Evaluation and improvement of stemmers for Serbian language

Petra Antić

Abstract – **In this paper the existing stemmers for Serbian language were evaluated based on the percentage of incorrect stems they produce. According to the obtained results, new rules were introduced to minimize the errors for three different word types. The evaluation after the improvement has showed that the new rules had the positive effects to the stemmer correctness for all three types.**

Keywords – **Stemmers, Serbian language, Error metric**

I. Introduction

Stemming is a low level task in natural language processing, whose goal is to map different word variations to the same form, called stem, by removing suffixes.

The corpora with pairs of words and their correct stems are not available most of the times, so authors mainly use manual ways for assesing the stemmer correctness. Milošević presented two methods [1]: in the first method, a news article was manually stemmed, and then the produced text was compared with the outcome of the stemmer applied to the same article. The other method used machine stemming as the first step, and then a person was reading the stemmed text. The stems were evaluated based on the possibility for the human reader to conclude the original meaning (no overstemming), and the ability of the stem to cover all morphological variations of its lemma (no understemming).

To the best of the author's knowledge, there are three publicly available stemming algorithms for Serbian and one for Croatian (which can also be applied to Serbian, given the similarities between these two languages). Two of them are by the authors Kešelj and Šipka [2] – the optimal and the greedy stemmer, and the third one is an improved version of the aforementioned greedy algorithm, given by Milošević [3]. They all employ a suffix subsumption approach, while the stemmer for Croatian, by the authors Ljubešić and Pandžić [4], relies on regular expressions.

In order to discover deficiencies of the existing stemmers with a goal to minimize them, but also to decrease human effort needed for this task and obtain more robust results, the author has defined an error metric expected to be applied to a Serbian language lexicon.

The next section covers the lexicon used for conducting the evaluation and the definition of the metric used. Section III presents the results obtained by the evaluation, while in Section IV the improvements and their effects are discussed. Finally, Section V summarizes the findings of this paper and gives directions for the future work.

Petra Antić is with the Faculty of Electronic Engineering, University of Niš, Aleksandra Medvedeva 14, 18000 Niš, Serbia, e-mail: petra.antic92@gmail.com.

II. Evaluation method

A. Serbian Language Lexicon

Stemmer evaluation was conducted on the Serbian language lexicon, srLex [5]. This is a flective lexicon, where each item is presented as the following 5-tuple: inflectional form, lemma (canonical form), morphosyntactic description, absolute frequency, frequency on million occurences (ženu, žena, Ncfsa, 15838, 0.028556). An inflectional form represents morphological variation of a lemma, such as a case for nouns or a tense for verbs. Morphosyntactic descriptions are given by MULTEXT-East Morphosyntactic Specifications for Serbian, Version 5 [6]. This lexicon contains 108,829 different lemmas, and 5,326,726 inflectional forms of these lemmas.

Before applying stemmers on the words from the lexicon, preprocessing was performed with a goal to reduce the amount of data and to remove the irrelevant data. The first step was the removal of the words whose occurrence frequency is less than 0.00002, as well as interpunction signs. This has reduced the number of items in the lexicon more than 10 times, which has enabled evaluation to be performed in reasonable time. Finally, only unique words were extracted from the first column, because stemmers do not differentiate between the different cases or the same word forms obtained from different word types, and proper nouns were removed, which gave a final number of 247,518 words for the stemmers' application and evaluation.

B. Stemmer Error Metric

The goal of stemming is to reduce a word to its base, which should be the same for all morphological variations of that word. This further means that a stem obtained from a word in a morphological form is expected to be the same as a stem obtained from that word's lemma. Since the used lexicon contains pairs of lemmas and their variations, it is possible to easily compare stems of these two words. If there is a difference between these stems, it can be concluded that there is an error in stemming.

Based on this idea, the following metric was defined – the percentage of items in the lexicon where a mismatch between stem of lemma and stem of its morphological variation exists. It is important to note that, while this metric can be used to locate the errors in stemming, it can not define overall stemmer correctness – even if the obtained stems are the same, it doesn't necessarily means they are correct. However, in most cases, if all of the variations are giving the same stem,

and there is no understemming, the stem can be accepted as correct.

The evaluation and the improvements were applied to Java implementation of the stemmers from the SCStemmers package [7], presented by Batanović [8].

III. Stemmer evaluation

With the aim to locate word types which are the most error prone, this metric was firstly calculated using all four stemmers on the main word types, shown in Table I.

