Statistical Named Entity Recognition for Hungarian – analysis of the impact of feature space characteristics

Richárd Farkas[1], György Szarvas[1]

[1] University of Szeged, Institute of Informatics
H-6720 Szeged, Aradi vértanuk tere 1.
{rfarkas, szarvas}@inf.u-szeged.hu

Abstract. In this paper we introduce a multilingual Named Entity Recognition (NER) system that uses statistical modeling techniques. The system identifies and classifies NEs in Hungarian texts, applying the Support Vector Classifier, Artificial Neural Network and C4.5 decision tree learning algorithms. We focused on building as large a feature set as possible, and applied statistical pre-processing methods for feature selection afterwards to fully exploit its potential. A segment of the Szeged Treebank was used for training and validation purposes which consisted entirely of news-wire articles taken from the Hungarian News Agency (MTI). Our best system achieved a 93.59% F measure at term-level, and 90.57% at the more established phrase-level evaluation.

Keywords: Named Entity Recognition, NER,feature selection, machine learning

1 Introduction

The identification and classification of proper nouns in plain text is of key importance in numerous natural language processing applications. In Information Extraction systems proper names generally carry important information about the text itself, and thus are targets for extraction. Machine Translation has to handle proper nouns and other sort of words in a different way due to the specific translation rules that apply to them. These two topics are in the focus of our research.

1.1 Related work

The task of categorizing proper names was introduced during the 90s as a part of the shared tasks in the Message Understanding Conferences [1]. The goal of these conferences was the identification and classification of proper nouns (like person, organization, location name), and phrases describing dates, time intervals, measures, quantities, and so on in texts collected from newspaper articles in English.

As a part of the Computational Natural Language Learning conference in 2002 and 2003 [5, 6] a shared task dealt with the development of such systems that were able to identify the more difficult classes defined in earlier approaches at MUCs simultaneously in two different languages. These less obvious categories they focused on were person, organization and place names, plus all proper nouns that did not belong to the three classes treated as miscellaneous proper names.

Research and development efforts in the last few years have focused on other languages, domains or cross-language recognition. Hungarian NER fits into this trend quite well, due to the special agglutinative property of the language.

There are some results on NER for the Hungarian language as well but all of them are based on expert rules [9], [12]. To our knowledge, no statistical models have yet been constructed for the Hungarian language.

In this paper we will briefly describe and characterize our Named Entity Corpus of business texts in Hungarian that was created at the University of Szeged, and introduce our current research and experimental results on Hungarian NER in newswire articles.

2 The corpus

To train and test our NER model on Hungarian texts, we decided to use a sub-corpus of the Szeged Treebank[1] [2] which contains business news articles from 38 NewsML[2] topics ranging from acquisitions to stock market changes or the opening of new industrial plants. We annotated this collection of texts with NE labels that followed the current international standards as no other NE corpus of reasonable size is available for Hungarian. The data can be obtained free of charge for research purposes[3].

One important characteristic of our data is the domain specificity of the corpus. The Hungarian texts we that we have consist of short newspaper articles from the domain of economy, and thus the *organization* class dominates the other three in frequency. This difference undoubtedly makes NER an easier problem on the Hungarian text, while the special characteristics of the Hungarian language (compared to English) like agglutinativity or free word order usually makes NLP tasks in Hungarian very difficult. Thus it is hard to compare the results for Hungarian with other languages but achieving similar results in English (the language for which the best results have been reported so far) is still a remarkable thing.

The annotation procedure of the corpus consisted of several phases where two independent annotators tagged the data and discussed problematic cases later on. In the final phase all the entities that showed some kind of similarity to one that was judged inconsistent were collected together from the corpus for a review by the two annotators and the chief annotator. The resulting corpus had an inter-annotator agreement rate of 99.89% and 99.77% compared to the annotations made by the two linguists on their own [7].

This corpus, then, is completely equivalent to other corpuses used on the CoNLL-2002 and CoNLL-2003 conferences, both in format and annotation style (the same classes are labeled). We hope that this will make both cross-language comparison and the use of the corpus in developing NER systems more straightforward.

Table 1.: Corpus details

	Tokens	Phrases
Non-tagged tokens	200.067	–
Person names	1.921	982
Organizations	20.433	10.533
Locations	1.501	1.294
Misc. entities	2.041	1.662

[1] The project was carried out together with MorphoLogic Ltd. and the Hungarian Academy's Research Institute for Linguistics

[2] See http://www.newsml.org/pages/docu_main.php for details.

