Adaptive, Multilingual Named Entity Recognition in Web Pages

Georgios Petasis[1], Vangelis Karkaletsis[1], Claire Grover[2], Benjamin Hachey[2], Maria-Teresa Pazienza[3], Michele Vindigni[3], Jose Coch[4]

Abstract. Most of the information on the Web today is in the form of HTML documents, which are designed for presentation purposes and not for machine understanding and reasoning. Existing web extraction systems require a lot of human involvement for maintenance due to changes to targeted web sites and for adaptation to new web sites or even to new domains. This paper presents the adaptive, multilingual named entity recognition and classification (NERC) technologies developed for processing web pages in the context of the R&D project CROSSMARC. The evaluation results demonstrate the viability of our approach.

1 INTRODUCTION

A number of systems have been developed to extract structured data from web pages. Such systems commonly include a set of wrappers that extract the relevant information from multiple web sources and a mediator that presents the extracted information in response to the users' requests. Most of the existing systems use delimiter-based approaches, which have proven to be very efficient with rigidly structured pages but they are not applicable to descriptions written in free text and they suffer from maintainability problems due to changes in supported web sites or the addition of new ones.

These problems were the motivation for the European funded R&D project CROSSMARC[5], which applies state-of-the-art language engineering, machine learning and ontology-based tools and techniques to develop technology for Web information retrieval and extraction. The system's approach to information extraction (IE) relies on a pipeline of three components: a *NERC component*, a *demarcator* and a *fact extraction (FE)* component. NERC identifies domain-specific named entities in pages from different sites; the demarcator groups the identified entities into products/offers inside the page. Then, FE identifies domain-specific facts, i.e. assigns domain-specific roles to some of the entities identified by the NERC component.

Although NERC is a familiar task within the IE research community, our work advances the state of the art as it presents a thorough evaluation of three different NERC technologies (from rule-based to hybrid to machine learning) with different adaptation strategies on two thematic domains, across four languages. Additionally, the selection of thematic domains (laptop offers, job offers) which involve a great variety of entity types (compared for instance to news articles used in several NERC applications), along with the fact that web documents are processed instead of raw text, raised several significant and interesting challenges both for implementation and for evaluation.

[1] Institute of Informatics and Telecommunications, NCSR "Demokritos", {petasis, vangelis}@iit.demokritos.gr

[2] Division of Informatics, University of Edinburgh, {grover, bhachey}@ed.ac.uk

[3] D.I.S.P., Universita di Roma "Tor Vergata", {pazienza, vindigni}@info.uniroma2.it

[4] Lingway, Jose.Coch@lingway.com

[5] http://www.iit.demokritos.gr/skel/crossmarc.

2 THE MULTILINGUAL NERC

Our multi-lingual IE integrates four language-specific sub-systems which operate as autonomous processors. A proxy mechanism, the Information Extraction Remote Invocation, was developed for interfacing with the IE sub-systems. This module takes the XHTML pages produced by the web pages collection system and routes them to the corresponding language-specific IE sub-system according to their language (see Fig. 1).

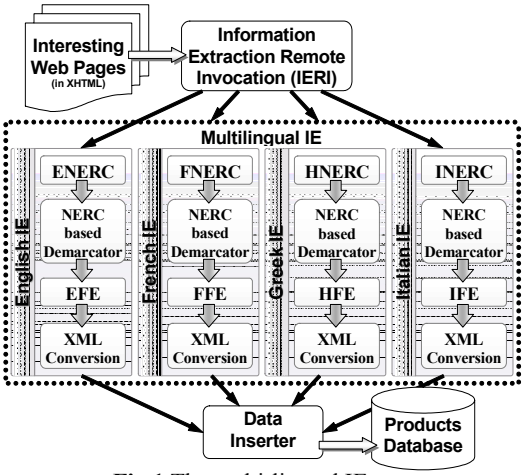

Fig.1 The multi-lingual IE system.

2.1 English NERC Customization

The ENERC module takes a machine learning approach to named entity recognition formulating it as a word tagging problem. Two named entity taggers have been developed using the C&C tagger [1] and the openNLP maximum entropy software. The customization strategy is a matter of annotating new training material and then training the classifier.

