Enlarging the Croatian Morphological Lexicon by Automatic Lexical Acquisition from Raw Corpora

Antoni Oliver*, **Marko Tadić****

*GRIAL Group - Universitat de Barcelona
G.V. Corts Catalanes 585 Barcelona Catalonia (Spain)
aoliverg@uoc.edu

**Filozofski fakultet - Sveučilište u Zagrebu
Ivana Lučića 3, 1000 Zagreb, Croatia
marko.tadic@ffzg.hr

Abstract

This paper presents experiments for enlarging the Croatian Morphological Lexicon by applying an automatic acquisition methodology. The basic sources of information for the system are a set of morphological rules and a raw corpus. The morphological rules have been automatically derived from the existing Croatian Morphological Lexicon and we have used in our experiments a subset of the Croatian National Corpus. The methodology has proved to be efficient for those languages that, like Croatian, present a rich and mainly concatenative morphology. This method can be applied for the creation of new resources, as well as in the enrichment of existing ones. We also present an extension of the system that uses automatic querying to Internet to acquire those entries for which we have not enough information in our corpus.

1. Introduction

There are several ways to deal with the complexity of the inflection of morphologically rich languages in NLP systems. One of possible solutions is to produce an inflectional lexicon covering at least 20,000-30,000 lemmas with their words-forms, preferably generated in an automatic way. This type of language resource could help in different treatments of corpora such as POS-tagging/lemmatising, lexicon building, information and term extraction, etc. The Croatian Morphological Lexicon (HML) (Tadić and Fulgosi, 2003) which has been collected at the Institute of linguistics, University of Zagreb has been filled up until now with more than 1.5 million word-forms generated from 12,000 nouns, 7,700 verbs and 5,500 adjectives. The HML is fully MULTEXT-East v 2.1 (Erjavec, 2001) conformant and thus can be treated with any tool that expects lexica in this format. In table 2 we can see an example of an entry of the Croatian Morphological Lexicon. The enlargement of the HML, as any other inflectional lexicon, is still a time and human-power consuming task, since there are no tools that help in expanding the HML automatically. For this reason we are testing an automatic lexical acquisition methodology that has been proven to be efficient for other highly inflective languages.

abeceda	abeceda	Ncfsn
abecede	abeceda	Ncfsg
abecedi	abeceda	Ncfsd
abecedu	abeceda	Ncfsa
abecedo	abeceda	Ncfsv
abecedi	abeceda	Ncfsl
abecedom	abeceda	Ncfsi
abecede	abeceda	Ncfpn
abeceda	abeceda	Ncfpg
abecedama	abeceda	Ncfpd
abecede	abeceda	Ncfpa
abecede	abeceda	Ncfpv
abecedama	abeceda	Ncfsl
abecedama	abeceda	Ncfsi

Table 1: Example of entries of the Croatian Morphological Lexicon

The automatic lexical acquisition methodology needs two basic sources of information: the morphological rules and a raw corpus. Morphological rules are expressed in a morphological stripping formalism and they are automatically derived from the existing Croatian Morphological Lexicon. Then, rules are grouped by paradigms and only those paradigms with a given degree of productivity, that is, with at least a given number of stems associated, will be used in the lexical acquisition process. The paradigms with less associated stems can be used to create a list of irregular word forms. This list of irregular words along with a list of words belonging to non-inflectional and closed categories are used to filter out these words from the acquisition process. The lexical acquisition methodology is based in the co-appearance of different word forms of the same paradigm in the corpus. We have also developed an extension of this system that uses queries to Internet to solve those cases that present ambiguity between two or more paradigms.

We first applied this methodology for Russian (Oliver et al., 2003) achieving a precision of 95.53% and a recall of 62.32% (F_1=75.43) without querying to Internet and a precision of 92.02% and a recall of 77.47% (F_1=84.12) using Internet to solve the ambiguous acquisitions. In the experiments for Russian, morphological rules were developed by hand and we used Internet only to discriminate between options in ambiguous acquisitions. In the experiments conducted for Croatian, we introduced two innovations: rules are automatically derived from the existing morphological lexicon and the queries to Internet are used to discriminate between options and to acquire new lexical entries. In this way we are able to acquire lexical information about word forms not present in the original corpus.

