/europarl/

2 RELATED WORK

WSD has attracted the attention of the research community for long. It is a tricky issue and needs resources that define the semantic relationships between words. In the last twenty five years various research activities have been undertaken to build large repositories that combined the description of semantic concepts with their relationships. Two efforts worth mentioning here are the Cycorp Cyc project [16] and the lexical semantic database WordNet [19]. Both approaches use a number of predicates to define relationships between concepts, such as "concept A is an instance of concept B" or "concept A is a specific case of concept B." WordNet also defined the notion of *synsets*, which defines a semantic concept through all relevant synonyms, *e.g.*, {mercury, quicksilver, Hg}.

The original version of the WordNet covered only the English language but the effort has been replicated for other languages as well [24]. Yet all these efforts have been handcrafted, rather than automatically generated and are monolingual in nature. Even though they are highly comprehensive, they require a major, sustained effort to maintain and update.

[9] used word alignment in an unsuperised manner to create pseudo-translations which were used for sense tagging of the parallel corpora. They used WordNet as the sense inventory of English. Firstly they aligned each French word with one or more words in English in each sentence. Then to create synsets they looked at the alignment of each French word with all corresponding translations in English in the whole corpus. In order to narrow down the number of combinations they used WordNet to identify nominal compounds, such as $honey\_bee$ and $queen\_bee$. WordNet was also used to manually assign sense tags to words in the subset of the corpus used for evaluation. They found the performance of their approach comparable with other unsupervised approaches.

Interest in the use of parallel corpora for unsupervised WSD has grown recently [12, 15]. In both cases, the use of multilingual synsets is discussed together with various ways of reducing their number.

3 MULTILINGUAL SYNSETS

Creating multilingual synsets is at the core of this project. Naturally emanating from word alignment in parallel corpora, they make a crucial link between semantics in the original bilingual corpora and the

English	German	French	Greek
resumption of	wiederaufnahme	reprise de	επανάληψη της
session	sitzungsperiode	session	συνσδου
adjourned on friday	erkläre am freitag	interrompue vendredi	διακοπεί παρασκευή
like once again	nochmals	renouvelle	ξανά
thank you	vielen dank	merci	ευχαριστώ
shall do so gladly	will :un gerne	ferai volontiers	πράξω ευχαρίστως

Figure 1. Examples of Synsets.

Figure 1 gives a few examples of the synsets. As can be seen many synsets are phrases rather than words. In the example one synset is comprised of four words "shall do so gladly".

Multilingual synsets help in disambiguating the senses of a word. Translating the English word 'bank' with the French 'banque' suggests two possible meanings: a financial institution or a collection of a particular kind (e.g., a blood bank), as these words share both meanings, but eliminating the English meaning of a 'river bank'. Increasing the number of languages could gradually remove all ambiguity, as in the case of {EN: bank, FR: banque, NL: bank}. Insofar these lists of words specify a single semantic concept, they can be seen as WordNet-like synsets that makes use of words of several languages, rather than just one. The greater the number of translations in this multilingual WordNet, the clearer the meaning, yet, one might object, the fewer the number of such polyglots, who could benefit from such translations. However, these multilingual synsets can also be useful in a monolingual context, as unique indices that distinguish the individual meanings of a word.

When annotating parallel corpora with lexical semantics, the multilingual synsets become the sense tags and the parallel corpora are tagged with corresponding tags in a single unsupervised process. The idea is as simple as it is elegant: assuming we have a word-aligned parallel corpus with n languages, annotate each word with a lexical semantic tag consisting of the n-tuple of aligned words. As a result, all occurrences of a given word in the text for language $\mathcal{L}$ are considered as having the same sense, provided they correspond to (are tagged with) the same multilingual synset.

Two great advantages of this scheme are that it is completely unsupervised, and the fact that, unlike manually tagged corpora using WordNet, all words in the corpus are *guaranteed* to have a corresponding multilingual synset.

