Proc. of IS 2005, the 8th International Multiconference on in Information Society, 11-17 October 2005, Ljubljana pp. 195-198
Conference on Data Mining and Data Warehouses (SiKDD 2005)

LEARNING RULES FOR MORPHOLOGICAL ANALYSIS AND SYNTHESIS OF MACEDONIAN NOUNS

Aneta Ivanovska[1], *Katerina Zdravkova*[1], *Sašo Džeroski*[2], *Tomaž Erjavec*[2]
[1]Institute of Informatics, Faculty of Natural Sciences and Mathematics
Arhimedova 5, 1000 Skopje, Macedonia
[2]Department of Knowledge Technologies, Jozef Stefan Institute
Jamova 39, 1000 Ljubljana, Slovenia

ABSTRACT

This paper presents a machine learning approach to morphological analysis and synthesis of Macedonian nouns. For training and testing we used the nouns originating from Orwell's "1984". The paper presents experimental results of using the learned rules in the process of analysis, and in the process of noun formation. Training was performed with the whole set of Macedonian nouns from "1984" and tested by 10-fold cross-validation. All the potential nouns forms generated by the learning rules were compared with 275000 Macedonian noun forms. The accuracy of 92-97% is encouraging to apply the same approach to all categories of Macedonian words.

1 INTRODUCTION

Although morphological rules for noun formation in Macedonian have exhaustively been studied by linguists [1,2] for decades, they have been recently systematized [3]. The initial aim of the research presented here was to define morphosyntactic descriptions (MSDs) of Macedonian in line with Multext-East [5] and then implement them over all the words originating in Orwell's "1984". This process was not straightforward because the translation of the book didn't exist in electronic version [7].

The process of converting the printed version into a text file, the assignment of all the word-forms into appropriate grammatical categories, and their presentation into triplets (word-form, lemma, MSD) is also presented.

Particular attention in the paper is paid to noun analysis and formation, based on a machine learning approach. It has been performed with the machine learning system Clog [8]. This paper presents the process of machine learning of Macedonian headword and noun forms in more details (Fig. 1). Section 2 deals with the preprocessing of the printed version of the novel "1984" and its conversion into an electronic dictionary of word forms. Section 3 explains MSD tagging of Macedonian nouns. Section 4 describes the preparation of annotated nouns for training with Clog, training the rules for analysis and synthesis of the word-forms of the lemmas, and testing the accuracy of the generated rules. Section 5 presents experimental results of noun analysis and formation. Conclusion discusses the results and directions for future work.

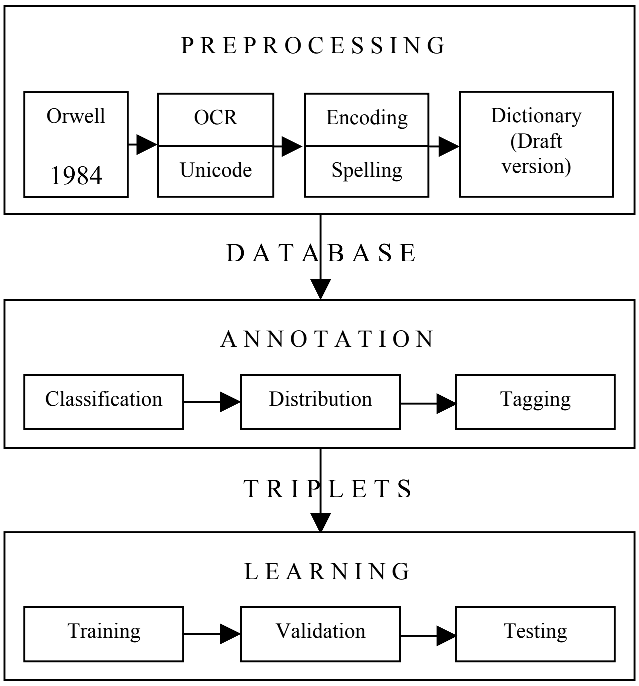


Figure 1: *Learning system diagram.*

2 PREPROCESSING OF "1984"

The Macedonian translation of Orwell's "1984" appeared rather late [7], but in spite of this fact, it didn't exist in an electronic form. There were two solutions to this initial problem: to type it from beginning to end, and simultaneously correct all the spelling errors, or to scan it. The second solution was found more appropriate, particularly because at that moment, no supplementary problems were noticeable.

