Annotating Web pages for the needs of Web Information Extraction applications

Georgios Sigletos, Dimitra Farmakiotou, Kostas Stamatakis,
Georgios Paliouras, Vangelis Karkaletsis
Institute of Informatics and Telecommunications, N.C.S.R. "Demokritos",
P.O. BOX 60228, Aghia Paraskevi, GR-153 10, Athens, Greece
+302106503215
{sigletos, dfarmak, kstam, paliourg, vangelis}@iit.demokritos.gr

ABSTRACT
This paper outlines our approach to the creation of annotated corpora for the purposes of Web Information Extraction, and presents the Web Annotation tool. This tool enables the annotation of Web pages from different domains and for different information extraction tasks providing a user-friendly interface to human annotators. Annotated information is stored in a representation format that can easily be exploited.

1. Introduction
Web *Information Extraction* (IE) systems rely on *patterns* that extract relevant information from a domain specific collection of documents and fill the slots of a predefined template. These patterns can be created either manually by knowledge engineers (*knowledge-based)* or automatically by exploiting *machine-learning* (ML) techniques. For the latter, the term *wrapper-induction* [6] is often used. In both cases, a domain specific collection of Web pages (*corpus*) annotated by domain experts, with information relevant to the extraction task, is required.

This paper presents a domain-independent methodology for annotating Web pages *specifically* for the needs of different IE applications and the corresponding tool that was implemented in the context of the CROSSMARC research project [3].

2. Our approach to Web IE
The CROSSMARC (CROSS-lingual Multi Agent Retail Comparison) project aims to employ state-of-the-art language engineering and ML techniques to achieve commercial strength technology for multilingual IE from Web pages. In CROSSMARC we follow the MUC (Message Understanding Conferences) [7] methodology dividing the IE task into a named entity recognition (NERC) task and a fact extraction (FE) task:

1. *Named Entity Recognition*: deals with the identification of specific *names, numerical, temporal expressions, etc.* (e.g. company, capacity, date).
2. *Fact extraction*: deals with the assignment of specific *roles* to some of the named entities identified in the previous step (e.g. laptop manufacturer, hard disk capacity).

Both the NERC and FE tasks are *multi-lingual*, i.e. they are performed by systems implemented for each of the four different languages of the project (English, French, Greek, Italian). *Cross-linguality* is also supported, i.e. IE is performed in one language and the results are presented in other languages. This is achieved exploiting a common *domain ontology*. The organization of the CROSSMARC ontology has been designed to be flexible enough to be applied to different domains without changing the overall structure; for this reason we planned a four layered ontological architecture: Template Model, Domain Model, Instance Layer, Lexical Layer.

The approach followed in CROSSMARC differs from *wrapper induction*, where ML techniques are used for generating delimiter-based rules for IE that do not use linguistic constraints. Furthermore, *multi-linguality* and *cross-linguality* issues have not received much attention by the wrapper induction community. The development and evaluation of IE systems requires annotated corpora. In the context of CROSSMARC, the need for consistently annotated Web pages with named entities and facts in more than language and domain, has been the motivation for the development of CROSSMARC Web page Annotation tool. The Web Annotation tool facilitates the annotation process for different languages by handling the corresponding encoding schemes. It has already been used for the annotation of named entities and facts in web pages containing laptop descriptions and it is currently being used for the annotation of web pages containing job offers, in the four different languages of the project [4].

3. Architecture of the Web Annotation tool
The Web Annotation tool is implemented in Java and it has been successfully tested on Windows and Sparc/Solaris platforms. The tool takes as input a domain specific collection of locally stored Web pages and three text files:

1. A Named Entity XML DTD describing *names*, *numerical, temporal expressions,* etc.
2. A Fact Extraction XML SCHEMA describing the *facts* assigned to named entities.
3. A domain ontology describing relations between domain facts and named entities.

The first two files contain information relevant to the NERC and FE tasks respectively (named entities and fact labels, definitions, restrictions, etc.). The domain ontology is encoded using an XML schema. For the maintenance and checking of the ontology, we use the Protégé 2000 system [8] and from this we export the ontology to XML. The ontology XML file contains information necessary for the annotation of normalized named entity values.

WWW 2003, May 20-24, 2003, Budapest, Hungary.
ACM xxx.

Corpora annotated for normalization are useful for supporting cross-linguality in the context of CROSSMARC. Adaptation of the tool to the annotation requirements of a new domain is accomplished with the provision of a named entity DTD, a Fact Extraction Schema, and Ontology for the new domain.