4 SYNSET GENERATION AND WSD

In order to generate the synsets we needed the word-aligned corpora. The Europarl corpus was taken. It was pre-processed, which included among other steps, tokenization of text, lowercasing, removal of empty lines and the removal of XML-tags. After pre-processing a paragraph aligned parallel corpus was obtained. English corpus was used as the pivotal one. All these were fed to GIZA++[3], a standard and freely available tool for word alignment. For alignment, pairwise corpora were fed into GIZA++ (German with English, French with English, and Greek with English). Thus the output of GIZA++ were pair-wise aligned parallel corpora with markings indicating which words in the target language aligned with which words in English. It might be the case that one word in one language aligns with more than one words in another or it aligns with nothing. Only the aligned words were of any use while generating synsets from the aligned corpora.

For actual synset generation from the aligned corpora we designed our own algorithm, which links two or more words in one language together if they align with the same word in another language. The process had to be carried out simultaneously for all the four languages, so as no useful information is lost.

The algorithm links the words of the pivotal language (PL) into phrases and maps all words of the non-pivotal languages to one of these phrases. The array a[1..N] serves to store in the field a[i] the number of the phrase to which word i in the pivotal language belongs. Initially, all PL words are assumed to belong to different phrases (i.e., they form a phrase on their own). Two or more PL words $a[j], ...a[j+k]$ are placed in the same group if there is a word in another language, which is aligned with all of them. This information is stored by assigning the same phrase number to $a[j], ..., a[j+k]$. The array t is used to store information about the word alignment between each non-PL and the PL. The assignment t[l,i]:= k represents the fact that the i-th word in non-PL l was aligned with the k-th word in the PL.

Subsequently, each synset is spelt out by producing a phrase in the pivotal language (consisting of one or more PL words with the same phrase number) and extracting for each non-PL language all the words that point to a PL word in that group: this final step is straightforward, and due to space limitations is not shown in Table 1.

```
Data Structures:

int N % number of words in the PL
int M % number of non-PLs
int array a[1..N] int array t[1..N,1..M]

Initialize:

for i=1 to N do a[i] := i

Form phrases:

 for l=1 to M
 |  L := number of words in lang.l
 |  for i=1 to L
 |  | if word i in lang.l is aligned
 |  |   with word j in the PL
 |  | then t[l,i] := j
 |  |   elseif word i in lang.l is aligned
 |  |       with words j,j+1,j+k in the PL
 |  |   then
 |  |       t[l,i] :=j
 |  | for z=1 to k do
 |__|___a[j+z] := a[j]
```

Table 1. Synset Generation Algorithm.

The WSD task was later performed by assigning indices to the synsets that would indicate common entries. Thus any two synsets with exactly the same words/phrases in all the four languages got

[3] http://fjoch.com/GIZA++.html

the same indices. This information was later used to sense tag the original English corpus. Thus each phrase in the corpus was replaced by index number of the synset at that position in the corpus. This sense tagged corpus is rich in semantic information, which needs to be evaluated.

Part of Speech (POS) is an extra bit of useful information that can be used for WSD [3, 14]. POS tags of the neighbors of the target word help in narrowing down the meanings of the word. We used Brill Tagger [2] to assign POS tags to individual words in the English phrases in the synsets.

The approach described here produces a large number of what we would call 'proto-synsets' - for a corpus of more than 1.8 million words, there are more than 1.5 million such synsets, yet a little more than 500,000 are unique, by virtue of their bearing different words/phrases in all the four languages. Their number can be reduced and their composition—brought closer to what one would expect to see in a hand-crafted dictionary in the following two ways: firstly, through the identification and merger of proto-synsets only varying in word forms corresponding to the same lexical entry (e.g., flight-X-Y-Z, flights-X-Y-Z); secondly, through the merger of proto-synsets in which the differences are limited to words that are synonyms in the given language (e.g., car-auto-*automobile* vs car-auto-*voiture*). These two approaches are addressed in the following two sections.

