Generating a Morphological Lexicon of Organization Entity Names

Nikola Ljubešić, Tomislava Lauc, Damir Boras

Faculty of Humanities and Social Sciences
Ivana Lučića 3, Zagreb, Croatia
{nljubesi,tlauc,dboras}@ffzg.hr

Abstract

This paper describes methods used for generating a morphological lexicon of organization entity names in Croatian. This resource is intended for two primary tasks: template-based natural language generation and named entity identification. The main problems concerning the lexicon generation are high level of inflection in Croatian and low linguistic quality of the primary resource containing named entities in normal form. The problem is divided into two subproblems concerning single-word and multi-word expressions. The single-word problem is solved by training a supervised learning algorithm called linear successive abstraction. With existing common language morphological resources and two simple hand-crafted rules backing up the algorithm, accuracy of 98.70% on the test set is achieved. The multi-word problem is solved through a semi-automated process for multi-word entities occurring in the first 10,000 named entities. The generated multi-word lexicon will be used for natural language generation only while named entity identification will be solved algorithmically in forthcoming research. The single-word lexicon is capable of handling both tasks.

1. Introduction

Information extraction (IE) is the task of deriving structured factual information from the text (Gaizauskas and Wilks, 1998). Natural language generation (NLG) is the task of generating natural language utterances from structured information representing the reverse process to information extraction.

One of information extraction tasks deals with identifying names of entities in unstructured or partially structured text. This task is called named entity identification (NEI). It is a necessary step in determining the relationships between entities and attributes of interest (in case of organization, named entity attributes like address, number of employees, solvency, capital value etc.) as well as between entities (e.g. joint business activity) which is called named entity recognition (NER). NER locates and classifies atomic elements in the text into predefined categories such as names of persons, organizations, locations, etc. (Nadeau and Sekine, 2007).

There are two main research approaches in the field of NER: the approach based on stochastic methods and deterministic approach. In stochastic approaches the named entity models are trained on a large amount of manually annotated data. The disadvantage of this approach is the acquisition bottleneck, i.e. the need for large amounts of manually annotated data. Deterministic methods consist of manually crafted rules mainly written in form of regular expressions, i.e. finite-state automata and transducers (Bekavac and Tadić, 2007). The disadvantage of this approach is the complexity of producing hand-crafted rules that require a full understanding of the problem for a given language. The challenge for empirical methods in NLP is to continue to match the demand for automatization by developing additional natural language learning techniques. These techniques replace manual coding efforts with automatically trainable components that make it increasingly faster and easier to build accurate and robust information extraction systems in new domains or languages (Cardie, 1997). NEI is still most often based on deterministic methods combining lists of named entities with finite state automata, i.e. transducers.

Considering the NLG task, there are two main approaches: the template-based approach that maps its non-linguistic input directly to the linguistic surface structure and the real, or standard approach that uses less direct mapping between the input and the surface form. The template-based approach is used more often since it is simpler and generates more accurate, but less diverse results. It is based on predefined templates with gaps that are filled based on database information (Reiter and Dale, 2000).

When dealing with highly inflected languages such as Croatian, tasks relating to NEI and template-based NLG become more complicated (Bekavac and Tadić, 2007). In the NEI task all possible forms for a given named entity have to be known for them to be recognized in text. In the template-based NLG task a specific gap expects a specific lexeme form. The most common approach comprises building a lexicon that contain all forms of required lexemes. Such lexicons are commonly called morphological lexicons (Tadić and Fulgosi, 2003). In general, the morphological lexicon generation task deals with creating a database that associates an inflected word form to a set of tuples containing a lemma and a feature set for identification and a lemma to a set of tuples containing an inflected word form and a feature set for generation (Kržak and Boras, 1985).

The rich nominal inflection in Croatian includes seven cases, which results in 14 suffixes concerning singular and plural. In poorly inflected languages such as English, NLP tasks are usually backed up by simple morphological normalizers. In cases of Slavic languages like Croatian there are often no freely available resources that could provide a morphological treatment of named entities because of their complexity.

An additional problem with highly inflected languages is that there is often no agreement between prescriptive linguists concerning the way specific named entities are inflected and that prescriptive work in general is not unique yielding in various rules being applied to same problems.

