Development of an English-Macedonian Machine Readable Dictionary by Using Parallel Corpora

Martin Saveski[1] and Igor Trajkovski[2]

[1] Staffordshire University, Faculty of Computing, Engineering and Technology, College Road, Stoke-on-Trent, Staffordshire, UK
saveski.martin@gmail.com
[2] Ss. Cyril and Methodius University, Faculty of Electrical Engineering and Information Technologies, Rugjer Boshkovik bb, PO Box 574, Skopje, Macedonia
itrajkovski@feit.ukim.edu.mk

Abstract. The dictionaries are one of the most useful lexical resources. However, most of the dictionaries today are not in digital form. This makes them cumbersome for usage by humans and impossible for integration in computer programs. The process of digitalizing an existing traditional dictionary is expensive and labor intensive task. In this paper, we present a method for development of Machine Readable Dictionaries by using the already available resources. Machine readable dictionary consists of simple word-to-word mappings, where word from the source language can be mapped into several optional words in the target language. We present a series of experiments where by using the parallel corpora and open source Statistical Machine Translation tools at our disposal, we managed to develop an English-Macedonian Machine Readable Dictionary containing 23,296 translation pairs (17,708 English and 18,343 Macedonian terms). A subset of the produced dictionary has been manually evaluated and showed accuracy of 79.8%.

Keywords: machine readable dictionary, parallel corpora, word alignment, filtering word alignments

1 Introduction

The dictionaries are one of the most powerful reference tools that we use in our everyday lives. They are beneficial both in the process of learning a language and its everyday use. In the past all dictionaries had been in printed form. However, with the rapid growth of the technology, the need for dictionaries in digital form has tremendously increased. The process of digitalizing the existing traditional dictionaries is long, cumbersome, and requires a lot of resources. Moreover, the problem of usage of the traditional electronic dictionaries is that translations of some words are not given in explicit format (word-to-word or word-to-phrase) but with direct translation of sentences containing the word to sentences in the target language. In this case, it is hard to automatically find the translation of the word. Machine

readable dictionaries, on the other hand, have exact translation, or mapping, of given a word (phrase) to a word (phrase).

The Natural Language Processing community has greatly benefited from the presence of large amount of text provided in different languages in the form of parallel and comparable corpora. These kind of textual collections have been extensively used to automatically extract bilingual lexicons for a wide variety of applications. This potential has been most recognized by the researchers in the field of Machine Translation where the statistical approaches have dominated the grammatical, rule-based techniques. Due to this trend a large number of free and open source tools for processing parallel corpora have been developed.

The main objective of this study is by making use of the available parallel corpora and the open source Statistical Machine Translation tools to develop an English-Macedonian Machine Readable Dictionary (EN-MK MRD).

The remainder of this paper is organized as follows. In the next section we provide a short overview of the related work after which we explain our methodology and the experiments conducted. In sections 4 and 5, we evaluate the results of the experiments, and discuss the pros and cons of our approach and ideas for future work.

2 Related Work

The idea of using existing resources (parallel corpora) to produce a bilingual Machine Readable Dictionaries (MRD) is not new. As mentioned in the introductory section, it origins from the studies which introduced the techniques of using parallel corpora and statistical methods for the purpose of Machine Translation. We are aware of many studies which have successfully applied this technique and resulted with satisfactory outcomes. In the remainder of this section, we outline some of the attempts found in the literature and considered as most interesting.

Due to the low processing and storage capabilities the early attempts relied on smaller corpora and consequently resulted with small size MRDs. Most notable is the early work of Tiedemann J. in [2], where Swedish-English and Swedish-German dictionaries have been extracted. Similar study was conducted by Velupillai S. and Dalianis H. [3] who created 10 pairs of parallel corpora of Nordic Languages (Swedish, Danish, Norwegian, Icelandic and Finnish) which contained on average less than 80,000 words per language pair. The results reported for some of the language pairs have been very successful and reached accuracy of 93.1%.

However, studies which adopted methodology most similar to ours are the attempts to develop Greek-English and Chinese-English dictionaries. Charitakis K. in [1] developed a Greek-English MRD of 1,276 entries and achieved accuracy of 61.7%. On the other hand, the experiments conducted by Hao-chun Xing and Xin Zhang in [4] resulted in Chinese-English dictionary of 2,118 entries with accuracy of 74.1%. Our study differs from these two mainly in the size of the parallel corpus and the MRD extracted.

