Lemmatization with reversed dictionary and fuzzy sets

Alexander Gashkov[1,*], *and Mariia* Eltsova[2]

[1]Perm State Institute of Culture, 614000, 18 Gazeta Zvezda str., Perm, Russia
[2]Perm National Research Polytechnic University, 614990, 15 Bukireva str., Perm, Russia

Abstract. This paper deals with the problem of lemmatization of unknown words in Russian and German. For this purpose, the improved analogy method is used. The analogy method being built around reverse dictionary is very efficient and simple to realize. Adopting fuzzy sets for improving the analogy method is described in this paper. The fuzzy set is a good tool for modelling continual properties of a natural language. It can be implemented for lemmatization as a convenient tool to summarize information that is obtained from different sources. The paper contributes to solving the problem of increasing the accuracy of unknown words analysis.

1 Introduction

The quick and thorough natural language processing (NLP) is becoming indispensable because of the growing information flow and knowledge overabundance as a result of this process. One of significant part of NLP is lemmatization that is widely used in many NLP tasks, such as parsing, machine translation, abstracting etc. Lemmatization for highly inflected languages (like Russian) is one of the basic, indispensable steps in a NLP pipeline. However, the lemmatizers, which have been developed to date, are limited by the fact that they do not possess the high enough accuracy while analyzing the unknown words. Under unknown words we consider words missing in the lexicon or training set. The proposed paper concentrates on developing the lemmatizer which will be able to analyze the unknown words in Russian. This lemmatizer will significantly increase the analysis accuracy and offer possibility to trade precision for recall.

Many significant works on lemmatization published during the last 20 years describe this process for European Languages: English, Czech, German, Spanish, Finnish etc. [1, 2]; there are few works for “exotic” languages, e.g. Syriac [3] and minority languages [4]. In some tasks, a resulting lemma is required not by itself but as an index to lexicological resources for further analysis. J. Kanis and L. Müller worked on a lemmatizer from a Full Form - Lemma (FFL) training dictionary and with lemmatization of words unseen in the FFL dictionary, i.e. out-of-vocabulary (OOV) words (missing full forms, unknown words, and compound words). They obtained the accuracy about 75.1 % (approximated from precision and recall) [1].

[*] Corresponding author: gashkov@dom.raid.ru

In 2015, T. Müller, R. Cotterell, A. Fraser, and H. Schütze presented LEMMING, a modular log-linear model that jointly models lemmatization and tagging as well as supports the integration of arbitrary global features. They achieved the accuracy from 67.8% to 95.2 % for different languages [2].

There are many works concerning Russian language processing [5-6 etc.], but the most relevant to our research is the article which presents the results of independent evaluation of Russian morphological parsers [6]. The accuracy obtained for 8 parsers which could be used as lemmatizers varies from 4% to 78.7%, according to the evaluation [6]. So it becomes evident that lemmatizers for Russian should be improved. Therefore, we studied the work of G. G. Belonogov [7], in which he describes the analogy method, its applications in computer linguistics and evaluates it. According to G. G. Belonogov, this method demonstrates the precision of 99% by analyzing both known and unknown words. He used this method for tagging (determination of words' grammatical attributes) but not for lemmatization. The method suggested in the mentioned work was selected as a main method for our purpose.

The aim of this work is to develop an approach which allows to improve the accuracy of lemmatization of Russian words.

2 The analogy method

In this paper, we present how to apply for the lemmatization the analogy method which is modified by the usage of fuzzy sets. We assume that this combination can significantly improve the analysis quality. Moreover, this approach can be applied to other inflected languages.

The main feature of analogy method is that it uses reverse dictionary to determine word's grammatical categories as follows:

1. The algorithm performs lookup for a word which should be analyzed;
2. If the word is found, the grammatical categories from dictionary are returned and the algorithm terminates;
3. Otherwise, the algorithm looks for a place for unknown word in a reversed order;
4. The unknown word takes grammatical categories of the first word after newly inserted word (the only exception is that the insertion point was the last word), the algorithm terminates.

As G. G. Belonogov mentioned, there is no need to keep full dictionary but only the most important words where the set of categories changes should be preserved [7].

To adopt this algorithm for lemmatization we made some changes to reverse dictionary and to the algorithm. There are two new fields added into the dictionary for each word form: 1) N – number of letters from the right side of word form to be deleted and 2) pseudo-inflection. We called those two fields together as rule. To restore lemma the algorithm deletes the last N letters of word form given and adds the pseudo-inflection to the rest of the word form. The algorithm was changed in such a way that it takes in consideration not only the next word but the previous one too.

There are many large ranges of words with same rules in reverse dictionary, e. g. a part of Russian dictionary that was used in experiments. The Table 1 contains a fragment of Russian reverse dictionary used in experiments in this paper.