Word type with the most errors are found to be pronouns, for which each stemmer made an error for almost a half of the word pairs. For adjectives, verbs and numbers, at least one stemmer gave acceptable results, however, it was decided to make further analysis for adjectives and verbs, and check whether some of their subtypes show significant errors in stemming.

The next step was calcuating the metric for subtypes of verbs, adjectives and pronouns. The idea was to extract the subtypes for which all of the stemmers make error of 0.5 or more, and use those subtypes as candidates for further analysis and improvement. Given the large number of possible subtypes, only the extracted candidates are shown in Tables II-IV. The rows in all tables in this paper are labeled with morphosyntactic descriptions used in srLex lexicon [6].

Table I
Percentage of different stems for word and lemma for main word types

	KSG	KSO	M	LjP
P	**0.61**	**0.59**	**0.59**	**0.48**
Q	0.12	0.12	0.12	0.12
A	**0.46**	**0.26**	**0.57**	**0.33**
R	0.28	0.17	0.23	0.18
C	0.02	0.02	0	0.02
S	0.07	0.09	0.06	0.06
V	**0.2**	**0.15**	**0.4**	**0.39**
M	0.62	0.45	0.52	0.17
N	0.21	0.18	0.33	0.07

A. Prounouns

For pronouns, two the most error prone groups are demonstrative (Pd) and indefinite pronouns (Pi), as well as some forms of personal (Pp), interrogative (Pq), reflexive (Px) and possessive (Ps) pronouns, with metric results shown in Table II.

The first problem is with short demonstrative pronouns (*taj*, *ovaj*), where two types of errors are present: either the word ending is completely removed, leaving the stem of only one letter (*toga-t*), or no change was made at all.

The other problem is related to demonstrative pronouns of quality, such as *takav/onakav/ovakav*, which show presence of a voice change of absent A, and are stemmed to base *takv* for every case form except nominative, which gets stemmed to *takav*.

Table II
Percentage of different stems for word and lemma for pronouns

	KSG	KSO	M	LjP
Pd	0.73	0.72	0.79	0.66
Pi	0.62	0.61	0.51	0.5
Pp2	0.69	0.69	0.5	0.69
Pq-	0.63	0.63	0.63	0.5
Px-n	0.79	0.5	0.79	0.5
Ps2mp	0.9	0.5	0.9	0.6

An additional problem are volatile vowels [8] which can be present in some case forms at the end of the word: *takvog*(*a*). Their presence or absence doesn't change a word in a grammatical or a semantical way, and its usage depends only on the writing style.

Interogative pronouns show errors similar to demonstrative ones, with short forms being stemmed to only one letter (*koga – k*), but also errors caused by different bases for nominative and other cases (*šta – čega*). This difference in base is also a cause of errors for personal pronouns (*ti – tebe, on-njemu*).

B. Verbs

Table III shows errors for verb subtypes, and it can be seen that all of them are the forms of auxiliary verbs (Va), which, because of their irregular nature, have a mismatch for almost all cases.

Table III
Percentage of different stems for word and lemma for auxiliary verbs

	KSG	KSO	M	LjP
Vam	1	1	1	1
Var	0.93	0.93	0.57	0.7
Vae	0.5	0.5	0.5	1

Different authors had different approaches regarding auxiliary verbs. For example, the stemmer by authors Kešelj and Šipka does not have special rules for this word type, so there are stemming examples such as *biti-b*. Milošević uses a dictionary of mappings for processing these verbs, where every form of an auxiliary verb is mapped to its infitive form. However, this dictionary does not cover all the possible forms, so there are still mismatches such as *budite-bud* and *biću-biti*. Finally, the authors Ljubešić and Pandžić have the third approach – they consider auxiliary verbs to be stopwords and ignore them during stemming. Still, there are also incosistencies in this approach, because word *bude* gets stemmed to form *bud*.

C. Adjectives

For adjectives, subtypes with the highest rate of errors are comparative (Agc) and superlative (Ags) forms, with the

results shown in Table IV. High values for superlative are expected, given that none of the stemmers are removing prefixes and prefix *naj-* always remains in the stem.

TABLE IV
PERCENTAGE OF DIFFERENT STEMS FOR WORD AND LEMMA FOR ADJECTIVES

	KSG	KSO	M	LjP
Agc	0.75	0.83	0.93	0.62
Ags	0.99	0.99	0.99	1

The analysis of the cases where a mismatch for a comparative exists, shown that one of the problems is the inconsistency in stemming – for some words, suffix for the comparative is removed, while for the others it is not. Then, all of the stemmers are making a mistake when iotation voice change occurs (*mlad– mlađi*), where only *-i* of the suffix *-ji* is removed, while stem keeps the changed consonant (*đ*). Another problem is also caused by a voice change, in this case the absent A, when the letter A is missing in the stem of a case form (*ružn*), while it is present in the stem of lemma (*ružan*). For examples such as *spontaniji – spotanij* only a partial removal of suffix *–iji* can be noticed.