Therefore, there are many ways in which some named entities are inflected in practice.
In this paper we describe our approach in generating the morphological lexicon of organization entity names which will be used for business news text generation and NEI for information extraction and business intelligence.

2. Analyzing the problem

The aim of this research is to generate a language resource - a morphological lexicon of organization entity names. The primary resource is a list of entities which consists of 263,772 organization entity names given in basic form (nominative case). This list is obtained from the Institute for Business Intelligence (ZAPI, 2008). The named entities in the list are ranked by relevance that is calculated through criteria like number of employees, frequency of occurrence in corpora, business performance etc.
The list contains single-word and multi-word named entities. In this research they are discussed separately because of different approaches in generating their word forms. Additionally, there is a third group of named entities containing the "dash" character whose use is not persistent in differentiating between a dash and a hyphen.
The use of the final resource is twofold:

1. template-based NLG, i.e. business news generation from database information
2. NEI for information extraction, i.e. business intelligence.

One of the main problems of these tasks is the inflectional complexity of the Croatian language. Namely, the nominal inflection in Croatian has seven cases and two genders. An additional problem for NEI is the fact that there is more than one way to inflect an organization entity name. For all occurrences of a named entity to be identified, the resource has to include all possible forms from all possible paradigms. For NLG the most preferred inflectional paradigm is used.
One of the reasons why most organization entity names can be inflected with more than one paradigm is because it is difficult to distinguish between an acronym and a proper noun. It is because some acronyms eventually become treated as proper nouns (e.g. "INA", "INA-e" or "Ina", "Ine"). Furthermore, grammar rules are often disregarded in everyday use, especially when they are related to named entities which consist of non-Croatian words or words of non-Croatian origin (e.g. "Techware" inflected like "Techwarea", "Techware-a" and "Techwara" where last two are wrong). The rule of thumb is that these words should be inflected as Croatian words, but there are many special cases and additionally prescriptive grammar rules can be quite vague in some cases.
The general grammar rule says that foreign names written in Latin alphabet are written originally but it is worth only for the nominative case. Since foreign personal names are inflected using Croatian inflectional paradigms there are mixed attributes concerning Croatian and other languages.
A few rules are provided here (Babić et al., 2002) to picture the complexity of the task of assigning an inflectional paradigm to a specific lemma, especially if taking into account differences towards everyday use:

- foreign names ending with nonaccentuated "-o" are inflected as Croatian names, where "o" is removed (e.g. "Crosco", "Crosca" and "Meiso", "Meisa")
- foreign names ending with accentuated "-o" are inflected as Croatian names without deletion and if the accentuation data is not present, both, this and the previous rule have to be applied (e.g. "Meiso", "Meisoa" and "Meiso", "Meisa")
- foreign names (masculinum) ending in unspoken "-e", do keep that "e" in the inflection (e.g. "Trade", "Tradea", and "Commerce", "Commercea") but in everyday use "Commerce" is more often inflected as "Commerca" than "Commercea"
- foreign names ending with "-i", or "-y" get the consonant "j" inserted between two vocals in the inflected form ("Dioki", "Diokija" and "Sony", "Sonyja") but in everyday use "Sonya" or "Sony-a" is as frequent as "Sonyja"
- in Romance personal name ending with "-ca " that are spoken as "-ka", "c" changes to "k" ("Veronica", "Veronike" and "Propublica", "Propublike") but this rule is almost never applied in everyday use

