Automatic Acquisition of a Slovak Lexicon from a Raw Corpus

Benoît Sagot

INRIA-Rocquencourt, Projet Atoll,
Domaine de Voluceau, Rocquencourt B.P. 105
78 153 Le Chesnay Cedex, France

Abstract. This paper presents an automatic methodology we used in an experiment to acquire a morphological lexicon for the Slovak language, and the lexicon we obtained. This methodology extends and refines approaches which have proven efficient, e.g. for the acquisition of French verbs or Croatian and Russian nouns, adjectives and verbs. It only relies on a raw corpus and on a morphological description of the language. The underlying idea is to build all possible lemmas that can explain all words found in the corpus, according to the morphological description, and to rank these hypothetical lemmas according to their likelihood given the corpus. Of course, hand-validation and iteration of the whole process is needed to achieve a high-quality lexicon, but the human involvement required is orders of magnitude lower than the cost of the fully manual development of such a resource. Moreover, this technique can be easily applied to other languages with a rich morphology that lack large-coverage lexical resources.

1 Introduction

Among the different resources that are needed for Natural Language Processing tasks, the lexicon plays a central role. It is for example a prerequisite to any wide-coverage parser. However, the development or enrichment of a large and precise lexicon, even restricted to morphological information, is a difficult task, in particular because of the huge amount of data that has to be collected. Therefore, most large-coverage morphological lexicons for NLP concern only a few languages, such as English. Moreover, these lexicons are usually the result of the careful work of human lexicographers who develop them manually over years, and for this reason they are often not freely available.

The aim of this paper is to show that this is not the only possible way to develop or enrich a morphological lexicon, and that this process can be automatized in such a way that the needed human labor is drastically reduced, at least for categories that have a rich morphology.[1] The only requirements are a

[1] We do not consider here closed classes like prepositions, pronouns or numerals, because they can be easily described manually and/or because they don't have a rich morphology, if any.

raw corpus and a morphological description of the language. This makes it possible to build morphological lexicon in relatively little time for languages that received less attention until now, for example because they are spoken by less people and/or because they are not supported by a large NLP community. We applied our methodology to Slovak language.

The idea of learning lexical information (with or without manual validation) is not new. It has been successfully applied, among other tasks, to terminology [1], collocations [2] or sub-categorization properties [3]. All assume the availability of a morphological lexicon. But to our knowledge, very few work has been published on automatic acquisition of morphological lexicons from raw corpus. Experiments have been conduced by Oliver and co-workers on Russian and Croatian [4, 5] to acquire or enlarge lexicons of verbs, nouns and adjectives. Independently, Clment, Sagot and Lang [6] have published the methodology they used to acquire a lexicon of French verbs. This paper is an extension of these methods for at least three reasons. First, we do not take into account only inflectional morphology, but also derivational morphology, which allows a better precision and recall as well as the acquisition of derivational relations in the lexicon. Second, we use a morphological description that is more powerful than the purely concatenative morphology used in previous works. Third, our algorithm relies on a very simple but rigorous probabilistic model.[2]

The main idea is that the acquisition of the lexicon of a corpus in a given language can be achieved by the iteration of a three-step loop:

1. Given the morphological description of the language, build all possible lemmas that can possibly explain the inflected forms found in the corpus,
2. Rank these possible lemmas according to their likelihood given the corpus,
3. Validate manually best ranked lemmas.

In the remainder of this paper, we will describe these steps, our morphological description of Slovak and our current results, with an emphasis on step 2.

2 Slovak morphology

Like most other Slavic languages, and contrary to English or French, Slovak is an inflected language. This means that nouns and adjectives (among others) are inflected according to their gender and number, but also to their grammatical function or to the preposition that governs them (case). This inflection is mostly realized by changing the ending of the word according to its inflectional class (or paradigm), but the stem itself can be affected. The latter occurs in particular for some feminine and neuter nouns in their genitive plural form. For example, *žena* ("woman"), in which the stem is *žen-*, has the genitive plural form *žien*.

