Extraction and Analysis of Proper Nouns in Slovak Texts

Radovan Garabík
Ľ. Štúr Institute of Linguistics
Slovak Academy of Sciences
Bratislava, Slovakia
garabik@kassiopeia.juls.savba.sk

Radoslav Brída
Ľ. Štúr Institute of Linguistics
Slovak Academy of Sciences
Bratislava, Slovakia
brida@korpus.sk

1 Introduction

Unknown named entity recognition in inflected languages faces several specific problems – the first and foremost is that the entities themselves are inflected[1] (Dvonč et al., 1966) leading to a problem of identifying word forms as belonging to the same lexeme, and also the problem of finding correct lemma. In this article we analyse the distribution of word forms for proper nouns in Slovak and describe an algorithm for their automatic extraction and lemmatisation.

The task of lemmatisation and morphological annotation of flective (and more specifically, Slavic) languages is reasonably researched and developed (Hajič, 2004). Since we cannot expect a morphological database (data relating lemmata to inflected word forms and their grammatical tags) to cover all or almost all the words present in the corpus (*especially* proper names that keep appearing depending on who or what has become a hot topic in mass media), using a well tuned guesser can improve the accuracy of lemmatisation and tagging.

Common sense says that named entities (proper names in particular) behave differently from common names, which translated into information theory terms means that the information about whether a word is a proper name is not independent from the information about its morphology paradigm. This means we can use the information about proper names to decrease the entropy of inflections, which is good because it helps the guesser choose between the possible lemmata and morphological tags.

2 Datasets

We denote Levenshtein distance (Левенштейн, 1965) between two words l and w by $\rho(l, w)$. Since a typical Slovak noun has up to 12 different word forms (two numbers, six cases – the vocative is rare), and the inflection is mostly realized by changing the suffix and root vowel alteration, we can expect the overall distance between lemma and its word forms to be not only bounded from above, but also have a regular distribution (roughly speaking, the less typical the suffix length, the less likely is such a word form to appear).

We used the morphological database of Slovak language (Garabík and Šimková, 2012; Karčová, 2008; Garabík, 2007), which contains (at the time of writing) complete morphological information of 35 009 nouns (lemmata), out of which 1031 are proper nouns. We randomly divided the database into two parts, the training set and the evaluation set, ensuring that about 90% of both common and proper nouns is present in the training set. The evaluation set contained 101 lemmata and 694 unique word forms for proper nouns.

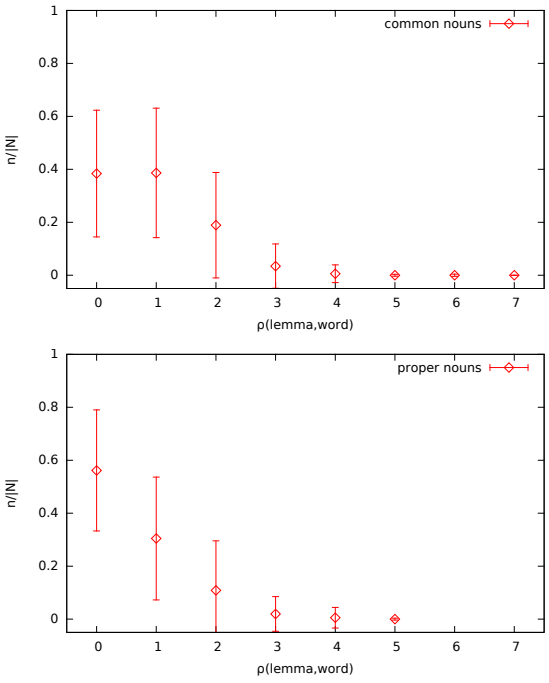

Figure 1: Distribution of word forms according to their Levenshtein distance from lemma.

[1] e.g. for the lemma *Galileo*, genitive would be *Galilea*, dative *Galileovi* etc.

In Vito Pirrelli, Claudia Marzi, Marcello Ferro (eds.): *Word Structure and Word Usage*. Proceedings of the NetWordS Final Conference, Pisa, March 30-April 1, 2015, published at http://ceur-ws.org

	†				‡		‡	‡	‡	†	†		
...	Toska	Toskala	Toskalu	Toskánec	Toskánska	Toskánske	Toskánsko	Toskánskom	Toskánsku	Toske	Tosky	Toso	...
	10	33	28	20	221	11	110	20	304	15	26	16	

Table 1: Alphabetic list of proper noun candidates, with number of occurances in the corpus. Note the extracted lemmata/lexemes $Toska^{\dagger}$ (La Tosca), $Toskánsko^{\ddagger}$ (Tuscany), as well as unrelated $Toskala$, $Toso$ and related $Toskánec$ (inhabitant of Tuscany) and $Toskánske$ (Tuscan, adjective).