Although, we are aware of studies which collected and sentence aligned English-Macedonian parallel corpora [5], we do not know of any attempts for building a large bilingual EN-MK dictionary by using the already available resources.

3 Methodology

By presenting the experiments conducted, in the remainder of this section we discuss our methodology and explain each of the stages included. The main tool used in the experiments is the *Uplug* system. Namely, Uplug provides collection of tools for linguistic corpus processing, word alignment, and term extraction from parallel corpora. The system has been designed by Tiedemann J. in order to develop, evaluate, and apply approaches to generation of translation data from bilingual text [10]. Most importantly, the system is a modular-based platform and therefore each component can be extended or modified without affecting the system pipeline as a whole.

3.1 Small Scale Experiment

In order to test whether the statistical methods are applicable for producing EN-MK MRD a small scale experiment was conducted. For the purpose of the experiment the *KDE4* parallel corpus has been used. This corpus is part of *OPUS (Open Source Parallel Corpus)* [6] collected from the localization files of KDE, which is an open source software package containing a wide variety of applications for communication, education, and entertainment. The whole corpus contains *399,597* EN-MK tokens i.e. *71,046* sentences, where all localization files were tokenized, sentence aligned, and stored in xml files in a format suitable for word alignment with Uplug. After the execution of the advanced word alignment module of Uplug a list of *62,565* EN-MK word alignments was produced. However, many entries contained noise and incorrect translations. Since, manual evaluation of the results was not possible, radical filtering was applied to retain only the meaningful translations. Thus, all word alignments which occurred less than three times or contained punctuation or numbers were removed, resulting in a MRD with *5,228* entries and accuracy of *~70%*. These results were satisfying and encouraged further experiments with larger corpus to be done.

3.2 Large Scale Experiment

For the purpose of the second experiment the data produced by South European Times (SETimes - `http://www.setimes.com/`) news website was used. This website publishes daily news for all countries in south-eastern Europe and Turkey. Most importantly, the content of the website is available in ten languages including English and Macedonian. However, unlike KDE4 this corpus was not available in any preprocessed form so a lot of preprocessing had to be done to transform it in a form suitable for applying the Uplug modules. The whole process is depicted in figure 1 and further explained in the remainder of this section.