5 EDIT DISTANCES

For the purposes of refinement, merging of synsets based on their syntax and semantics is the crucial bit. Morphemes could be both inflectional and derivational. In inflectional morphemes the meaning is not changed. Hence both *dog* and *dogs* have the same meaning and *dogs* is an inflection of *dog*. In derivational morphemes, however, the meaning might change. Thus *unhappy* is derived from *happy*, yet they are antonyms of each other. Both inflectional and derivational morphemes could be reflected in edit distance measures, which hide a lot of useful information. Yet, due to their lack of clear distinction between inflectional and derivational morphemes, can only be used as auxiliary information along with POS tags and synonymy relationships.

While edit distances, specially small distances, could be a good guide to differentiating the syntactic from the semantic. For derivational morphology, since semantics if important, synonymy detection would be a better tool to refine the proto-synsets, which is discussed in the next section.

Edit distance measures the minimum number of edit steps required to convert one string into another [11],[13],[17]. The only three operations allowed are *insertion* of a character from the first string, *deletion* of a character from the first string, or *substitution/replacement* of a character in the first string with a character in the second string. Thus *dogs* has an edit distance of 1 with *dog*, since only a deletion of 's' would suffice for conversion. There might be more then one ways to conversion, hence the minimum edit distance is a more useful measure.

We divided the synsets into two groups. The first group contained all the synsets with frequency one, based on the English phrase. The other group contained synsets which have frequency more than one, based on their English phrase. Pair-wise edit distances were measured between every two synsets that shared the English phrase. This information would be used in future to determine which two synsets should be merged.

6 SYNONYMY DETECTION

Synonymy is a relationship between words which makes them inter-substitutable. Yet [7] says that "natural languages abhor absolute synonyms just as nature abhors a vacuum." Absolute synonymy is rare and restricted mostly to technical terms [10]. Near-synonyms are of greater significance and are very similar but not completely inter-substitutable or identical.

According to [23] a common approach to synonymy detection is distributional similarity. Thus synonymous words share common contexts, and thus they could be inter-substituted without changing the context. They showed that use of multilingual resources for extraction of synonyms had higher precision and recall as compared to the monolingual resources.

Turney [22] used PMI-IR (Pointwise Mutual Information and Information Retrieval) to determine the synonymy between two words. The algorithm maximizes Pointwise Mutual Information [6, 5], which in turn is based on co-occurrence [18].

We can use the above ideas to detect synonymy between the words/phrases for a given language, then merge the multilingual proto-synsets that only vary in this respect. Similarly, we can apply similarity measures to 4-tuples, e.g., if the words/phrases in all but one langugage are the same, or a number of alternatives for some languages appear together in several permutations, e.g., car-auto-auto, car-auto-voiture, automobile-auto-auto, automobile-auto-voiture, we can consider them as synonyms.

7 EVALUATION

Evaluation of the results need to be carried out and some preprocessing steps have been accomplished, like preparing the data for evaluation, and creating clusters. Yet the clusters need to be compared. Yet ascertaining the veracity of the claim that cues from other languages help in disambiguating the pivotal language, is not a trivial task.

Due to lack of a gold standard and other benchmarks for this study, we had to define our own gold standard, with which our own results could be compared. Thus the original corpus tagged with chapter tags, considered to be classes, and speaker tags corresponding to individual documents, can be considered as the gold standard. The Europarl parallel corpus, available in many European languages, detail the European parliamentary proceedings. The translations of the Europarl proceedings demonstrate a great level of consensus among the human annotators.

In order to evaluate the results we have clustered documents in the original English corpus without the sense tags and with the sense tags as explained above. The next step is to determine if assigning sense tags to the English corpus using the proto-synsets improve clustering in any way over the clusters produced by the corpus without the sense tags. Thus the clusters formed in each case need to be compared with the gold standard classes and each cluster needs to be paired with the class that shares the most number of documents with it.