2. Acquisition system

2.1. Goal of the acquisition system

The main goal of the acquisition system is to automatically acquire a list, as complete as possible, of word forms with their associated lemma and morphosyntactic information. For example, word form:*ženom*; lemma: *žena*; Part-Of-Speech: *Ncfsi*[1]; all expressed as ženom:žena:Ncfsi.

2.2. Components of the acquisition system

2.2.1. 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 `om:a:Ncfsi` may express the entry ženom:žena:Ncfsi. In previous experiments for Russian rules were develope by hand following the most productive models in the Zalizjak's dictionary (Zaliznjak, 1977). In the experiments for Croatian rules have been automatically derived from the Croatian Morphological Lexicon using a very simple supervised algorithm. This algorithm works as follows:

1. The common substring beginning from the left between the word form and the lemma is calculated and considered as a stem. For example: word form: *ženom*; lemma: *žena*; stem: *žen*.
2. The form ending (`FE`) is obtained taking out the stem from the word form: ženom – žen = om.
3. The lemma ending (`LE`) is obtained taking out the stem from the lemma: žena – žen = a.
4. The morphological description (`Desc`) is known (Ncfsi in the example).
5. Now we have all the components (`TF:TL:Desc`) to write the rule: `om:a:Ncfsi`.

While the morphological rules are derived from the morphological lexicon they are also classified into paradigms, and for each paradigm we store the following information:

(1) a:ncf
(2) a:ncfpg;a:ncfsn;ama:ncfpd;...;om:ncfsi;u:ncfsa
(3) j:z:c:n:m:d:v:r:t:ž:l:b:č:ć:š:s:g:p:k:f:h
(4) 3796

1. `LE:POS`: Lemma ending and Part of Speech.
2. The list of `FE:Desc` (form endings and morphosyntactic descriptions), including the lemma ending. For superlative adjectives, formed by the prefix *naj* the form ending includes the prefix between brackets: `[naj]ija:axsfsnx`.
3. The list of possible morphological contexts, that is, the list of letters than can precede the endings.
4. The productivity of the paradigm, that is, the number of associated stems.

Some of the derived paradigms are classified as regular, following a criterion of productivity. In these experiments our criterion is to classify as regular the paradigms required to cover the 95% of the morphological lexicon if they have at least 5 associated stems. After the selection of the regular paradigms we expand them with the morphological contexts. In table 2 we can see the number of extracted, regular and regular after expansion paradigms for each Part of Speech.

POS	Extracted	Regular	Expanded R.
N	385	74	377
NCM	241	54	255
NCF	98	11	76
NCN	46	9	46
A	187	12	96
V	555	115	398
Total	1,127	201	871

Table 2: Number of paradigms for each Part of Speech: extracted, regular and regular after expansion.

2.2.2. Irregular word list

Irregular words are also excluded from the acquisition process. In previous experiments for Russian this list was developed by hand. Now we can automatically develope a list of irregular words with the paradigms considered as irregular.

2.2.3. Wordlists of non–inflectional categories and closed categories

The words belonging to non–inflectional and closed categories are excluded from the acquisition process. We have manually constructed lists of words of such categories.

2.2.4. Corpus

In our experiments we used a subsection of the Croatian National Corpus (HNK) of newspaper texts totalling 45 milion words. For the acquisition methodology the corpus is reduced to a list of the word forms appearing in it.

2.3. Acquisition methodology

The lexical acquisition methodology can be divided in 4 steps: splitting the word forms of the corpus and grouping by possible paradigms; acquisition by comparison between the paradigms and the groups of word forms; creation of the file of unsolved entries and querying to Internet to solve those entries for which we don't have enough information in the corpus.