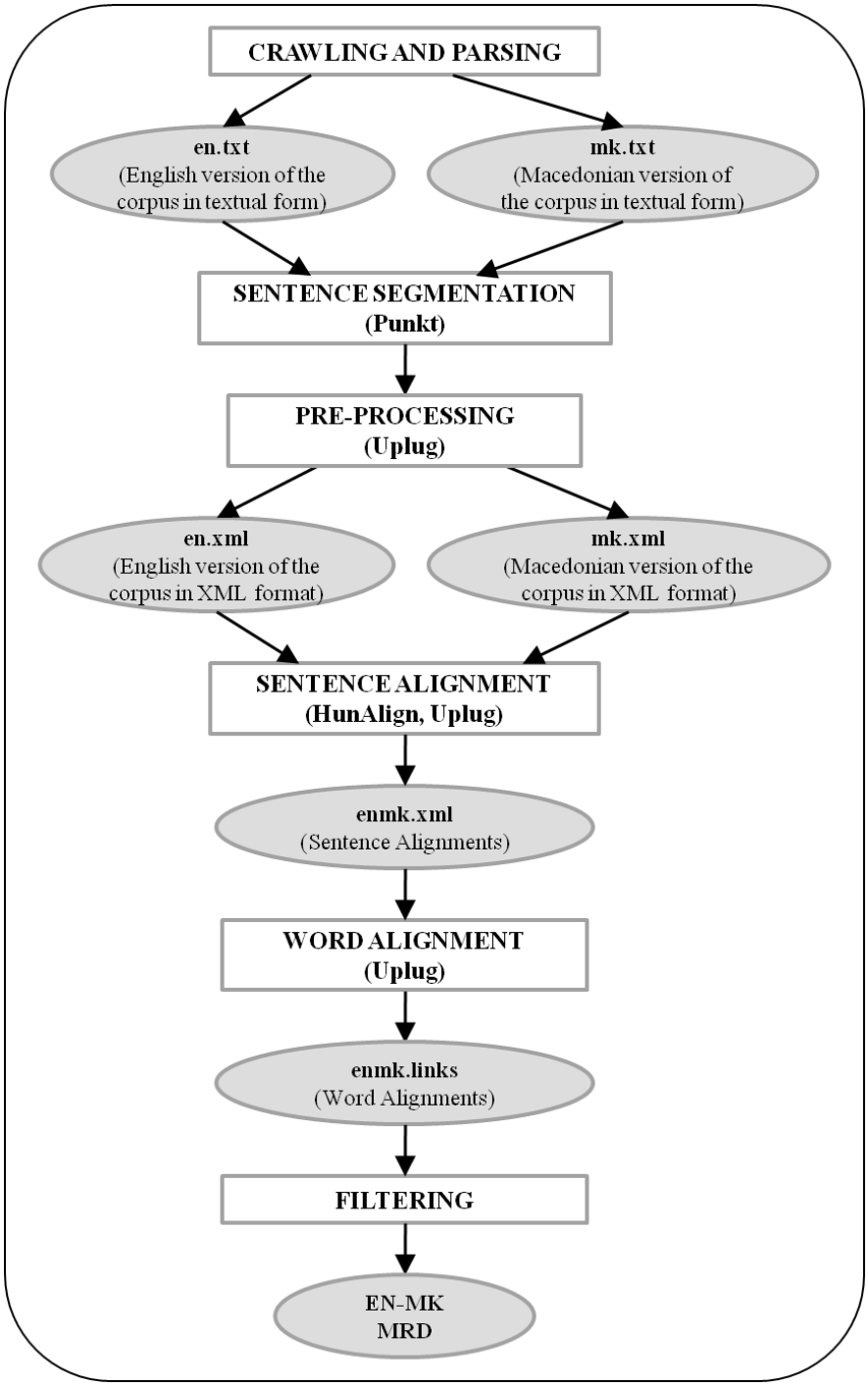

Fig. 1. The process of producing MRD from parallel corpora.

3.2.1 Crawling and Parsing

Since the only source was the official website of SETimes, the first step was to develop a simple crawler and parser. The purpose of the crawler was to collect the URLs of each article and to download the article in both languages. Afterwards, the parser was used to extract the article's text from the HTML code and to remove all unnecessary characters. Finally, the articles were stored in two text files, one for each language, where one line represented one article. The content of these files was manually verified to ensure that the article in the n^{th} line in the first file corresponds to the translated article in the second. The articles which were missing in one language were removed from both files.

3.2.2 Sentence Segmentation

The next step was to segment each article in sentences. Although, Uplug includes module for sentence segmentation, this module relies on simple rules and did not produce satisfactory results. Instead, *Punkt* was considered [7]. Punkt is a computer program which implements a language-independent unsupervised algorithm for sentence boundary detection. Understood intuitively, it is based on the assumption that a large number of ambiguities in the determination of sentence boundaries can be eliminated once abbreviations have been identified [7]. Punkt is open source, available through the Python *NLTK* (Natural Language Toolkit) [8] and could be easily applied to the collected corpora. To further facilitate the process of sentence segmentation, all articles that included paragraph HTML tags were first segmented on paragraphs and then sentence segmented. After this step it could be concluded that the whole corpus contains *28,980* articles i.e. *294,693* sentences per language.

3.2.3 Pre-Processing

Once the corpus was sentence segmented the Uplug pre-processing module was applied to allow the corpus to be further processed with other Uplug modules. The pre-processing module tokenizes the text and converts the text files in XML format by using basic markup for each paragraph, sentence, and word.

3.2.4 Sentence Alignment

Next, the sentence alignment module was applied. The purpose of this module is to link all sentences in one file to the corresponding translation sentences in the other. Uplug contains several sentence alignment modules. After experimenting with each, it was concluded that the module which uses *HunAlign* [9] showed most satisfying results. HunAlign is a language independent module which aligns sentences in bilingual texts by combining the so-called length-based and dictionary-based approaches. In the first pass of the corpus, HunAlign uses the sentence-length information to make a rough alignment of the sentences and to build a dictionary