Standard measures of detecting the goodness of clusters can be employed to see if after WSD the new clusters have more in common with the classes. [1] have compared clusters obtained from the English, Bulgarian parallel corpora. Their results show that smaller number of clusters provide better mapping between clusters of the two languages, with a high degree of purity. With 10 clusters, they obtained 100% mapping.

Different measures could be adopted to ascertain the goodness of the clusters. One such measure is the Davies-Bouldin Index (DBI)

[8]. DBI takes into account both the intra-cluster and inter-clusters distances to ascertain the quality of clusters produced. The intra-cluster distances can be measured using distances between individual documents and the cluster centroid, defined in (1), where Q_k is the cluster k, $x_i \in Q_k$, c_k is the center of the cluster, $k \leq K$ and N_k is the number of documents in the cluster.

$$d_{centroid}(Q_k) = \frac{\sum_i \parallel x_i - c_k \parallel}{N_k} \quad (1)$$

For inter-cluster distances is basically the distance between the centroids of the two clusters, for which it is being measured, defined in (2), where c_k is the centroid of the cluster Q_k and c_l is the centroid of the cluster Q_l.

$$d_{between} = \parallel c_k - c_l \parallel \quad (2)$$

Thus DBI is defined as:

$$DBI = \frac{1}{K}\sum_{k=1}^{K} max_{l \neq k}\{\frac{d_{centroid}(Q_k) + d_{centroid}(Q_l)}{d_{between}(Q_k, Q_l)}\} \quad (3)$$

The aim is to minimize the DBI, which essentially means reducing the intra-cluster distances and increasing the inter-cluster distances. This is an internal criterion for measuring the quality of clusters and might not yield effective results in an application.

Other measures that could be employed are Purity, Precision, Recall and F-score. They can be used for comparing the results of clustering with the gold standard, an external measure. Purity, defined in (4), indicates how many of the documents in a cluster are correctly assigned a class, where K is the set of clusters, C is the set of classes, N is the number of documents, W_k is a particular cluster and C_j is a particular class and $|w_k \cap c_j|$ denotes the number of documents in cluster k that belong to a certain class. It is measured as a sum of individual cluster purities.

$$purity(K, C) = \frac{1}{N}\sum_k \max_j |w_k \cap c_j| \quad (4)$$

Precision, defined in (5), defines the fraction of documents in the cluster C, which is also in the class L.

$$precision(C, L) = \frac{|C \cap L|}{|C|}, C \in C_{ALL}, L \in L_{ALL} \quad (5)$$

Recall, defined in (6), is the fraction of the documents in class L that is also in cluster C.

$$Recall(C, L) = \frac{|C \cap L|}{|L|}, C \in C_{ALL}, L \in L_{ALL} \quad (6)$$

F-Score [20], [21], [4] defined in (7), combine both Precision and Recall with equal weight to each.

$$F - Score(C, L) = \frac{2 * Precision(C, L) * Recall(C, L)}{Precision(C, L) + Recall(C, L)} \quad (7)$$

8 CONCLUSION

The value of this approach is in its use of unsupervised techniques that do not require an annotated corpus. In this way, *all* words are guaranteed to be tagged with a synset, which is not often the case with other approaches. This has been done on a large dataset with more than 1.8 million words. WSD of such a large corpus is valuable even if the additional benefits of the lexical resource produced are not considered.

REFERENCES

[1] R. Alfred, D. Kazakov, M. Bartlett, and E. Paskaleva, 'Hierarchical aggolemrative clustering for cross-language information retrieval', *International Journal of Translation*, **19**(1), 139–162, (2007).

[2] E. Brill, 'A simple rule-based part of speech tagger', in *Proceedings of the Third Conference on Applied Natural Language Processing*, pp. 152–155, (1992).