When the user completes the annotation of a page, the annotations are saved in a simple text file, using the same encoding as the corresponding Web page. Annotations are stored in separate lines, each containing information about the *starting* and *ending* offset of the annotation in the page, the *named entity* label, the *fact*, the *normalized value according to the ontology* and the *product description* that the annotation belongs to. A special mechanism was implemented for *mapping* the offsets of the selected text in the viewer of the tool, to the corresponding offsets in the source code of the Web page, and vice versa. The form of the annotation file allows the easy transformation of the annotations to different formats. In practice, annotations have been transformed to TIPSTER [9] and XML format according to the project's needs. Existing annotation files can be reloaded to the tool allowing the user to edit annotations.

4. Using the Web Annotation tool

The tool provides a user-friendly interface to the human annotator, which is depicted in Figure 2:

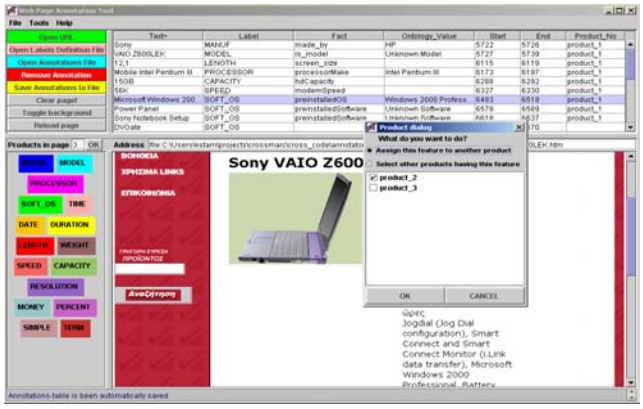

Figure 2. The graphical interface of the tool

Annotating *named entities* requires selecting text and clicking the appropriate label button on the left panel of the tool. Named entity annotations are added to the table on the top of the tool. By clicking on the appropriate row of the table, the corresponding text is highlighted on the page viewer. Furthermore, by clicking the appropriate label button, all existing annotations of the same label, i.e. the same named entity type, are highlighted on the page viewer. In order to assign a *fact* to a named entity, the user must click the left mouse button on the corresponding "Fact" column in the annotations table. A special fact dialog appears containing both the fact and a list of ontology values –from the domain ontology- for the corresponding fact. Furthermore, the user may assign product information to each fact annotation, in case a page contains more than one product description. Distinguishing between different product descriptions is an important issue during fact annotation.

5. Related Work

A variety of tools have been developed for the annotation of Web pages, including [1,2,10]. Some of them are extensions of existing browsers, such as [10]. Rather than a general-purpose annotation tool, the departure of our tool is that it is targeted to the annotation of corpora specifically for the needs of Web IE applications. Our tool supports *multi-lingual* annotation by exploiting different encoding schemes and *cross-linguality* by allowing the assignment of domain ontology values to certain fact labels. Furthermore, the tool can be easily adapted to new domains. As far as we know, other Web IE systems (e.g. [5, 6]) do not report any tool that supports similar capabilities.

6. Future Work

We plan to semi-automate the annotation process exploiting a combination of linguistic and machine-learning based techniques. Finally, we also plan to evaluate the user-friendliness of the tool.

7. ACKNOWLEDGMENTS

CROSSMARC is an R&D project of the IST Programme of the European Union (IST 2000-25366). We would like to thank the annotation teams of Edinburgh University, VeltiNet, Roma Tor Vergata University, as well as the annotation team of our laboratory for their comments and suggestions.

8. REFERENCES

[1] *Amaya*, http://www.w3.org/Amaya.

[2] *Annotea*, http://www.w3.org/2001/Annotea/.

[3] CROSSMARC, http://www.iit.demokritos.gr/skel/crossmarc/

[4] Grover C., McDonald S., Gearailt D.N., Karkaletsis V., Farmakiotou D., Samaritakis G., Petasis G., Pazienza M.T., Vindigni M., Vichot F., Wolinski F.. *Multilingual XML-based Named Entity Recognition for E-Retail Domains*. Proc. Of the LREC –2002, Las Palmas, May 2002.

[5] Knoblock C., Minton S., *The Ariadne approach to Web Information Integration*, IEEE Intelligent Systems 13(5), September/October 1998.

[6] N. Kushmerick. *Wrapper Induction for Information Extraction*. PhD Thesis. Dept. of Computer Science. U. of Washington. 1997

[7] *MUC-7,* http://www.itl.nist.gov/iaui/894.02/related_projects/muc.

[8] Protégé , http://protege.stanford.edu/index.html

[9] TIPSTER, http://cs.nyu.edu/cs/faculty/grishman/tipster.html

[10] Röscheisen M., Winograd T., Paepcke A., *Content Rating and Other Third-Party Value-Added Applications for the WWW*. D-Lib Journal. CNRI, August, 1995.