Vocabulary extension by paradigm prediction

Attila Novák
(Supervisor: Dr. Gábor Prószéky)
novak.attila@itk.ppke.hu

***Abstract*—Morphological analysis and generation are important tasks in natural language processing systems, especially in the case of morphologically complex languages. Computational morphologies often consist of a lexicon and some rule component, the creation of which requires various competences and considerable effort. Such a description, on the other hand, makes an easy extension of the morphology with new lexical items possible. Most freely available morphological resources, however, contain no rule component. They are usually based on just a morphological lexicon, containing base forms and some information (often just a paradigm ID) identifying the inflectional paradigm of the word, possibly augmented with some other morphosyntactic features. The aim of the research presented in this paper was to prepare an algorithm that makes the integration of new words into such resources similarly easy to the way a rule-based morphology can be extended. This is achieved by predicting the correct paradigm for words, which are not present in the lexicon. The supervised learning method described in this paper is based on longest matching suffixes and lexical frequency data, and is demonstrated and evaluated for Russian.**

***Keywords*-morphology; paradigm prediction; Russian**

I. Introduction

Morphological analysis is an important task in any natural language processing chain, preceding any further analysis of texts. It is also unavoidable in information retrieval, or indexing algorithms, where the lemma of words are to be used in order to have a robust representation of the information present in the documents.

Large-scale computational morphologies are usually created using a morphological grammar formalism that minimizes the amount of information necessary to include in the source lexicon about each lexical item by providing some rule-based method of formalization of the morphological behavior of words. This allows an easy extension of the morphology with new lexical items. This approach also gives the creator of the morphology complete control over the quality of the resource. Building rule-based morphological grammars, however, requires threefold competence: familiarity with the formalism, knowledge of the morphology, phonology and orthography of the language, and extensive lexical knowledge. Many morphological resources, on the other hand, contain no explicit rule component. Such resources are created by converting the information included in some morphological dictionary to some simple data structures representing the inflectional behavior of the lexical items included in the lexicon. The representation often only contains base forms and some information (often just a paradigm ID) identifying the inflectional paradigm of the word, possibly augmented with some other morphosyntactic features. With no rules, the extension of such resources with new lexical items is not such a straightforward task, as it is in the case of rule-based grammars. However, the application of machine learning methods may be able to make up for the lack of a rule component. In this paper, we intend to solve the problem of predicting the appropriate inflectional paradigm of out-of-vocabulary words, which are not included in the morphological lexicon. The method is based on a longest suffix matching model for paradigm identification, and it is showcased with and evaluated against an open-source Russian morphological lexicon.

Table I
THE SIZE OF EACH TEST SET

	rare	average	frequent
Number of words	3970	36917	9633

II. Training and test data

In the experiments described in this paper, we used the LGPL-licensed open-source Russian morphology available from www.aot.ru [1]. The core vocabulary of this morphology is based on Zaliznyak's morphological dictionary [2]. It contains 174 785 lexical entries, each of which are classified into one of 2 767 paradigms. For the evaluation of the performance of the paradigm assignment algorithm, we also used the frequency distribution of Russian lemmas, taken from Serge Sharoff's Russian internet frequency list.[1]

The morphological lexicon was then separated into training and test sets in three different settings based on lemma frequencies. First, rare words were separated from the lexicon. These are the ones occurring at most 10 times in the internet corpus. As the frequency list includes words with at least 8 occurrences, this was the lower limit. In the second setting, the middle range words were separated for testing, i.e. the ones that occur at most 100 times. In the third case, the most frequent words were considered, which correspond to a frequency value of at least 1000. The training set in each case was the complement of the training set with regard to the whole lexicon. The size of each set is shown in Table I.