2.1 Slovak language

The Slovak language is a Slavic (and therefore Indo-European) language that is the official language of the Slovak Republic. Its closest relative is the Czech

[2] This is already the case in [6], but the model presented here seems more convincing.

language. Both languages coexisted during a long period within former Czechoslovakia. For this reason, and because of the proximity of these languages, most Slovak understand Czech, and people wishing to learn "the" language spoken in Czechoslovakia learned Czech. Consequences of this are for example that Slovak language is under-represented among language manuals[3], and that it received less attention than other Slavic languages such as Czech or Russian. The only big project concerning Slovak in computational linguistics is the Slovak National Corpus [7], which is a highly valuable resource. Because we think that having only one resource for a given language is not necessarily satisfying, we decided not to use this resource and the information it contains. However, we of course intend in the near future to compare our lexicon to this corpus.

2.2 Description of Slovak morphology

As already mentioned in the Introduction, automatic lexical acquisition of a morphological lexicon from a raw corpus strongly relies on morphological knowledge. Moreover, this knowledge has to be represented and used in a symmetrical way, in the sense that we want to be able to *inflect* lemmas (associated with an inflectional class) but also to *ambiguously un-inflect* forms found in the corpus into all possible lemmas that might explain them. Moreover, the morphological description of the language must be written by hand, and therefore in a reasonable format. It must also be exploited in a very efficient way, since we want to deal with big corpus, and therefore with a big amount of hypothetical lemmas.

Our description of the Slovak morphology[4], inspired among others from [8] and validated by a native speaker of the language, is represented by an XML file that contains three main kinds of objects, that will be described successively:

1. letters, and the classes they belong to,
2. fusion rules that model the interaction of the final letters of a stem and the initial letters of an ending,
3. inflectional classes.

The list of letters deserves no special comment. We associate to each of these letters a list of the phonetic classes of the phoneme they denote. We use six classes: consonants, soft consonants, non-soft consonants, vowels, long vowels (including diphtongs), short vowels.

The second kind of information we described about Slovak morphology is a set of fusion patterns that describe the interaction between the final letters of a stem and the initial letters of an ending. This allows to model with a

[3] For example, and to our knowledge, Slovak is today the only official language of a European country for which no manual in French language is available.

[4] It is important to state here that this morphological description is not the main point of the paper. Any other morphological description, possibly better or more justified from a linguistic point of view, could be used. The only requirement is that this description must be able to give all inflected forms of a given lemma, as well as all possible lemmas having a given form in their inflectional paradigm.

reasonable amount of inflectional classes phenomena that can be explained by standard classes, provided fusion patterns are used. Let us take an example. If a stem ending in `t'` like *kost'* ("bone") gets an ending beginning in `i`, like in our example the ending `i` of the locative singular, then the result is not `-t'i-` (here **kost'i*) but `-ti-` (here *kosti*). Therefore, we can describe a pattern `t'i` → `ti`. An other example is the plural genitive *žien* of *žena* mentioned above: we decide that the ending is in this case `-`, and we add the following fusion pattern, where `\c` means "any letter of the class of consonants": `e\c-` → `ie\c`. We also defined the special operator `$` that means "end of word", and the special class `\*` that matches any letter. An example that uses these operators is the pair of patterns `bec\*` → `bc\*` and `bc\$` → `bec`, which allows to model the alternance between the `-bec` form of some stems when they get an empty ending and the `-bc` form of the same stems when the ending is non-empty (e.g., *vrabec*, *vrabca*,...). Both patterns are needed since we need our morphological description to be used in both directions: from a lemma to its forms, using the first rule, and from a form to its possible lemmas, using the second rule.

