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

1 Introduction

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 not belong to that three classes treated as miscellaneous proper names.

In this paper we shortly describe and characterize a 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 Named Entity Recognition (NER) in newswire articles.

2 The corpus

The Named Entity Corpus for Hungarian is a sub corpus of Szeged Treebank [2], which contains full syntactic annotations done manually by linguist experts[1]. This group of texts has been annotated with Named Entity class labels following the annotation standards used on CONLL conferences.

Table 1.: Corpus details

	Tokens	Phrases
Non-tagged tokens	200067	–
Person names	1.921	982
Organizations	20.433	10.533
Locations	1.501	1.294
Misc. entities	2.041	1.662

3 Experimental results

The authors carried out some previous research [3] on the classification accuracy of NER on Hungarian business news texts which showed that different machine learning methods (Artificial Neural Network, Support Vector Classifier, Decision Tree Classifier) can achieve near or above 90% in F measure regarding term accuracy of recognizing the four NE classes.

Our recent research efforts focused on enriching the embedding feature space with further attributes of various nature, ranging from orthographic parameters to values describing sentence level information. We have collected 225 different properties for each token in the analyzed text and use them to set up supervised learning models that can predict effectively whether the token in question is part of an an entity name, and if so, to which of the four category does it most likely belong to. 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), and thus proper preprocessing of data can benefit both in terms of decreased training time and improved performance. We used the well-known statistical chi-squared test to evaluate attributes and discard those that do not hold significant information for the decision task.

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

In this paper we 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 scenario described in [3].

Our best model obtained 93.32% F value term accuracy using all of the 225 features collected which is a 1.51% improvement to the results discussed in [3]. With using feature selection we can achieve an accuracy of 93.59% using 135 features that shows 1.14% improvement in performance compared to our best previous model that worked on a reduced feature space. Using the phrase accuracy evaluation script of the CONLL conferences, our models show 90.57% F measure (with 225 features) and 90.99% F measure using a reduced feature space.

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

The features that proved to be the most useful and informative to the task were frequency information collected from the Szószablya [4] frequency dictionary, type of the first letter of the word (upper/lower case), word is a valid first name, contains uppercase letter inside the word, case and POS codes, word length, position in the sentence, one of the consecutive words is a company type (like "Ltd.", "Co.", etc.), the word is the name of a country.

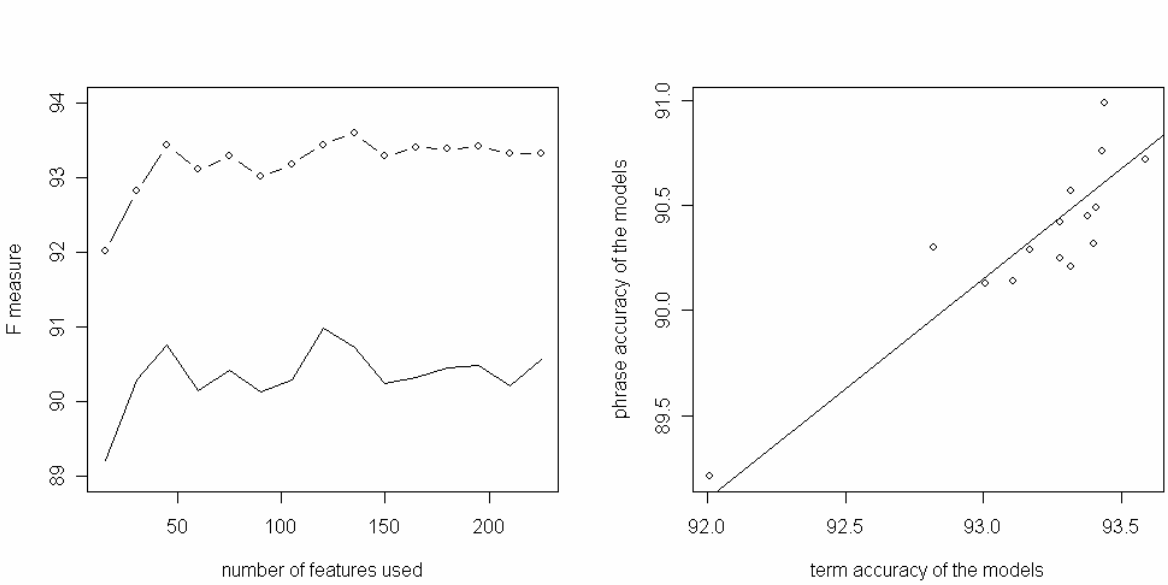

Figure 1.: Change of prediction (term and phrase) accuracy with growing feature space dimension and the (quasi linear) relationship of the two measure

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