Named-Entity Recognition from Greek and English Texts

Vangelis Karkaletsis, Georgios Paliouras, Georgios Petasis,

Natasa Manousopoulou and Constantine D. Spyropoulos

Software and Knowledge Engineering Laboratory
Institute of Informatics and Telecommunications,
N.C.S.R. "Demokritos",
Tel: +301-6503197, Fax: +301-6532175
e-mail: { vangelis, paliourg, petasis, costass, natasa}@iit.demokritos.gr

Abstract

Named-entity recognition (NER) involves the identification and classification of named entities in text. This is an important subtask in most language engineering applications, in particular information extraction, where different types of named entity are associated with specific roles in events. In this paper, we present a prototype NER system for Greek texts that we developed based on a NER system for English. Both systems are evaluated on corpora of the same domain and of similar size. The time-consuming process for the construction and update of domain-specific resources in both systems led us to examine a machine learning method for the automatic construction of such resources for a particular application in a specific language.

Keywords: named-entity recognition, information extraction, machine learning

1 Introduction

Today's overload of information, particularly through the World Wide Web, makes difficult the user's access to the right information. The situation becomes even more difficult due to the fact that a lot of this information is in different languages. Therefore, it is important to apply an information process that will extract from all that volume of information only the facts that match the user's interests, and allow the user to access facts written in a different language. Information Extraction (IE) technology can meet these requirements, since unlike what happens with information retrieval and filtering technology, in IE the user's interests are on specific facts extracted from the documents and not just on the documents themselves. Some documents may contain the requested keywords but be irrelevant to the user's interests. Working with specific facts instead of documents provides users with information that is more relevant to their interests [Gaizauskas & Wilks 1997].

The IE systems developed so far, extract, in most cases, fixed information from documents in a fixed language. However, in order for the IE technology to be truly applicable in real life applications, IE systems need to be easily customisable to new domains and languages. The IE task involves typically two sub-tasks: the recognition of named entities (e.g. persons, organisations, locations, dates) involved in an event and the recognition of the relationships holding between named entities in that event (e.g. personnel joining and leaving companies in management succession events). A named entity (NE) is a phrase, which serves as a name for something or someone. According to this definition, the phrase in question must by a noun phrase (NP). Clearly, not all NPs are named entities. Named-entity recognition (NER) involves two tasks: identification of NPs that are NEs and classification of NEs into different types, such as organization and person names.

A typical NER system involves the exploitation of a lexicon and a grammar, which need to be updated when the system is customised to a new domain. The lexicon is a set of gazetteer lists, containing names that are known beforehand and have been classified into named-entity types. The grammar is used to recognise named entities that are not in the gazetteer lists or they occur in more than one gazetteer list. Manual construction of these resources is a very time-consuming process and it is therefore worth examining methods that could automate their construction for a particular application in a specific language. Automated knowledge acquisition, with the use of machine learning techniques, has recently been proposed as a promising solution to this and other similar problems in language engineering.

In this paper, we present the prototype GIE NER system we developed for the Greek language. This is based on the English VIE system of Sheffield University [Humphreys *et al.* 1997] and was developed in the context of the research project GIE[1] (Greek Information Extraction). We also examine, in both languages, the use of the learning method C4.5 [Quinlan 1993] for the automated acquisition of NER grammars when moving to a new domain. More specifically, section 2 presents related work. Section 3 presents the prototype GIE system. GIE and VIE are evaluated and compared on corpora of the same domain and of similar size. The customisation in different domains, for English and Greek texts, is presented in Section 4. Section 5 discusses the development of GIE based on the VIE system and draws conclusions on the usability of the learning method in the NER task.

2 Related Work

Recent progress in IE technology is due to the increase in available resources such as machine readable dictionaries and text corpora, in computational power and processing volume as well as the development of Language Technology techniques that can be applied in practice. This progress is proved from the results of Message Understanding Conferences – MUCs where several IE systems are evaluated (see MUC Website in http://www.muc.saic.com/). Named entity recognition is one of the evaluation tasks, on which also the best results are achieved, proving that this technology is mature.

[1] GIE (Greek Information Extraction) is a bilateral project between NCSR "Demokritos" (GR) and Univ. of Sheffield (GB), funded by the Greek General Secretariat of Research & Technology and the British Council.

The identification of named entities in a corpus along with their classification as persons, organisations, etc., can be useful not only as the first stage of a complete IE system, but also for other tasks, such as indexing of documents, maintenance of data bases containing information for the identified persons, organisations, etc.

