Exploiting Learning in Bilingual Named Entity Recognition

Georgios Petasis

Institute of Informatics and Telecommunications, N.C.S.R. "Demokritos"

Tel: +301-6503158, Fax: +301-6532175, e-mail: petasis@iit.demokritos.gr

Introduction

Named-entity recognition *(NERC)* is the identification of proper names in text and their classification as different types of named entity. A typical NERC system consists of a lexicon and a grammar. The lexicon is a set of gazetteer lists, containing names that are known beforehand and have been classified into named-entity types, such as persons, organisations, locations etc. The grammar is used to recognise named entities that are not in the gazetteer lists. Both of these resources seem to heavily depend on both domain and language. The manual adaptation of these two resources to a particular domain or language is time-consuming and in some cases impossible, due to the lack of experts. As a result, automatic acquisition of resources from corpora is highly desirable.

In this paper we examine the use of the learning method C4.5 [Quinlan, 1993] for the automated acquisition of NERC grammars when moving to a new domain. The method was evaluated on management succession events in English and Greek texts. For English, a part of the MUC-6 corpus [DARPA 1995] was used, whereas for Greek we used corpus provided by the company "Advertising Week" (*http://www.addweek.gr*). Although management succession events involve several named-entity types, person and organisation names are more difficult to identify and classify. For this reason, our study focuses on these two entity types.

Related Work

Nymble [Bikel *et al.* 1997], Alembic [Vilain & Day 1996, Day *et al.* 1998] and AutoLearn [Cowie 1995], as well as the approach presented here, are examples of systems exploiting supervised learning techniques. Nymble [Bikel *et al.* 1997] uses statistical learning to acquire a Hidden Markov Model (HMM) that recognises named entities in text. NERC in Alembic [Vilain & Day 1996] is based on a rule learning approach introduced in Brill's work on part-of-speech tagging [Brill 1993]. The AutoLearn system [Cowie 1995] is based on a decision tree learning algorithm, named ID3 [Quinlan 1991]. [Cuchiarelli *et al.* 1998] present an unsupervised learning algorithm to classify the unknown named entities (i.e. those named entities that the NERC system identified as such but didn't manage to classify) in Italian texts.

Experimental Results

The English corpus contains 461 organisation and 373 person instances and the Greek corpus 425 organisation and 262 person instances. In our study we used the VIE NERC system for English [Humphreys *et al.*, 1997] and the Greek NERC system GIE [Karkaletsis *et al.*, 1998]. Both systems involve the following modules: tokeniser, sentence splitter, part-of-speech tagger, gazetteer-list lookup, and named-entity parser. Different tokenisers and sentence splitters are used for the two languages. The Brill tagger [Brill, 1993] is used for part of speech tagging in both languages. A new set of part-of-speech tags was specified for Greek in order to take into account issues such as gender, number, etc. The English gazetteers consist of 2559 organisations, 94 company designators, 135 organisation keywords, 476 persons and 163 person titles. The Greek gazetteers consist of 475 organisations, 19 company designators and 842 persons. A bottom-up chart parser is used for named-entity parsing in both languages. NERC grammars have been hand-written for both languages.

In this work, we aim to speed-up the customisation of the above-described systems, by learning domain-specific NERC grammars. The learning algorithm used for this purpose is C4.5. The algorithm requires the training data to be provided in a particular format, which is common in most work in symbolic machine learning. Each organisation and person instance is represented by a feature vector. Two features are used for each word: its gazetteer tag, if it has one, and its part of speech. The feature vector consists of 13 words: 9 words for the NE phrase plus the two adjacent words on each side of the phrase. Therefore, each vector consists of 26 features, 13 part-of-speech and 13 gazetteer tags. In addition to the training examples corresponding to person and organisation NE phrases, a number of negative, i.e., non-NE, example phrases are constructed from the data. This is needed, in order to capture the dual nature of the NERC task, namely the identification *and* classification of NE phrases. The negative examples in our study are constructed using all *noun phrases* that are not NE phrases. Two experiments were done, one for each language. In each experiment, C4.5 was asked to construct decision trees that distinguish between three classes: *person*, *organisation* and *non-NE.* The two pruning parameters of C4.5 (pre- and post-pruning) were varied to give different tree sizes. Performance on the NERC task was evaluated at each tree size. In order to gain an unbiased estimate of the performance of the system on unseen data, 10-fold cross-validation was performed at each level of tree pruning, at each different tree size. According to this evaluation method, the dataset is split into ten, equally-sized subsets and the final result is the average over ten runs. In each run, nine of the ten subsets of the data are used to construct the named-entity recogniser and the tenth is held out for the evaluation. The measures that were chosen for the evaluation are those typically used in the language engineering literature: *recall* and *precision*.

