Automatic Construction of Wordnets by Using Machine Translation and Language Modeling

Martin Saveski*, Igor Trajkovski†

* Faculty of Computing, Engineering and Technology,
Staffordshire University,
College Road, Stoke-on-Trent, Staffordshire, UK
saveski.martin@gmail.com

† Faculty of Electrical Engineering and Information Technologies,
Ss. Cyril and Methodius University,
Rugjer Boshkovik bb, PO Box 574, Skopje, Macedonia
itrajkovski@feit.ukim.edu.mk

Abstract

WordNet is one of the most valuable lexical resources in the Natural Language Processing community. Unfortunately, the benefits of building a WordNet for the Macedonian language have never been recognized. Due to the time and labor intensive process of manual building of such a lexical resource, we were inspired to develop a method for its automated construction. In this paper, we present a new method for construction of non-English WordNets by using the Princeton implementation of WordNet as a backbone for their construction along with Google's translation tool and search engine. We applied the new method for construction of the Macedonian WordNet and managed to develop a WordNet containing 17,553 words grouped into 33,276 synsets. However, the method in consideration is general and can also be applied for other languages. Finally, we report the results of an experiment using the Macedonian WordNet as a means to improve the performance of the text classification algorithms.

Avtomatska izdelava wordneta z uporabo strojnega prevajanja in jezikovnega modeliranja

Wordnet velja za enega najbolj uporabnih leksikalnih virov na področju računalniške obdelave naravnega jezika, vendar za makedonščino še ne obstaja. Ker je ročna izdelava tovrstnega vira izjemno dolgotrajna in draga, smo se odločili za gradnjo z avtomatskimi pristopi. V prispevku predstavljamo metodo za izdelavo wordneta v izbranem ciljnem jeziku, pri čemer izhajamo iz angleškega Prinecton WordNeta, za generiranje sinsetov pa uporabimo dvojezični slovar, Googlov spletni strojni prevajalnik in iskalnik. Čeprav je na ta način mogoče izdelati wordnet za kateri koli jezik, smo v pričujoči raziskavi generirali makedonski wordnet, ki vsebuje 17.553 besed oz. 33.265 sinsetov. Izdelan wordnet tudi preizkusimo na sistemu za avtomatsko klasifikacijo besedil in s tem preverimo njegovo uporabnost v praksi.

1. Introduction

WordNet (Fellbaum, 1998) is a lexical database for the English language. It groups the English words into sets of cognitive synonyms (synsets) which represent different concepts. Each synset contains a gloss (explanation of the concept captured by the synset) and links to other synsets, which define the place of the synset in the conceptual space.

The public release of the Princeton WordNet (PWN), encoding the English language inspired many researchers around the world to develop similar lexical resources for other languages. As of today, there have been more than sixty WordNets built worldwide for more than fifty languages[1]. Moreover, WordNet had become an ideal tool and source of motivation for researchers from various fields. A plethora of applications which use WordNet have been developed including: word sense disambiguation, text categorization, text clustering, query expansion, machine translation, and many others.

Unfortunately, this potential has never been utilized for the Macedonian language and other than traditional lexical resources, such as dictionaries and lexicons, we are not aware of any current large lexical resources such as WordNet ontology for the Macedonian language.

Although, the manual construction of such lexical resource is most accurate, as far as linguistic soundness is concerned, it requires a lot of time and resources. Therefore, we have developed a method for automated construction of WordNets by using the PWN as a backbone for the construction and Google's translation tool and search engine.

The method is based on the assumption that the conceptual space modeled by PWN is not depended on the language in which it is expressed. Furthermore, we assume that the majority of the concepts exist in both languages, the source and target language, but only have different notations. Given that the conceptual space is already represented in English by the PWN, our goal is to find the corresponding concept notations in the target language by finding the proper translations of the synset members. However, we are aware of the fact that the WordNet produced by our method will be strongly influenced by the effectiveness in which PWN conceptualizes the world. Moreover, we are aware that PWN is not a perfect lexical resource and that all of its mistakes and drawbacks will also be inherited in the WordNet that is produced. Even if a lot of the parts of the produced WordNet remain in English, we believe that it will still be valuable for many WordNet applications in the target language, as a result of the WordNet structure.

