Automatic Lexical Acquisition from Raw Corpora: An Application to Russian

Antoni Oliver
IN3
U. Oberta de Catalunya
aoliverg@uoc.edu

Irene Castellón
GRIAL Group
Lingüística General - UB
castel@lingua.fil.ub.es

Lluís Màrquez
TALP Research Center
LSI, UPC
lluism@lsi.upc.es

Abstract

This paper presents a methodology for the automatic acquisition of lexical and morpho–syntactic information from raw corpora. The system uses information about the inflectional morphology declared by rules and is based on the co–occurrence of different forms of the same paradigm in the corpus. A direct application of this methodology gives very poor precision rates due to rule interaction between paradigms. We present a rule analysis algorithm that solves this problem, giving quite better precision rates, although recall decreases dramatically. Finally, we investigate some techniques to raise the recall, achieving recall rates around 67% with a precision of 92%.

1 Introduction

The implementation of different NLP applications requires a lot of lexical information. In particular, the construction of word–form lists usually requires a lot of human effort. In this paper we present a method for the automatic acquisition of lexical and morpho–syntantic information from raw corpora. The main goal is to acquire a complete list of word–forms with the associated morpho–syntactic information, expressed with tags that follow the Multext recommendations (Véronis and Khouri, 1995; Erjavec, 2001) (e.g., form: *мостом*; lemma: *мост*; POS/TAG: *NCMSI*; all expressed as *мостом:мост:NCMSI*).

This methodology could be very useful for those languages that do not have such word–form lists available, and suitable to enrich existing word–form lists.

The knowledge about morphology is expressed by means of rules. Some previous works used a symilar formalism, for example (Sanfilippo, 1990) for English and Italian. Other works on slavic morphology are (Sheremetyeva et al., 1998; Mikheev and Liubushkina, 1995) for Russian, and (Tadić, 1994) for Croatian.

The system searches for co–occurrences of different forms of the same paradigm in the corpus to infer the associated lemma and morpho–syntactic characteristics. Experiments of this methodology were firstly carried out for Croatian (Oliver et al., 2002). In this paper we present, on the one hand, an extension for Russian that includes an algorithm for the analysis of the interaction between rules of different paradigms, and, on the other hand, a technique for improving recall.

Our task is similar to the learning of word–category for unknown words, as in (Mikheev, 1996a; Mikheev, 1996b). Nevertheless, in our methodology almost all words are unknown and more detailed morpho–syntactic information, as well as the lemma, is acquired.

This paper is organized as follows: section 2 presents the main components of the acqui-

sition system. In section 3 the acquisition methodology and its main drawbacks are explained. Section 4 deals with rule ambiguity and describes the rule analysis algorithm. Section 5 focuses some techniques for improving the acquisition process and in section 6 the results obtained in the experiments are presented. Section 7 concludes and outlines some lines of future research.

2 Components of the acquisition system

The system is composed by several analysis modules. First, non–inflectional and irregular forms are dealt with by means of a predefined list. In the second module, the decomposition of the rest of the forms is made by morphological rules. The source of word forms is a raw corpus.

2.1 Wordlists of non–inflectional categories and closed categories

The words that belong to non–inflectional and closed categories are excluded from the acquisition process. We have manually constructed lists of words of such categories. In table 1 we list the number of elements of these classes.

Category	Elements
pronouns	1,037
numerals	706
prepositions	123
conjunctions	87
interjections	185
particles	105
adverbs	1,389

Table 1: Number of elements of the list of non–inflectional and closed categories

As can be seen, adverbs are provisionally included in this list, but they cannot be considered as a closed class. In the near future, we are planning to include derivative rules for the most productive processes of adverb formation from other categories.

2.2 Irregular word list

Irregular words are also excluded from the acquisition process. These words are declared in a list that includes all forms with the associated lemma and morpho–syntactic information. This list is currently being developed by hand. Our criterion is to write all forms of the irregular words included in the 5,000 most frequent words (Sharov, 2001).