Generating all possible forms in case of single-word named entities is rather feasible ("CROSCO", "CROSCO-a" or "Crosco", "Crosca" or "Crosco", "Croscoa"), but in case of multi-word named entities this does not apply. If a multi-word named entity has two tokens that can be inflected in more than one way, the final list of possible word forms is the Cartesian product of the paradigm sets. If additionally the variable word order is taken into account, the task of generating all possible word forms for multi-word named entities becomes almost impossible.
That is why in case of single-word entities a supervised learning approach is used. The first 4,987 named entities are tagged manually and a statistical linear interpolation model is trained. All remaining named entities are tagged automatically which leaves the task of generating the lexicon to a trivial generation algorithm. In case of multi-word entities, the first 5,013 named entities are tagged automatically, retagged manually, generated automatically and checked manually in this four-step process.
Before solving the single-word and multi-word problems the "dash" problem is solved by manually correcting the 1,628 in the first 10,000 named entities that have the "dash" character. The problem with the "dash" character lies in its inconsistent use (e.g. "ŽUPANIJA KRAPINSKO - ZAGORSKA-ŽUPAN" should be written as "ŽUPANIJA KRAPINSKO-ZAGORSKA - ŽUPAN" if not "Župan Županije krapinsko-zagorske"). This manual correction is undertaken only on the first 10,000 named entities because only these are likely to be used in the NLG task. For NEI, multi-word named entities are not generated because of the reasons stated before. Possible single-word named entities with the dash character under the rank 10,000 are handled as multi-word named entities.

3. Single-word named entities

As stated before, the single-word problem is solved with a supervised learning approach. The first 4,987 single-word named entities are tagged manually with one or more of the 39 possible paradigms. During that process 68 paradigm combinations are defined. The first assigned paradigm is the preferred one which is very important for NLG since in NLG only that paradigm is used. Since the information about the preferred paradigm is irrelevant for the rest of the named entities, the 68 paradigms are reduced to 59 (the category "51" with the paradigm "5" as the preferred one and "1" as the non-preferred equals the category "15").

Most models for classifying lexemes into morphological categories are supervised and are based on n-grams. Because of the sparse data problem in natural language processing there is a need for combining evidence from different size n-grams. Two basic techniques are commonly used: linear interpolation and smoothing by redistributing a part of the probability mass to unseen n-grams (Dagan et al., 1997). The method applied in this research uses linear interpolation. It was introduced in (Samuelsson, 1996) and used in (Brants, 2000) and is called linear successive abstraction. The values calculated in the model are conditional probabilities of a specific tag t given the last m letters of an n letter word. The algorithm combines that conditional probability $P(t|l_{n-m+1}...l_n)$ (shorter $P(t|l_{o+1}...l_n)$) with the conditional probabilites of more general contexts $P(t|l_{n-m+2}...l_n), P(t|l_{n-m+3}...l_n), ...P(t)$ (shorter $P(t|l_{o+2}...l_n), P(t|l_{o+3}...l_n)$...). The recursion formula is

$$P(t|l_{o+1}...l_n) = \frac{\hat{P}(t|l_{o+1}...l_n) + \Theta_i P(t|l_{o+2}...l_n)}{1 + \Theta_i} \quad (1)$$

for $o = n + i$ and $i = m...0$ using the maximum likelihood estimates $\hat{P}$ from frequencies in the training set, weights Θ_i and the initialization

$$P(t) = \hat{P}(t) \quad (2)$$

The maximum likelihood estimate for a suffix of length i is derived from the training set by

$$\hat{P}(t|l_{n-i+1}, ...l_n) = \frac{C(t, l_{n-i+1}, ...l_n)}{C(l_{n-i+1}, ...l_n)} \quad (3)$$

where $C()$ is the count function.

The weights proven to get best results are standard deviations of unconditioned maximum likelihood estimates of n-grams in the training set (Samuelsson, 1996) by

$$\Theta_i = \sqrt{\frac{1}{N}\sum_{j=1}^{N}(\hat{P}(l_{n-i+1}, ...l_n) - \bar{P})^2} \quad (4)$$

$$\bar{P} = \frac{1}{N}\sum_{j=1}^{N}\hat{P}(l_{n-i+1}, ...l_n) \quad (5)$$

Besides the model, morphological lexica of general language, personal names and settlements are used in the decision process. For 33.20% of all named entities longer than three characters (e.g. "Zvijezda") and 51.19% of latter parts of named entities not ending on "e", "i" or "u" (e.g. "Tehnocentar", "centar"), entries are found and these named entities are inflected like the lexemes in the lexica. The above decisions are based on manual observation of possible results of the method. Such decisions result in absolute precision of the results.

At this point named entities found in lexica are removed from the dataset which shrinks down to 2,410 data points. The remaining dataset is divided into a 9/10 training and validation set and a 1/10 test set (2,169, ie. 241 data points). Since the parameter m has to be empirically tuned, holdout validation with 100 iterations is used to estimate the parameter value more accurately.

