Influence of Word Normalization on Text Classification

Michal Toman[a], Roman Tesar[a] and Karel Jezek[a]
[a] *University of West Bohemia, Faculty of Applied Sciences, Plzen, Czech Republic*

In this paper we focus our attention on the comparison of various lemmatization and stemming algorithms, which are often used in nature language processing (NLP). Sometimes these two techniques are considered to be identical, but there is an important difference. Lemmatization is generally more utilizable, because it produces the basic word form which is required in many application areas (i.e. cross-language processing and machine translation). However, lemmatization is a difficult task - especially for highly inflected natural languages having a lot of words for the same normalized word form. We present a novel lemmatization algorithm which utilizes the multilingual semantic thesaurus Eurowordnet (EWN). We describe the algorithm in detail and compare it with other widely used algorithms for word normalization on two different corpora. We present promising results obtained by our EWN-based lemmatization approach in comparison to other techniques. We also discuss the influence of the word normalization on classification task in general. In overall, the performance of our method is good and it achieves similar precision and recall in comparison with other word normalization methods. However, our experiments indicate that word normalization does not affect the text classification task significantly.

Keywords: lemmatization, classification, EuroWordNet, stemming, word normalization.

1 INTRODUCTION

This paper deals mainly with the comparison of various word normalization techniques used in many NLP areas. Two different normalization approaches are usually distinguished – stemming and lemmatization. Both techniques produce a normalized form. However, there are important differences. Lemmatization replaces the suffix of a word with a different one or removes the suffix of a word completely to get the basic word form (lemma). On the other hand, word stemming does not usually produce a basic form, but only an approximation of this form (called stem or generally normalized form). For example, the words *calculate*, *calculating* or *calculated* will be stemmed to *calculat*, but the normalized form is the infinitive of the word: *calculate*.

Even if lemmatization is a more difficult way to word normalization, in some cases it can be beneficial. It produces the basic word form which is required in many application areas. Lemmatization is a challenging task especially for highly inflected languages. Our work aims at this problem. In this contribution it is presented our novel lemmatization method compared with other word normalization methods. Our method is based on the EuroWordNet (EWN) thesaurus, which represents a multilingual database of words and relations between them for most European languages. The EWN-based approach we present transforms text into the language independent form. Thanks to the internal EWN relationships it is possible to consider the synonymous words to be the full equivalents in other text processing steps (e.g. searching, classification, disambiguation). We describe the algorithm in detail and compare it with other widely used algorithms for word normalization on two different corpora written in different languages. Our language selection includes both morphologically simple (English) and complicated (Czech) languages. They consists of articles obtained from press agencies, thus the type of documents in both datasets is similar.

For the comparison the text stemmed and lemmatized by various algorithms was used in the classification task. We used the multinomial Naive Bayes (NB) classifier in our classification tests. In the section 3 we describe our experiments with NB classifier on different corpora.

The comparison of our method with other normalization approaches is presented at the end of this article. The aim of our tests is not only to compare our lemmatizer with other ones, but we also want to determine the influence of word normalization upon the classification task.

2 EUROWORDNET THESAURUS

This section describes the main principle of the EWN-based algorithm. The comparison of our method with the other normalization algorithms can be found in section 3.

Thesaurus EuroWordNet (EWN) plays a central role in our approach. EWN together with Ispell[8] transforms word into its basic form (*lemma*). Moreover, each word is connected with the EWN thesaurus node – *synset*, which represents a set of synonymous words.

2.1 EuroWordNet Thesaurus

EuroWordNet thesaurus can be applied in many NLP areas. It is a multilingual database of words and relations for most European languages. It contains sets of synonyms - synsets - and relations between them. A unique index is assigned to each synset. It interconnects the languages through an inter-lingual-index in such a way that the same synset in one language has the same index in another one. Thanks to the EWN-based approach, it is possible to perform additional techniques in the processing e.g. query expansion, cross-language information retrieval, and word sense disambiguation as shown in many papers[9][10].

In order to use EWN, it is necessary to assign an EWN index to each term of a document. To accomplish that, words must be firstly transformed into basic forms (words must be normalized). Lemmatization is the only possible way how to transform the word to obtain the basic form – lemma.

2.2 Lemmatization

Lemmatization transforms words into their basic forms. EWN-based lemmatization belongs to the group of dictionary lemmatization algorithms. A dictionary creation can be considered as the most difficult part of the EWN-based approach. We proposed a method for the lemmatization dictionary building based on the use of EWN thesaurus and Ispell dictionary. The lemmatization dictionary was created by extraction of word forms using the Ispell utility. We were able to generate all existing word forms from the stem stored in the Ispell dictionary. The dictionary contains stems and attributes specifying the possible suffixes and prefixes, which were applied to stems in order to derive all possible word forms. We assume that one of the derived forms is a basic form (lemma). In order to recognize the basic form we looked for the corresponding lemma in EWN. A fuzzy match routine [6] can be optionally enabled for searching lemmas in EWN thesaurus, which helps especially in the case of highly inflected languages.

