RULE-BASED NAMED ENTITY RECOGNITION FOR GREEK FINANCIAL TEXTS

Dimitra Farmakiotou[♣], *Vangelis Karkaletsis*[♣], *John Koutsias*[♣], *George Sigletos*[♣], *Constantine D. Spyropoulos*[♣] *and Panagiotis Stamatopoulos*[♦]

[♣]Institute of Informatics and Telecommunications, N.C.S.R "Demokritos"
15310 Aghia Paraskevi, Athens, Greece
{dfarmak, vangelis, jkoutsi, sigletos, costass}@iit.demokritos.gr

[♦]Department of Informatics, University of Athens
TYPA Buildings, Panepistimioupolis, 157 71 Athens, Greece
T.Stamatopoulos@di.uoa.gr

ABSTRACT

The identification and classification of proper names (named entity recognition) is considered an important task in the area of Information Retrieval and Extraction. A typical named entity recognition (NER) system mainly consists of a lexicon and a grammar. When moving to a new domain, these lexical resources should be customised, either manually or exploiting machine learning techniques. In this paper, we present a NER system based on hand crafted lexical resources. The system is part of a Greek information extraction system and was tested on a Greek corpus of financial news with satisfactory results.

Keywords: information extraction, named entity recognition, pattern matching

1. INTRODUCTION

Information Extraction (IE) is the task of automatically extracting information of interest from unconstrained text creating a structured representation of this information. An IE task involves two main sub-tasks: the recognition of the named entities involved in an event and the recognition of the relationships holding between named entities in that event. Named-entity recognition (NER) involves the identification of proper names in text and their classification as different types of named entity, e.g. persons, organizations, locations. NER is evaluated as a separate task at the international evaluation conferences for IE (Message Understanding Conferences – MUC [4]). NER is not only an important subtask in IE [8] but also in lexical acquisition for the development of robust natural language processing systems [5]. Moreover NER may prove fruitful for tasks such as indexing of documents and maintenance of data bases containing information for the identified named entities.

The lexical resources that are typically included in a NER system are a lexicon, in the form of gazetteer lists, and a grammar, responsible for recognizing the entities that are either not in the lexicon or appear in more than one gazetteer lists. Existing NER systems belong into the following types:

- Systems based on hand crafted grammars and gazetteers. Typical examples are LaSIE I [7], LaSIE II [9] and FASTUS [1]
- Systems exploiting Machine Learning techniques for the automatic acquisition of NER lexical resources. MENE [3] and Nymble [2] are examples of such systems.
- Systems combining the two previous approaches like the LTG system of the University of Edinburgh [11].

In this paper, we describe a rule-based Greek NER system with hand crafted grammars and gazetteers. It forms a part of a larger Greek information extraction system, being developed in the context of the R&D project MITOS[1]. The NER system is described in Section 2. The results of its application in a corpus of financial news are discussed in Section 3. The paper concludes presenting our future plans in Section 4.

[1] MITOS (EPET II – 1.3 – 102) is an R&D project on information filtering, extraction and data mining, funded partially by the Greek government. MITOS partners include NCSR "Demokritos" (coordinator), Athens Univ. of Economics & Business, Univ. of Pireaus, Univ. of Patras, KNOWLEDGE S.A., SENA, Kapa-TEL.

2. THE NER SYSTEM

The NER system involves modules for linguistic preprocessing, named entity identification and classification.

2.1 Linguistic Preprocessing

Preprocessing involves tokenization, sentence splitting, part of speech tagging, stemming and matching against lists of known names (gazetteer lookup) in the body of the text.

The tokenizer identifies tokens tagging them with a token-type tag. The token-type tag set encodes graphological information (e.g. the token is comprised of lower case Latin characters).

A separate module for sentence splitting has been developed since financial news texts contain a great number of abbreviations with periods. These abbreviated forms may belong to proper names, common nouns or adjectives. Sentence boundaries are detected by means of selected tokens and token-type sequences.

For part of speech tagging, a machine learning based tagger is used [14]. This tagger has been trained and tested on Greek financial texts with an accuracy of 95%. The tag set used includes not only part of speech information, but also morphological features (gender, number, tense).

Once the text has been annotated with part of speech tags, a stemmer is used. The stemmer converts all lexical units to lowercase and unstressed, and removes the inflectional suffixes of Greek nouns and adjectives. The aim is to reduce the size of the gazetteer lists and the NER grammar for the following reasons:

- Nouns and adjectives are the two parts of speech that are more likely to be part of a proper name.
- Capitalization may occur in different forms, e.g. initial, for all the characters, for some of the characters.
- Greek is an inflected language so nouns and adjectives have different forms in different cases, numbers and genders.

So a lowercase, unstressed stem of a noun is used instead of all the forms this noun may have in different cases and numbers.

A gazetteer lookup module is used next. This identifies names by searching pre-stored lists of known proper names (gazetteers). The gazetteers of organization, person and location names contain stemmed forms and have been compiled from WWW sites and an annotated training corpus. The total number of names contained in the gazetteers is 3059. This size is rather small compared to the gazetteers compiled for English NER systems Indicatively, the gazetteers of LTG [12] and LaSIE-II [9] contain 45.000 and 23.000 names respectively.