The loss function used is

$$L(P_a, P_m) = 1 - \frac{C(P_a \cap P_m)}{C(P_m)} \quad (6)$$

where P_a are the paradigms assigned by the model, P_m the manually assigned paradigms and $C()$ the count function. Two examples of the loss function would be $L('15','159') = 0.33$ and $L('15','1') = 0.0$. Overgeneration is not penalized since this data will be used for NEI only and all the 59 categories consist of paradigms that correspond in the category of number and gender (1, 5 and 9 are all masculine, singular) which makes an increase in the probability of a homonymy clash with another lexeme very low. It should be stressed that in this research accuracy equals recall while precision is neglected because the overgeneration problem is disregarded.

At this point most frequent errors are manually checked and, since overgeneration is disregarded, two general expansion rules are introduced:

1. mutually expand paradigms 1 ('TVIN', 'TVIN-a') and 5 ('Tvin', 'Tvina')
2. mutually expand paradigms 2 ('INA', 'INA-e') and 3 ('Ina', 'Ine')

For example, by these rules the class "5" is expanded to class "15" and class "2" to class "23".

The basic task is to find the optimal value of the parameter m which determines the length of the longest n-gram observed. As mentioned before, in this research accuracy is identical to recall. Accuracy of the holdout validation process regarding the value of m without using the lexicon or applying the expansion rules is maximal for $m = (2, 3, 4)$ reaching values of $a = (0.9469, 0.9458, 0.9465)$. The obtained data shows that a lot of information is contained in just the last character of a word (for $m = 1$, $a = 0.9125$). When unigram and digram information is combined, there is a low increase in accuracy (3.77%) in respect to using just the unigram information. In case of m greater than 2, no significant advance is obtained.

Table 1 contains the accuracy measures regarding the value of m with expansion rules applied. The increase in accuracy regarding $m = 1$ and $m = 2$ in this case is even smaller (1.51%). The maximum accuracy is obtained with $m = 3$. Therefore, the value of m is set to 3.

m	accuracy
1	0.9526
2	0.9670
3	**0.9688**
4	0.9674
5	0.9653
6	0.9678
7	0.9671
8	0.9692
9	0.9647
10	0.9624

Table 1: Accuracy regarding *m* with rules applied

Figure 1 depicts accuracy regarding the size of the training dataset. Accuracy behaves typically log-linear and increases rapidly up to the training set size of 700 points after which the increase starts to drop gradually. At the dataset size used in this research (1,952 for training during validation) the slope is still positive which indicates that a larger training set could provide even better results.

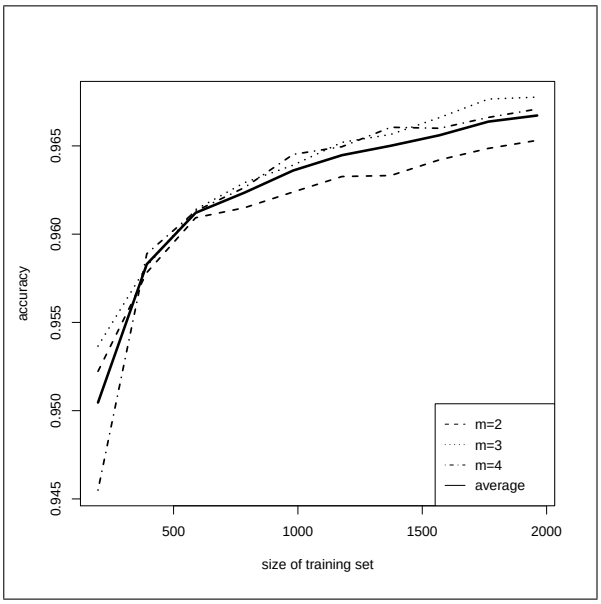

Figure 1: Accuracy regarding the size of the training set

For $m = 3$ and with expansion rules applied, accuracy is 96.88%. In case of using morphological lexica, accuracy rises up to 98.54% since 2,321 of the data points in the training set are found in the lexicon ($(2169 * 0.9688 + 2321)/4488$). When the method is applied on the test set, an accuracy of 97.33% is achieved on entities not found in the lexicon. With 51.19% of entities found in lexica, the final accuracy of the method on the test set is 98.70%.