based on this alignment. In the second pass, it uses the produced dictionary to realign the sentences. Furthermore, HunAlign includes one-to-many and many-to-one alignments, which allows the errors made in the sentence segmentation stage to be corrected with proper sentence alignment. The result of this step is an XML file containing the sentence links and the alignment certainty of each link. Sample output is shown in figure 2.

```
…
<linkGrp targType="s" toDoc="setimes/mk.xml"
   fromDoc="setimes/en.xml">
 <link certainty="3.64407" xtargets="s1.1;s1.1" id="SL2" />
 <link certainty="3.374068" xtargets="s1.2;s1.2" id="SL3" />
 <link certainty="1.819944" xtargets="s1.3;s1.3" id="SL4" />
 <link certainty="4.003576" xtargets="s1.4;s1.4" id="SL5" />
 <link certainty="11.63679" xtargets="s1.5;s1.5" id="SL6" />
…
```

Fig. 2. Sample output of the Uplug (HunAlign) sentence alignment module.

3.2.5 Word Alignment

Once the sentences were aligned the word alignment module was applied to the corpus. Word alignment refers to the process of linking corresponding words and phrases in the aligned sentences. For this purpose Uplug has three different modules: basic, tagged, and advanced. Since, part-of-speech tagger for the Macedonian language was not available at our disposal to achieve best results we used the advanced word alignment module. This module includes several sub-modules which run in the following order:

1. **Basic Clues**: computes basic alignment clues using association measures,
2. **Giza-word-refined**: runs GIZA++ in both alignment directions and converts the lexical probabilities to the clue aligner format,
3. **Dynamic Clues**: learns clues from the "refined" combination of both Viterbi alignments,
4. **Gizaclue-word-prefix**: takes only the three initial characters of each token and runs GIZA++ in both directions and converts probabilities to clues,
5. **Link**: clue alignment using basic clues, GIZA++ clues, and learned clues,
6. **Dynamic Clues**: learns clues from previously aligned data,
7. **Link**: clue alignment using all clues (basic, giza, learned),
8. The last three steps are repeated 3 times. [10]

Clue alignment refers to incorporating several knowledge resources (clues) in the process of word alignment. This module is the result of extensive research and experiments conducted in [10].

The output of this step is an *XCES* XML file [11] which includes the word links and the certainty of each alignment. Figure 3, shows sample output of this file, where each word link element has a certainty, lexical pair, and xtragets (link word ids) attributes.

```
…
<linkGrp targType="s" toDoc="setimes/mk.xml"
  fromDoc="setimes/en.xml">
 <link certainty="3.64407" xtargets="s1.1;s1.1"id="SL2">
   <wordLink certainty="0.04366786" lexPair="week;недела"
     xtargets="w1.1.9;w1.1.13" />
   <wordLink certainty="0.02486187" lexPair="prize;награда"
     xtargets="w1.1.7;w1.1.9" />
   <wordLink certainty="0. 03209486"
     lexPair="mayor;градоначалникот" xtargets="w1.1.2;w1.1.2" />
…
```

Fig. 3. Sample output of the Uplug advanced word alignment module.

To produce more readable output the *xces-to-text* Uplug module was applied. As figure 4 shows, the result is a text file containing all word alignments and their frequency of occurrence. As expected, the conjunctions occur most frequently.

44352	and	и	12950	in	во
24692	the	на	12605	serbia	србија
24538	in	во	11708	bih	бих
22182	with	со	11401	also	исто така
21006	eu	еу	11209	that	дека
14615	is	е	10430	kosovo	косово
13927	will	ќе	9378	turkey	турција
13091	on	ти	9352	the	на
12984	he	тој	8833	as	како

Fig. 4. Sample output of the xces-to-text Uplug module.

3.2.6 Filtering

Due to the errors made in the previous stages of processing the corpus, the word alignments contain a lot of noisy and incorrect translations which need to be excluded. The process of filtering the word alignments consists of two stages, where each stage includes several rules. All alignments which occurred less than 3 times were considered as a noise produced by the word alignment module and were excluded prior to applying the filtering rules.

The first stage considers each of the terms in the word alignments pairs as one string. The following rules apply:

- If one of the terms in the pair is an empty string, than the word alignment is considered invalid and is excluded.
- If the English term contains an alphabetical character and the Macedonian term does not contain a Cyrillic character, or vice verse, the word alignment is excluded as well.
- If both terms do not contain letters, then the pair is considered as numeric pair and is removed.
- If one term contains digit, while the other does not, the pair is also excluded.

