Demonstration of the CROSSMARC System

Vangelis Karkaletsis[†], Constantine D. Spyropoulos[†], Dimitris Souflis[°], Claire Grover[*], Ben Hachey[*], Maria Teresa Pazienza[◇], Michele Vindigni[◇], Emmanuel Cartier[‡], José Coch[‡]

[†]Institute for Informatics and Telecommunications, NCSR "Demokritos"
{vangelis, costass}@iit.demokritos.gr

[°]Velti S.A.
Dsouflis@velti.net

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

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

[‡]Lingway
{emmanuel.cartier, Jose.Coch}@lingway.com

1 Introduction

The EC-funded R&D project, CROSSMARC, is developing technology for extracting information from domain-specific web pages, employing language technology methods as well as machine learning methods in order to facilitate technology porting to new domains. CROSSMARC also employs localisation methodologies and user modelling techniques in order to provide the results of extraction in accordance with the user's personal preferences and constraints. The system's implementation is based on a multi-agent architecture, which ensures a clear separation of responsibilities and provides the system with clear interfaces and robust and intelligent information processing capabilities.

2 System Architecture

The CROSSMARC architecture consists of the following main processing stages:

- Collection of domain-specific web pages, involving two sub-stages:
  - domain-specific web crawling (focused crawling) for the identification of web sites that are of relevance to the particular domain (e.g. retailers of electronic products).
  - domain-specific spidering of the retrieved web sites in order to identify web pages of interest (e.g. laptop product descriptions).
- Information extraction from the domain-specific web pages, which involves two main sub-stages:
  - named entity recognition to identify named entities such as product manufacturer name or company name in descriptions inside the web page written in any of the project's four languages (English, Greek, French, Italian) (Grover et al. 2002). Cross-lingual name matching techniques are also employed in order to link expressions referring to the same named entities across languages.
  - fact extraction to identify those named entities that fill the slots of the template specifying the information to be extracted from each web page. To achieve this the project combines wrapper-induction approaches for fact extraction with language-based information extraction in order to develop site independent wrappers for the domain examined.
- Data Storage, to store the extracted information (from the web page descriptions in any of the project's four languages) into a common database.
- Data Presentation, to present the extracted information to the end-user through a multilingual user interface, in accordance with the user's language and preferences.

As a cross-lingual multi-domain system, the goal of CROSSMARC is to cover a wide area of possible knowledge domains and a wide range of conceivable facts in each domain. To achieve this we construct an ontology of each domain which reflects a certain degree of domain expert knowledge (Pazienza et al. 2003). Cross-linguality is achieved with the lexica, which provide language specific synonyms for all the ontology entries. During information extraction, web pages are matched against the domain ontology and an abstract representation of this real world information (facts) is generated.

As shown in Figure 1, the CROSSMARC multi-agent architecture includes agents for web page collection (crawling agent, spidering agent), information extraction, data storage and data presentation. These agents communicate through the blackboard. The Crawling Agent defines a schedule for invoking the focused crawler which is

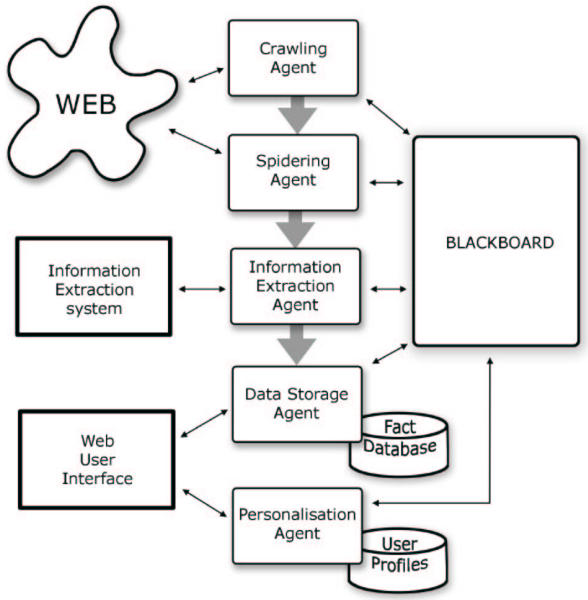

Figure 1: Architecture of the CROSSMARC system

written to the blackboard and can be refined by the human administrator. The Spidering Agent is an autonomous software component, which retrieves sites to spider from the blackboard and locates interesting web pages within them by traversing their links. Again, status information is written to the blackboard.

The multi-lingual IE system is a distributed one where the individual monolingual components are autonomous processors, which need not all be installed on the same machine. (These components have been developed using a wide range of base technologies: see, for example, Petasis et al. (2002), Mikheev et al. (1998), Pazienza and Vindigni (2000)). The IE systems are not offered as web services, therefore a proxy mechanism is required, utilising established remote access mechanisms (e.g. HTTP) to act as a front-end for every IE system in the project. In effect, this proxy mechanism turns every IE system into a web service. For this purpose, we have developed an Information Extraction Remote Invocation module (IERI) which takes XHTML pages as input and routes them to the corresponding monolingual IE system according to the language they are written in. The Information Extraction Agent retrieves pages stored on the blackboard by the Spidering Agent, invokes the Information Extraction system (through IERI) for each language and writes the extracted facts (or error messages) on the blackboard. This information will then be used by the Data Storage Agent in order to read the extracted facts and to store them in the product database.

3 The CROSSMARC Demonstration

The first part of the CROSSMARC demonstration is the user-interface accessed via a web-page. The user is presented with the prototype user-interface which supports menu-driven querying of the product databases for the two domains. The user enters his/her preferences and is presented with information about matching products including links to the pages which contain the offers.

The main part of the demonstration shows the full information extraction system including web crawling, site spidering and Information Extraction. The demonstration show the results of the individual modules including real-time spidering of web-sites to find pages which contain product offers and real-time information extraction from the pages in the four project languages, English, French, Italian and Greek. Screen shots of various parts of the system are available at `http://www.iit.demokritos.gr/skel/crossmarc/demo-images.htm`

Acknowledgments

This research is funded by the European Commission (IST2000-25366). Further information about the CROSSMARC project can be found at `http://www.iit.demokritos.gr/skel/crossmarc/`.

References

C. Grover, S. McDonald, V. Karkaletsis, D. Farmakiotou, G. Samaritakis, G. Petasis, M.T. Pazienza, M. Vindigni, F. Vichot and F. Wolinski. 2002. Multilingual XML-Based Named Entity Recognition In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2002)*.

A. Mikheev, C. Grover, and M. Moens. 1998. Description of the LTG system used for MUC-7. In *Seventh Message Understanding Conference (MUC–7): Proceedings of a Conference held in Fairfax, Virginia, 29 April–1 May, 1998*. `http://www.muc.saic.com/proceedings/muc_7_toc.html`.

M. T. Pazienza, A. Stellato, M. Vindigni, A. Valarakos, and V. Karkaletsis. 2003. Ontology integration in a multilingual e-retail system. In *Proceedings of the Human Computer Interaction International (HCII'2003), Special Session on "Ontologies and Multilinguality in User Interfaces*.

M. T. Pazienza and M. Vindigni. 2000. Identification and classification of Italian complex proper names. In *Proceedings of ACIDCA2000 International Conference*.

G. Petasis, V. Karkaletsis, G. Paliouras, I. Androutsopoulos, and C. D. Spyropoulos. 2002. Ellogon: A new text engineering platform. In *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC 2002)*.