III. Features affecting the paradigmatic behavior of Russian words

When attempting to predict the inflectional paradigm for Russian words, certain grammatical features of the lexical item need to be known in order to have a good chance of guessing right. Lemma and part of speech are obviously necessary features, although part of speech can be guessed

[1] http://corpus.leeds.ac.uk/frqc/internet-ru.num

A. Novák, "Vocabulary extension by paradigm prediction,"
in *PhD Proceedings Annual Issues of the Doctoral School - 2014*, T. Roska, G. Prószéky, P. Szolgay, Eds.
Faculty of Information Technology and Bionics, Pázmány Péter Catholic University.
Budapest, Hungary: Pázmány University ePress, 2014, pp. 145-148.

from the lemma for adjectives and verbs with rather good confidence. Nevertheless, we assumed these to be known, as these properties of words are present in any dictionary.

For nouns, a number of additional features (gender, countability and animacy) play a role in determining the morphosyntatctic feature combination slots which make up the paradigm of the given lemma. There are also nouns, which are undeclinable. Of these features, gender is indicated for each headword in any dictionary, and undeclinable nouns are also usually marked as such. Certain abstract, collective and mass nouns (and, in the aot resource, also many proper names) lack plural forms, while there are also pluralia tantum, which have no singular.

Animacy affects the nominal paradigm in a manner that does not influence the actual set of possible word forms. However, there is a case syncretism in Russian, which depends on animacy. For animate nouns, plural accusative coincides with genitive (for masculine nouns, the same applies also to singular). For inanimate nouns, on the other hand, the form of accusative matches that of the nominative. This difference is still present in the case of homonyms, where one of the senses of the word is animate, and another form is inanimate. Note, however, that the animacy feature, although it is present in the aot lexicon, is not generally made explicit in other dictionaries, because a human user can infer this information from the meaning of the word. We thus have not used this information. Similarly, the set of valid morphosyntactic feature combinations for verbs depends on verbal aspect and transitivity/reflexivity. Thus, these properties need to be known for verbs, and, indeed, they are listed in dictionaries. E.g. nontransitive verbs lack passive participles; verbs of perfective aspect lack present participle forms; and many verbs of imperfect aspect lack past participial (especially passive) forms. The adverbial participial forms a verb may assume also depend on aspect (and also on other idiosyncratic lexical features).

Defectivities of the adjectival paradigm, e.g. the lack of short predicative forms and synthetic comparative and superlative forms depends on semantic and other, seemingly idiosyncratic, features of the lexeme. E.g. relational adjectives usually lack these forms. Such properties, however, were not made explicit in the aot lexicon, neither are they present in normal dictionaries, so we did not use any lexical features for adjectives beside part of speech.

Thus, when defining the feature set for predicting inflectional paradigms of words, we assumed that the lemma and the lexical properties mentioned above: part of speech, gender, verb type, etc., are known. However, some morphological characteristics relevant from the aspect of inflection cannot be derived neither from a simple dictionary, nor from the surface form of a word.

The other set of features we used are n-character-long suffixes of the lemma for various lengths n. The maximum suffix length is a parameter of the algorithm. It was set to 10 in the experiments reported in this paper. In order to exploit this information, a suffix model is created based on the lexicon. An illustration of how this model including both the endings and the lexical features is generated is shown in Figure 1.

мумиё[N.n.*.-];prd:25	мумиё|n*[N.n-25]
остриё[N.n.-];sfx:ё;prd:1709	остри#ё|n[N.n-1709]
бабьё[N.n-];sfx:ё;prd:210	бабь#ё|ns[N.n-210]
дубьё[N.n-];sfx:ё;prd:210	дубь#ё|ns[N.n-210]
свежевьё[N.n-];sfx:ё;prd:210	свежевь#ё|ns[N.n-210]
цевьё[N.n.-];sfx:ьё;prd:1433	цев#ьё|n[N.n-1433]
жнивьё[N.n];sfx:ё;prd:1103	жнивь#ё|n[N.n-1103]
суровьё[N.n];sfx:ё;prd:210	суровь#ё|ns[N.n-210]
мостовьё[N.n];sfx:ё;prd:210	мостовь#ё|ns[N.n-210]

Figure 1. A portion of the suffix model. The format of the right column is: `lem#ma|lex-features[PosTag-paradigmID]`, where `ma` is a required ending of the lemma for all items in the paradigm identified by `paradigmID`.