Languages with a rich flex are more difficult to be processed in general. We used a Czech morphological analyzer [7] to overcome this problem. Thanks to this module we obtained a further improvement of our method. The English lemmatization is relatively simple, thus it is possible to use basic lemmatization algorithms with satisfying results. We implemented lemmatization modules for Czech and English languages, but the main principle remains the same also for other languages. However, language specific processing steps (morphological analysis, disambiguation) are needed in some areas to achieve better results.

2.3 Indexing

A unique index is assigned to each synset. It interconnects the languages through an inter-lingual-index in such a way that the same synset in one language has the same index in another one. Thus, words are indexed identically in all languages on their semantic basis according to the affiliation to the synset. Other NLP tasks (e.g. cross-language information retrieval) can take advantage of the assignment. With EWN, completely language independent semantic aware processing and storage can be carried out.

3 EXPERIMENTS

As mentioned before, lemmatization can be utilized in many text processing tasks. In this paper we focused our attention on text classification. We were interested not only in the comparison of our proposed algorithm with other similar approaches, but we also wanted to generally examine the influence of word normalization and stop-words removal on the text classification. There exist many classification algorithms. Among them, the multinomial Naive Bayes classifier [2] is widely used in different areas and despite of its simplicity it achieves outstanding results. The decision to use this classifier for our experiments was also supported by the fact that other researchers employed the Naive Bayes classifier for text classification on different lemmatized, stemmed or non-preprocessed datasets, thus we have the possibility to compare our observations and results with them. We consider the standard metrics for the classification performance evaluation, namely micro-F1 and macro-F1 values. As usual, micro-F1 value is computed for all documents over all document categories. Macro-F1 measure represents the averaged value determined from F1 values computed for each classification category separately. For completeness, we present also precision (p) and recall (r) measures. The standard definition of these basic measures can be found, e.g., in [3]. Furthermore, we also report the statistical significance of our results using the McNemar's test [4] considering the p-value of 0.05.

The datasets we used for classification are described in section 3.1. Because these text datasets are not

standardized, we used the 4-cross fold validation technique [4] to ensure the correctness of our results. Both of them were always firstly preprocessed by various word normalization approaches and then classified – always with and then without stop-words. The word normalization approaches we utilized are described in section 3.2.

Two different setups of EWN-based lemmatizer were used in our experiments. The first one can be considered as a simple lemmatization approach. It means that each word was replaced with a corresponding lemma and indexed without taking into account any semantic information from EWN. The second approach denoted in results as *EWN lemmatization – indexes* fully incorporates EWN, thus words are indexed using the EWN indexes uniquely identifying each synset.

3.1 Datasets

For our experiments we used two datasets. The first of them consists of 8000 documents in English selected from Reuters Corpus Volume 1 dataset and it contains 6 different categories. English is an example of morphologically simple language. The second corpus contains 8000 documents in the Czech language which is morphologically complex and contains 5 categories thematically similar to categories in the English corpus. The Czech dataset was created from documents provided by Czech News Agency. For both corpora, all numbers were removed and the letters of all words were put in lower case. For the stop-words removal we used the stop-list available at [5] which consisted of 388 words.

3.2 Word Normalisation Algorithms

We examined 6 different word normalization approaches – Lovins and Iterated Lovins, Paice, Porter's stemmer, EWN-based lemmatization with and without using indexes. Porter's stemmer processes the word in different steps, e.g. plural –s, and past tense –ed removal, y to i replacement, endings removal. Lovis and Paice methods are based on transformation rules and their constraints. A non-processed corpus was used as a baseline for the classification task.

3.3 Results

As can be seen from the results for the English dataset presented in Figure 1 the number of words in dataset generally corresponds to the classification accuracy. The more words the higher accuracy. The only exception is the stemming algorithm Iterated Lovins, where the number of unique words is only 28896. Although Porter's stemmer, Paice, Lovins and EWN lemmatization produced more words, the classification accuracy we achieved was lower.