The systems participating in MUCs are required to process texts, identify the pieces of text that are relevant to the domain, and fill templates which contain slots for the events to be extracted and the entities involved. Information analysts design the template structure and fill manually the templates which are then used in the evaluation. The domain areas examined so far in the MUCs are the following: Navy messages (MUCK 1987, MUCK-II 1989), news for terrorist attacks (MUC-3 1991, MUC-4 1992), company news (MUC-5 1993, MUC-6 1995), launches of air missiles (MUC-7 1998).

The main measures used for the evaluation of NER systems are **recall** and **precision**. Recall measures the number of items of a certain type correctly identified, divided by the total number of items of this type in the training data. Precision is the ratio of the number of items of a certain type correctly identified to all items that were assigned that particular type by the system. NER modules represent the most mature IE technology. The best score obtained in the NER task in MUC-6 was 92% recall and 95% precision [DARPA 1998].

NER involves the exploitation of gazetteers and named-entity grammars, which need to be updated when the system is customised to a new domain. The exploitation of learning techniques to support the customisation of NER systems has recently attracted a lot of attention. Machine learning techniques are subdivided into two broad categories: *supervised* and *unsupervised*. Supervised learning techniques require the existence of training examples that have been hand-tagged with the correct class. On the other hand, unsupervised techniques assume that the correct classification of the training examples is not known and classify the examples according to their common characteristics. Supervised methods are more expensive than unsupervised ones, in terms of the time spend to pre-process the training data. However, the additional information included in supervised data leads usually to a better classification system.

Nymble [Bikel *et al.* 1997], Alembic [Vilain & Day 1996, Day *et al.* 1998], AutoLearn [Cowie 1995], NYU [Sekine 1998], as well as the approach presented here, are examples of systems exploiting supervised learning techniques for NER systems. Nymble [Bikel *et al.* 1997] uses statistical learning to acquire a Hidden Markov Model (HMM) that recognises named entities in text. NER 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]. The NYU system [Sekine 1998] is based on the same supervised learning algorithm with our approach for named-entity recognition from Japanese texts. Cuchiarelli *et al.* [1998] present an unsupervised learning algorithm to classify the unknown named entities (i.e., those named entities that the NER system identified as such but didn't manage to classify) in Italian texts.

3 Named-Entity Recognition in GIE

The named entity recogniser in GIE is based on the VIE English NER system developed at the University of Sheffield [Humphreys *et al.* 1997] using the GATE language engineering platform [Cunningham *et al.* 1996]. Both systems involve the following modules: tokenizer, sentence splitter, part-of-speech tagger, gazetteer-list lookup, and named-entity parser (see Figure 1). 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 the gender for nouns and adjectives, number for adjectives and verbs, etc. The language specific resources of the gazetteer-list look up module and the named-entity parser were replaced by the corresponding Greek resources. A bottom-up chart parser is used for named-entity parsing in both languages. Gazetteers and grammars have been hand-written for both languages.