IV. STEMMER IMPROVEMENTS

All four stemmers have shown advantages and disadvantages, so it was not obvious based only on the results which stemmer should have been chosen for the improvements. However, by comparing their implementations, the one by Ljubešić and Pandžić has proven to be the most appropriate for the expansion. Its biggest advantage is that, by using regular expressions, suffix removal is dependent on the letter combination which precedes it. Also, it is structured in a way that each rule pattern covers more suffixes, so it is easier to deduct which type of word a rule covers. In addition, the number of the rules is the smallest – it has 70 rules, opposed to 285 in the stemmer by Milošević, or even 1000 and over 17000 rules in the greedy and the optimal version of stemmer by Kešelj and Šipka.

One of the proposed improvements was to add a new structure for post-transformations, which will solve problems with different bases of an inflection and a lemma, or problems with base transformations due to iotation changes. It was implemented as a two level hash map, where keys on the first level are the regular expressions for the word start, and keys on the second level are the regular expressions of the word end. The value obtained based on these two keys is post transformation which should be applied.

In addition to this new structure, new rules were also introduced for all word types classified as the most error prone.

A. New Rules

Since almost all pronoun subtypes showed problems with stemming, and a number of existing pronouns is limited and many are different only in prefixes (*kakav – ikakav – nekakav*), it was decided to define special rules for all pronouns. These rules would be placed on the beginning of the list, to be applied before more general rules.

Personal pronouns were stemmed in a way to keep the gender (*njemu – on, njoj – ona*), while other, adjective-like pronoun types, were stemmed to the male gender. Using this approach, twenty new rules were added.

For defining rules for verbs, the author has used rules for building simple verb forms. Using the rules instead of the dictionary was chosen for its structure and smaller posibility to exclude some of the forms. By Klajn's recommendation [9], and in complience with lemmas in the lexicon, forms of the verb *jesam* were reduced to the base *biti*. Using this aproach, six new rules were added.

Rules for adjectives included four irregular verbs stated in the grammar books (*dobar, zao, mali, veliki*), as well as three adjectives known to be the only ones with suffix *–ši* (*lep, lak, mek*). In addition, a rule was created to try to solve the problem of iotation in the adjective's base in the comparative form. This rule included the usage of post tranformation to transfrom the changed letter of the base (*đ* in *mlađ*) to its original letter (*d*). Another rule was added to ensure the consistency in removing suffix *-iji* and its inflections.

B. Results of Improvement

TABLE V
PERCENTAGE OF DIFFERENT STEMS FOR WORD AND LEMMA FOR PRONOUNS AFTER THE IMPROVEMENTS

	LjP	A
P	0.48	0.06
Pp	0.61	0.15
Pq	0.31	0.19
Ps	0.18	0
Pd	0.66	0.02
Px	0.41	0
Pi	0.5	0.1

After introducing the new rules, the error metric was recalculated for the relevant word subtypes, and the results are shown in Tables V-VII.

There is a significant decrease in errors for all subtypes of pronouns, which can be seen in Table V. The errors that are still present are mostly dependent on:

- The lexicon structure – there are words which are not common in Serbian, such as *tko* instead of *ko*; personal pronouns are given in the lexicon with lemma in the male gender, and an assumption was made that for these pronouns the gender should be preserved.
- The overlapping forms for a different gender and number (e.g. *one* can be genitive, 2nd person, singular, female gender, or acusative, 3rd person, plural, male gender).

Improvements are noticeable also for auxiliary verbs (Table VI), but the problems similar to those with the pronouns still

remain: the lexicon contains some iecavic forms which are not covered by the rules, and there is also overlapping with other word types (*je* can be shortened version of verb *jesam*, but also a shortened version of the personal pronoun *ona* in acusative singular form).