The third set of information we built is the set of inflectional classes. For each class, we list its name, the ending that has to be removed from the lemma to get the stem, and (when needed) a regular expression that stems using this inflectional class have to match. To exemplify the latter point, we say that verbs in `-at'`/`-iam`/`-ajú` (like *merat'*) have stems that must end with a soft consonant. Each inflectional class contains a set of inflected forms defined by their ending and a morphological tag, supplemented, if needed, by an other regular expression that the stem must match. This allows to merge into one inflectional class two paradigms that differ only on a few forms in a way that can be predicted from the stem.[5] Classes also contains derived lemmas, defined by their ending and their inflection class. For example, the inflectional class of regular `-at'` verbs have (for the moment) two possible derivations, namely the associated noun in `-anie` and the associated adjective in `-aný`.

3 Automatic acquisition of the lexicon

As mentioned in the introduction, we iterate a three-step loop as many times as wanted. The three steps are the generation and the inflection of all possible lemmas, the ranking of these possible lemmas, and a partial manual validation. Each step takes into account the information given by the manual validator during previous steps. We shall now describe these steps the one after the other. The probabilistic model we developed that underlies step 2 is described in the corresponding paragraph.

[5] For example, we have only one inflectional class for regular `-at'` verbs. The form-level regular expression checks if the last vowel of the stem is long or short, thus allowing to decide between the endings `-ám`, `-áš`,..., and the endings `-am`, `-aš`,.... Indeed, infinitive, participle and 3rd person plural endings are identical, as well as derived lemmas (see below).

3.1 Generation and inflection of all possible lemmas

For our experiments, we used a relatively small corpus of 150,000 words representing 20,000 different words. This corpus includes texts produced by the European Union (including the project of Constitutional Treaty) and free-of-use articles found on the Internet (both scientific and journalistic style are represented). The first step is to remove from the corpus all words that are present in a hand-crafted list of words belonging to closed classes (pronouns, some adverbs, prepositions, and so on).

After the extraction of the words of our corpus and their number of occurrences, we need to build all hypothetical lemmas that match the morphological description of Slovak language and have among their inflected form at least one word which is attested in the corpus. We then need to inflect these hypothetical lemmas to build all their inflected forms (we call "lemma" a canonical form with the name of its inflection class[6]). To achieve these goals, we developed a script that reads our morphological description and turns it into two programs. The first one can be seen as a non-deterministic morphological parser (or ambiguous lemmatizer), and the second one as an inflecter. In a few dozens of seconds, the first program generates 73,000 hypothetical lemmas out of the 20,000 different words of the corpus. These lemmas are then inflected by the second program in a few other dozens of seconds, thus generating more than 1,500,000 inflected forms associated with their lemma and a morphological tag.

3.2 Ranking possible lemmas

At this point, our goal is to rank the hypothetical lemmas we generated in such a way that the best ranked lemmas are (ideally) all correct, and the least ranked lemmas are all erroneous. Therefore, we need a way to model some kind of plausibility measure for each lemma. We have chosen to compute the likelihood of each lemma given the corpus. Since we do not have the required information to do so directly, we use a fix-point algorithm according to the following model.

We consider the following experiment: we choose at random in the corpus a token (i.e. one occurence of an inflected form, hereafter simply "form"). The probability to have chosen a given form f is $P(f) = occ(f)/n_{tot}$, where $occ(f)$ is the number of occurrences of f in the corpus and n_{tot} the total number of words, both known. Let us call $\mathcal{L}_f$ the set of all hypothetical lemmas that have f as one of their inflected forms. We denote by $P(l)$ the probability that the token we chose is an inflected form of l, and by $P(f|l)$ the probability to have chosen an occurrence of f given the fact that we have chosen an inflected form of the lemma l. We then have the following equality (for the first iteration of the fix-point algorithm, $P(l)$ is initialized to an arbitrary value P_0, typically 0.1):

$$P(f|l) = \frac{P(f) - \sum_{l' \in \mathcal{L}_f, l' \neq l} P(l')P(f|l')}{P(l)}.$$

[6] Hence, a same canonical form can come from several different lemmas, provided they do not belong to the same inflectional classes.