[3] R. Bruce and J. Wiebe, 'Word-sense disambiguation using decomposable models', in *Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 139–146, (1994).

[4] W. chiu Wong and A. Fu, 'Incremental document clustering for web page classification', Japan, (2000).

[5] K.W. Church, W. Gale, P. Hanks, and D. Hindle, *Using Statistics in Lexical Analysis*, chapter In Lexical Acquisition: Using On-Line Resources to Build a Lexicon, edited by Uri Zernik, 115–164, Lawrence Erlbaum, 1991.

[6] K.W. Church and P. Hanks, 'Word association norms, mutual information and lexicography', in *Proceedings of the 27th Annual Meeting of the Association of Computational Linguistics (ACL)*, pp. 76–83, (1989).

[7] A.D. Cruse, *Lexical Semantics*, Cambridge University Press, Cambridge, UK, 1986.

[8] D. L. Davies and D. W. Bouldin, 'A cluster separation measure', *IEEE Transactions and Pattern Analysis and Machine Intelligence*, **1/2**, 224–227, (1979).

[9] M. Diab and P. Resnik, 'An unsupervised method for word sense tagging using parallel corpora', in *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 255–262, (2002).

[10] P. Edmonds and G. Hirst, 'Near-synonymy and lexical choice', *Computational Linguistics*, **28**(2), 105–145, (2002).

[11] D. Gusfield, *Algorithms on Strings, Trees and Sequences*, Cambridge University Press, Cambridge, UK, 1997.

[12] D. Kazakov and A.R. Shahid, 'Unsupervised construction of a multilingual Wordnet from parallel corpora', (2009).

[13] J.B. Kruskal, 'An overview of sequence comparison: Time warps, string edits, and macromolecules', *SIAM Review*, **25**(2), 201–237, (1983).

[14] Y.K. Lee and H.T. Ng, 'An empirical evaluation of knowledge sources and learning algorithms for word sense disambiguation', in *Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing - Volume 10*, pp. 41–48, (2002).

[15] E. Lefever and V. Hoste, 'Semeval-2010 task 3: Cross-lingual word sense disambiguation', in *Proceedings of the Workshop on Semantic Evaluations: Recent Achievements and Future Directions*, (2009).

[16] D. B. Lenat, 'Cyc: A large-scale investment in knowledge infrastructure', *Communications of the ACM*, **38:11**, 33–38, (1995).

[17] V.I. Levenstein, 'Binary codes capable of correcting, insertions and reversals', *Sov. Phys. Dokl.*, **10**, 707–710, (1966).

[18] C.D. Manning and H. Schutze, *Foundations of Statistical Natural Language Processing*, The MIT Press, 1999.

[19] G. A. Miller, 'Five papers on Wordnet', *Special Issue of International Journal of Lexicogrphy*, **3:4**, (1990).

[20] M. Steinbach, G. Karypis, and V. Kumar, 'A comparison of document clustering techniques', in *In KDD Workshop on Text Mining*, (2000).

[21] A. Strehl. Relationship-based clustering and cluster ensembles for high-dimensional data mining. PhD thesis.

[22] P. Turney, 'Mining the web for synonyms: Pmi-ir versus lsa on TOEFL', in *Proceedings of the Twelfth European Conference on Machine Learning*, pp. 491–502, (2001).

[23] L. Van der Plas and J. Tiedemann, 'Finding synonyms using automatic word alignment and measures of distributional similarity', in *Proceedings of ACL/COLING 2006*, (2006).

[24] P. Vossen, *Eurowordnet: A Multilingual Database with Lexical Semantic Networks*, Kluwer Academic Publishers, 1998.

Author Index

Proceedings of AISB `11: Learning Language Models from Multilingual Corpora
Dimitar Kazakov and George Tsoulas (eds.)
ISBN 978-1-908187-05-5

Published by the Society for the Study of Artificial Intelligence and the Simulation of Behaviour
Printed by the University of York, York, UK