The reminder of this paper is organized as follows: in the next section we provide a short overview of a related work after which we will describe our approach and methodology for the construction of the Macedonian WordNet. In sections 3 and 4, we present the results and

[1] http://www.globalwordnet.org/gwa/wordnet_table.htm

explain the usage of the WordNet in practical applications. Lastly in section 5, we discuss the pros and cons of our approach and ideas for future work.

2. Related Work

Due to the time consuming and labour intensive process of manual construction of WordNet, many automated and semi-automated construction methods have been proposed. This section provides a short overview of the methods for automated construction found in literature and considered most interesting.

An attempt to build a Macedonian WordNet was previously made by Aleksandar Pechkov in a scope of coursework. However, none of the deliverables from this study are publicly available.

Fišer D. and Sagot B. (2008) used a multilingual parallel corpus to construct Slovene (SloWNet) and French (WOLF) WordNets. They have PoS tagged, lemmatized, sentence, and word aligned the corpus in order to produce five multilingual lexicons which included French and four multilingual lexicons which included Slovene. Apart from Slovene and French, WordNets for the other languages (Romanian, Czech, and Bulgarian) have already been built and linked to PWN as part of the BalkaNet project (Tufis, 2000). Next, each of the lexicon entries produced is assigned a synset id from the WordNet of the corresponding language. Finally, the intersection of the synset ids of the entries is computed and assigned as a synset id to the Slovene and French words in the lexicon entry.

Changki L. and JungYun S. (2000), for the purpose of construction of Korean WordNet, define the problem of WordNet construction quite differently than the other methods discussed in this section. Namely, each Korean word is mapped to a list of English translations, each of which is expanded with the PWN synsets in which it belongs. Thus, the problem of WordNet construction is defined as finding the adequate English synset for a given Korean word. The authors propose six heuristics: maximum similarity, prior probability, sense ordering, IS-A relation, word match, and co-occurrence. Most interesting was found the word match heuristic which assigns a score to a given candidate synset according to the portion of overlapping words in the English dictionary definition of the Korean word and the English synset gloss and usage examples. Finally, in order to make a final decision, the heuristics are combined by using decision tree learning, where manually mapped senses are used as training data.

Barbu E. and Mititelu B. (2007) developed four other heuristics for automated construction of the Romanian WordNet. Namely, the intersection, WordNet Domains, IS-A relation, and dictionary definitions heuristics were proposed. The last two were found very similar to the IS-A relation and word match heuristics mentioned in the previous paragraph. More attention was paid to the intersection and WordNet Domains heuristics. The second makes use of the WordNet Domains project, which linked the PWN synsets with a set of 200 domain labels from the Dewey Decimal library classification. By using a collection of domain classified documents, all Romanian words in the EN-RO dictionary are labeled with the same domain labels as in WordNet Domains. Thus, when translating a source synset only, the translation candidates which match the synset domain are considered. These experiments proved to be very interesting since they were evaluated against the manually constructed Romanian WordNet and a formal measure of their performance was given.

3. Methodology

3.1. The Approach

Given the assumptions mentioned in the introductory section, the problem of automated construction of the Macedonian WordNet can be formulated as follows: Given a synset from PWN, the method should find a set of Macedonian words which lexicalize the concept captured by the synset.

The first step is by using an English – Macedonian (EN-MK) machine readable dictionary (MRD) to find the translations of all words contained in the synset. These translations are called *candidate words*. Since not all English words have Macedonian translations or are not contained in the MRD, for quality assurance it is assumed that if more than 65% of the words contained in the synset can be translated, then the concept captured by the synset can be expressed with a subset of the candidate words. Thus, the performance of the method is strongly influenced by the size and quality of the MRD used. For this reason, we have spent a lot of time and effort building a large and accurate in-house-developed MRD (Saveski, 2010). The MRD contains 181,987 entries i.e. 61,118 English and 79,956 Macedonian unique terms, where each English word is mapped into a set of Macedonian translations grouped by part of speech. The synsets which did not contain enough known words were skipped and retained in English.

However, not all of the candidate words reflect the concept represented by the synset. Therefore, a subset of words must be selected.

Let that the original synset contain ***n*** English words:

$w_{1,} \ldots, w_{i,} \ldots, w_n$,

and the word w_i has ***m*** translations,

$cw_{1,} \ldots, cw_m$, in the MRD.