The Bayes formula allows us then to compute the probability to have chosen an inflected form of the lemma l given the fact that we have chosen an occurrence of f, i.e., the probability for f to come from the lemma l:

$$P(l|f) = \frac{P(l)P(f|l)}{\sum_{l' \in \mathcal{L}_f} P(l')P(f|l')}.$$

This gives directly the probability that we have chosen a form f coming from l:

$$P(f \wedge l) = P(f)P(f|l).$$

Let us define the probability $\Pi(l)$ (very different from $P(l)$) that l is a valid lemma.[7] We then introduce the *odd* of the lemma l, defined by

$$O_l = \frac{\Pi(l)}{1 - \Pi(l)}.$$

It is well known that the Bayes formula can be expressed as a formula on odds in the following way: learning a new information i (here, the fact that f is – or is not – attested in the corpus) multiplies the odd of the hypothesis "the lemma l is valid" by the odds ratio $OR_l(f)$ defined by:

$$OR_l(f) = \frac{P(i \text{ if } l \text{ is valid})}{P(i \text{ if } l \text{ is not valid})}.$$

This has to be done for each possible form of l, and not only its attested forms. If f is attested in the corpus, the previous formula becomes

$$OR_l(f) = \frac{\sum_{l \in \mathcal{L}_f} P(f \wedge l)}{\sum_{l' \in \mathcal{L}_f, l' \neq l} P(f \wedge l)}.$$

If it is not, we need to evaluate the probability of *not* finding the inflected form f given the corpus, both if l is and is not valid, since the odds ratio is the ratio between these two probabilities. But as can be easily seen, this ratio is equal to the probability of having not chosen the form f given the fact that we have chosen an inflected form of lemma l. To compute this, we use the probability that the chosen form ends with a given ending, given the inflection class of its lemma (this is done thanks to $P(f|l)$ and the related morphological information). For space reasons, we will not give the (simple) details of this computation.

Once having computed all odds ratios, we just need to assume that the original odd (knowing nothing about the corpus) of each lemma is $O_l^0 = 1$ (i.e., $\Pi^0(l) = 1/2$), except if it is an already validated lemma. We then have the odds of each lemma given the corpus by computing the product of O_l^0 by all odds ratios of the form $O_l(f)$, where f is an inflected form of l. These odds are in fact slightly modified, in order to take into account the presence of prefixes that

[7] Of course, if some lemmas have already been validated, e.g., if one starts from a non-empty lexicon, then $\Pi(l) = 1$ for all these lemmas.

are productive derivational morphology mechanisms. For example, the odds of lemmas `urobit'` and `robit'` (with their common inflectional class) are mutually augmented, in order to take into account the fact that they co-occur and the fact that `u-` is a valid prefix.

At this point, we can compute the probability $\Pi(l) = O_l/(1 + O_l)$ that l is valid. If we denote by $\mathcal{F}_l$ the set of all inflected forms of l, we can define the number of occurrences of l by $occ(l) = \sum_{f \in \mathcal{F}_l} occ(f).P(l|f)$. We then have a new way to compute $P(l)$, by saying that $P(l) = occ(l).\Pi(l)/n_{tot}$. The latter formula allows to iterate anew the whole computation, until convergence.

After the last iteration (in practice, we do 15 iterations), lemmas are ordered according to the probability that they are valid. Lemmas that have a probability equal to 1 are ordered according to $occ(l)$. When appropriate, we associate to lemmas their derived lemmas.

3.3 Manual validation

The manual validation process is performed on the ordered list of lemmas generated at the last step. The aim of this step is to classify the best-ranked lemmas in one of the following classes:

- valid lemmas, that are appended to the lexicon,
- erroneous lemmas generated by valid forms (i.e., by verbal, nominal or adjectival forms that have to be associated in the future to another lemma),
- erroneous lemmas generated by invalid forms (i.e., by forms that are either not verbal, nominal or adjectival, or that are misspelled; such forms have to be filtered out from the corpus during the next iteration of the complete process).