4. Multi-word named entities

For purpose of creating this lexicon only the first 5,013 multi-word named entities are generated. All other 192,667 are not generated because of three reasons stated previously:

- there is no fully automated way of generating word forms like in the case of single-word entities
- it is not possible to generate all possible forms of multi-word entities as they can occur in text because of more possible paradigms for some tokens and variable word order
- only the first 10,000 named entities are expected to be used in text generation (5,013 multi-word entities) whilst NEI of multi-word entities will be solved on-the-fly

For every multi-word named entity (except in case of indeclinabilia) there is a noun phrase in nominative that is inflected. The rest of the named entity remains unchanged (e.g. "Ministarstvo vanjskih poslova Republike Hrvatske", "Ministarstva vanjskih poslova Republike Hrvatske"). In some cases the multi-word entity consists only of the inflected noun phrase (e.g. "Filozofski fakultet", "Filozofskog fakulteta").
The process of generating all forms for the first 5,013 multi-word named entities is implemented in four steps:

1. automated tagging of inflected noun phrases with help of existing morphological resources
2. manual correction of the tagging process with additional tagging concerning the letter case and paradigm categories for unknown tokens
3. automated generation of word forms
4. manual correction of the generated output

The automated tagging of inflected noun phrases is done in the manner that tokens are tagged with tags "a", "n" or "b" regarding the possibility of the token being an adjective ("a"), a noun ("n") or an adjective or noun ("b" as both") in nominative, singular or plural. After the tagging the first pattern matching the python-like regular expression `r'[ab]*n'` is tagged as the inflected noun phrase. After the fourth step, this method proves to be in 1,315 cases (54.93%) completely accurate for named entities whose normal form is not changed by human annotators (2,394, ie. 47.76%) and partially accurate (finding at least part of the noun phrase without tokens outside the noun phrase) in 2,100 cases (87.72%).
While manually correcting the output of the tagger human annotators changed the basic form of the named entity in 2,619 cases (52.24%) (e.g. "ZADARSKA ŽUPANIJA ŽUPANIJSKO POGLAVARSTVO" into "ŽUPANIJSKO POGLAVARSTVO ZADARSKE ŽUPANIJE") and annotated their letter case with following tags:

- for whole named entities
  - a - just first token title case
  - b - first and last token title case
  - c - all tokens title case
  - d - like the original

- for a specific token in the named entity
  - 0 - title case
  - 1 - lower case
  - 2 - upper case
  - 3 - like the original

Through this method "HRVATSKE ŠUME" is tagged with the tag "a" which generates the output "Hrvatske šume" and "ŽUPANIJSKO POGLAVARSTVO ZADARSKE ŽUPANJIJE" with the tag sequence "0101" generating the output "Županijsko poglavarstvo Zadarske županije". Human annotators also correct the index range pointing to the location of the inflected noun phrase ("ŽUPANIJSKO POGLAVARSTVO ZADARSKE ŽUPANIJE" has the human readable index range "12" since the first two tokens are inflected) and tag every inflected token not defined in the used morphological lexica with a paradigm tag from the previous section.

Based on the data provided in the previous step, all word forms are generated automatically. The output of that step is given to human annotators again who check the generator output and correct 2,471 out of 30,078 records (8.22%).

Lemmata in the final version of the resource are completely identical in only 223 cases (4.45%) and, when ignoring the letter case, in 2,394 cases (47.76%). In 4,659 cases (92.94%) the number of tokens is identical to the original.

Based on the 5,013 multi-word named entities and their generated word forms, Table 2 shows the probability of inflecting a specific token for named entities of length between 2 and 6 tokens. The data confirms the intuition used in the first step that the inflected noun phrase is located rather at the beginning of the named entity. This only does not hold for named entities of length 2 where more often only the second token is inflected than only the first one.