TABLE VI
PERCENTAGE OF DIFFERENT STEMS FOR WORD AND LEMMA FOR AUXILIARY VERBS AFTER THE IMPROVEMENTS

	LjP	A
Vam	1	0
Var	0.7	0.13
Vae	1	0.5

For comparative forms of adjectives there is only a slight decrease in errors, which can be seen in Table VII. The reasons for those results are various:

- Some adjectives do not form a comparative of its full positive form, but from shortened from (*dub-ok – dub-lji*). In this cases, because of *-ok* and similar suffixes not consistently covered by the rules, there is still a mismatch even if comparative is correctly stemmed.
- The absent A is still an unsolved problem, given that it is very hard to differentiate between situations where it is present and where it is not (e.g. *smotan – smotaniji* and *verovatan – verovatniji*: by simply looking at the suffixes and letters which are preceding the suffix, a difference can not be noticed, so this type of a problem should be approached in a more sofisticated way).
- Some of the comparative forms are already covered by a rule for female nouns ending with the suffix *-ija* (e.g. *galija*), so that suffix is not removed for them. Again, a more sofisticated approach would be needed to differentiate between these two cases.

TABLE VII
PERCENTAGE OF DIFFERENT STEMS FOR WORD AND LEMMA FOR ADJECTIVES AFTER THE IMPROVEMENTS

	LjP	A
Agc	0.62	0.56
Ags	1	1

The final recalculation of the error metric was conducted for all word types, in order to evaluate the impact of the new rules on the whole word set. The results are given in Table VIII, where we have excluded, for the clarity, word types without a changed result of the error metric.

The sigificant improvement can be noticed for pronouns in general. For adjectives and verbs, the improvements are minor, since only the specific word types were targeted – however, it can be concluded that the additional errors were not created by the new rules. Interesting side effect can be seen for adverbs, where a small decrease in error metric was caused due to the equivalent forms for the adverbs and the adjectives in the neutral gender. Only the nouns have suffered the increase of the error metric, most probably due to the introduced iotation rules for adjectives.

TABLE VIII
PERCENTAGE OF DIFFERENT STEMS FOR WORD AND LEMMA FOR MAIN WORD TYPES AFTER THE IMPROVEMENTS

	LjP	A
Total	0.238557	0.236956
P	0.478708	0.057269
A	0.334851	0.329862
R	0.175959	0.174632
V	0.395296	0.391382
N	0.066764	0.074437

V. CONCLUSION

This paper has presented the common errors of existing stemmers for Serbian language and the attempt to minimize those errors. For particulat word types there were noticeable improvements, and for others there are remaining problems that need attention. However, the efficiency of the improved stemmer was not evaluated in a real application. One of the following steps would be to apply this stemmer in a task such as information extraction or sentiment analysis, and verify if the introduced changes are showing better results for that task.

REFERENCES

[1] N. Milošević, "Sentiment Analysis of Sentences in Serbian language", Master's Degree Thesis. School of Electrical Engineering, University of Belgrade, Belgrade, Serbia, 2012. ("Mašinska analiza sentimenta rečenica na srpskom jeziku")

[2] V. Kešelj and D. Šipka, "A Suffix Subsumption-Based Approach to Building Stemmers and Lemmatizers for Highly Inflectional Languages with Sparse Resources," INFOtheca, vol. 9, no. 1–2, pp. 23a–33a, 2008.

[3] N. Milošević, "Stemmer for Serbian language." arXiv 1209.4471, 2012.

[4] N. Ljubešić, D. Boras, and O. Kubelka, "Retrieving Information in Croatian: Building a Simple and Efficient Rule-Based Stemmer," in INFuture2007: Digital Information and Heritage, Zagreb, Croatia: Department for Information Sciences, Faculty of Humanities and Social Sciences, 2007, pp. 313–320.

[5] N. Ljubešić, F. Klubička, Ž. Agić, and I.-P. Jazbec, "New Inflectional Lexicons and Training Corpora for Improved Morphosyntactic Annotation of Croatian and Serbian," 10th International Conference on Language Resources and Evaluation (LREC 2016), Conference Proceedings, pp. 4264–4270, Portorož, Slovenia, 2016.

[6] MULTEXT-East Morphosyntactic Specifications, Version 5. http://nl.ijs.si/ME/V5/msd/html/msd-hr.html#msd.R-hr

[7] SCStemmers – GitHub repository. https://vukbatanovic.github.io/SCStemmers/

[8] V. Batanović, B. Nikolić, and M. Milosavljević, "Reliable Baselines for Sentiment Analysis in Resource-Limited Languages: The Serbian Movie Review Dataset," 10th International Conference on Language Resources and Evaluation (LREC 2016), Conference Proceedings, pp. 2688–2696, Portorož, Slovenia, 2016.

[9] I. Klajn. *Serbian language Grammar*. Beograd, Zavod za učenike i nastavna sredstva, 2005. (*Gramatika srpskog jezika)*