2.3 Morphological rules

Morphological rules are implemented following a morphological stripping formalism (Alshawi, 1992). These rules are converted into Perl regular expressions at running time. The rules are of the form `FE:LE:Desc`, where: `FE` stands for the form ending, `LE` for the lemma ending, and `Desc` for morphological description. For example, the generic rule

```
ом::NCMSI
```

can express the entry `мостом:мост:NCMSI`. As it can be observed, this rule expresses a null lemma ending. By using Perl regular expressions we can describe the lemma ending with more precision. An example can be seen in the following rule:

```
([^аяеэиыоуюькгхжшщчцй])ом:\1:NCMSI
```

where '^' means the complementary set of symbols written between square brackets and '\1' is a variable representing the symbol matched by the regular expression between brackets. Regular expressions allow to express other complex morphological phenomena, such as vowel alternation. For example, the rule:

```
ль([^аяеэиыоуюькгхжшщчцй])а:ле\1:NCMSG
```

can express an entry as `льва:лев:NCMSG`.

Table 2 shows the number of rules manually developed for each part of speech.

The rules have been developed following the most productive models in (Zaliznjak, 1977). The high number of rules corresponding to verbs is due to the fact that some forms, as participles, are declinable. The rules corresponding to declined forms are derived automatically from the rule that expresses the base

Category	Rules
Nouns	565
Adjectives	219
Verbs	12,038
TOTAL	12,822

Table 2: Number of rules developed for each category

form. Not all the rules will be used in the acquisition process. Those rules expressing alternative endings equal to other endings in the paradigm are left out. For example, the rule

```
([^аяеэиыоуюькгхжшщчцй])а:\1:NCMPN
```

that expresses the alternative plural nominative in *а* for the masculine nouns, is left out because the `FE` '*а*' is equal to that of the genitive singular.

In execution time, rules are transformed into Perl substitutions, for example the rule

```
([^кгхжшщчц])у:\1а:NCFSA
```

is transformed into the substitution

```
s/([^кгхжшщчц])у/\1а/
```

This expression means "change the final `у`, if the preceding letter is not on the list `{к,г,х,ж,ш,щ,ч,ц}` and replace this `у` by an `а`". This substitution allows the formation of the lemma of a feminine noun from its singular accusative.

2.4 Corpus

A 16,000,000 word corpus has been compiled from newspapers and literary texts[1]. The corpus has been automatically segmented into sentences and no other kind of linguistic information has been added.

3 Basic acquisition methodology

The acquisition methodology is described in figure 1. The corpus is seen as a list of all the word–forms appearing in it. For example, let us consider that the corpus

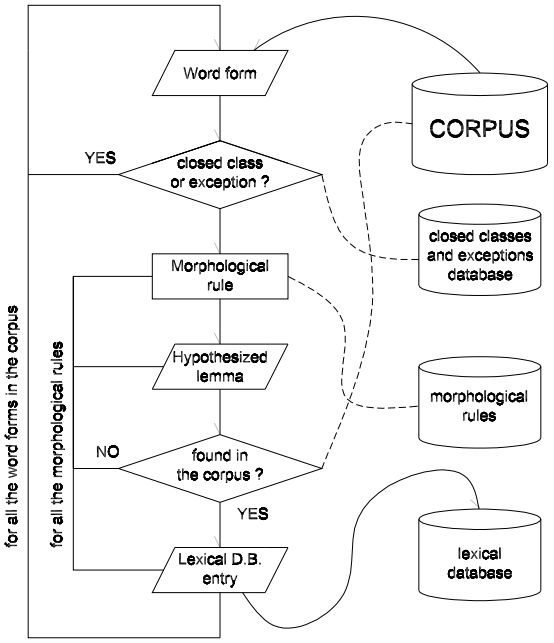

Figure 1: Basic acquisition methodology

is formed by all the word–forms of the paradigms of the lemma *мост*, that is: *мост, моста, мостах, мостам, мостами, мосте, мостом, мостов, мосту* and *мосты*. The algorithm takes one form, for example *мостом*, and verifies if the form is an exception or a word belonging to a closed class. In this case, it is neither one, so the algorithm goes to the next step and searches for an applicable morphological rule. In the example one applicable rule is

```
([^аяеэиыоуюькгхжшщчцй])ом:\1:NCMSI
```

For the word–form *мостом* the lemma *мост* is hypothesized by applying the rule. The algorithm then verifies if the hypothesized lemma appears in the corpus. If so, the algorithm constructs a lexical entry with the form, the associated lemma, and the morpho–syntactic information of the applied rule. In our example *мостом:мост:NCMSI*.