This manual validation step can be performed very quickly, and without any in depth linguistic knowledge. We asked a native speaker of Slovak, who has no scientific background in linguistics, to perform this task. The only preparation needed is to learn the names of the inflectional categories. Once several dozens or hundredths of lemmas are validated this way, the whole loop is started anew.

4 Results and perspectives

Using this method, and after a few iterations of the whole loop (including 2 hours only of cumulated validation time), we have acquired in a few hours only a lexicon of Slovak language containing approximately 2,000 lemmas generating more than 50,000 inflected forms (i.e., 26,000 different tokens[8]). These forms cover 74% of the attested forms of the corpus that have not been ruled out manually (like prepositions, adverbs, particles, pronouns, and so on). By construction, the precision is 100% since our lexicon is manually validated.[9]

[8] Indeed, a same token can be the inflected form of several lemmas, or more frequently several inflected forms of the same lemma but with different morphological tags.

[9] Figures given here concerns the current state of the lexicon. As said later on, we go on acquiring this lexicon, and these figures will be higer very soon.

While preliminary[10], these results are very promising, especially if the short validation time is taken into account. First, they show the feasibility of a process of automatic lexical acquisition, even on a relatively small corpus. This method only relies on the fact that Slovak has a rich morphology. Therefore, it can be applied easily to any language (or category in a language) for which one has a morphological module that can be used in both manners (from lemmas to forms and from forms to hypothetical lemmas). Second, they have led to a Slovak lexicon that will be made freely available on the internet in the near future, under a free-software license. While not yet wide-coverage, this lexicon is interesting for at least two reasons: it contains information on derivational morphology (prefixes, nominalizations and adjectivizations of verbs), and it contains real-life words found in the corpus that may be absent from standard dictionaries, as for example *korpusový*, adjectivization of *korpus* ("corpus").

Of course, we are still going on in the validation process and iteration of the whole loop. We also want to increase the size of our corpus, both to raise the precision of the process and to acquire a more varied lexicon.

Acknowledgment

We would like to thank very warmly Katarína Mať ašovičová, native speaker of Slovak, who has been our validator during the acquisition process described here.

References

1. Daille, B.: Morphological rule induction for terminology acquisition. In: Proceedings of the 18th International Conference on Computational Linguistics (COLING 2000), Saarbrucken, Germany (2000) 215–221
2. Dunning, T.: Accurate methods for the statistics of surprise and coincidence. Computational Linguistics **19** (1993) 61–74
3. Briscoe, T., Carroll, J.: Automatic extraction of subcategorization from corpora. In: Proceedings of the Fifth Conference on Applied Natural Language Processing, Washington, DC (1997)
4. Oliver, A., Castellón, I., Màrquez, L.: Use of internet for augmenting coverage in a lexical acquisition system from raw corpora: application to russian. In: IESL Workshop of RANLP '03, Bulgaria, Borovets, Bulgaria (2003)
5. Oliver, A., Tadić, M.: Enlarging the croatian morphological lexicon by automatic lexical acquisition from raw corpora. In: Proceedings of LREC'04, Lisbon, Portugal (2004) 1259–1262
6. Clément, L., Sagot, B., Lang, B.: Morphology based automatic acquisition of large-coverage lexica. In: Proceedings of LREC'04, Lisbon, Portugal (2004) 1841–1844
7. Jazykovedný ústav Ľ. Štúra SAV: Slovenský národný korpus (Slovak National Corpus). URL: http://korpus.juls.savba.sk (2004)
8. Pečiar, Š. *and others*: Pravidlá Slovenského Pravopisu. Vydavateľstvo Slovenskej Akadémie Vied, Bratislava (1970)

[10] In particular, the corpus we used could be much bigger. This should be the case in our future work on this topic.