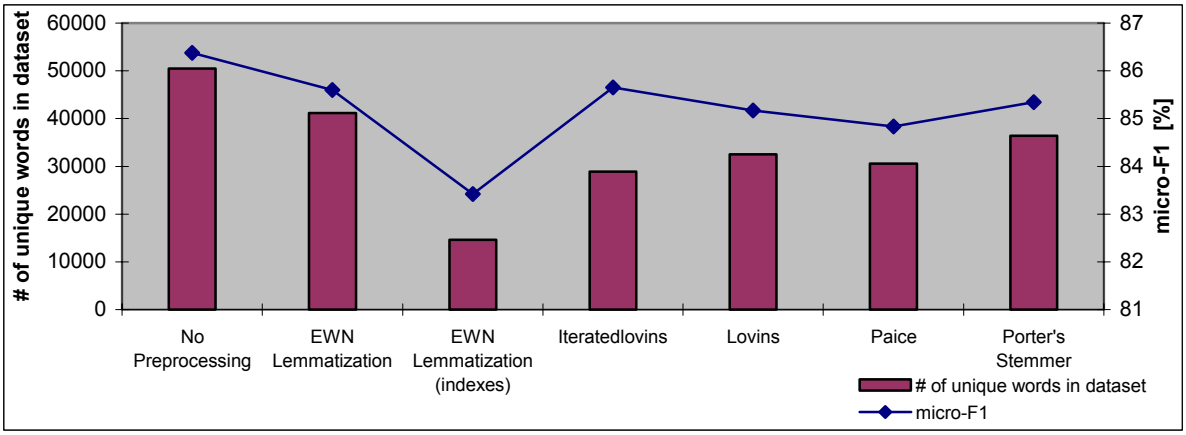

Fig. 1 Number of words in English dataset and the obtained classification accuracy for various word normalization approaches (stop-words removed)

Even if the performance of EWN-based method is slightly worse than other methods, it can be considered a promising algorithm taking into account 50% decrease of words count in the lemmatized corpora. Anyway, due to the reduced amount of information the classification cannot perform as well as in the case of other methods and 2% micro-F1 measure decrease occurs. This issue is related to the EWN structure, because the thesaurus does not cover the whole natural language and a lot of words are missing. We expect that by enlarging the thesaurus the higher classification accuracy of the method will be achieved.

For this dataset, EWN lemmatization and Iterated Lovins performed best and when stop-words were

removed the decrease in classification accuracy was not statistically significant. Especially in the case of Iterated Lovins the reduction of unique words in the dataset was noticeable.
All results also for the case when stop-words were not removed from the English dataset can be found in Table 1.

Table1. The results of classification obtained for the English dataset when various stemming and lemmatization algorithms were applied

	Stopwords Removed	P	R	Micro-F1	Macro-F1	Number of unique words in dataset
No Preprocessing	No	82.55	90.22	86.21	84.06	50813
	Yes	82.55	90.60	86.38	84.26	50494
EWN-Lemmatization	No	81.41	89.92	85.44	83.32	42047
	Yes	81.35	90.33	85.60	83.49	41151
EWN-Lemmatization (indexes)	No	77.10	90.29	83.18	80.33	14725
	Yes	77.36	90.51	83.42	80.59	14625
Iteratedlovins	No	79.56	89.32	84.16	81.69	33244
	Yes	80.80	91.14	85.66	83.58	28896
Povine	No	79.77	89.38	84.29	81.86	36806
	Yes	80.57	90.35	85.17	83.03	32517
Paice	No	80.88	89.52	84.98	82.47	40126
	Yes	80.01	90.30	84.84	82.70	30540
Porters Stemmer	No	80.47	90.33	85.11	82.87	36636
	Yes	80.66	90.62	85.35	83.17	36421

For the Czech dataset (see Figure 2) we can notice that the algorithmic lemmatization does not work very well. Although the index-based EWN lemmatization reduced the number of words much more, the classification accuracy was very similar.

Czech language is especially complex. It has a rich set of conjugation and declension rules, thus it is a challenging task to create an algorithmic method. Even if the affixes are removed carefully, the precision is not very high. A dictionary-based method is more appropriate as can be seen from Figure 2.