4 Rule ambiguity and rule analysis algorithm

In the example above everything worked fine because the hypothesized lemma is in fact a

[1]Thanks to Огонек, Правда and Библиотека Максима Мошкова for letting us use their texts

lemma, but a simple occurrence in the corpus does not assure that what the system hypothesizes is actually a lemma. It may be any other form of the paradigm. Let us consider one example of this situation using the word–form *мосту*. By applying the rule

```
([^кгхжшщчц])у:\1а:NCFSA
```

the algorithm hypothesizes the lemma *моста*, which exists in the corpus. However, this is not a lemma but the singular genitive of *мост*. Therefore, the algorithm wrongly creates a lexical entry: *мосту:моста:NCFSA*.

This fact shows us that interferences between the morphological rules exist and that rules cannot be straightforwardly applied in the acquisition process.

One simple approach is to set up a minimal number of rules that must be applied to validate a lemma. In other words, to set a minimum number of times the lemma is hypothesized. In the first prototypes for the Croatian language this number was empirically set to 3, achieving a precision of 88.9%. However, no figures of recall were available.

Another approach consists of previously identifying ambiguous and non–ambiguous rules in order to devise incremental application procedures that guarantee no interferences between rules.

All the rules with equal `FE` and `LE` (the rules that form the lemma) are ambiguous, because in the acquisition process the hypothesized lemma is equal to the form and it validates itself. For the rest of the rules the analysis is more complex because there is an interaction between endings and roots and because up to three paradigms can be implied in the ambiguity. For example, let us consider the following three rules:

```
ущего:ать:VIPP0SMGA,VIPP0SNGA
ать:атить:VIM02S0
рущего:ереть:VIPP0SMGA,VIPP0SNGA
```

And the following forms: *трущего* (present participle of *тереть*, singular genitive, masculine or neutre) and *тратъ* (imperative of *тратить*). The algorithm will wrongly acquire the entry:

**трущего:тратъ:VIPP0SMGA,VIPP0SNGA*

applying the first rule.

An efficient algorithm to analyze the rules and classify them as either ambiguous or non–ambiguous is necessary. First approach to this algorithm considered a list of embedded *if-then-else* conditions, with ad hoc conditions for the Russian language. This approach gave us good results, but nevertheless we preferred to develop a more general algorithm that could be applied to other languages (at least to languages with a concatenative morphology). The strategy of the work consists of converting the rule analysis into a controlled acquisition by using a pseudo–form–list. The comparison of the results of that acquisition with the pseudo–form–list lets us classify rules as either ambiguous or non–ambiguous. The full process is explained in the following 6 points:

1. *Rule expansion* Rules expressing morphological contexts are expanded to all their possibilities. For example:
   `([^аяеэиыоуюькгхжшщчцй])ом:\1:NCMSI`
   is expanded to the following rules:

   ```
   ъом:ъ:NCMSI   фом:ф:NCMSI   том:т:NCMSI
   сом:с:NCMSI   ром:р:NCMSI   пом:п:NCMSI
   ном:н:NCMSI   мом:м:NCMSI   лом:л:NCMSI
   зом:з:NCMSI   дом:д:NCMSI   вом:в:NCMSI
   бом:б:NCMSI
   ```

2. *Rule filtering with final letter n–grams* The expansion of the rules leads to a multiplication of the number of rules. Some of these rules cannot be applied because the combinations of final letters expressed by the rules are not found in the language (or at least are not found in the corpus). After rule expansion, rules are filtered using the set of letter *n*–grams extracted from the word–form endings in our corpus. By this process a large number of rules that cannot be applied are eliminated, and make the rest of the process faster.

In the filtering process of the example above, the rule ъом:ъ:NCMSI was eliminated because the ending ъом is not found in the corpus. No matter if the ending actually exists, as if it does not exist in our corpus the rule is not applicable.

3. *Creation of the pseudo–form–list* In this step a list of pseudo–word–forms is created with the associated pseudo–lemma and the morpho–syntactic information. This list has the form CODE+TF:CODE+TL:POS/TAG. The code is related to the part of speech. For example, following the same example above, the following pseudo–word–form list would be created:

NCMфом:NCMф:NCMSI	NCMтом:NCMт:NCMSI
NCMсом:NCMс:NCMSI	NCMром:NCMр:NCMSI
NCMпом:NCMп:NCMSI	NCMном:NCMн:NCMSI
NCMмом:NCMм:NCMSI	NCMлом:NCMл:NCMSI
NCMзом:NCMз:NCMSI	NCMдом:NCMд:NCMSI
NCMвом:NCMв:NCMSI	NCMбом:NCMб:NCMSI

4. *Acquisition process with the pseudo–forms* After the creation of the pseudo–form–list an acquisition process is performed with the pseudo–forms and the remaining rules. This acquisition process will give correct and incorrect results. An example of a correct acquisition is NCMтом:NCMт:NCMSI, and an example of an incorrect one is NCMту:NCMта:NCFSA. With each acquisition, the rule used to do it is stored. This information will be useful for the next step.