Since the MRD has no means of differentiating between word senses, the set of translations of w_i ($cw_1 \ldots cw_m$) will contain the translations of all senses of the word w_i. It is a task of the method to determine which of these words, if any, correspond to the concept captured by the synset. Stated in this way, the problem of translating the WordNet synsets is essentially a *word sense disambiguation* (WSD) problem.

This is not very encouraging because WSD is still an open problem, but nevertheless gives us some pointers which may help in determining the best candidate words. Throughout the history of Artificial Intelligence, many approaches and algorithms have been proposed to solve the problem of WSD. Dagan I. and Itai A. (1994) stated that by using the word sense dictionary definition and a large textual corpus, the sense in which the word occurs can be determined. In other words, the words in the dictionary definition of the word sense tend to occur in the corpus more often, closely to the word in question, when the word is actually in the sense defined, and less often when the word represents other senses.

In terms of the problem of WordNet construction, this means that if the synset gloss can be translated, it will give us a good approximation of which of the candidate words are most relevant for the synset in question. Since manual translation of the glosses is not possible (translating the glosses is equivalent to translating the PWN), the English-to-Macedonian machine translation tool available through Google on the Web was chosen to be used. Although the Google EN-MK translation tool was not extremely accurate at the time of conducting this study, its performance was good enough to capture the meaning of the gloss. From the observations, it was concluded that the most common mistakes made by the translation tool were inappropriate selection of the genre and case of the words. However, this does not affect the use of the gloss translation as an approximation of the correlation between the candidate words and the synset.

The next crucial element for applying the statistical WSD technique is a large Macedonian textual corpus. Although, we are aware of some small textual corpora, mostly newspaper archives available on the Web, any attempt of collecting a large, domain independent corpus is not known to exist. Using a small and domain dependent corpus may significantly affect the performance of the method. On the other hand, collecting a large textual corpus from scratch requires a lot of time and resources, which were not available for this study. Therefore, an alternative method for measuring the correlation between the translated gloss and the candidate words was considered.

Namely, the *Google Similarity Distance* (GSD) proposed in (Cilibrasi & Vitanyi, 2007), calculates the correlation between two words/phrases based on the Google result counts returned when using the word, phrase, and both as a query. Most importantly, the result of applying the GSD is a similarity score between 0 and 1 representing the semantic relatedness of the candidate word and the translated synset gloss. The GSD is calculated for each candidate word and then the words are sorted according to their similarity.

Next, the candidate words are selected based on the following two criteria:

1. the words must have GSD score greater than: 0.2,
2. the words must have GSD score greater than: 0.8 * the maximum GSD score among the candidates.

The first criterion ensures that the words exceed minimum correlation with the gloss translation while the second makes discrimination between the words which lexicalize the concept captured by the synset and those that do not. The coefficients in both criteria were determined experimentally.

Finally, the words selected are included in the resulting Macedonian synset while the other candidate words are considered as not lexicalizing the concept captured by the synset. Figure 1 depicts the method explained in this section.

3.2. Google Similarity Distance

Google Similarity Distance (GSD) is a *word/phrase semantic similarity distance* metric developed by Rudi Cilibrasi and Paul Vitanyi proposed in (Cilibrasi & Vitanyi, 2007). The measure is based on the fact that words and phrases acquire meaning from the way they are used in the society and from their relative semantics to other words and phrases. The World Wide Web is the largest database of human knowledge and contains context information entered by millions of independent users.