IV. Creation of the suffix model

A suffix trie is built of words input to the training algorithm in the form shown in the right column of Figure 1. The lemma is decorated with the following features (from right to left):

- The tag in brackets consists of two parts: part of speech (and, in the example below: gender) is followed by the appropriate paradigm ID from the aot database; the two are separated by a hyphen. This is the information to be predicted by the algorithm for unknown words. After processing the training data, terminal nodes of the suffix trie link to a data structure representing the distribution (relative frequency) of tags for the given suffix.
- A suffix following a vertical bar is attached to the end of the lemma. This represents the available lexical knowledge about the lexical item in an encoded form.[2]
- Some paradigms are restricted to lemmas ending in a specific suffix. There is a hash mark at the beginning of the suffix of the lemma that is required by the given paradigm ID to be valid. The given paradigm ID is not applicable to words not having that ending. E.g. all lemmas in paradigm 1433 must end in *ъё*.

V. Ranking

The suffix-trie-based ranking algorithm that we used was inspired by the suffix guesser algorithm used in Brants' TnT tagger to estimate the lexical probability of out-of-vocabulary words ([3]). However, that model did not prove to perform well enough in this task. So we modified the model step-by-step until we arrived at a model that turned out to be simpler, yet to perform much better. The paradigms are predicted by assigning a score to each paradigm for each word. Then, the higher this score is for a paradigm tag for a certain word, the more probable it is that the word belongs to that paradigm. We select the top-ranked paradigm to be the predicted inflectional class.

The score for each paradigm in the case of a word is calculated for all suffixes of the word, including the lexical properties, from shortest to longest. More formally, for all tags, the rank is calculated iteratively according to Formula 1.

$$rank^{i+1}[tag] = sign \times len\_sfx \times rel\_freq + rank^{i}[tag] \quad (1)$$

[2]n: neuter noun, *: undeclinable, s: singular only

where

$sign$	is negative if the suffix is shorter than the minimal suffix required by the paradigm
len_sfx	is the length of suffix w/o lexical properties
rel_freq	is the relative frequency of tag for the suffix
$rank^i[tag]$	is divided by len_sfx if $len\_sfx > 1$ is negated if $sign > 0$ and $rank^i[tag] < 0$ before calculating $rank^{i+1}[tag]$

The applied ranking score clearly prefers the most frequent paradigm for the longest matching suffix.

VI. Evaluation

Evaluation of the ranking algorithm was performed for the four different test sets described in Section II. These are rare words (LT10), average words (LT100), and frequent words (MT1000). We used standard evaluation metrics for measuring the performance of our method. *First-best accuracy* measures the ratio of having the correct paradigm ranked at first place. This reflects the ability of the system to automatically classify new words to paradigms. In addition, the accuracy values for 2^{nd} to 9^{th} ranks were also calculated. *Recall* is the ratio of having the correct paradigm in the set of the first ten highest ranked candidates. Following the metrics used by [4], precision was calculated as *average precision at maximum recall*, i.e. $1/(1+n)$ for each word, where n is the rank of the correct paradigm. This measures the performance of the ranking algorithm. As it might be the case that paradigm prediction is used to aid human classification, this metric reflects the ratio of noise a human must face with when verifying the results. Finally, *f-measure* is the harmonic mean of precision and recall.

In order to measure the advances in the performance, two baselines were created. The first one uses Brants' suffix guesser model ([3]) instead of the longest suffix matching method. This model uses a θ factor to combine tag probability estimates for endings of different length in order to get a smoothed estimate. θ is set as the standard deviation of the probabilites of tags. First, the probability distribution for all suffixes is generated from the training set, then it is smoothed by successive abstraction according to Formula 2.

$$P(t|l_{n-i+1},...l_n) = \frac{\hat{P}(t|l_{n-i+1},...l_n) + \theta_i P(t|l_{n-i},...l_n)}{1+\theta_i} \quad (2)$$

for $i = m...0$, with the initial setting $P(t) = \hat{P}$, where

$\hat{P}$	are maximum likelihood estimates from the frequencies in the lexicon
θ_i	weights are the standard deviation of the unconditioned maximum likelihood probabilities of the tags in the training set for all i

The other baseline assigns the most frequent paradigm identifier to each word based on its part of speech. The results of these baselines compared to our system are shown in Table II. As expected, the second baseline, choosing the most frequent tag, has a very low accuracy, however, our longest suffix method outperforms the first baseline as well. A key difference between the two models is that Brants' model assigns more weight to unconditioned tag distributions and ones conditioned on shorter suffixes than those conditioned on longer ones. This is just the other way round in the longest suffix algorithm.