The book was published in Macedonian Cyrillic script that can easily be transliterated into Latin. However, neither ISO Latin 2, which was used for other Slavic languages in Multext-East, nor its extension ISO Latin 5, could cope with Macedonian characters ќ and ѓ. Therefore, Unicode/UTF-8 encoding was chosen as the most suitable. Unexpectedly, the selection of Unicode encoding became the first problem. Namely, many Latin characters, such as b (Cyrillic v), c (Cyrillic s), p (Cyrillic r), and s (Cyrillic dz) were recognized as characters from the English alphabet. The character r was recognized as Cyrillic g. In the printed version of the novel, capital O was typed as number 0, the hyphen as minus, and the character ` (existing in words such as `рбет (spine)), as quotation mark ‘.

All incorrect characters were replaced. Then, correction of spelling errors started. After this stage, the electronic version of the book was ready for further processing.

In order to obtain the full “1984” dictionary, all sentences were first converted into a set of words, and multiple occurrences of the same 98846 word form were deleted from it. This approach was afterwards found wrong, because out of context, many word forms belonging to several different categories were deleted. Such examples are, for example, plural forms of the nouns врата = врати (doors), града = гради (breasts), забрана = забрани (prohibition), which are at the same time verb bases врати (to return), гради (to build), and забрани (to forbid). The deleted words were returned back in the dictionary to be later classified into other categories.

The preprocessing stage ended with a dictionary with 16836 different word forms. In order to enable annotation, they were converted into a database divided into 11 tables according to the grammatical category of the word (Fig 2.).

Category PoS	Occurrences	Attributes	Values
Noun	5466	5	15
Verb	4565	8	23
Adjective	3952	6	19
Pronoun	78	10	34
Adverb	606	2	7
Preposition	28	2	3
Conjunction	64	2	4
Numeral	70	15	17
Interjection	4	11	2
Abbreviation	16	21	1
Particle	3	2	9

Table 1: *Occurrence of word forms and their MSDs.*

3 ANNOTATIONS OF THE WORDS

Annotation of the words was made with Multext notation [6]. According to this notation, each word form is associated with a morphosyntactic description (MSD) presented as a packed string. Its first character, always uppercase, represents the part-of-speech (grammatical category). It is followed by a list of character values corresponding to the part-of-speech attributes.

The Macedonian language has 11 word categories, 84 attributes, and 134 values (Table 1). The number of possible combinations of these has not been determined yet.

3.1 Word classification

Almost 60% of all the words were automatically classified according to their inflexions. Unclassified words were manually distributed. The process was not completely bug-free, so final adjustment was done during MSD tagging.

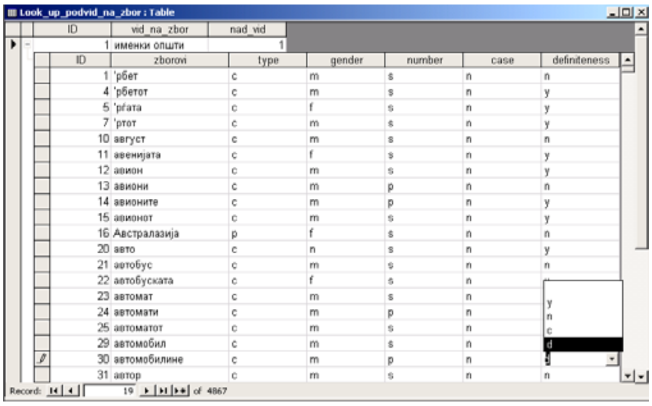
In parallel with the manual word classification, automatic rule-based classification of Macedonian words was also attempted. With the current accuracy of about 80%, this automatic system was found useless for Orwell’s novel.

3.2 Annotation of Macedonian nouns

Macedonian nouns have 5 attributes: type (common, proper), gender (masculine, feminine, neuter), number (singular, plural, count), case (nominative, vocative, oblique) and definiteness (no, yes, close, distant).

Three of these 15 values are language specific: the case oblique, which represents the remaining of former genitive, dative, and accusative and always has a unique form: Иван - Ивана, ешко - ешка. Definiteness is expressed by three forms: the suffixes от, та, то, те express definiteness of nouns independently of their location, suffixes ов, ва, во, ве express definiteness of nouns which are close, while suffixes он, на, но, не express definiteness of nouns which are distant.

The dictionary database enables the association of attributes and values, the addition, deletion, and replacement of the values, and manual correction of wrongly associated values (Fig 2, right corner).