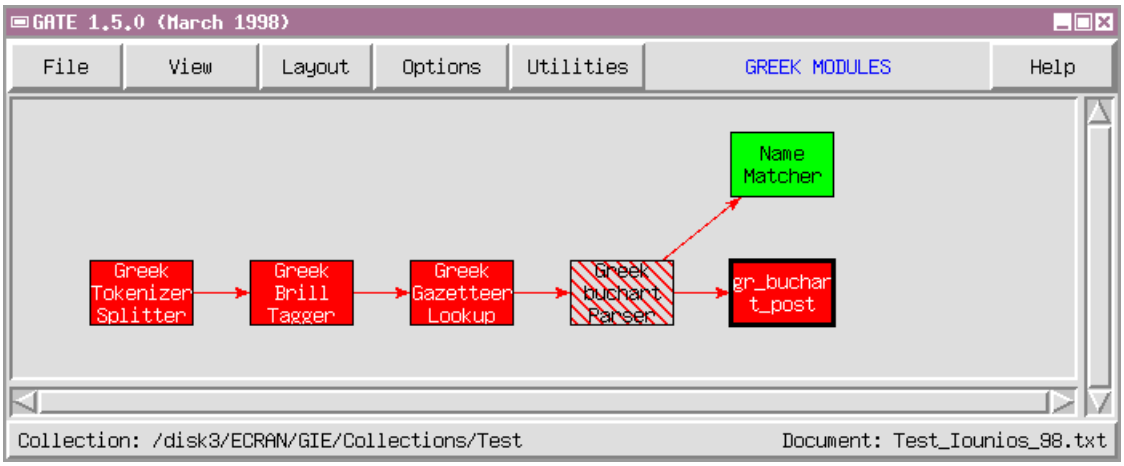

Figure 1. Modules of GIE Named Entity Recogniser

We used Greek corpora from two different sources in order to train and then evaluate our system. The first one contained news articles from the Greek company *"Advertising Week" (http://www.adweek.gr)* on "management succession events" (personnel leaving or joining companies for the period from 1/96 until 12/98). The corpus size was about 65,000 words. Part of this corpus (about 36.000 words) was hand-tagged in order to be used for our experiments. The second corpus is a general-theme hand-tagged corpus, which was provided by the Wired Communications Laboratory (WCL) Laboratory of Patras University. The size of this corpus is about 125.000 words.

3.1 Tokeniser, Sentence Splitter, Part-of-Speech Tagger

We developed a new tokeniser and a new sentence splitter for the Greek language. The tokeniser accepts raw text as input and produces a list of tokens and their boundaries. The sentence splitter uses the tokens produced to generate a list of sentences. Both use a set of rules in order to identify the tokens and the sentences respectively.

The Brill tagger [Brill 1993] is used for part of speech tagging in both languages. We specified a new tag set for the Greek language in order to cover features that do not occur in English, i.e., cases for nouns, adjectives and verbs, mood for verbs, etc. For efficiency reasons the Greek tag set is rather limited, containing only 58 tags. The original tag set used by the Brill tagger for the English language contains 48 tags. Due to our interest in examining the learning procedure of the Brill tagger under different thematic domains, we used the domain specific corpus of *"Advertising Week"* and the general-theme corpus provided by the WCL Laboratory of Patras University. We trained the Brill tagger over the two Greek corpora and found its performance to be around 95% for both of them [Petasis *et al.* 1999]. This result shows that the performance of Brill tagger does not significantly depend on the corpus domain, at least when applied to the Greek language. Therefore, porting the tagger to different domains should require minimal effort. However, the performance of Brill tagger for Greek is slightly lower compared to English. Our main aim for the future is to try to improve its performance, by trying to isolate the difficulties and solve the problems that arise in the context of the Greek language.

3.2 Gazetteer Lookup

This module attempts to identify phrases and keywords related to named entities, as defined for the management succession task (persons, organisations, locations, dates). This is done by searching a series of pre-stored lists (gazetteers) of organisations, locations, date forms, currency names, etc.

In order to use that module for the Greek language, we had to create Greek gazetteers. 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 sample of them is presented in Table 1.

Persons	Organizations	Locations
Φραγκίσκος	Σκάι 100,4 FM	Ουκρανίας
Τράγκας	ΣΚΑΪ 100,4 FM	Ουγγαρίας
Τονιά	ΣΚΑΙ	Ουγγαρία
Σόφη	Ποπ-Κορν	Ολλανδίας
Σόνια	Ποπ&Ροκ	Ολλανδία
Σωτήρης	Ποπ Κορν	Νοτίου Ελλάδος
Σίσσυ	Μελωδία FM 100	
Σίμος	Μελωδία	
Σίλια	ΜΕΛΩΔΙΑ FM 100	**Titles**
Σέργιος	Ι.Γ. Δραγούνης & Υιοί ΑΕ	Σύμβουλος Media
Λία	Ι.Γ. Δραγούνης & Υιοί	Σύμβουλος Marketing
Λένα	Ι.Γ. Δραγούνης	Πρόεδρος Διοικητικού Συμβουλίου
Ιφιγένεια	Ι. Γ. Δραγούνης & Υιοί ΑΕ	Διεύθυνση Στρατηγικού Σχεδιασμού
Ισίδωρος	Ι. Γ. Δραγούνης	Διεύθυνση Επικοινωνίας
Ελισάβετ	Flash 96.1	Διεύθυνση Διαφήμισης
Ελεονόρα	Flash 96,1 FM	Διεύθυνση Marketing
Ελεάνας	Flash 961	Senior Product Manager
Ελεάνα	Flash 9,61 FM	Senior Media Planner
	Flash 9,61	

Table 1. Sample of Greek Gazetteers

The construction of Greek gazetteers for the domain of management succession presented a lot of problems. The names of persons and organisations occurring in the relevant corpus are actually bi-lingual, i.e., there are several English names especially in the case of organisations (see Table 1). This means that we actually have to maintain two sets of gazetteers: one for Greek and one for English names. We also found out that several of the names are composed from both English and Greek characters. Another problem concerns the different writings of the same name, such as the name of the radio station "Flash" in Table 1. There is also the problem of Greek proper nouns declension ("Ελεάνα" in nominative and "Ελεάνας" in accusative) and with accented characters, which are not used all the times. We plan to develop a module that will be responsible for identifying these cases and resolve them.