Table II
First-best accuracy of full tags achieved by the longest suffix match algorithm, Brants' model, and by assigning the most frequent paradigm tag

	Longest suffix	Brants' model	Most frequent tag
LT10	0.9274	0.6269	0.3433
LT100	0.9174	0.6148	0.3386
MT1000	0.8087	0.5687	0.3287

Table III
Evaluation of each test set for the ranked results

	LT10		LT100		MT1000	
	full	equip	full	equip	full	equip
#1	0.8924	**0.9274**	0.8750	**0.9174**	0.7416	**0.8087**
#2	0.0614	0.2322	0.0685	0.2278	0.0684	0.2371
#3	0.0168	0.2090	0.0223	0.2201	0.0314	0.2435
#4	0.0057	0.1518	0.0078	0.1452	0.0168	0.1900
#5	0.0035	0.1692	0.0037	0.1723	0.0090	0.2165
#6	0.0015	0.1884	0.0019	0.1683	0.0083	0.1697
#7	0.0000	0.1871	0.0012	0.1836	0.0032	0.1562
#8	0.0005	0.1400	0.0011	0.1496	0.0043	0.1418
#9	0.0010	0.1095	0.0007	0.1573	0.0017	0.1078
precision	0.9329	0.9538	0.92195	0.9481	0.8067	0.8550
recall	0.9841	0.9876	0.9832	0.9875	0.8872	0.9158
f-measure	0.9578	**0.9704**	0.9516	0.9674	0.8450	0.8843

The tags assigned to paradigms and syntactic features define a very sophisticated classification of words. However, some of the features that distinguish two different paradigms are not relevant from the aspect of their inflectional behavior, such as the subtype of a non-inflecting adverb. Also some paradigm differences are irrelevant from the point of view of a pure lemmatization task, because they do not affect the set of word forms in the paradigm. To see how the algorithms perform in that task, equivalence classes of paradigms were generated, and a prediction was considered correct if the set of inflected forms generated by the predicted paradigm was identical to the set of word forms generated by the correct paradigm. Of the 2 767 different paradigms, 921 non-unique paradigms could be collapsed into 283 equivalence classes. Table III shows the results for each setup, where columns 'full' and 'equip' correspond to full tag and equivalence class evaluations respectively. Note that the values in the 'equip' columns do not sum up to 1, since, in many cases, two or more paradigms on the list of top-ranked paradigm candidates would generate the same set of inflected word forms for a given lexical item.

As the numbers show, our system performs best on rare words, while it achieved the worst results on very frequent words. This is not very surprising, as irregular words tend

Table IV
FIRST-BEST ACCURACY OF FULL TAG PREDICTION IN THE CASE OF ALL TYPES OF WORDS, NOUNS, VERBS AND ADJECTIVES

	ALL	NOUNS	VERBS	ADJECTIVES
LT10	0.9166	0.9547	0.8158	0.8665
LT100	0.9038	0.9489	0.8114	0.8381
MT1000	0.7678	0.8594	0.6884	0.5991

to be frequent words, while rare words have regular inflectional behavior. Correctly predicting the exact paradigm of an unknown personal pronoun or an irregular verb is indeed a rather difficult task. Since our aim was to extend existing morphological lexicons, and such resources already contain the most frequent words of the language, the results obtained for rare words are the ones which are relevant for our task.

Also note that beside similar recall values, precision and first-best accuracy are significantly higher when equivalent paradigms are collapsed. The prediction algorithm works reasonably well for extending resources for tasks that do not require full morphological analysis such as indexing for information retrieval or dictionary lookup.