In the first experiment, C4.5 was applied on the MUC-6 data, asked to learn decision trees that can distinguish between person names, organisation names and noun phrases that do not belong to either of these categories (*non-NE*). As a basis for comparing the results in the experiments we can use the performance of the manually constructed set of rules in the VIE NERC system [Humphreys *et al.* 1997]. The results of this system on the data are shown in Table 1. Also, Table 1 presents the results of the experiment for organisation and person phrases. Each percent is the average of the 10 values acquired in the corresponding 10-fold cross-validation experiment. Because of the limited space of this paper, only the results that correspond to tree sizes with 200 nodes are presented here. The results of the experiment are better than those of the manually constructed rules. Similar to the manually constructed NERC system, the performance for organisations is lower than that for persons and the decision tree performs significantly better than the manually constructed system. Interestingly, recall and precision are at similar levels, both for organisations and persons, which indicates robustness in the performance of the task.

	Recall (o)	*Precision (o)*	*Recall (p)*	*Precision (p)*
Manually Constructed NERC System	69.25%	83.42%	84.97%	92.50%
Trained NERC System	89.93%	81.40%	91.31%	92.86%

Table 1: Performance of the NERC systems on the whole dataset (English text).

The same experiment was repeated for Greek text of similar size and type. The task is again the distinction between person names, organisation names and other noun phrases, but it is considerably more difficult than in the English-text experiment, especially due to the small size of the organisation gazetteer lists that are used. The results of a manually constructed NERC grammar on the data are shown in Table 2. In the same table, the results of the experiment in Greek text for the two different types of named entity are also presented. These results correspond to trees having 200 nodes, as was the case for the English corpus. Clearly the results are worse than in English, especially for organisations due to the small coverage of the gazetteer lists, although they are somewhat better than the ones taken from the manually constructed rules. Particularly worrying are the recall results for organisations, which are around 60%, almost 20 pp. lower than in the English text. Precision for organisations is at similar levels as for the English language, i.e. around 80%. In the case of persons, the results are better. In comparison with the English experiment, recall for persons is lower, around 77%, while precision is higher, around 93%.

	Recall (o)	*Precision (o)*	*Recall (p)*	*Precision (p)*
Manually Constructed NERC System	40.4%	57.3%	77.0%	88.8%
Trained NERC System	64.1%	83.4%	77.1%	93.8%

Table 2: Performance of the NERC systems on the whole dataset (Greek text).

Conclusions

We have applied a popular machine learning technique to the NERC task and have shown that it can do better than equivalent manually constructed tools. The set of recognition rules generated by the learning method is comprehensible and intuitive. More importantly, the approach was shown to be insensitive to the change of language from English to Greek. However, there is still much to be desired from such an approach and the results of this paper indicate that more can be delivered. Our main aim for the future is to improve the results of our approach when the available linguistic resources, in particular gazetteer lists, are limited, which may involve the use of a different learning algorithm, or even the development of a new one that will be more suited to the particular problem.

References

[DARPA 1995] Defense Advanced Research Projects Agency. Proceedings of the Sixth Message Understanding Conference (MUC-6), Morgan Kaufmann.

[DARPA 1998] Defense Advanced Research Projects Agency. Proceedings of the Seventh Message Understanding Conference (MUC-7), Morgan Kaufmann.

[Bikel *et al.* 1997] Bikel, D.M., Miller, S., Schwartz, R., and Weischedel, R. “Nymble: a High-Performance Learning Name-finder”. In *Proceedings of the Fifth Conference on Applied Natural Language Processing (ANLP-97),* Washington, D.C., pp. 194 – 201.

[Brill 1993]. Brill, E. “A corpus-based approach to language learning”. *PhD Dissertation*, Univ. of Pennsylvania.

[Cowie 1995] Cowie, J. “Description of the CRL/NMSU System Used for MUC-6”. In [DARPA 1995].

[Cuchiarelli *et al.* 1998] Cuchiarelli, A., Luzi, D., and Velardi, P. “Automatic Semantic Tagging of Unknown Proper Names”. In *Proceedings of COLING-98,* Montreal.

[Day *et al.* 1998] Day, D., Robinson, P., Vilain, M., and Yeh, A. Description of the ALEMBIC system as used for MUC-7. In [DARPA 1998].

[Humphreys *et al.* 1997] Humphreys, K., Gaizauskas, R., Cunningham, H., and Azzam, S. VIE Technical Specifications. Department of Computer Science, University of Sheffield.

[Karkaletsis *et al*., 1998] Karkaletsis, V., Spyropoulos, C.D., and Petasis, G. “Named Entity Recognition from Greek texts: the GIE Project”. In *“Advances in Intelligent Systems: Concepts, Tools and Applications”*, ed. S.Tzafestas, Kluwer Academic Publishers.

[Quinlan 1991] Quinlan, J.R. “Machine Learning: Easily Understood Decision Rules”. In *Computer* Systems that Learn, eds. Weiss, S.M. and Kulikowski, C.A., Morgan Kaufmann.

[Quinlan, 1993] Quinlan, J. R., C4.5: Programs for machine learning, Morgan-Kaufmann, San Mateo, CA, 1993.

[Vilain & Day 1996] Vilain, M., and Day, D. “Finite-state phrase parsing by rule sequences”. In *Proceedings of COLING-96*, vol. 1, pp. 274-279.