ID	vid_na_zbor	nad_vid
1	именки општи	1

ID	zborovi	type	gender	number	case	definiteness
1	'рбет	c	m	s	n	n
4	'рбетот	c	m	s	n	y
5	'рѓата	c	f	s	n	y
7	'ртот	c	m	s	n	y
10	август	c	m	s	n	n
11	авенијата	c	f	s	n	y
12	авион	c	m	s	n	y
13	авиони	c	m	p	n	n
14	авионите	c	m	p	n	y
15	авионот	c	m	s	n	y
16	Австралазија	p	f	s	n	n
20	авто	c	n	s	n	y
21	автобус	c	m	s	n	n
22	автобуската	c	f	s	n	
23	автомат	c	m	s	n	
24	автомати	c	m	p	n	
25	автоматот	c	m	s	n	
29	автомобил	c	m	s	n	
30	автомобилине	c	m	p	n	d
31	автор	c	m	s	n	n

Figure 2: *Table of all the nouns and their MSDs.*

MSD tagging was automatically performed, and manually polished. Common nouns appearing in the middle of the sentence were separated from proper nouns according to the initial letter. The gender of definite nouns was determined from the definite suffixes, the gender of indefinite nouns according to the last character. The number of definite nouns was concluded by the definite suffixes, while the plural of neuter nouns and the count plural of masculine nouns was concluded according to the final characters (иња, and а). The case was set to be nominative, except for indefinite masculine and feminine nouns that ended with о (vocative) and masculine and neuter nouns that ended with а (oblique). Definiteness was completely determined by the suffixes.

In parallel with the annotation, lemmas were added in a separate column. At the beginning, it was intended to perform automatic analysis, but the number of rules [3], and the former use of a machine-learning tool indicated that manual analysis of "1984" dictionary should be easier and more accurate.

Addition of lemmas was best opportunity to polish the remaining spelling errors, and correct the exclusions of the rules used for MSD tagging.

4 PREPARATIONS OF LEARNING RULES

After adding the columns with the lemmas in the table of all the nouns and their MSDs, the table had to be transformed in a format suitable for running Clog. Consequently, the table was transformed into a document in which every noun was in a separate row, together with its lemma and its MSD, separated with a TAB space. The rows of the document were of the form:

word-form <TAB> lemma <TAB> MSD

where the word-form was the word as it appears in the running Orwell's text. Before this document was further used in the process of analysis, it had to be transliterated. The document was transliterated from Unicode/UTF8 to Latin2. The format of the document remained the same (triplets: word-form, lemma, MSD). Then, the process of analysis started. The analysis was divided into two parts: training of the rules and testing the accuracy of the rules.

The rules were trained with Clog, on the whole set of Macedonian nouns. Clog was run separately for each MSD, once for analysis and once for synthesis. The triplets, where each triplet was an example of analysis of the form *MSD (orth, lemma)* from the training set, were used. *MSD (orth, lemma)* is a relation, or predicate that consists of all pairs (word-form, lemma) that have the same morphosyntactic description. *Orth* is the input, and *Lemma* is the output argument. A set of rules had to be learned for each of the MSD predicates. For every MSD predicate there could be a set of rules, and a set of exceptions from the rules. An example of the rules and the exceptions for morphological analysis are given in Figure 3.

Exceptions:

raspravii -> rasprava
strui -> struja
race -> raka
noze -> noga
boi -> boja

Rules:

*sti -> *st
*ii -> *ija
id*i -> id*ja
*i -> *a

Figure 3: *Morphological analysis exceptions and rules for common nouns of feminine gender plural.*

During the process of training, several mistakes were noticed in the set of words used for training. Those were mostly spelling errors, but they induced wrong rules, that cannot be used later. The spelling errors were easy to notice, and after they were fixed, the training was performed again. The number of erroneous rules decreased.

The second part of the process of analysis was the testing of the induced rules. At first, the testing was performed on the whole set of words that was used for the training. The accuracy of the rules was 100%, because the rules were induced from exactly the same set of words.

Therefore, to test the real accuracy of the rules, 10-fold cross validation was performed on the set of words. The ten sets were created by a random choice of words, and they consisted of approximately the same number of words (around 500).

There was an important numeric variable, which influenced the 10-fold cross-validation, namely the minimum number of examples each trained MSD should have. By default, that number was set to 100. Changing that number caused slight differences in the accuracy of the rules. If that value were very high, the accuracy would have been smaller. The same happened when that number was very small. Generally, with a minimum number of 100, the average accuracy was around 96-97%.

The next thing that has been done is the opposite process of analysis - synthesis and generation of the word-forms from the lemmas.

The first thing that was done, was rearranging the document with the triplets of the nouns. The columns with the lemma and the word-form were switched, so the triplets looked like this:

lemma <TAB> word-form <TAB> MSD

The process of synthesis is very similar to the process of analysis, so the activities for training and testing the rules are approximately the same. After the rearrangement of the document with the words, transliteration from Unicode/UTF8 to Latin2 has been made.