Table IV shows the first-best accuracy results for all words, nouns, verbs and adjectives separately. In this table, instead of full tag agreement, only the paradigm identifiers were considered. The exact paradigm of verbs and adjectives turned out to be more difficult to guess than that of nouns, due to semantic factors and stress variation as explained in the next section of this paper.

VII. ERROR ANALYSIS

The most frequent confusions of the longest suffix algorithm for infrequent words are due to failure to correctly predict

- whether an adjective has synthetic comparative forms
- whether a *-ние*-final abstract noun has an alternative *-нье* spelling
- whether a noun has a second genitive form (used in partitive constructions)
- stress in past passive participles of certain verb classes (this results in an *е*∼*ё* contrast not normally reflected in orthography)
- whether an adjective has synthetic superlative forms
- stress in short and comparative forms of certain adjectives (this results in an *е*∼*ё* contrast not normally reflected in orthography)
- whether a non-inflecting noun can be interpreted as plural
- whether an imperfective verb has past passive participle forms
- optional stress variation across the paradigm
- whether an adjective has short predicative forms

Except for stress-related issues and semantically motivated or idiosyncratic defectivity, incorrect forms are very rarely predicted by the algorithm. Humans would probably make similar mistakes for words they do not know, especially if they do not know the meaning of the word either. The system sometimes highlights inconsistencies in the original aot data that even the authors, who are not native or even advanced speakers of Russian can identify as errors, e.g. that while the name of the energy company *Кубаньэнерго* is categorized as lexically non-plural, the similarly formed *Сахалинэнерго* does not have this property.

When looking at errors the algorithm makes when applied to frequent words, we find that the types of errors are similar. Nevertheless, failure to predict superlatives, comparatives, second genitives or special locative forms is much more prevalent for this data, as a much higher proportion of very frequent words have these "irregular" forms.

The most frequent errors of Brants' original suffix guesser algorithm, on the other hand, include absurd errors that would not be made even by beginning learners of Russian. This is due to overemphasizing distributions conditioned on shorter suffixes over those on longer ones. The top-ranked candidate paradigm is often totally inapplicable to words having the ending the given lexical item has, such as the paradigm of *-ский*-final adjectives to *-ный*-final ones (the most frequent error of that algorithm for infrequent words).

VIII. CONCLUSION

In this article, we presented and evaluated a suffix-trie-based supervised learning algorithm capable of predicting inflectional paradigms for words based on the ending of their lemma and some basic lexical properties. The algorithm can be used to automatically extend the vocabulary of computational morphologies lacking an independent rule component. The experiments were demonstrated for Russian, however, with minimal adaptation the tool can be used for any language provided there is a morphological resource available. Moreover, we assumed that a dictionary with some lexical features is also available, thus such features could be used for disambiguating paradigm candidates. The results showed that our method performs above 90% in all the different setups, achieving the best performance on relatively rare words, which are good candidates of being absent in the original lexicon.

We found that assigning more weight to distributions conditioned on longer suffixes than on shorter ones yields much better prediction performance, not only in terms of the number of exact predicted paradigm matches, but especially when taking into account what sorts of errors the system makes. While the baseline suffix guesser algorithm often proposes paradigms inapplicable to the given lexical item, our algorithm makes errors that arise due to the lack of lexical semantic information. Humans would make similar errors in similar situations.

REFERENCES

[1] A. V. Sokirko, "Morphological modules at the site www.aot.ru," in *Dialog'2004*, 2004.

[2] A. A. Zaliznyak, *Russian grammatical dictionary – Inflection*. Moskva: Russkij Jazyk, 1980.

[3] T. Brants, "Tnt - a statistical part-of-speech tagger," in *Proceedings of the Sixth Applied Natural Language Processing (ANLP-2000)*, Seattle, WA, 2000.

[4] K. Linden, "Entry generation by analogy – encoding new words for morphological lexicons," in *Journal Northern European Journal of Language Technology*, 2009, pp. 1–25.