[3] http://www.inf.u-szeged.hu/~hlt/index.html

3 Our NER system for Hungarian

In this section we describe the major characteristics (features used, learning methods, feature selection algorithms) of our system.

3.1 Feature set

Initial features. We employed a very rich feature set for our word-level classification model, describing the characteristics of the word itself along with its actual context (a moving window of size four). Our features fell into the following main categories:

- gazetteers of *unambiguous NEs* from the train data: we used the NE phrases which occur more than five in the train texts and got the same label more in than 90% of the cases,
- *dictionaries* of first names, company types, denominators of locations (mountains, city) and so on: we collected 6 lists from the Internet
- *orthographical features*: capitalization, word length, common bit information about the word form (contains a digit or not, has an uppercase character inside the word and so on). We collected the most characteristic character level bi/trigrams from the train texts assigned to each NE class,
- *frequency information*: frequency of the token, the ratio of the token's capitalized and lowercase occurrences, the ratio of capitalized and sentence beginning frequencies of the token,
- *phrasal information*: NP codes and forecasted class of few preceding words (we used online evaluation),
- *contextual information*: POS codes (we used codes generated by our POS tagger for Hungarian instead of the existing tags from the Szeged Treebank), sentence position, trigger words (the most frequent and unambiguous tokens in a window around the NEs) from the train text, the word is between quotes and so on.

3.2 Learning algorithms

C4.5 decision tree: C4.5 (Quinlan, 1993) is based on the well-known ID3 tree learning algorithm. It is able to learn pre-defined discrete classes from labelled examples. The result of the learning process is an axis-parallel decision tree. This means that during the training, the sample space is divided into subspaces by hyperplanes that are parallel to every axis but one. In this way, we get many n-dimensional rectangular regions that are labelled with class labels and organized in a hierarchical way, which can then be encoded into the tree. Splitting is done by axis-parallel hyper-planes, and hence learning is very fast. One great advantage of the method is time complexity; in the worst case it is $O(dn^2)$, where d is the number of features and n is the number of samples.

Artificial Neural Networks (ANN): Since it was realized that, under proper conditions, ANNs can model the class posteriors (Bishop, 1996), neural nets have become evermore popular in the Natural Language Processing community. But describing the mathematical background of the ANNs here is beyond the scope of this article. Besides, we believe that they are well known to those who are acquainted with pattern recognition. In our experiments we used the most common feed-forward multilayer perceptron network with the backpropagation learning rule.

Support Vector Machines (SVM): The well-known and widely used Support Vector Machines (Vapnik, 1996) is a kernel method that separates data points of different classes with the help of a hyperplane. The created separating hyperplane has a margin of maximal size with a proven optimal generalization capacity. Another significant feature of margin maximization is that the calculated result is independent of the distribution of the sample points. Perhaps the success and popularity of this method can be attributed to this property.

3.3 Feature selection

A general characteristic of learning models in Natural Language Processing is that the features available are far from uncorrelated (and some might even be totally uninformative to the target variable), thus the proper preprocessing of data can be beneficial both in terms of decreased training time and improved performance.

Chi-squared statistic: We used the well known chi-squared statistics to measure the association between individual features and the target attribute (that is, the term's NE label). This statistical method computes how strong the dependency is by comparing the joint distribution and the marginal distributions of the examined feature and the target variable. This way attributes are ranked based on their individual relevance and this enables us to discard insignificant features automatically.