Fig. 1 displays the distribution of known common (top) and proper (bottom) nouns, summed and normalized through all the nouns in the training set. Vertical error bars display the standard deviation for the given distance of word form from the lemma. From the graphs, we derive several conclusions – proper nouns are "less inflected", higher ratio of them is in the basic form (lemma), and the maximum distance is $\rho = 7$ for common nouns (nouns with greater distance are those with very irregular declension, e.g. *človek* → *ľudia* "human/humans") and $\rho = 5$ for proper nouns. Distributions of common and proper nouns from the evaluation set match those from the training set, so there appears to be some difference between common and proper nouns globally. However, categorising single nouns using these differencies between distributions is not reliable.

3 Extracting Candidates

Our algorithm extracts plausible candidates for proper nouns (those beginning with a capital letter but not at the beginning of a sentence, together with some additional filters) and for each candidate, it considers the set of words with $\rho \leq 5$. This would require calculating the Levenshtein distance between all pairs of words in the set and the complexity would be $O(n^2)$, which is unacceptable for corpus sized inputs. Unfortunately, Levenshtein distance is a metric but cannot be used to make an ordered set out of a list of words (in particular, it cannot be used to define an ordering binary relation $\leq$).

However, a trick can be applied – in a lexicographically ordered list of words (see Table 1) we need to look only at some interval around the word; word forms from beyond the interval are very unlikely to belong to the same lexeme. The complexity will be $O(Cn)$; where C is the (constant) size of the interval. This means that for some of the nouns not all word forms will be covered; especially for the shorter ones, where there is a higher probability that many unrelated words will be within the interval. Empirically we estimated the reasonable interval width to be 2000 words – increasing it above this number does not improve the accuracy anymore and the speed is acceptable. It should be noted that this interval is *not* a width of the context of the concordance – this is an interval in the lexicographically ordered set of proper noun candidates extracted from a given text, e.g. from a novel if we want to extract the whole inflectional paradigms of (new, unknown) proper nouns from the novel, or indeed from the whole corpus, if we aim to augment a morphological database.

We formally describe a Levenshtein edit operation $e = (o, i_s, i_d)$ – a triple of operation type o, position i_s in the source string s and position i_d in the destination string d, where operation type o is one of *replace*, *insert* or *delete*. For *replace* or *insert*, the replacement/new character is taken from the destination string d.

Sequence of edit operations $q = (e_1, e_2, e_3, ...)$, together with the destination string d, when applied to a string $s \in \mathbb{S}$ defines a mapping $f_{q,d} : \mathbb{S}_{q,d} \mapsto \mathbb{S}$, where $\mathbb{S}_{q,d}$ and $\mathbb{S}$ are sets of strings.[2]

If we denote by t a morphological tag for a given word form w, then for a lexeme with a lemma l a tuple (w, t) unambiguously refers to one inflected word form and its grammatical categories. We can then construct a sequence of edit operations leading from l to w, denoted by $q(l, t)$.

For each proper noun from the training set, we precompute the functions $f_{q(l,t),l}$ (this can be improved by dividing the nouns into categories based on their declension rules and using only one noun from each category), to get the sequence of operations leading from the lemma to the tuple (w, t) of the word form and morphological tag. Then, for each extracted word, we apply the functions $f_{q(l,t),l}$ to every word from the abovementioned interval and the word with greatest coverage (sum of the frequencies of generated word forms within the interval) is declared the lemma to the extracted word. Of course, this maximum can be attained by more than one word, especially if the lexeme is incom-

[2] It is not possible to define the function f for every source string, since some of the operations might not be applicable to the given strings.

plete. We assume that at least the most common inflectional paradigms (used for proper nouns) are present in the training set.

word forms	[%]	number of lemmata assigned per word form: correct	all
100	18.9	0	1
4	0.8	0	2
418	79.2	1	1
6	1.1	1	2
Σ 528	100.0		

Table 2: Number of automatically assigned lemmata per word form.

4 Evaluation

We used the algorithm to extract proper nouns from the Slovak National Corpus, version *prim-6.1-public-all*[3], of the size 829 million tokens, and evaluated the results on the proper nouns from the evaluation set. The percentage of correctly automatically assigned lemmata is shown in Table 2 – we see that 79.2% word forms had been assigned a unique lemma, which was also the correct one, while 18.9% had been assigned a unique, but incorrect lemma[4].