3.3 Named-Entity Chart Parser

The parser is a modification of the Gazdar and Mellish bottom-up chart parser [Gazdar & Mellish 1989]. It applies a named-entity grammar to construct proper noun phrases. In the named-entity grammar of the English IE system, there are 189 rules for organisations, persons, locations, temporal and number expressions. The following information is taken into account by the grammar: part-of-speech tags of the words in the NE and close to it, gazetteer tags for the words in the NE and close to it, punctuation.

We had to create new rules for the Greek language, excluding most of the English rules. This was due to the nature of the Greek language and of the specific corpus. For instance, several of the English rules for organisations are based on the existence of a company designator (i.e. Ltd., Co.), which is not used in the Greek corpus. Another example is the case of person names, where there are English rules based on person title (i.e. Mr., Mrs). However such titles are rare in the Greek corpus.

It seems that in the case of Greek corpus, the named entity rules should be mainly based on the existing gazetteer tags. A rich gazetteer is thus required in order to improve the named entity parser results. However, it is difficult to create and maintain such rich gazetteers. New rules should be included in order to identify named entities and classify them as persons, organisations, etc. For this purpose we constructed manually a set of grammar rules and performed some first evaluation tests, the results of which are presented in the following section..

3.4 Evaluation results

As a basis for evaluating the results of the GIE NER system, we compared its performance with that of the VIE system [Humphreys *et al.* 1997] for the identification of organisations (o) and persons (p), since the general consensus is that person and organisation names are more difficult to identify and classify. Both systems were evaluated on management succession events. For English, we used a part of the MUC-6 corpus [DARPA 1995], whereas for Greek we used part of the corpus provided by "Advertising Week". The English MUC-6 corpus contains 461 organisation and 373 person instances and the Greek corpus of "Advertising Week" 425 organisation and 262 person instances. The results of the two systems on the data are shown in Table 2.

Recall (o)	*Precision (o)*	*Recall (p)*	*Precision (p)*
69.25%	83.42%	84.97%	92.5%

VIE-English corpus

Recall (o)	*Precision (o)*	*Recall (p)*	*Precision (p)*
40.4%	57.3%	77.0%	88.8%

GIE-Greek corpus

Table 2: Performance of VIE and GIE NER systems

VIE results are significantly lower than the aggregate results presented for the various systems participating in MUC-6 and MUC-7 [DARPA 1995, 1998]. This is due to the difficulty in identifying person and organisation names. In both languages, the results are better for persons than for organisations. Person names are shorter and are usually either included in the gazetteers, or preceded by a person title. This fact makes their identification easier than for organisation phrases, which can be lengthy and may consist of words of various parts of speech and gazetteer types.

The results for VIE are significantly higher than the results of GIE. This is mainly due to the limited size of Greek gazetteers (especially in the case of organisations), the existence of several English names in the Greek texts, as well as the limited set of Greek grammar rules used by the named-entity parser.

We have to note again that gazetteers and grammars have been hand-written for both languages. Manual construction / update of these resources is a very time-consuming process. That's why we decided to examine methods that could automate the construction of such resources for a particular application in a specific language. Automated knowledge acquisition, with the use of machine learning techniques, has recently been proposed as a promising solution to this and other similar problems in language engineering. In the next Section we examine the use of the learning method C4.5 [Quinlan 1993] for the automated acquisition of NER grammars when moving to a new domain. The method was evaluated on both English and Greek corpora.

4 Domain Customisation of the Named-Entity Recogniser for English and Greek Texts

Our objective is to minimise human effort in the customisation of a NER system to a given domain and examine this in two different languages. A NER system usually includes a grammar consisting of tags that are assigned by examining part-of-speech and syntactic properties of the words in a phrase and looking up the gazetteer lists. In our study we used the VIE NER system for English [Humphreys *et al.* 1997] and the GIE NER system for Greek. Our study focuses on two entity types (person and organisation) as it was the case for the manually constructed grammars.