2.2 Named Entity Identification

The identification of named-entities involves the detection of their boundaries, i.e. the start and the end of all the possible spans of tokens that are likely to belong to a named entity. Identification involves three stages: initial delimitation, separation and exclusion.

The stage of initial delimitation involves the application of very general patterns. These patterns are combinations of a limited number of words, selected types of tokens (e.g. tokens consisting of capital characters), special symbols and punctuation marks. The possible name spans delimited at this stage may include more than one names, e.g. *IBM και* (and) *Dell Computers*, non-named-entities and named entities, e.g. *Πρόεδρος Κλίντον* (President Clinton), as well as non named-entities, e.g. *TV*.

At the separation stage, possible named entities that are likely to contain more than one named entity or a named entity attached to a non named entity, are detected and attachment problems are resolved. As in English the conjunction *και* (and) and the equivalent of the English *of* genitive article *του* (male or neutral singular), *της* (female singular), *των* (plural) are ambiguous as to whether they are within name boundaries or not, e.g. *Εθνική Τράπεζα της Ελλάδος* (National Bank of Greece), *ΟΗΕ και ΝΑΤΟ* (UN and NATO). Certain appositives like *ΥΠΕΘΟ* (abbreviation for minister of Economics) are capitalized and therefore included in the possible name span, these cannot be excluded from the first stage of initial delimitation because they are often ambiguous and may refer to named entities in different texts, for instance, *ΥΠΕΘΟ* is also an abbreviation for the Ministry of Economics. Words contained in a possible name span, keywords from the context and gazetteer lookup annotations are used for determining whether an initially delimited possible name can be separated.

Capitalized words, which are common nouns (e.g. I.X. that means “private car”) or proper names not belonging to the named entities specified (e.g. names of financial products), are excluded from the list of possible named-entities during the ex-

clusion stage of named-entity identification. There are two criteria for exclusion from the possible named entity list: context and being part of a "killer" list. Suggestive context for exclusion is common names that refer to products, services or artefacts. "Killer" list includes capitalized abbreviations of common nouns, financial terms, capitalized person titles when not ambiguous and nouns commonly found in names of products, artefacts and services.

2.3 Named-Entity Classification

Once the possible named entities have been identified, classification begins. Classification involves three stages: application of rules, gazetteer-based classification, and partial matching of classified named-entities with unclassified ones.

Classification rules are hand crafted. They take into account both internal and external evidence [13], i.e. the words and symbols that comprise the possible name and the context in which it occurs. Classification rules that rely on internal evidence (e.g. company designators or first names) are applied before rules that rely on external evidence such as appositives or certain keywords preceding or succeeding a possible name. In case of ambiguity, both internal and external evidence are used for the classification. The named entities classified first are organizations because they may include person or location names (organizations are often named after their founder or the place they are located in). Their full form usually includes corporate designators such as Bank, Corp., Inc. It must also be noted that in news texts, organization names are often introduced in their full form while their truncated form is used for all the subsequent references to them. As for international well-known organizations (e.g. United Nations) only their abbreviated form (e.g. U.N.) is sufficient for the human reader. Person names are classified after organizations. Person titles and appositive phrases are mainly used for their classification. Location names are classified last since the annotations of the gazetteer are the basis for their classification. Moreover references to locations in this text genre include mainly names of countries, cities or places within Greece, so appositive phrases referring to locations are rare and the form of location names is not suggestive.

Partial matching is the last stage of NER in the body of the text. At this stage classified names are matched against unclassified ones. This stage aims at the recognition of the truncated or variable forms of names. At this stage, for instance, the organization name *ABN AMRO Bank* is matched against the unclassified name *ABN AMRO* which is the truncated form of the first.

A heuristic similar to partial matching is implemented for the recognition of named-entities in the title. The titles of the articles we use for training and evaluation purposes are upper-case. This feature deters the use of the pos tagger, as well as the detection of named-entity boundaries based on capitalization. Names recognized in the text are matched against consecutive tokens of the title for the delimitation and classification of named-entities. A shortcoming of this heuristic is that it does not allow the recognition of names which do not appear or are not recognized in the text. Fortunately, there are few such cases.

3. THE EXPERIMENT

For the purposes of the experiment presented in this paper we used corpora of financial news provided by the Greek company Kapa-TEL. A 30.000 word corpus was used for training and a 140.000 word corpus was used for testing. Both corpora have been manually tagged with annotation tools developed for this purpose for three categories of named-entities, namely organizations, persons and locations. The guidelines followed for their tagging are similar to the ones used at the Named Entity evaluation task at MUC-7 [4].

System performance has been measured using Precision, Recall and F-measure. Recall measures the number of items of a certain named entity type correctly identified, divided by the total number of items of this type. Precision is the ratio of the number of items of a certain named entity type correctly identified to all items that were assigned that particular type by the system. F-measure combines Recall (R) and Precision (P) using the formula ((2xPxR)/(P+R)).