2.2 Hellenic NERC Customization

The HNERC module is built entirely on machine learning, as it uses a combination of several machine learning techniques. Its architecture can be decomposed into four subsystems. The first subsystem performs lexical pre-processing, such as tokenisation, sentence boundary identification, part-of-speech tagging and

gazetteer lookup. The second subsystem, viewing NERC as a word tagging problem it operates over word tokens and applies five independent taggers, and a simple majority voting process. The third subsystem, views NERC as a classification problem of phrases that possibly are names of entities. It operates over phrases identified with an automatically induced grammar from the training corpus and uses a decision tree classifier to recognise entities. The fourth subsystem combines the results of the 2nd and 3rd subsystems, and performs some basic filtering over their results.

2.3 French NERC Customization

The FNERC team has developed a customization methodology, which uses a tool based on machine learning techniques to help the human expert adapt the existing FNERC to a new domain. The idea here is that it is very tedious and time-consuming for a human to write easy rules, but it is very hard for a totally automatic system to write difficult rules: the solution is thus a semi-automatic customization tool. Based on the human-annotated corpus, the machine-learning module produces a first version of human-readable rules plus several useful lists of examples and counter-examples to possible rules and relevant contexts. The human expert then modifies the rule set appropriately.

2.4 Italian NERC Customization

The INERC component has a modular architecture with processing elements being general and reusable in new domains. Customization can be restricted to the knowledge bases (i.e. domain ontology, the Italian lexicon and terminology) used by the system.

INERC features a pipeline of linguistic processors driven by a set of XSLT transformations which provide the control strategy to browse into the different document sections and separate the layout specific information from the textual content. The pipeline includes tokenization, terminological analysis, lexicon lookup, numeric entity recognition and ontology-driven entity classification. The customization method which has been implemented for INERC involves a statistically driven process of generalization from the annotated material to increase coverage of the observed (linguistic) phenomena. This information is then used to automatically tune the lexical resources to new domains, building a set of (possibly partial) entries that bear strong semantic evidence of target categories.

3 RESULTS

3.1 Experimental Setting

We specified a common methodology for the collection of the necessary training and testing corpora for each domain and each language, which allows us to make a comparison of the results of the four language-specific NERC components. This methodology is comprised of two parts. First, we identify interesting characteristics of product descriptions and collect relevant statistics from product descriptions for at least 50 different sites per language. In the second part of the collection process, we gather pages and create training and testing materials with a representative distribution according to the identified domain statistics.

The next stage is corpus annotation. We devised a common methodology and developed an annotation tool. The annotation is based on a set of guidelines developed for the specific domain. Corpora of web pages for the two domains of the project were collected and annotated for each language using the above methodologies and the annotation tool. Table 1 provides the total number of named entities included in the testing corpus along with the number of offerings (in parentheses) per language and domain.

Table 1. Laptop and Job offer corpora counts: entities (offerings).

	1st Domain	*2nd Domain*
English	5111 (423)	842 (110)
French	2400 (204)	1738 (166)
Hellenic	1759 (136)	757 (128)
Italian	3296 (267)	1170 (156)

3.2 Evaluation

In evaluating the mono-lingual NERC systems we follow the standard practice in the IE field of comparing system output against the hand-annotated gold-standard and measuring *precision* and *recall* for each category of named entity. The standard way to incorporate precision and recall into a single score is to compute *f-measure*. The IE community generally reports the harmonic mean which weights recall and precision equally (i.e. F = 2*(recall*precision)/(recall+precision). Table 2 shows f-measure scores across all categories of named entity for each of the four systems in the two domains of the project.

Table 2. Overall NERC Evaluation Results (f-measure)

	1st Domain	*2nd Domain*
ENERC	0.73	0.59
FNERC	0.77	0.75
HNERC	0.86	0.68
INERC	0.82	0.77

4 CONCLUDING REMARKS

The move from textual domains to web pages has been complex and challenging. Web pages differ from more standard text types in terms of both content and presentation style. These differences can affect the performance of standard NLP techniques. The project partners have ported their NERC technologies to web pages taking into account the different document genres. Our approach compares favourably with other methods of information extraction from Web pages, such as wrapper induction, because it is not site-specific and it can be used on pages with irregular formats which have not been seen before in the training material. Our multi-lingual system has considered adaptability a key design point and is rapidly extensible both to new languages and to new domains. Comparing the NERC systems across the two domains, the general conclusion to be drawn is that the machine learning approaches are capable of performing as well as the rule-based ones if enough domain-specific resources or training material is available. For the question of customization, given the difficulties associated with gathering training material, an interactive approach, where machine learning assists the human developer of rule sets, appears to be a very promising route to investigate.

5 REFERENCES

[1] J. R. Curran & S. Clark. 2003. Language Independent NER using a Maximum Entropy Tagger. CoNLL 2003.