We aim to speed up the customisation of the two NER systems, by learning domain-specific NER grammars. The learning algorithm used for this purpose is C4.5 [Quinlan 1993]. 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. As an example of the way in which NE phrases are coded into feature vectors consider the following Greek phrase:

… θα έχει *ο Φώτης Μπόμπολας*, Διευθύνων …

where the person phrase is shown in italics. The vector corresponding to this phrase is the following:

```
[RP, NOTAG, VBF, NOTAG, DDT, NOTAG, NNPM, person, NNPM, NOTAG, ?, ?, ?, ?, ?, ?, ?, ?, ?, ?, ?,
?, COMMA, NOTAG, JJM, NOTAG]
```

where the part-of-speech tags are to be interpreted as follows: `RP`: particle, `VBF`: verb in future tense, singular, `DDT`: definite determiner, `NNPM`: proper noun masculine, `JJM`: adjective masculine. The gazetteer tags appearing in the example are: `person`, `NOTAG`. The word *Φώτης* appears in the list of persons and is therefore assigned the tag `person`.

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 NER 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.

The experiments presented in this section have two goals:

- To examine the feasibility of constructing a named-entity recogniser from tagged training data, using C4.5.
- To examine the applicability of this approach to two different languages: Greek and English. In addition to the linguistic differences between English and Greek named entities, the problem of constructing a NER grammar for Greek is made more difficult due to the shortage of resources. The small size of the organisation lists is particularly important.

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 NER 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.

4.1 English Named-Entity Recognition

C4.5 was applied to the MUC-6 data, learning 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 NER system [Humphreys *et al.* 1997] (see Table 2).

Figures 2 and 3 present the results of the experiment for organisation and person phrases. Each point in the graph is the average of the 10 values acquired in the corresponding 10-fold cross-validation experiment. Similar to the manually constructed VIE NER system, the performance for organisations is substantially lower than that for persons. The decision tree performs significantly better than the manually constructed system. This is an unexpected and very encouraging result for the use of machine learning in the construction of NER systems.

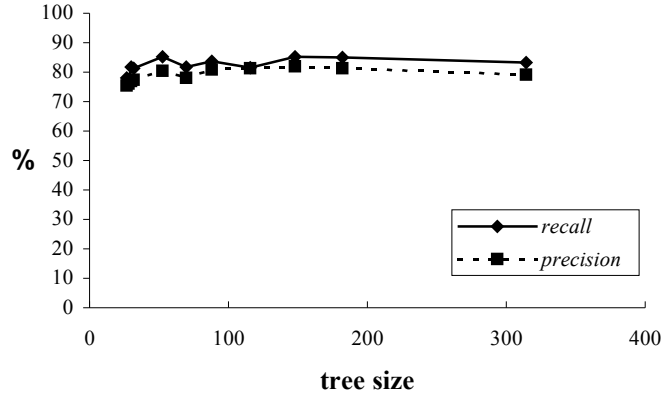

Figure 2. Results for the organisation phrases, in the English text

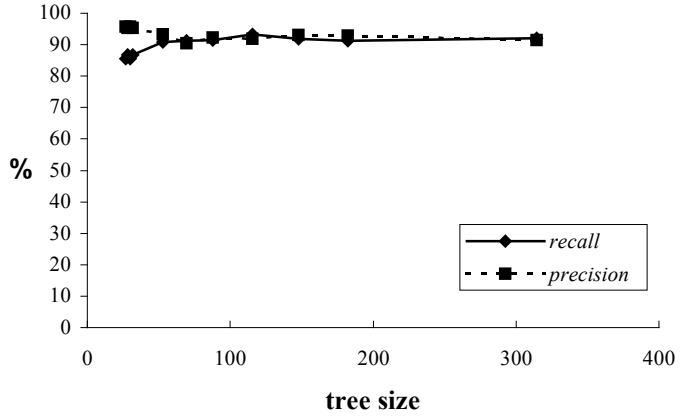

Figure 3. Results for the person phrases, in the English text

Interestingly, recall and precision are at similar levels, both for organisations and persons. Furthermore, they do not seem to be affected significantly by changes in the size of the decision tree, with the exception of very small trees, i.e., below 50 nodes. Both this characteristics indicate robustness in the performance of the NER task.

4.2 Greek Named-Entity Recognition

The same experiment was repeated for Greek text of similar size and of the same domain. 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. Figures 4 and 5 present the results of the experiment in Greek text for the two different types of named entity. Clearly the results are worse than in English, especially for organisations due to the reduced size of the gazetteer lists. Despite that, they are again better than those of the manually constructed rules.