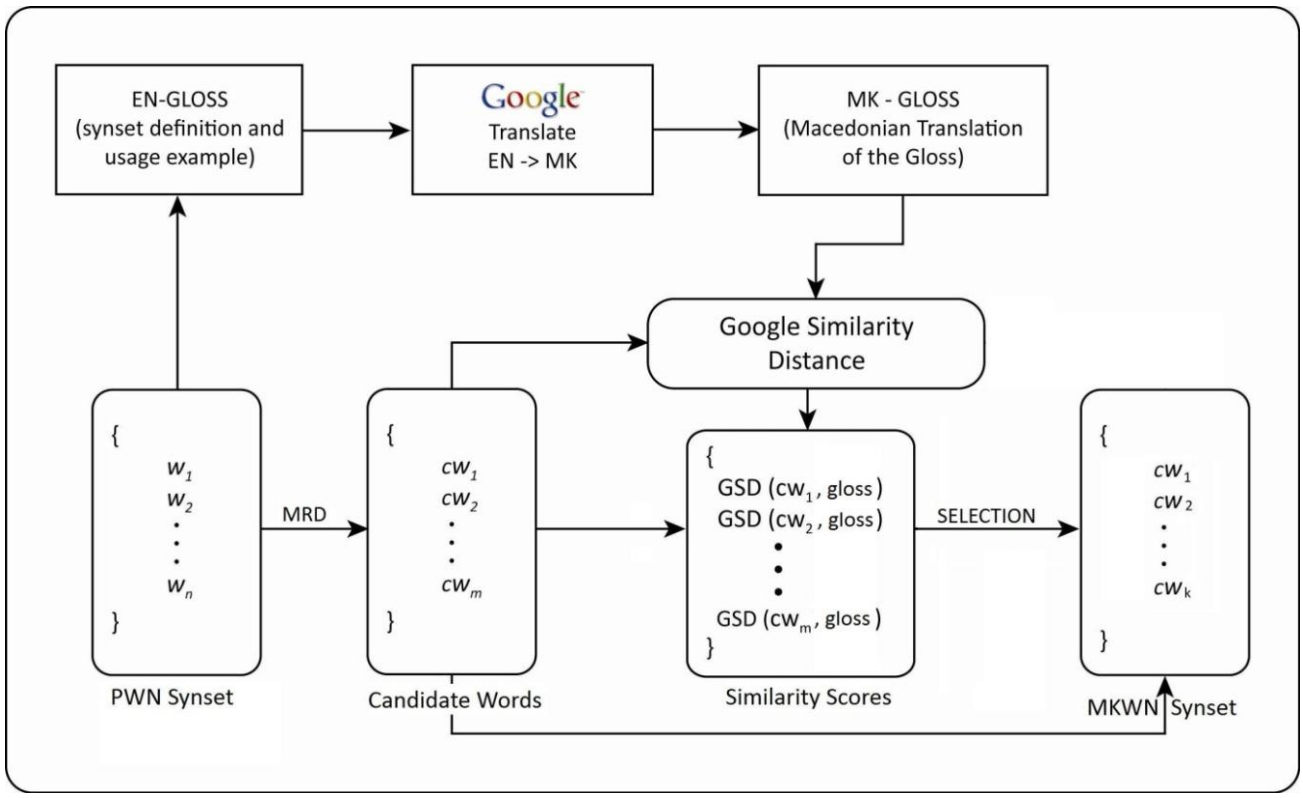


Figure 1. The Google Similarity Distance Method,
(*n*: dimension of the PWN synset, *m*: the number of candidate words, *k*: dimension of the resulting synset)

The authors claim that by using a search engine, such as Google, to search this knowledge, the semantic similarity of words and phrases can be automatically extracted. Moreover, they claim that the result counts of the words in question estimate the current use of the words in the society. As defined in (Cilibrasi & Vitanyi, 2007), the normalized Google Similarity Distance between words/phrases *x* and *y* is calculated as:

$$GSD(x,y) = \frac{\max\{\log f(x),\ \log f(y)\} - \log f(x,y)}{\log N - \min\{\log f(x),\ \log f(y)\}}$$

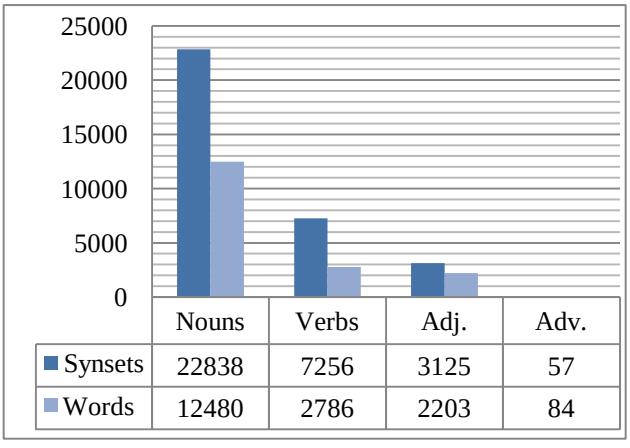
where *f(x)* and *f(y)* denote the result counts returned for *x* and *y*, respectively, and *f(x, y)* denotes the result count when both *x* and *y* are included in the query. The normalization factor *N*, can be chosen but has to be greater than the maximum result count returned. In our case, *f(x)* is the result count returned when the candidate word is included in the query, *f(y)* is the result count of the gloss translation, and *f(x, y)* is the result count when both are included in the query.