5. *Verification and detection of ambiguous rules* At this point, the result of the acquisition can be compared with the expected result (the pseudo–form–list itself). Since the rule used for each acquisition has been stored, those rules leading to acquisition errors can be now classified as ambiguous.

6. *Creation of the list of non–ambiguous rules* The rules not leading to any acquisition error will be classified as unambiguous.

5 Improved acquisition methodology

By the rule analysis algorithm it is possible to know which rules will lead to acquisition errors and which will not. The improved acquisition methodology consists of two steps:

- Acquisition with non–ambiguous rules. The lemmas acquired in this step will be correct with a high reliability.
- Acquisition with ambiguous rules and validation of the acquisition with the lemmas acquired in the first step.

In fact, for practical reasons, we made the two steps jointly in a single acquisition process with all the rules (storing the rule used for each acquisition). All the acquisitions made with unambiguous rules were validated and also those acquisitions made with ambiguous rules but with a lemma acquired with an unambiguous rule.

6 Experimental evaluation

A test corpus has been built only with regular and known forms, large enough and with a distribution of lemmas and forms as real as possible. All the forms are known so the result of each experiment can be evaluated automatically. The test corpus was built as follows: A word–form list was created with all the forms of 78,519 lemmas, totalling 1,247,202 forms. Each of these forms was included only if it occurs in our 16,000,000 word corpus of Russian texts. The result is a corpus of 232,770 regular word forms corresponding to 43,543 lemmas. This corpus is, in fact, a word–form list.

6.1 Results using all rules

The simple application of the acquisition process with no rule filtering gives the following figures of precision and recall:

Precision	34,65 %
Recall	85,25 %

As we have already mentioned, the low figures for precision are mainly due to the interaction of rules of different paradigms. The results

presented in this section, can be taken as baseline for precision and upper bound for recall in this experimental setting.

6.2 Results with rule analysis

Running the acquisition process after having applied the rule analysis algorithm gives the following results:

Precision	93.49 %
Recall	38.52 %

With rule analysis a significant improvement of the precision, but a very important decrease in recall have been obtained. The fall in recall is due to the high amount of ambiguous rules. These results can be taken as a baseline for recall and a upper bound for precision. In table 3 baseline and upper bound for precision and recall are summarized.

	Baseline	**Upper Bound**
Precision	34.65 %	93.49 %
Recall	38.52 %	85.25 %

Table 3: Baseline and upper bound for precision and recall in the experimental setting

The results given so far are calculated in an alphabetic basis, that is, all the process is done with all the words beginning with a given letter and repeated for each letter.

It is also interesting to observe, that the different parts of speech have different behaviours. The next figures show the percentages of nouns, adjectives and verbs in the corpus and in the result of the acquisition:

Category	**Corpus**	**Acquisition**
Nouns	46.67 %	18.53 %
Adjectives	28.38 %	9.75 %
Verbs	24.94 %	71.73 %

Table 4: Part of speech distribution in the acquisition

As we can see, the category of verb is the most acquired part of speech. This result can be explained by the fact that the verb is the part of speech which has more forms per lemma and also because nouns and adjectives have more ambiguous rules between them.

6.3 Improvement of recall

Some previous experiments showed that the smaller the corpus, the higher the recall. This is mainly due to the fact that the rule analysis uses the final n–grams of the word forms present in the corpus. If the corpus is small, the amount of ambiguous rules is small too, so the recall improves. To treat the whole corpus as a collection of smaller corpora where higher recall could be achieved, the acquisition process was repeated alphabetically in groups of two and three initial letters. The results are presented in table 5.

Using this approach, significant improvements in recall were obtained, with a slight drop in precision (p: 91.93%; r: 67.19%). We think that the still low results of recall can be explained mainly by two factors:

- A great amount of morphological rules are ambiguous for the acquisition task. The rule analysis in an alphabetic basis gives, on average, 32.96 % of unambiguous rules.
- Not all the forms of a paradigm are present in the corpus. Besides, for some paradigms, the lemma is not present. The methodology searches for the hypothesized lemma in the corpus, but if it is not present no acquisition will be made for the paradigm.