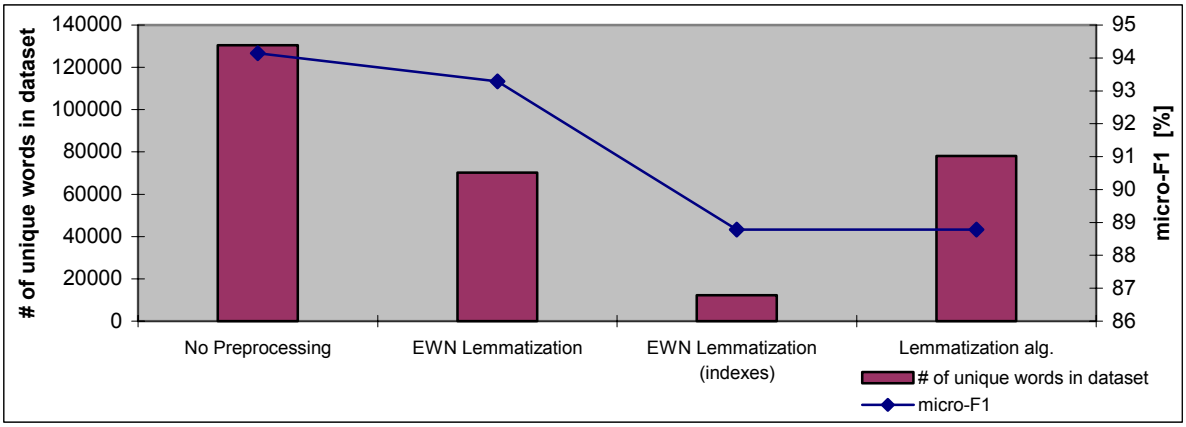

Fig. 2 Number of words in Czech dataset and the obtained classification accuracy for various word normalization approaches (stop-words removed)

Similarly to the English dataset, the best accuracy was achieved when no pre-processing was applied on the Czech corpora. Even if the corpus created with EWN-based algorithm is 10-times smaller than a non-processed corpus, the micro-F1 measure doesn't prove a significant decrease. This is caused mainly by the EWN and a semantic aware indexing.

Table2. The results of classification obtained for the Czech dataset when various stemming and lemmatization algorithms were applied

	Stopwords Removed	P	R	Micro-F1	Macro-F1	Number of unique words in dataset
No Preprocessing	No	94.65	93.44	94.04	71.30	130778
	Yes	93.79	94.49	94.14	74.21	130428
EWN-Lemmatization	No	91.50	95.32	93.37	73.89	70403
	Yes	90.91	95.80	93.29	75.20	70289
EWN-Lemmatization (indexes)	No	81.51	97.74	88.89	70.84	12182
	Yes	81.35	97.72	88.79	70.71	12224
Lemmatization alg. Hodek	No	91.69	81.99	86.57	46.81	78176
	Yes	89.20	88.41	88.78	56.00	78051

3 CONCLUSION

Our experiments indicate that the influence of word normalization on text categorization is negative rather than positive for both morphologically complex and morphologically simple languages. In some cases, lemmatization and stemming slightly improved the macro-F1 value (see Table 1 and Table 2), but the improvements were not statistically significant. On the other hand, stop-words removal improved in most cases the classification accuracy, but also in this case the results were not statistically significant. However, stop-words removal reduces the dimension of classified documents and compresses the classification time.

The best preprocessing approach for text classification seems to be only to apply stop-words removal and to omit word normalization. The decrease in classification accuracy when word normalization was applied was usually noticeable and often statistically significant.

When word normalization for English is needed the Porter's stemmer is the most appropriate algorithm. It has no significant influence on the processing precision and it decreases the corpus dimension to 72 %.

On the contrary, a dictionary method is better for morphologically rich languages. EWN-based algorithm without transformation to EWN indexes performs well on the Czech language. We expect that the increasing quality of EWN will raise the performance of the EWN-based method. As discussed in section 2.1, EWN structure can be used in other NLP tasks, which is a significant advantage of our approach.

REFERENCES

[1] http://www.illc.uva.nl/EuroWordNet/

[2] McCallum A., K. Nigam. 1998. A Comparison of Event Models for Naive Bayes Text Classification. In *AAAI/ICML-98* Workshop on Learning for Text Categorization, Technical Report WS-98-05, AAAI Press, pp. 41-48.

[3] Yang Y., Pedersen J.O. 1997. A Comparative Study on Feature Selection in Text Categorization. Proceedings of the Fourteenth International Conference on Machine Learning, pp.412-420.

[4] Dietterich, T.G. 1998. Approximate Statistical Test for Comparing Supervised Classification Learning Algorithms. Neural Computation, Vol. 10, no.7, pp. 1895-1923.

[5] http://fog.bio.unipd.it/waishelp/stoplist.html

[6] Eugene W. Myers. 1986. An O(ND) Difference Algorithm and Its Variations, Algorithmica Vol. 1, pp. 251-266.

[7] Jan Hajic. Morphological Analyzer, http://quest.ms.mff.cuni.cz/pdt/Morphology_and_Tagging/Morphology/index.html

[8] Ispell. http://fmg-www.cs.ucla.edu/fmg-members/geoff/ispell.html

[9] Toman M., Jezek K.: 2005. Document Categorization in Multilingual Environment, ELPUB 2005, Proceedings of the 9th ICCC International Conference on ElectronicPublishing.

[10] Toman M., Steinberger, J., Jezek, K. 2006. Searching and Summarizing in Multilingual Environment, ELPUB 2006