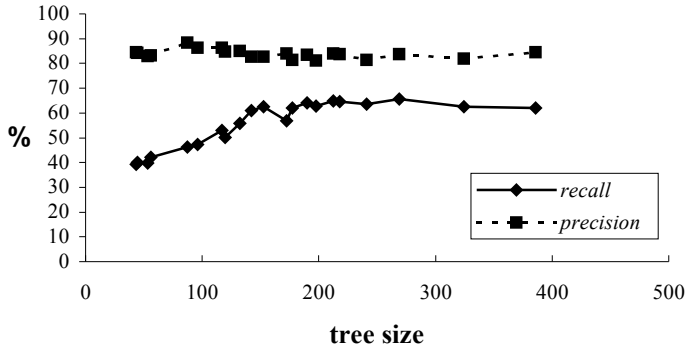

Figure 4. Results for the organisation phrases, in the Greek text

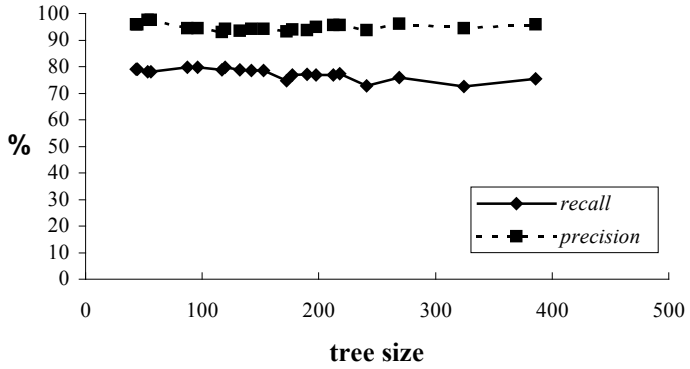

Figure 5. Results for the person phrases, in the Greek text

Particularly worrying are the recall results for the most difficult of the two entity types, i.e., organisations. Recall of organisations starts off at just above 40% and stabilises around 60% for trees above 150 nodes. Even at that level it is almost 20 pp. lower than in the English text. Precision for organisations does not fluctuate significantly for different tree sizes and is at a similar level as for English text, i.e., around 80%. For person names the results are better. In comparison to the English text, recall of person names is lower, around 80%, while precision is higher, around 95%. Overall the results are not discouraging, given the limited use of linguistic resources in this experiment.

5 Concluding Remarks and Further Work

The customisation of language engineering tools in new languages and domains is becoming essential, as the need for automatic text analysis and information extraction in various, unrelated domains increases. NER is an important component of most of these tools. Therefore, it is worth putting effort into speeding up the customisation of NER systems to new languages and domains.

We developed the Greek GIE NER system based on the English system VIE. The language engineering platform of GATE, where VIE was developed, facilitated the addition of new Greek modules (tokeniser, sentence splitter) and the customisation of English modules (training the part of speech tagger, replacing the language specific resources of the gazetteer lookup and the named-entity parser). We evaluated both systems on texts of similar size and of the same domain (management succession events) and compared them. The performance of VIE is significantly higher than that of GIE. This is mainly due to the limited size of the resources available for the Greek language. However manual enrichment of resources is a very time-consuming process. That's why we decided to examine methods that could automate the construction of such resources for a particular application in a specific language.

We applied a popular machine learning technique to the construction of NER grammars and shown that it can do better than equivalent manually constructed tools. 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 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. This 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.

6 References

[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.

[Cunningham *et al.* 1996] Cunningham, H., Wilks, Y., Gaizauskas, R., GATE - a General Architecture for Text Engineering, In Proceedings of 16th Conference on Computational Linguistics (COLING'96), 274-279, 1996.

[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.

[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].

[Gaizauskas & Wilks 1997] Gaizauskas, R., Wilks,Y. Information Extraction beyond Document Retrieval, University of Sheffield, Dept. of Computer Science, CS-97-10, 1997.

[Gazdar & Mellish 1989] Gazdar G. and Mellish C, 1989. Natural Language Processing in Prolog. Addison-Wesley, 1989.

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

[Petasis *et al.* 1999] Petasis G., Paliouras G., Karkaletsis V. Spyropoulos C.D. "Resolving Part-Of-Speech Ambiguity in the Greek Language Using Learning Techniques". In Proceedings of ACAI'99 Workshop on "Machine Learning in Human Language Technology", Chania, Greece, July 1999.

[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.

[Sekine 1998] Sekine, S., NYU: Description of the Japanese NE system used for MET-2. In [DARPA 1998].

[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.