To evaluate the effect of the second cause one more acquisition experiment was performed with an artificially modified corpus. The modification consisted of the inclusion of all the lemmas of the forms present in the test corpus. The inclusion of all the lemmas gives a slight increase of recall of about 6 %, confirming that the main cause for the low recall is the great amount of ambiguous rules.

7 Conclusions and future work

We have presented a methodology for the automatic acquisition of lexical and morpho–

	alfab. 1 letter		alfab. 2 letters		alfab. 3 letters	
	Precision	Recall	Precision	Recall	Precision	Recall
Nouns	91.65	33.29	90.95	51.78	88.29	64.57
Adjectives	98.23	18.13	98.42	39.56	96.62	64.65
Verbs	93.77	73.98	93.65	74.75	93.59	75.46
Overall	93.49	38.52	93.38	53.65	**91.93**	**67.19**

Table 5: Results of the experiments for each part of speech

syntantic information. This methodology has been applied to a large corpus of Russian. Results for precision are quite good, but the main drawback of the methodology is a relatively low recall.

A new acquisition method that does not need the presence of the lemma in the corpus to validate the acquisition is currently under progress. This enhaced algorithm will acquire an entry even if the lemma is not present in the corpus. Also, to validate some of the acquisitions made with ambiguous rules, the context of occurrence in the corpus will be used. In some cases, the surrounding words can be tagged with the information acquired so far, and this information can be used to confirm a hypothesis or to reject it. This approach will be specially useful for nouns and adjectives, which are indeed the categories with lower recall.

In the experiments conducted so far, morphological rules have been written by hand based on traditional grammars as (Zaliznjak, 1977) for Russian and (Barić et al., 1995) for Croatian. We plan to develope some algorithms to learn the most productives paradigms from the raw corpus, as in (Goldsmith, 2001). These algorithms will allow to learn also the most productive derivative processes.

Future work includes testing these methodologies in other languages, such as Croatian, Spanish and Catalan.

Acknowledgments

This investigation is partially supported by the projects INTERLINGUA (Universitat Oberta de Catalunya and IN3 – IR266) and HERMES (TIC 2000–0335–C03–02)

References

H. Alshawi, editor. 1992. *The Core Language.* MIT Press.

E. Barić, M. Lončarić, D. Malić, S. Pavešić, M. Peti, V. Zečević, and M. Znika. 1995. *Hrvatska gramatika.* Školska Knjiga, Zagreb.

T. Erjavec. 2001. Specifications and notation for multext-east lexicon encoding. Technical report, Multext-East/Concede.(http://nl.ijs.si/MTE/V2).

J. Goldsmith. 2001. Unsupervised learning of the morphology of a natural language. *Computational Linguistics*, 27(2):153–198.

A. Mikheev and L. Liubushkina. 1995. Russian morphology: An engineering approach. *Natural Language Engineering*, 1(3):235–260.

A. Mikheev. 1996a. Unsupervised learning of word-category guessing rules. In *Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics (ACL-96), Santa Cruz, USA.*

A. Mikheev. 1996b. Learning part-of-speech guessing rules from lexicon: Extension to non-concatenative operations. In *Proceedings of the 16th International Conference on Computational Linguistics (COLING'96)*, pages 227–234, University of Copenhagen. Copenhagen. Denmark.

A. Oliver, I. Castellón, and L. Màrquez. 2002. Adquisición automática de información léxica y morfosintáctica a partir de corpus sin anotar: aplicación al serbo-croata y ruso. *Procesamiento del Lenguaje Natural*, 29:97–104.

A. Sanfilippo. 1990. Morphological analyzer for english and italian. Technical report, ESPRIT BRA 3030 Acquilex WP4.

S.A. Sharov. 2001. Chastotnyi slovar. www.artint.ru/projects/frqlist.asp.

S. Sheremetyeva, W. Jinm, and S. Niremburg. 1998. Rapid deployment morpology. *Machine Translation*, 13:239–268.

M. Tadić. 1994. *Računalna obradba morfologije hrvatskoga književnog jezika.* Ph.D. thesis, Sveučilište u Zagrebu, Filozofski fakultet. Zagreb.

J. Véronis and L. Khouri. 1995. Etiquetage grammatical multilingue: modèle. Technical report, MULTEXT Project, http://www.lpl.univ-aix.fr/projects/multext/LEX/LEX2.html.

A.A. Zaliznjak. 1977. *Grammaticheskii slovar russkogo jazyka. Slovoizmenenie.* Izdatelstvo "Russkii jazyk" Moskva.