The best overall results (Figure 1) have been achieved for location and organization names, 0.869 and 0.824 respectively, whereas results for person names are the lowest ones. The low Recall for persons is mainly due to the fact that 33% of person names in the evaluation corpus are not Greek. Although the overall performance (F-measures) for these types of named entities is lower than the highest performance (0.94) achieved in MUC-7 for English texts [13], it is quite satisfactory compared to relevant work for Greek texts [6,10].

	Organization	Person	Location
Precision	0.898	0.875	0.905
Recall	0.842	0.765	0.756
F-measure	0.869	0.816	0.824

Figure 1

More than 30% of the mistakes in the recognition of organization names are due to spelling and preprocessing errors (see Fig. 2). The preprocessing modules that contribute most to errors are the sentence splitter and the part of speech tagger. Wrong output from the sentence splitter results in wrong identification of name boundaries. Incorrect tagging of capitalized nouns affects the output of the stemmer and the gazetteer lookup module.

Errors due to	**Organization**	**Person**	**Location**
Ambiguity	6%	42%	40%
Other	59%	31%	32%
Preprocessing	14%	17%	14%
Spelling	21%	10%	14%

Figure 2

4. CONCLUSIONS

The NER system presented in this paper is based on hand crafted lexical resources and was tested on a Greek corpus of financial news with satisfactory results. We are currently working on the improvement of the preprocessing module in order to reduce its negative effects to system's performance. Our next step is to combine the hand crafted NER system with one created using machine learning techniques. The final NER system will be integrated at the complete IE system, which is currently being developed in the context of the MITOS project.

5. REFERENCES

[1] Appelt D., Hobbs J. R., Bear J., Israel D., Kameyama M., Kehler A., Martin D., Myers K. and M. Tyson (1995), SRI International FASTUS system MUC-6 Test Results and Analysis. *Proc. of the 6th Message Understanding Conference.*

[2] Bikel D., Miller S., Schwartz R. and R., Weischedel (1997) Nymble: A High-Performance Learning Name-finder. *Proc. of the 5th Conference on Applied Natural Language Processing*.

[3] Borthwick, J. Sterling, E. Agichtein and R. Grishman (1998) "NYU: Description of the MENE Named Entity System as Used in MUC-7" *Proc. of the 6th Message Understanding Conference.*

[4] Chinchor, N. (1997) *MUC-7 Named Entity Task Definition*, v.3.5. http://www.muc.saic.com

[5] Coates-Stephens, S. (1992) *The Analysis and Acquisition of Proper Names for Robust Text Understanding*. PhD thesis, Dept. of Computer Science, City University, London.

[6] Demiros, I. Boutsis, S., Giouli, V., Liakata, M., Papageorgiou, H. and S.Piperidis, Named Entity Recognition in Greek Texts. *Proc. of the 2nd International Conference on Language Resources and Evaluation (LREC-2000)*, Athens, pp.1223-1228.

[7] Gaizauskas, R., Wakao, T., Humphreys, K., Cunningham, H., and Y., Wilks, "University of Sheffield: Description of the LaSIE System as used for MUC-6". *Proc. of the Sixth Message Understanding Conference (MUC-6)*. Morgan Kaufmann, 1995, pp. 207-220.

[8] Grishman, R. Information Extraction: Techniques and Challenges (1997) in Pazienza M-T. ed. *Information Extraction: A Multidisciplinary Approach to an Emerging Information Technology, SCIE-97, Frascatti*, Italy, July 1997, pp. 10-26.

[9] K. Humphreys, R. Gaizauskas, S. Azzam, C. Huyck, B. Mitchell, H. Cunningham, Y. Wilks (1998), University of Sheffield: Description of the LaSIE-II System as Used for MUC-7. *Proc. of the 7th Message Understanding Conference.*

[10] Karkaletsis V., Paliouras G., Petasis G., Manousopoulou N., Spyropoulos C. (1999), Named-Entity Recognition from Greek and English Texts. *Journal of Intelligent and Robotic Systems 26, pp. 123-135*.

[11] Mikheev A., Grover C. and Moens M. (1998), Description of the LTG System Used for MUC-7" *http://muc.saic.com/proceedings/muc_7_toc.html*

[12] Mikheev A., Grover C. and Moens M. (1998), Named Entity Recognition Without Gazetteers *Proc. of the 7th Message Understanding Conference*

[13] McDonald, D. (1996), Internal and External Evidence in the Identification and Semantic Categorization of Proper Names. In B. Boguraev & J. Pustejovski (eds.) *Corpus Processing for Lexical Acquisition.* MIT Press, pp 21–39.

[14] G. Petasis, G. Paliouras, V. Karkaletsis, C.D. Spyropoulos and I. Androutsopoulos, Using Machine Learning Techniques for Part-of-Speech Tagging in the Greek Language, *Proc. of the 7th Hellenic Conference on Informatics*, Ioannina, Greece, 1999.