Here, the similarity distance is defined by using Google as a search engine, but is applicable with any search engine which returns aggregated result counts. The authors observed that the distance between words and phrases measured in different periods of time is almost the same. This shows that the measure is not influenced by the growth of the index of the search engine and therefore it is stable and scale invariant.

One possible drawback of the method is that it relies on the accuracy of the result counts returned. The Google index changes rapidly over time and the result counts returned are only estimated. However, linguists judge that the accuracy of the Google result counts is trustworthy enough. In (Keller & Lapata, 2003) it is shown that web searches for rare two-word phrases correlated well with the frequency found in the traditional corpora, as well as with human judgment of whether those phrases were natural.

3.3. Comparison with the Intersection Heuristic

In order to evaluate the results of our method, we also applied the Intersection heuristic proposed in (Barbu & Mititelu, 2007), and we have compared the results produced by both methods. The results of applying this heuristic, when compared with the manually produced WordNet, showed most successful results during the experiments for automated construction of the Romanian WordNet. As reported by the authors, on a selected subset of synsets an error rate of only 2% has been achieved.



	Nouns	Verbs	Adj.	Adv.
Synsets	22838	7256	3125	57
Words	12480	2786	2203	84

Figure 2. The size of the produced WordNet

After applying this heuristic for the construction of the Macedonian WordNet, we found out that **45%** of the synsets produced by the GSD method contained exactly the same words. However, because the two methods rely on different rules for translating the synsets, each succeeds to translate different subsets of PWN.

The last step of the construction of the MWN was to combine the synsets produced by both methods in order to produce a single WordNet. Namely, the synsets which could be translated by using both methods but did not result with the same words were produced by using the following rules. If the synset could be translated by using only the monosemous-word rule of the Intersection heuristic (Barbu & Mititelu, 2007), then the synset is produced by applying the GSD method. On the other hand, if the intersection rule of the Intersection heuristic is applicable, then the synset is produced by applying that rule. The rules are based on the fact that the GSD method and the intersection rule of the Intersection heuristic are more restrictive than the monosemous word translation rule.

Figure 2 shows the number of words and synsets produced by combining both methods, grouped by part of speech. It is important to note that all words included in the WordNet are lemmas.

4. Results and Evaluation

4.1. Using the MWN for Text Classification

The most common practices for evaluation of the quality of the automatically built WordNets are manual verification of the synsets produced (e.g. (Changki & JungYun, 2000)) or their comparison with the synsets of the manually developed WordNets, if such exist (e.g. (Barbu & Mititelu, 2007)). Although the manual verification of the synsets developed during the automatic construction of the Macedonian WordNet would be the most accurate and objective evaluation, it would require a lot of time and resources and thus is not an option. Also, as previously mentioned, a manually developed WordNet for the Macedonian language is not available and thus there is no golden standard against which we can evaluate the WordNet produced.

However, it is important to note that the initial objective of this study was not to develop a WordNet which will be a perfect lexical resource, but rather to develop a resource which will give us the opportunity to include semantics in the already developed techniques for Machine Learning (ML) and Natural Language Processing (NLP). Therefore, it was considered that it is much more suitable to evaluate the WordNet developed by its performance in a particular NLP/ML application and by the possible improvements that its usage may allow.

Namely, we were interested in how the use of the Macedonian WordNet will influence the performance of the text classification algorithms. This is only one of the plethora of applications of WordNet. However, it was considered mainly because the performance of the classification algorithms can be measured unambiguously and compared easily.

4.2. The Experiment

The first step towards defining a method for measuring the semantic similarity between two text documents using WordNet is to define how the distance between two WordNet synsets can be measured. We have adopted the *Leacock-Chodorow (LCH)* (1998) and *Wu and Palmer (WUP)* (1994) conceptual distance measures. The LCH measure defines the distance of the concepts (synsets) in terms of the number of nodes between the two synsets in the hierarchy while the WUP measure is based on the number of arcs between the synsets. For more information and comparison of the measures the interested reader can consult (Budanitsky, 1999) and (Budanitsky & Hirst, 2001). Next, since one word can be found in many synsets, we have extended the synset distance measures to word-to-word level. Namely, the distance between two words is defined as the minimum distance (maximum similarity) between the synsets where the first word was found and the synsets where the second word was found. Finally, by using the method defined in (Mihalcea et al., 2006), we extended the semantic word-to-word similarity measure to text-to-text semantic similarity. This measure combines the metrics of word-to-word similarity and *word specificity* (inverse document frequency - *idf*) into a single measure which can be used as an indicator of the semantic similarity of two texts.