This was followed by the process of training the rules for the synthesis of word-forms from the lemmas. Again, some mistakes were found and corrected, and the training was repeated. An example of the rules produces can be seen in Figure 4.

Same as in the process of analysis, 10-fold cross validation was performed. The accuracy of the rules is slightly smaller, than the accuracy we have seen for the analysis.

The next step was the generation of all the word-forms for every lemma. The motivation for this was producing a lexicon for all the word-forms appearing in Orwell's „1984". First, it was decided what combination of MSDs one lemma can have. For example, if the gender of the lemma is feminine, then the word-forms of the lemma can have one of these MSDs: Npfsnn, Ncfsnn, Ncfsny, Ncfpny, Npfsvn, Ncfsvn, Ncfsnc, Ncfpnc, Npfson, Ncfson, Ncfsnd, Ncfpnd, Npfpnn, Ncfpnn.

Exceptions:

kolenica -> kolenicite
dete -> decata
zivotno -> zivotnite
bebe -> bebinjata

Rules:

po*e -> po*injata
*ce -> *cata

Figure 4: *Morphological synthesis exceptions and rules for common nouns of neutral gender single.*

Since the MSDs of the possible word-forms of one lemma were known, the induced rules from the process of synthesis could be used to generate all the word-forms of the lemmas. A document of all the triplets: lemma-MSD-word-form, was generated, where every combination of a lemma and a word-form was in a different row (Fig 5.).

vladetel	Ncmsnn	|vladetel|
vladetel	Ncmsvm	|!?!|
vladetel	Ncmson	|!?!|
vladetel	Ncmsny	|vladetelot|
vladetel	Ncmsnc	|vladetelov|
vladetel	Ncmsnd	|!?!|
vladetel	Ncmpny	|vladetelite|
vladetel	Ncmpnc	|???|
vladetel	Ncmpnd	|!?!|
vladetel	Ncmtnn	|vladetela|
vladetel	Ncmpnn	|vladeteli|

Figure 5: *Generated word-forms of the noun 'vladetel'*

The accuracy of the generation of word-forms is not yet known, since there is not a suitable document, in the right format, on which accuracy can be tested.

5 EXPERIMENTAL RESULTS

During the process of analysis and synthesis, measuring the accuracy of the rules has been made.
During the testing of the rules with 10-fold cross validation, the accuracy of every set of words has been calculated (Table 2).

Minimum number	0	10	50	100
Accuracy	96.97%	97.90%	97.29%	97.01%
Standard Deviation	0,71	0,46	0,50	0,47

Table 2: *Accuracy of the rules of analysis*

The total number of Prolog rules and exceptions for the analysis is 317. Here are the accuracies of the rules generated during the process of synthesis (Table 3):

Minimum number	0	10	50	100
Accuracy	94,18%	94,48%	94,61%	94,81%
Standard Deviation	2,72	2,43	1.29	0,18

Table 3: *Accuracy of the rules of synthesis*

The total number of Prolog rules and exceptions for the analysis is 364.

6 CONCLUSION

In the research presented in this paper, we have created Orwell's "1984" corpus, defined MSD specifications for Macedonian language, and manually annotated all the nouns form the created corpus. We have also presented the process of obtaining learning rules for the analysis and synthesis of Macedonian nouns.
We obtained an accuracy of more than 97% for noun analysis, and an accuracy of 94% for noun synthesis. It is encouraging to implement the same approach for other grammatical categories.
Further work will mainly focus on learning more nouns, preferably from the rule-based lexicon [3]. Furthermore, we intend to generate a complete lexicon for Orwell, and manually test at least for sample of forthcoming rule-based lexicon of Macedonian language.

References

[1] Koneski, B., Grammar of Standard Macedonian, Prosvetno delo, (first edition), in Macedonian, 1952
[2] Koneski, B. Grammar of Macedonian Language, Prosvetno delo,, in Macedonian, 2004
[3] Petrovski, A. About Macedonian Computational Dictionary, Proceeding of BCI2005, Ohrid, Macedonia, 2005.
[4] MULTEXT-EAST Web site (http://nl.ijs.si/ME/)
[5] Erjavec, T., Džeroski, S. Machine Learning of Morphosyntactic Structure: Lemmatizing Unknown Slovene Words, *Applied Artificial Intelligence* 18(1), 2004.
[6] Orvel, Dž, "1984", Detska radost, Skopje, in Macedonian, 1998.
[7] Manandhar, S., Džeroski, S., Erjavec, T. Learning Multilingual Morphology with Clog, Lecture Notes in Artificial Intelligence 1446, Springer, 1998.