1	2	3	4	5	6
0.4968	0.6425				
0.8154	0.6895	0.1785			
0.9242	0.8471	0.1027	0.0513		
0.9625	0.6370	0.2272	0.0492	0.0328	
0.9581	0.6492	0.1047	0.0366	0.0105	0.0314

Table 2: Probability of inflecting a token in a multi-word named entity for named entity length from 2 to 6

Table 3 shows the probability of the length of the inflected noun phrase. The data shows that 0.66% of the named entities are indeclinabilia. The most frequent length of the inflected noun phrase is 1, ie. 2. Inflected noun phrases longer than 2 tokens are rather rare (5.31%).

The data gathered from the 5,013 multi-word named entities will be very useful in developing the on-the-fly method of identifying multi-word named entities. Developing that algorithm is not part of the research covered in this paper.

length of NP	probability
0	0.0066
1	0.5192
2	0.4211
3	0.0501
4	0.0030

Table 3: Probability (p()) of the length of the inflected noun phrase (len(NP))

5. Conclusion and further work

This paper presents methods used in building a morphological lexicon of organization named entities for the Croatian language. The resource generation problem is divided into two subproblems - the single-word and the multi-word problem.

The single-word problem, being much simpler, is solved by annotating 4,987 named entities by hand and training a linear successive abstraction algorithm. The algorithm combines weighted evidence from different length endings using linear interpolation. Best results are obtained by using endings up to three characters. The results show a surprising amount of information encoded just in the last character. The algorithm, backed up by existing morphological language resources and two general hand-crafted rules is used to annotate the remaining 101,544 named entities. The method achieves accuracy of 98.70% on the test set.

The multi-word problem is solved only for the NLG task covering the first 5,013 named entities (10,000 all together). The remaining named entities are not included in this resource because of the complexity of the problem and the inability of generating all possible multi-word forms as they can occur in text. For the first 5,013 named entities a four-step method is used where two are automated while two require manual annotation. A big percentage of multi-word lemmata is changed by hand (52.24%, ie. 95.55% if the letter case is not ignored) which stresses the problem of low-quality primary data. The generated word forms show that the inflected noun phrase is mostly of length 1 or 2 and that it is mostly situated at the beginning of the named entity.

The generated resource covers fully single-word named entities for both tasks - NLG and NEI, whilst for multi-word entities only the NLG task is covered. Further research will be necessary to develop and optimize an algorithm for on-the-fly multi-word NEI. Data obtained from the 5,013 generated multi-word named entities will be very useful in the process of developing the algorithm.

6. References

S. Babić, B. Finka, and M. Moguš. 2002. *Hrvatski pravopis*. Školska knjiga.

B. Bekavac and M. Tadić. 2007. Implementation of croatian nerc system. In *Proceedings of the Workshop on Balto-Slavonic Natural Language Processing 2007*, Prague, Czech Republic.

T. Brants. 2000. Tnt – a statistical part-of-speech tagger. In *Proceedings of the 6th Applied NLP Conference, ANLP-2000*, Seattle, WA.

C. Cardie. 1997. Empirical methods in information extraction. *AI Magazine*, 18(4):65–80.

I. Dagan, L. Lee, and F. Pereira. 1997. Similarity-based methods for word sense disambiguation. In Philip R.

Cohen and Wolfgang Wahlster, editors, *Proceedings of the Thirty-Fifth Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics*, pages 56–63, Somerset, New Jersey. Association for Computational Linguistics.

R. Gaizauskas and Y. Wilks. 1998. Information extraction: Beyond document retrieval. *Journal of Documentation*, 54(1):70–105.

M. Kržak and D. Boras. 1985. Lexical database of the croatian literary language. *Informatologia Yugoslavica*, 17(3-4):223–242.

D. Nadeau and S. Sekine. 2007. A survey of named entity recognition and classification. *Linguisticae Investigationes*, 30(1):3–26, January.

E. Reiter and R. Dale. 2000. *Building Natural Language Generation Systems*. Cambridge University Press.

C. Samuelsson. 1996. Handling sparse data by successive abstraction. In *Proceedings of COLING-96*, Kopenhagen, Denmark.

M. Tadić and S. Fulgosi. 2003. Building the croatian morphological lexicon. In *Proceedings of the EACL2003 Workshop on Morphological Processing of Slavic Languages*, pages 41–46, Budapest. ACL.

ZAPI. 2008. Institute for business intelligence, http://www.zapi.hr.