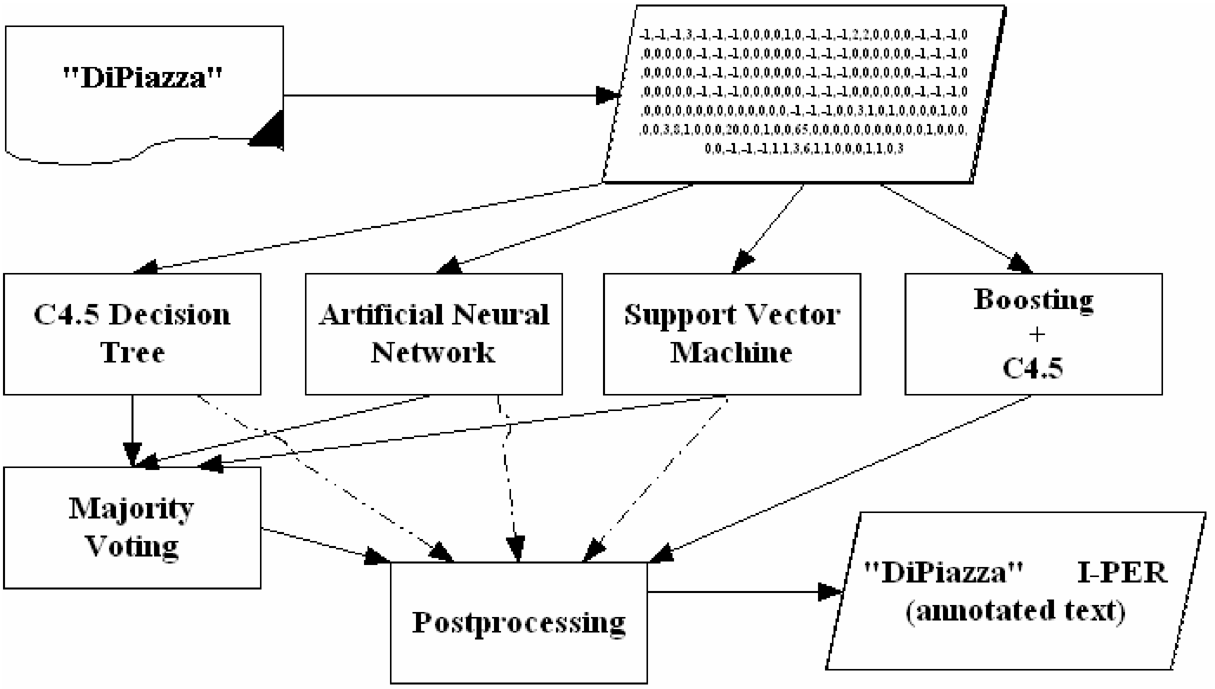

Figure 1: *Outline of the structure of our NER model.*

4 Experimental results

4.1 Metrics used

In this paper we will also evaluate our models using the phrase accuracy measures introduced in the CONLL conferences along with term accuracy used in our previous experiments. In every other aspect we followed the framework outlined in [3].

The values obtained from the two metrics for a certain prediction is illustrated in Figure 2b. It can be easily seen that the relationship between them is quasi linear.

4.2 Feature space

Our recent research efforts have focused on enriching the embedding feature space with additional attributes of all kinds, ranging from orthographic parameters to values describing sentence level information. We collected new features and finally had 225 different properties for each token in the text we analyzed and then used them to construct supervised learning models that could predict more effectively whether the token in question was part of an entity name, and if so, we determined which of the four category it most likely belonged to.

The features that proved to be the most useful and informative (according to their Chi-square score) for the given NER problem were frequency information collected from the Szószablya [4] frequency dictionary, type of the first letter of the word (upper/lower case), whether the word was a valid first name, whether it contained uppercase letters inside the word, case and POS codes, the word length, position in the sentence, whether one of the consecutive words was a company type (like ‘Plc.’, ’Ltd.’, ’Co.’, etc.), and whether the word was the name of a country.

4.3 Results

No independent results on Hungarian NER using this corpus have yet been published. The results here are compared to our previous results, which we think are the best that have been published so far [3].

Different machine learning methods (Artificial Neural Network, Support Vector Classifier, Decision Tree Classifier) can achieve an F-measure of 90% or more regarding accuracy of recognizing terms in one of the four NE classes [3].

Our best model with its enriched feature set achieved an Fmeasure of 93.32% in term accuracy using all of the 225 features collected – which is a 1.51% improvement on the results given in [3]. By using a feature selection procedure we can achieve an accuracy of 93.59% using 135 features. Figure 2a shows the classification accuracy with a growing feature space size and the relationship between the two different evaluation metrics used. It shows a 1.14% improvement in performance compared to our best previous model that worked on a reduced feature space. Applying the phrase accuracy evaluation script of the CONLL conferences, our models attained a 90.57% F-measure (with 225 features) and a 90.99% F-measure respectively using a reduced feature space.