During the experiment we compared the performance when using the following three similarity measures:

1. Semantic text similarity based on LCH synset similarity,
2. Semantic text similarity based on WUP synset similarity,
3. Cosine Similarity.

The Cosine Similarity is a classical approach for comparing text documents where the similarity between two documents is defined as the cosine of the angle between the two document vectors. This measure is used as a base line for comparison of the performance of the other two metrics.

In addition, we made use of the KNN - K Nearest Neighbors classification algorithm as a method which is easy to implement and allows the similarity measures to be compared unambiguously. To speed up the classification and improve the performance of the algorithm, during the training phase, we structured the data samples in an *inverted index* (Manning et al., 2008).

For the purpose of the experiment a corpus of Macedonian news articles was used. The articles are taken from the archive of the A1 Television Website published between January 2005 and May 2008. As table 1 shows, the corpus contains 9,637 articles i.e. 1,289,196 tokens classified in 6 categories.

Category	Articles	Tokens
Balkan	1,264	159,956
Economy	1,053	160,579
Macedonia	3,323	585,368
Sci/Tech	920	17,775
World	1,845	222,560
Sport	1,232	142,958
TOTAL	**9,637**	**1,289,196**

Table 1. A1 Corpus, size and categories

4.3. Results

Figure 3 compares the performance of the three similarity metrics by their F-Measure score. As seen in the figure, the LCH-semantic similarity fails to improve the performance of the Cosine Similarity metric. The main reason for the low performance of this measure is due to its inability to calculate the similarity between words with different part of speech. The WUP-semantic similarity metric, on the other hand, has improved classification performance and outperforms both the Cosine Similarity and LCH-semantic similarity metrics by ***6.7%*** and ***20.6%***, respectively. When compared to the Cosine Similarity, as a baseline, this metric manages to find more patterns in the text documents. This is especially evident in the documents from the Sci/Tech and Economy categories.

Although by doing this experiment we cannot argue about the validity of the WordNet produced, we can conclude that the information encoded by the WordNet is meaningful and accurately models the real world. Moreover, we have practically shown that the Macedonian WordNet can be used to include semantics in the existing ML and NLP algorithms and to improve their performance.

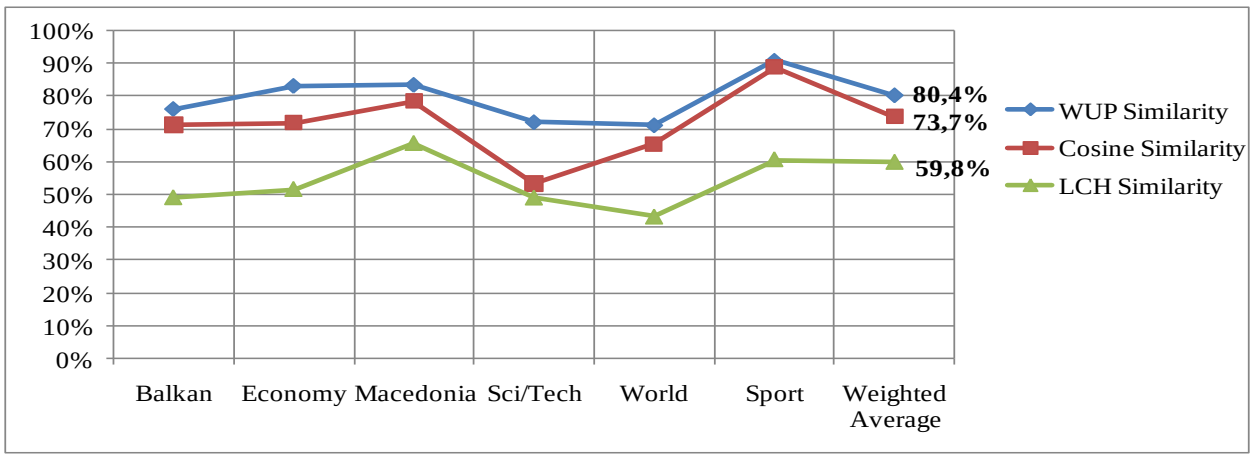


Figure 3. Comparison of all text similarity measures (F-measure)

5. Conclusion and Future Work

In this paper we have proposed a new method for automated construction of WordNets. The method relies on a bilingual dictionary and PWN, as a backbone for the construction and uses Google's machine translation and result counts to make a selection between candidate words. The method presented has been successfully applied for the construction of the Macedonian WordNet but can also be applied to other languages if machine readable dictionary and translation system are available. We have experimentally evaluated the accuracy of the produced WordNet. By using it as a mean to include semantics in the text classification algorithms, we have managed to improve the performance achieved by the standard techniques.