2.3.1. Splitting word forms and grouping by possible paradigms

The algorithm splits all the word forms in the corpus (in stem and ending) by all possible endings and they are

[1] Singular Femenine Common Noun in Instrumental

grouped by paradigms. To explain this step, let us consider that the corpus is formed by the words: *most, mostom, mosta, mostu, mostima, ženi, ženom ženu and ženama*. If we split these words by all possible endings and group them, the following result is obtained:

NCMA:most	most,mostom,mosta,mostu,mostima
NCFA:mosta	most,mostom,mosta,mostu
NCMA:žen	ženi,ženom,ženu
NCFA:žena	ženi,ženom,ženu,ženama

2.3.2. Acquisition by comparison

Once we have all stems and endings grouped the acquisition process can start. For each word in the corpus all possible divisions are found and the system chooses as the correct one the division that has more forms in the corpus. For example, if we take the word form *mostom* we observe that can be associated with the group NCMA:*most* or with the group NCFA:*mosta*, but the first option has five associated forms and the second only four, so we choose the first option. With this information we can associate the word form *mostom* with the lemma *most* and we can retrieve the morphosyntactic information from the morphological rules. Then a new entry can be created: mostom:most:Ncmsi. In the same way the word form *ženom* can be associated with the group NCMA:*žen*, with three associated forms, or with the group NCFA:*žena* that has four associated forms. This last option is taken as the correct one. It is worth noting that is not necessary for the lemma to be present in the corpus. We can acquire an entry even if the lemma does not occur, as it is the case for the word form *ženom* in the example above.

2.3.3. Creation of the file of unsolved entries

In the examples above everything worked fine because there were present in the corpus some word forms with endings belonging to one paradigm but not to the other (*mostima* and *ženama*). Let us consider now that the corpus is formed by the word forms: *most, mostom, mosta, mostu, žena, ženom and ženu*. After splitting word forms and grouping by possible paradigms we get the following results:

NCMA:most	most,mostom,mosta,mostu
NCFA:mosta	most,mostom,mosta,mostu
NCMA:žen	ženi,ženom,ženu
NCFA:žena	ženi,ženom,ženu

If the system tries to acquire the lemma and the morphosyntactic information associated with the word form *mostom*, two possible options are found: NCMA:*most* and NCFA:*mosta*, both with four associated word forms. Thus it is impossible to determine the correct option using the method from 2.3.2. The same happens if we try to acquire the associated information of the word form *ženom* because we find the two tied options NCMA:*žen* and NCFA:*žena* with the same number of associated word forms.

In cases like these our system generates a file of unsolved entries. In this file all the options for each unsolved entry are specified. In the examples above this file would show the following information:

word form: mostom	word form: ženom
opt. 1: mos,NCMA,t	opt. 1: že,NCMA,n
opt. 2: mos,NCFA,ta	opt. 2: že,NCFA,na

For each option this file gives the associated stem, the flexive group and the lemma ending. This file of unsolved entries is interesting because we can use it in next steps to try to solve it with a bigger corpus, or by Internet querying.

2.3.4. Querying to Internet

The unsolved entries are mainly due to the lack of certain forms of the paradigm in the corpus. If we increase enough the size of the corpus we would find the needed word forms to solve the ambiguities. Existing corpora or Internet search engines can be used to verify the existence of such word forms that can help us to discriminate between the different options.

To determine the forms that discriminate between options we generate all the forms corresponding to each option. The discriminating forms will be those present in one model but not in the others. Following the example above, to discriminate the options of the word form *most* we would generate all the forms corresponding to the option *mos*:NCMA:*t* (*most, mosta, mostu, moste, mostu, mostom, mosti, mostima*) and all the forms corresponding to the option *mos*:NCFA:*ta* (*mosta, mostu, moste, mosti, mosto, mostama*). The algorithm then composes the Internet queries from the non common word forms among the different options, that is, in our example the queries are composed from the word form *mostima*, corresponding to the first option, and the word forms *mosto* and *mostama* corresponding to the second option. In the example it would generate the query 'mostima' and the query 'mosto||mostama' (the symbol || means OR). The first query would return a greater number of documents than the second, so we would validate the first option. In a similar way, to discriminate between the options of the word form *ženom* we generate the queries 'ženima' and 'ženo||ženama'. In this case the second query would return a greater number of documents, validating the že:NCFA:na option. In the case of verbs there are a lot of discriminating forms, so we limit the number of forms to make the query, otherwise we would generate errors in the Internet search engine.

3. 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. In the test corpus there are the word forms of the corpus that are also present in the Croatian Morphological Lexicon. In these experiment we are not dealing with proper nouns and we are not using the irregular words list; we only use the wordlists of non–inflectional categories and closed categories. From these experiment some information can not be acquired, namely: animate-inanimate category for masculine nouns, the distinction between qualificative and possessive adjectives and the definite-indefinite information for adjectives.