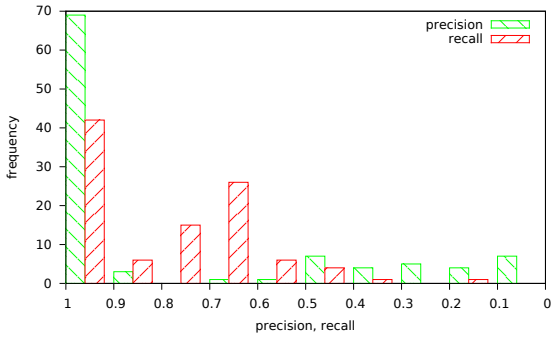

Figure 2: Histogram of precision and recall on automatically assigned word forms of the lexeme(s) for the evaluation data.

Figure 2 displays the precision and recall on word forms for proper nouns (i.e. how much of the lexeme has been extracted; the numbers are not weighted by the frequency of word forms in the corpus) from the evaluation set; we note that about 70 lexemes[5] have $precision \approx 1$; about 40 lexemes have $recall \approx 1$, and about 50 lexemes have $0.9 \gtrapprox recall \gtrapprox 0.6$, while only a small number of lexemes have lower precision. The lower recall is caused by insufficient data coverage – not all the word forms were present in the analysed corpus. The precision we obtained is excellent and the accuracy of automatic lemma assignment is good.

5 Augmenting Morphological Database

The abovementioned process was used to increase the number of proper nouns in Slovak morphological database. We used the extracted candidates from the *prim-6.1-public-all* corpus with a number of occurrences at least 100 (count of all possible word forms derived from a given lemma). We calculate the coverage of word forms for one lemma as $r = C(w,t)/C(g)$, where $C(w,t)$ is the number of generated tuples of word forms and their corresponding morphological tags, and $C(g)$ the number of grammar categories (usually 7 or 14; 7 cases including the vocative and one or two grammatical numbers, with many proper nouns present only in singular).

After removing generated word forms with no corpus evidence, the average coverage of word forms per lemma is $r = 0.84 \pm 0.23$, i.e. 84% of word forms is present in the corpus, 0.23 is the standard deviation of the coverage. Generated word forms still contain a lot of noise, therefore we also removed those word forms whose contribution to the number of occurrences of given lemma was less than 1% (it is rare for a grammatical case to have such a low percentage compared to other cases). After this, the coverage changed to $r = 0.75 \pm 0.24$, where again 0.24 is the standard deviation of the coverage. Then we manually proofread, corrected and filled in the word forms for the several hundred most frequent lemmata. After adding these words to the morphological database, we iterated the process, re-training the algorithm and generating another list of less frequent proper nouns.

6 Conclusion

The method has been used to improve the coverage of proper nouns in the Slovak morphological database and is used as a part of morphological guesser, providing candidate lemmata and morphological tags for unknown proper nouns, as part of the morphosyntactic analysis and part of speech tagging of the Slovak National Corpus.[6]

[3] http://korpus.juls.savba.sk/res.html

[4] For 13 word forms (2.5%) the correct lemma was not present in the interval of 2000 words.

[5] Since the number of proper nouns in our evaluation set was 101, these numbers are fortuitously almost identical to percentage.

[6] http://korpus.juls.savba.sk

References

[Левенштейн1965] Владимир Иосифович Левенштейн. 1965. Двоичные коды с исправлением выпадений, вставок и замещений символов. *Докл. АН СССР*, 4(163):845–848.

[Dvonč et al.1966] Ladislav Dvonč, Gejza Horák, František Miko, Jozef Mistrík, Ján Oravec, Jozef Ružička, and Milan Urbančok. 1966. *Morfológia slovenského jazyka*. Vydavateľstvo SAV, Bratislava, Slovakia, 1st edition. 895 p.

[Garabík and Šimková2012] Radovan Garabík and Mária Šimková. 2012. Slovak Morphosyntactic Tagset. *Journal of Language Modelling*, 0(1):41–63.

[Garabík2007] Radovan Garabík. 2007. Slovak morphology analyzer based on Levenshtein edit operations. In M. Laclavík, I. Budinská, and L. Hluchý, editors, *Proceedings of the WIKT'06 conference*, pages 2–5, Bratislava. Institute of Informatics SAS.

[Hajič2004] Jan Hajič. 2004. *Disambiguation of Rich Inflection (Computational Morphology of Czech)*. Karolinum, Charles Univeristy Press, Prague, Czech Republic.

[Karčová2008] Agáta Karčová. 2008. Príprava a uskutočňovanie projektu morfologického analyzátora. In Anna Gálisová and Alexandra Chomová, editors, *Varia. 15. Zborník materiálov z XV. kolokvia mladých jazykovedcov*, pages 286–292, Bratislava. Slovenská jazykovedná spoločnosť pri SAV – Katedra slovenského jazyka a literatúry FHV UMB v Banskej Bystrici.