A baseline tagger that selects complete unambiguous named entities appearing in the training data had an F-measure phrase accuracy of 70.99%. These results are slightly better than those published for different languages, which is probably due to the unique characteristics of business news texts where the distribution of entities is biased towards the organization class.

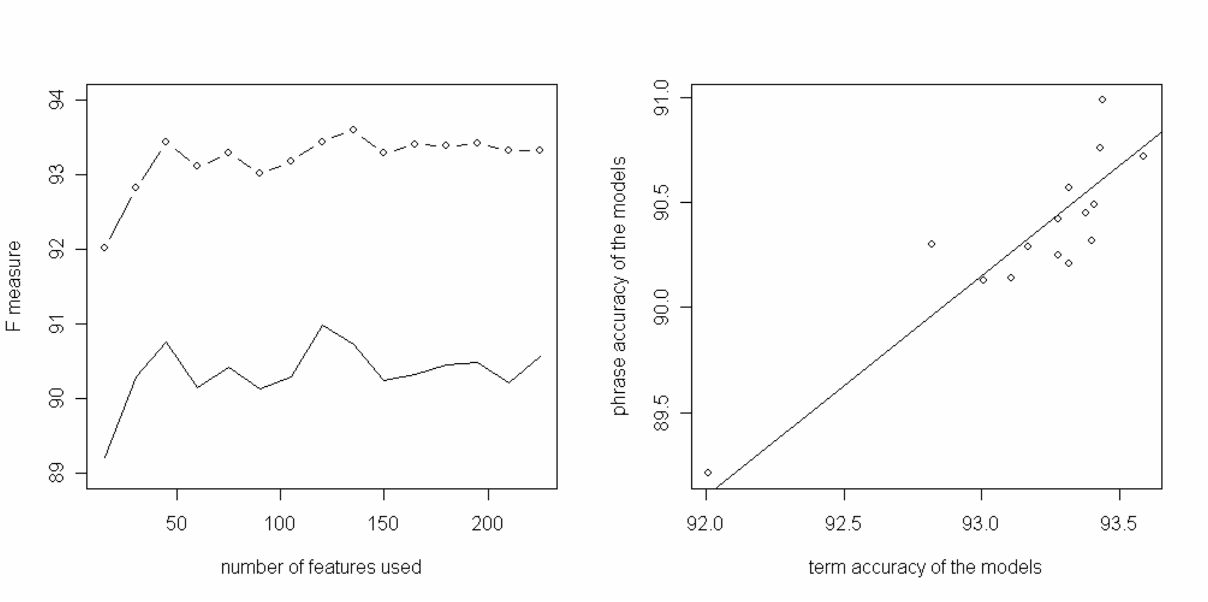

Figure 2: *How the prediction (term and phrase) accuracy depends on the growing feature space dimension and the (quasi linear) relationship of the two measures*

Bibliography

1. Chinchor, N.: MUC-7 Named Entity Task Definition, in Proceedings of Seventh Message Understanding Conference (MUC-7), Washington (1998) http://www.itl.nist.gov/iaui/894.02/related_projects/muc/proceedings/muc_7_toc.html
2. Csendes, D., Csirik, J., Gyimóthy, T.: The Szeged Corpus: A POS tagged and Syntactically Annotated Hungarian Natural Language Corpus in Proc. of TSD 2004, Brno, LNAI vol. 3206, pp. 41-49 (2004)
3. Farkas, R., Szarvas, Gy., Kocsor, A.: Named Entity Recognition for Hungarian Using Various Machine Learning Algorithms, Acta Cybernetica, accepted for publication (2005)
4. Halácsy P., Kornai A., Németh L., Rung A., Szakadát I., Trón V.: A szószablya projekt – www.szoszablya.hu. MSZNY 2003, 298–299, Szeged, Hungary (2003)
5. Tjong Kim Sang, E. F.: Introduction to the CoNLL-2002 Shared Task: Language-Independent Named Entity Recognition. In Proceedings of CoNLL-2002, pp. 155-158., Taipei (2002)
6. Tjong Kim Sang, Erik F. and De Meulder, Fien: Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition, in Proceedings of CoNLL-2003, 142-147, Edmonton (2003) http://cnts.uia.ac.be/signll/conll.html
7. György Szarvas, Richárd Farkas, László Felföldi, András Kocsor, János Csirik: A highly accurate Named Entity corpus for Hungarian, Proceedings of International Conference on Language Resources and Evaluation (2006)