You are free to use colour illustrations for the online version of the proceedings but any print version will be printed in black and white unless special arrangements have been made with the conference organiser. Please check whether or not this is the case. If the print version will be black and white only, you should check your figure captions carefully and remove any reference to colour in the illustration and text. In addition, some colour figures will degrade or suffer loss of information when converted to black and white, and this should be taken into account when preparing them.

Table 1. Fragment of Russian reverse dictionary.

Word form	**Rule**
коврига	0, ∅
квадрига	0, ∅
верига	0, ∅
блицкрига	1, ∅
интрига	0, ∅
настрига	1, ∅
пострига	1, ∅
расстрига	*0, ∅*

The Table 1 demonstrates that many words following each other may have the same rule. We call this a range cluster. The borders of clusters are places where rule changes. In the given example, five clusters are present. The largest cluster of 3 words begins with 'коврига' and ends with 'верига'.

An average cluster size can be used to predict quality of analogy method for language. Inversely metric α – a number of clusters per word is used further. We can presume that an unknown word will take a random place of dictionary. Firstly, we presume that the analyzed word does not generate a new cluster. There are two possible situations: the word will take place 1) between two words with identical rules (inside a cluster), 2) between two words with different rules. In the first case we consider that this word surely possesses the same rule that the words in a cluster.

In other case, we are not able to assume its rule. We can consider that in the second case the precision will make 50 % and the probability of incorrect identification will equal α/2.

Secondly, we presume that the analyzed word generates a new cluster. That occurs when the word form placed between forms which have rules different from it (our example with 'блицкрига' above). We can assume that the probability of this event equals α. Thus, the general probability (GP) of error is not less than

$$GP \geq 1 - (1 - 0{,}5\alpha)\,(1 - \alpha) \tag{1}$$

and the probability (P) that rule is identified correctly is less than

$$P < (1 - 0{,}5\alpha)\,(1 - \alpha) \tag{2}$$

If an analyzed word is placed into a cluster (a word before it and a word after it have the same lemmatization rule), the lemma of an unknown word is considered as exactly determined and is built according to that rule. For more complex cases, e.g. words belong to different clusters (the rules are different), there is an ambiguity, and two or more rules of lemmatization, can occur some problems. Therefore, we use fuzzy sets to represent the results. All rules are considered as equally probable and the membership function is calculated as 1/n, where n is total quantity. If the rule occurs in both neighbor words strings, its membership function is calculated as 2/n.

Let’s consider an example of lemmatization of the word form missing in the lexicon, 'боулинга' (genitive singular from 'боулинг' – ‘bowling’). It is placed between the word forms 'шиллинга' (shilling) and 'лемминга' (lemming) as it is presented in the Table 3.

Table 2. Fragment of Russian reverse dictionary.

Word form	Lemma	Rule
шиллинга	шиллинг	1, ∅
лемминга	лемминг	1, ∅

We stated above that this approach can be applied to other inflected languages. Let’s demonstrate this statement on the German example. It is easy to observe from the given example that there are many large ranges of words with same rules in reverse dictionary (e. g. a part of German dictionary that was used in experiments). There are present seven clusters in given example. The largest cluster of 5 words begins with 'geheimniskündendem' and ends with 'auslaufendem'.

Table 3. A fragment of German reverse dictionary.

Word form	Rule
dem	1,r
ehedem	0,∅
jedem	1,r
alledem	0,∅
trotzalledem	0,∅
nachdem	0,∅
fremdem	2,∅
niemandem	2,∅
jemandem	2,∅
tandem	0,∅
geheimniskündendem	2,∅
zündendem	2,∅
schrumpfendem	2,∅
laufendem	2,∅
auslaufendem	2,∅

On the example of the German unknown word ‘Holzschraube’, we will demonstrate the ambiguous lemmatization (Table 4).

Table 4. Fragment of German reverse dictionary.

Word form	Lemma	Rule
Schraube	Schraube	1, ∅
Taube	Taub/Taube	1, ∅; 0,∅

‘Holzschraube’ (wood screw) is placed in the reverse dictionary between Schraube (screw) and Taube (pigeon or deaf). The rule for Schraube “delete zero letters and add zero ending” generates lemma Schraube. Taube has two rules associated. The first rule generates lemma Taub and the second one – Taube. So, the algorithm gives for ‘Holzschraube’ two lemmas (Holzschraube/0,667 and Holzschraub/0,333).

3 A review of modern Russian lemmatizers

It is well-known that most lemmatizers, regardless of the language they analyze, can be divided into two general classes: with lexicon and without it. While the first demonstrate 100% accuracy of lemmatizing the known words, on one hand, they have some troubles dealing with unknown words. On the other hand, lemmatizers without lexicon analyze unknown words just like known but they rarely reach 100% accuracy.

We mentioned above the article by O. N. Lyashevskaya et all. [6], which independently evaluates some Russian morphological parsers and presents the accuracy of them. The Table 5 represents the data published in [6]. Parsers were evaluated anonymously, so their nicknames are also given in that table.

Table 5. The accuracy of Russians lemmatizers for unknown words.