We present two sets of results. First results (see table 3) correspond to the basic acquisiton process without querying to the Internet. As we can observe best results of prescision are obtained for nouns (P: 95.43%), namely for femenine nouns (P: 98.42%). Best results of recall are obtained for verbs (R: 69.50). Worse results, both of precision and recall are obtained for adjectives (P: 76.27%; R: 20.79%) (only worse results of recall are obtained for neuter nouns (R: 10.88%).

	Precision	**Recall**	**F1**
N	95.43	43.99	60.22
NCM	91.16	57.67	70.64
NCF	98.42	45.42	62.16
NCN	94.80	10.88	19.52
A	76.27	20.79	32.68
V	89.38	69.50	78.20
Global	**86.13**	**35.36**	**50.14**

Table 3: Results of the basic acquisition process

In table 4 we can observe the results of the acquisition process including the queries to the Internet. As we can observe we obtained a globalincrease of only 3 points. This is mainly due to the very bad results obtained for masculine nouns and adjectives. For these categories we didn't improve the results of recall and results of precision get worse. But for the rest of categories we improve the recall with a slight decrease of precision. For example, for femenine nouns the recall increased 14.49 points with a decrease of precision of only 0.15 points.

	Precision	**Recall**	**F1**
N	94.66	52.65	67.67
NCM	87.42	57.67	64.49
NCF	98.57	59.91	74.52
NCN	95.58	13.71	23.99
A	72.05	20.79	32.27
V	88.66	74.07	80.71
Global	**84.50**	**38.36**	**52.76**

Table 4: Results of the acquisition process with queries to the Internet

In our experiments we used Internet only to discriminate between different options for those entries for which we do not have enough information in the corpus. The process implies to verify the existence of several word forms not existing in our corpus by querying to Internet engines. Once we verify the existence of these word forms we can create new entries and acquire more lexical information. We have evaluated this possibility but only for a subset of the corpus due to the big amount of queries to the Internet needed to perform the whole experiment. The experiments where performed for the wordforms beginning with letters *a* and *n* and we obtained an global improvement of 7.22 points in recall.

4. Conclusions and future work

Comparing the results with those obtained for Russian (Oliver et al., 2003) we can observe that for the basic acquisition methodology we obtain worse results for Croatian (9.5 points less of precission and 26.96 points less of recall). The main differences are in figures for adjectives (23.19 points less of precission and 44.03 points less of recall). The rules used in the experiments for Russian were developed by hand following the most productive paradigms of the Zaliznjak's dictionary (Zaliznjak, 1977). In the experiments conducted for Croatian the rules were automatically extracted from the Croatian Morphological lexicon.

Obviously the Croatian adjectival system is either more complex than Russian or is the MSD system used here too complex for this kind of machine learning approach. The extremly large internal homography between different word-forms of Croatian adjectives – up to 22 different MSD readings in certain cases of the same phonological string accompained by the large number of regular word-forms (usually over 200) for each adjectival lemma – can be one of the reasons for the drop in the precision. One of possible ways to improve that would be the development of rules for adjectives by hand as they were developed for Russian in the previous experiment..

We could try a different approach in the future. Since the HML has been generated from the list of lemmas accompanied only by their stems and numbers of inflectional patterns, we could try to learn from that information. Given the association of the number of inflectional pattern and stem from the list of lemmas and generated word-forms from the same lemma, the system could induce the association between word-forms and number of inflectional pattern. This association can then be checked against Internet data.

Acknowledgments

This investigation is partially supported by the project SENSEM (ref: BFF2003-06456) and the i-project MZT 2002-100.

5. References

Alshawi, H. (ed.), 1992. *The Core Language Engine*. MIT Press.

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

Oliver, A., I. Castellón, and L. Màrquez, 2003. Use of internet for augmenting coverage in a lexical acquisition system from raw corpora. In *Workshop IESL*. RANLP 2003. Bulgaria.

Tadić, M. and S. Fulgosi, 2003. Building the croatian morphological lexicon. In *Workshop on Morphological Processing of Slavic Languages*. Association for Computational Linguistics. 10th Conference of The European Chapter of the EACL.

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