The second stage checks the single tokens in both terms. Prior to applying the rules the strings are tokenized and processed with a method which removes the leading and trailing non-alphabetic/non-Cyrillic characters.

- If the number of tokens in one of the terms is greater than 3, the pair is excluded. Phrases in the word alignments are unusual output of the word alignment module and therefore are considered as an erroneous behavior.
- If one of the terms contains stop word token, than the pair is considered invalid.
- Finally, the one-to-one word alignments were lemmatized. The English words were lemmatized by using the Princeton WordNet [16], while for the purpose of lemmatizing the Macedonian words the lexicon developed in [12] was used.

After applying the filtering rules the list of *46,875* word alignments was shortlisted to *23,296* translation pairs. This is the size of the extracted dictionary which includes *17,708* English and *18,343* Macedonian unique terms.

4 Results and Evaluation

Several methods for evaluating the quality of the automatically extracted dictionaries have been proposed in the literature. Dagan I. and Church W. in [13] and Fung P. and McKeown K. in [14], measure the accuracy of the extracted dictionary by measuring the increase in efficiency that can be observed when translators are using the dictionary. A similar scenario of practically evaluating the automatically extracted dictionary is when lexicographers use the extracted dictionary to extend the existing dictionaries. In this case, the quality of the extracted dictionary is measured by the number of entries added to the existing dictionary.

However, two most common techniques for evaluating the automatically extracted dictionaries are: (1) automatic comparison with existing MRD and (2) manual evaluation of a subset of the dictionary entries. However, the use of the first technique may result in inaccurate evaluation of the extracted dictionary. Non-standard translations, translations of collocations, technical terminology, etc. are often not found in standard dictionaries and as a consequence may produce misleading evaluation [15]. Therefore, we have decided to use the second technique for the purpose of evaluating the EN-MK dictionary extracted during the course of this study.

Namely, we selected a subset of 2000 entries of the extracted dictionary, which is 8.5% of the dictionaries entries, to be manually evaluated. The entries were uniformly selected form the dictionary, i.e. every ~14^{th} entry was taken, so that alignments which occurred most and less frequently are equally included. We believe that in this way we will get the most accurate and objective evaluation of the extracted dictionary. Each entry in the subset was given one of the following three scores:

- ***C – Correct***: The translation is correct and the both the English and Macedonian terms are in the same case and gender (e.g. *army – армија*).
- ***S – Somewhat Correct***: The translation captures the meaning of the English word – some will understand the original term, but quality of the translation is low, or the case and genre are incorrect (e.g. *vote – гласање, sell – продаде*).

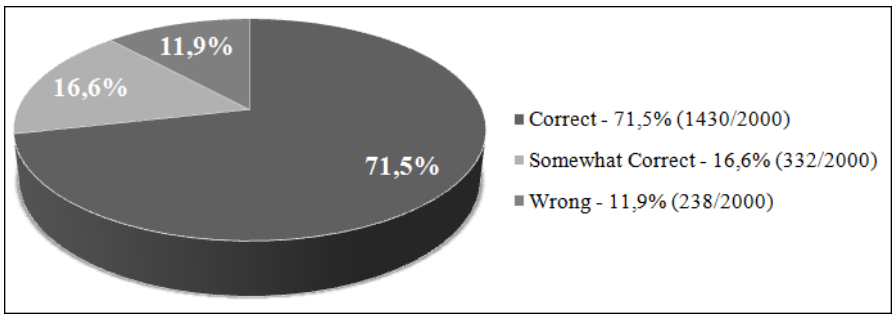

Fig. 5. Results of the manual evaluation of the extracted dictionary.

— ***W – Wrong***: The translation fails to capture the meaning of the English word (e.g. *back – поддржи*, *news – друг*, *wing – левичарски*, etc.).

The evaluation was performed by volunteers who are fluent speakers in both languages. Figure 5, shows the results of the evaluation.

In order to give a single measure of the accuracy of the dictionary we have combined the results by using the following formula [1]:

$$Accuracy = \frac{Correct\ Translations + 0.5 * Somewhat\ Correct\ Translations}{Number\ of\ Translations\ Evalvuated}$$

For example, if there are three translations, one is accurate, one is somewhat correct, and the last one is wrong, then the accuracy will be (1+0.5*1)/3=50%. By using this formula, we concluded that the accuracy of the extracted dictionary is ***79.8%***.

