V. Bobicev, T. Zidraşco

A LANGUAGE-INDEPENDENT METHOD OF WORD LEMMATIZATION

1. Introduction

According to *Collins English Dictionary*[1], to **lemmatize** means «to group together the inflected forms of a word for analysis as a single item».

While in English there is not much difference in wordforms, other, inflectionally richer, languages exhibit great variation in forms between the word in the text and its lemma.

One of such languages is Romanian. Method of Romanian words lemmatization is discussed in this paper. The main idea of our approach is to find the initial form of the word in text and attach it to the given word form without any additional processing of text.

A simplest way to implement lemmatization is to use a morphological dictionary where lemma is attached to each wordform. But with dictionary-based lemmatization an obvious problem appears: one needs a very large morphological dictionary. Besides it is known that a number of words in a language are ambiguous and the problem of ambiguity must be solved as well.

An example of lemmatizer for English, German, and Dutch is MBLEM[2]. Even though it uses TiMBL as our system does, its features are much more complicate and based on deep morphological word analysis.

---

[1] Collins English Dictionary. William Collins & Sons & Co. Ltd., London, 1991.

[2] *Van den Bosch A., Daelemans W.* Memory-Based Morphological Analysis // Proceedings ACL–99. 1999. P. 285–292.

Another approach[3] focused on word endings. Method proposed in the paper achieved 77,0% accuracy.

A statistics-based trigram tagger[4] was used to learn morpho-syntactic tagging and a first-order decision list learning system was used to learn rules for morphological analysis. The results reported in the paper were better (92%) but they were using much more (grammatical) information.

Most of the systems perform lemmatization as a subtask of the PoS tagging. Our aim was to obtain word lemma without any PoS information, dictionary, grammatical or morphological information about the word.

2. Experiments

The goal of the experiments reported in this paper is to see how well it is possible to perform word lemmatization in text with the minimum morphological information. Providing an additional method for lemmatization, the purpose of doing learning at such language-independent level is to supply a methodology for languages that have only few lexical and syntactic resources.

We used three different data sources for our experiments. First it was the electronic dictionary (further denoted as **dic**), containing about 90 000 word forms.

Second source was the Multext-east corpus[5]. For our purpose we used «1984» by Orwell translated to Romanian (further denoted as **1984**), morphologically annotated and manually corrected.

---

[3] *Plisson J., Lavrac N., Mladenic D.* A Rule Based Approach to Word Lemmatization // Proceedings of Information Socitey. 2004.

[4] *Erjavec T., Džeroski S.* Machine Learning of Morphosyntactic Structure: Lemmatising Unknown Slovene Words // Applied Artificial Intelligence. 2004. 18(1). P. 17–40.

[5] *Erjavec T.* Multext-East Version 3: Multilingual Morphosyntactic Specifications, Lexicons and Corpora // Proceedings of the 4th Intl. Conf. on Language Resources and Evaluation. LREC–04, ELRA. 2004.

The third source was Corpus of Supreme Court Decisions (further denoted as **Hot**) containing 410 law documents morphologically annotated.

For our experiments we used TiMBL[6] which implements memory based learning. To perform memory based learning TiMBL requires a training set of instances. Each instance consists of a number of features and a class these features predict.

In memory-based learning algorithms features influence the final result considerably. As mentioned in the introduction we do not want to use part of speech tagger or any other morphological or syntactic analyzers. Hence, we decided to use a certain number of final letters of the word as features. Our idea was that all word endings would be reflected in these features.

Firstly, we used the letters as is. It means that each letter of word ending became a feature in feature vector. Sample feature vector is presented below:

a , j , u , r , u , l , a j u r

The first six letters divided by column are the features, the last fragment presents the ending of lemma that must replace these six letters in the word to obtain lemma.

Secondly, we took the word endings of different length decreasing the number of letters in each feature starting with six letters and finishing with one. Sample features look like:

ajurul , jurul , urul , rul , ul , l , ajur

In the last experiment we added the previous word ending to the vector of features. Sample feature vector is presented below:

senină , u , r , o , a , s , ă , uros

The first feature in a sample is previous word ending.

---

[6] *Daelemans W., Zavrel J., van den Sloot K., van den Bosch A.* Timbl: Tilburg Memory-Based Learner. Version 4.0. Reference Guide. Tech. rep., University of Antwerpen, 2001.

We extracted from corpora sets of instances as described in previous section using small perl scripts. These sets were processed by TiMBL. In all our experiments we performed ten-fold cross-validation. The results of experiments with different number of letters look like follows.

Table 1. Lemmatization accuracy for three sources with letters as features

Num, of final letters	**dic**	**1984**	**Hot**
5 letters	58,0%	85,8%	90,3%
6 letters	42,3%	84,7%	90,8%
7 letters	26,9%	83,8%	90,6%
8 letters	16%	83,3%	90,6%
9 letters		82,6%	90,2%

The result of the second experiment with different length of fragments as features is presented in the table 2.

Table 2. Lemmatization accuracy for three sources with fragments as features

Num, of fragments	**dic**	**1984**	**Hot**
5 fragments	57,1%	83,6%	90,0%
6 fragments	41,5%	82,0%	90,5%
7 fragments	26,0%	80,4%	90,0%
8 fragments	16%	79,4%	89,7%
9 fragments		79,0%	91,3%

As the results of second experiment didn't show much improvement, we decided to stop at 6 letters as features.

The last experiment was performed with previous word ending added to existent features. It slightly improved the obtained result, 85,2% for 1984 and 92,2% for Hot.

As we considered our method language-independent we decided to try it on some other language. So we applied the method on

Bulgarian. We used the same corpus 1984 created during MULTEXT-EAST project for Bulgarian.

Not having any linguistic information about this language we made several experiments taking 3, 4, 5 and 6 final letters. An example of feature vector with six letters is presented below:

л , е , с , к , а , н , лескан

The results of the experiments are presented in the table 3.

Table 3. Lemmatization accuracy for Bulgarian

Num, of final letters	Accuracy
2 letters	61,6%
3 letters	72,5%
4 letters	83,0%
5 letters	86,1%
6 letters	83,9%

The result appeared to be very similar to the result obtained for Romanian. Comparing the results we found that 5 letters ending gives the best percentage – 86,1%.

3. Conclusion

In the paper a memory-based method of lemmatization is presented. The technique is new to our knowledge and its strongest advantage stands in its capability of lemmatization without any additional grammatical or morphological information. The fact that the resource requirements are so modest compared to other methods makes it easy to use for languages for which large linguistic resources are not available. No preprocessing steps are required and no tools or dictionaries are employed. The only requirement is a corpus of texts with lemmas attached to words.