Participant	**Accuracy**
Desert	78,7%
Beaver	70,7%
Burlywood	69,3%
Copper	62,7%
Lavender	61,3%
Shadow	56,0%
Snow	13,3%
Forest	4,0%

During the evaluation, the parsers had to analyze 75 unknown words, the best parsers use a context as an auxiliary means to increase the precision and accuracy of analysis. The Table 5 demonstrates that the spread is very high, and the overage value is 62%, which could not be considered as a satisfactory result.

4 Experiment

4.1 Datasets

The first thing which should be described here are the dictionaries. Russian reverse dictionary is auto-generated from Zaliznyak's Russian grammatical dictionary [8] by deriving all possible forms and their rules from words and sorting in reversed order. The reverse dictionary consists of ~1,450,000 entries coding ~1,500,000 lemmas.

The reverse German dictionary is auto-generated from the TIGER corpus [9] and has noticeable lesser entries (~84,400) as it was created for proof only.

The test set contains 500 random words from the Internet, both Russian and German, which are not present in corresponding dictionary.

We compare our lemmatizer to the best ones described above for Russian and to LEMMING for German.

4.2 The experiment

We use here the fuzzy set, which considers two elements of dictionary (the previous one and the next one) as described above. All 500 words were placed into a dictionary in reverse alphabetic order, then their lemmas were given automatically for each word according to its neighbored words in cluster (clusters). The experts checked if the automated placement and lemmatization were correct. If the previous neighbored word and the next neighbored word generated two different word forms, 0.5 was added to the right answer and 0.5 – to the error.

It is possible to use many neighbored words placed above and down the analyzed one. Therefore, it becomes possible to manage a precision-recall ratio: the higher the number of neighbored words is, the higher is the recall and, sometimes, the lower is the precision. But the increase of the number of neighbored words allows to reduce the influence of exceptions, e.g. for words which lemmatizing rules differ from those of both neighbored words.

4.3 Results

We compare the results given by the described lemmatizers with our one. The results for the Russian are given in the Table 6.

Table 6. Quality of lemmatization of Russian words.

Participant	Accuracy
Desert	78,7%
Beaver	70,7%
Our approach	84.0 %

The Table 6 makes it evident that the improved analogy method (our approach in the Table 6) gives the highest result for the Russian out of the three tested lemmatizers, though it does not use context.

In the Table 7, there are the experimental results for the German language.

Table 7. Quality of lemmatization of German words.

Participant	Accuracy
LEMMING	90.9 – 93 %
Our approach	86.7 %

The Table 7 demonstrates that the results obtained with improved analogy method are comparable with those of one of the best European lemmatizers for German (LEMMING). By comparing the results for improved analogy method from Table 6 and Table 7 (our approach in the Table 6 and Table 7), we can point to the fact that it has better results for the Russian and comparable results for the German. Another fact for the German should be mentioned: precision for the German language is a bit higher because there are a lot of compound words in German, which are placed into the same cluster of reverse dictionary as their main component. Therefore, the improved analogy method could be a good additional solution for the lemmatization of German words. Moreover, the German reverse dictionary used for this purpose can be much shorter than the Russian one.

5 Conclusions

We have introduced the approach for analyzing the unknown words that requires no training or rules and can be used for many different languages. The approach has been evaluated on two languages and shows accuracy from 84% to 86.7%.

Advantages of the method are as follows:

- The quality of lemmatization can be estimated without experiments only by reverse dictionary;
- The method is simple to realize;
- The method can be used for most inflected and for some analytic languages.

The disadvantage of this approach is that it requires a relatively large dictionary, but the dictionary can be effectively reduced according to the application.

The results show that the developing approach can be used as standalone lemmatizer or addition to other lemmatization methods.

References

1. J. Kanis, L. Müller, Automatic Lemmatizer Construction with Focus on OOV Words Lemmatization, *TSD*, **3658** (2005)
2. T. Müller, R. Cotterell, A. Fraser, H. Schütze, Joint Lemmatization and Morphological Tagging with LEMMING, *EMNLP* (2015)
3. P. McClanahan et al., A probabilistic morphological analyzer for Syriac, *EMNLP* (2010)
4. N. Green, P. Breimyer, V. Kumar, and N. F. Samatova, WebBANC: Building Semantically-Rich Annotated Corpora from Web User Annotations of Minority Languages, *NODALIDA*, **4** (2009)
5. I. Segalovich, A fast morphological algorithm with unknown word guessing induced by a dictionary for a web search engine, *MLMTA'03* (2003)
6. O. N. Lyashevskaya et all. NLP evaluation: Russian morphological parsers, Transl. Russian, *Dialog'2010*, (2010)
7. G. G. Belonogov, *Computer linguistics and perspective information technologies* (2004)
8. A. A. Zaliznyak, *Russian grammatical dictionary: inflection* (1977)
9. S. Brants, S. Dipper, S. Hansen, W. Lezius, G. Smith, *Proceedings of the workshop on treebanks and linguistic theories* (2002)