5 Conclusion and Future Work

The series of experiments reported in this paper, to our knowledge, are the first attempt to develop a large bilingual English-Macedonian dictionary by using purely statistical methods. We have conducted two experiments. The first, small scale, experiment proved that the technique of using parallel corpora to develop bilingual dictionaries is applicable and yields satisfactory results. This has encouraged us to conduct a second experiment with much larger corpus. By making use of the Statistical Machine Translation tools available at our disposal, we have processed the corpus in order to acquire a list of word alignments. This list has been further filtered to remove incorrect and noisy alignments and to acquire the final result of the experiment – the bilingual dictionary. The manual evaluation of a subset of the extracted dictionary resulted in an accuracy of *79.8%*. The extracted dictionary has been made available for public use on the Web through the following URL: `http://www.time.mk/trajkovski/tools/dict/`.

In the future, we plan to further study the process of filtering the word alignments. Namely, we believe that modeling the problem of filtering the word alignments as a

supervised learning problem will allow us to detect more incorrect translations. The frequency of the word alignments and the word and sentence alignment probabilities are good indicators of the accuracy of the word alignment and therefore can be used as features. On the other hand, the manually verified translations or the entries of exiting dictionaries found in the alignments can be used as a training data. We believe that by using this, more sophisticated, technique we will be able to improve the filtering of the word alignments and thus significantly increase the accuracy of the resulting dictionary.

References

1. Charitakis, K.: Using parallel corpora to create a Greek-English dictionary with Uplug. In: Nodalida (2007)
2. Tiedemann, J.: Automatical Lexicon Extraction from Aligned Bilingual Corpora. Master Thesis at University of Magdeburg (1997)
3. Velupillai, S., and Dalianis H.: Automatic Construction of Domain-specific Dictionaries on Sparse Parallel Corpora in the Nordic Languages. In: Coling (eds.) Workshop on Multi-source Multilingual Information Extraction and Summarization. Manchester (2008)
4. Hao-chun X., and Xin Z.: Using parallel corpora and Uplug to create a Chinese-English dictionary. Master Thesis at Stockholm University, Royal Institute of Technology (2008)
5. Stolic M., and Zdravkova K.: Resources for Machine Translation of the Macedonian Language. In: ICT Innovations Conference. Ohrid, Macedonia (2009)
6. Tiedemann, J.: News from OPUS - A Collection of Multilingual Parallel Corpora with Tools and Interfaces. In: Nicolov, N., Bontcheva K., Angelova G., Mitkov R. (eds.) Recent Advances in Natural Language Processing, vol. 5, pp. 237--248, Amsterdam (2009)
7. Tibor, K., and Strunk J.: Unsupervised Multilingual Sentence Boundary Detection. In: Computational Linguistics, vol. 32, num. 4 (2006)
8. NLTK - Natural Language Toolkit, http://www.nltk.org/
9. Varga D., et al.: Parallel corpora for medium density languages. In: Recent Advances in Natural Language Processing, pp. 590--596 (2005)
10. Tiedemann, J.: Recycling Translations - Extraction of Lexical Data from Parallel Corpora and their Application in Natural Language Processing. Doctoral Thesis at Uppsala University (2003)
11. XCES - Corpus Encoding Standard for XML, http://www.xces.org/
12. Petrovski, A.: Морфолошки компјутерски речник - придонес кон македонските јазични ресурси. Doctoral Thesis, Cyril and Methodius University. In Macedonian (2008)
13. Dagan, I., and Church, W.: Termight: Identifying and Translating Technical Terminology. In: Conference on Applied Natural Language Processing, pp. 34--40 (1994)
14. Fung, P., and McKeown K.: A Technical Word and Term Translation Aid using Noisy Parallel Corpora Across Language Groups. In: The Machine Translation Journal, Special Issue on New Tools for Human Translators, pp. 53--87 (1996)
15. Merkel, M., and Ahrenberg L.: Evaluating Word Alignment Systems. In: Second International Conference on Language Resources and Evaluation (LREC), pp. 1255--1261 (2000)
16. WordNet, http://wordnet.princeton.edu/