However, our method currently considers each candidate word independently, not taking into account the semantic relatedness which exists between some of the candidate words. In the future, we plan to investigate how the candidate words can be clustered (grouped) prior to assigning them to the synset. We want to consider how, based on the individual similarity between the words and the similarity of each word and the gloss, it can be determined which group is most suitable to express the concept captured by the synset. In this way, we can compensate for some of the possible mistakes made during both the translation of the gloss and the measuring the semantic similarity between the candidate word and the gloss. Moreover, the probability of incorrectly assigning a group of words to a synset is much lower than the probability of incorrectly assigning an individual word.

Next, we would like to repeat the text classification experiment by using larger corpus of text documents and to investigate whether this improvement in the performance will also be evident. Moreover, we are interested in how the use of more complex word-in-context-to-word-in-context similarity measure will influence the performance.

Finally, we plan to conduct similar experiments for other WordNet applications, such as text clustering and word sense disambiguation, and to apply this method for construction of other non-English WordNet.

6. References

Barbu, E. & Mititelu, B. V. (2007). Automatic Building of Wordnets. In Proceedings of *Recent Advances in Natural Language Processing IV*, John Benjamins, pp. 217--226, Amsterdam, 2007.

Budanitsky, A. & Hirst, G. (2001). Semantic distance in WordNet: An experimental, application oriented evaluation of five measures. In Proceedings of the *NAACL Workshop on WordNet and Other Lexical Resources*.

Budanitsky, A. (1999). *Lexical Semantic Relatedness and its Application in Natural Language Processing* [Online]. Accessed from: http://www.cs.toronto.edu/

Changki, L. & JungYun, S. (2000). Automatic WordNet mapping using word sense disambiguation. In Proceedings of the *Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*.

Cilibrasi, R. & Vitanyi, M. B. (2007). The Google Similarity Distance. In the proceedings of *IEEE Trans. on Knowl. and Data Eng.*, vol. 19, pp. 370--383.

Dagan, I. & Itai, A. (1994). Word Sense Disambiguation Using a Second Language Monolingual Corpus. In the proceedings of *Computational Linguistics 1994*, vol. 20, pp. 563--596.

Fellbaum, C. Et al. (1998). *WordNet: An electronic lexical database*. Cambridge: MIT press.

Fišer, D. & Sagot, B. (2008). Combining Multiple Resources to Build Reliable Wordnets. In Proceedings of *Text, Speech and Dialogue (LNCS 2546)*, Springer 2008, pp. 61--68, Berlin: Heidelberg.

Keller, F. & Lapata, M. (2003). Using the web to obtain frequencies for unseen bigrams. In the proceedings of *Computational Linguistics 2003*, vol. 29:3, pp. 459--484.

Leacock, C. & Chodorow, M. (1998). Combining local context and WordNet similarity for word sense identification. In Fellbaum C., *WordNet: An electronic lexical database*, pp. 265--283.

Manning, C. D., Raghavan, P. & Schütze, H. (2008). *Introduction to Information Retrieval*. Cambridge University Press.

Mihalcea, R., Courtney, C. & Strapparava, C. (2006). Corpus-based and Knowledge-based Measures of Text Semantic Similarity. In Proceedings of *American Association for Artificial Intelligence*.

Saveski, M. (2010). *Development of a WordNet Prototype for the Macedonian Language*. Bachelor (Hons) Thesis, Staffordshire University, UK.

Tufis, D. (2000). BalkaNet: design and development of multilingual Balkan wordnet. In Proceedings of the *Romanian Journal of Information Science and Technology*, vol. 7(1-2).

Wu, Z. & Palmer, M. (1994). Verb semantics and lexical selection. In Proceedings of the *Annual Meeting of the Association for Computational Linguistics*.