Cross-lingual Information Extraction from Web pages: the use of a general-purpose Text Engineering Platform

Georgios Petasis, Vangelis Karkaletsis and Constantine D. Spyropoulos

Software and Knowledge Engineering Laboratory,
Institute of Informatics and Telecommunications,
NCSR "Demokritos", GR-15310, Athens, Greece,
{petasis, vangelis, costass}@iit.demokritos.gr
http://www.iit.demokritos.gr/skel/

Abstract

In this paper we present how the use of a general-purpose text engineering platform has facilitated the development of a cross-lingual information extraction system and its adaptation to new domains and languages. Our approach for cross-lingual information extraction from the Web covers all the way from the identification of Web sites of interest, to the location of the domain-specific Web pages, to the extraction of specific information from the Web pages and its presentation to the end-user. This approach has been implemented in the context of the IST project CROSSMARC. The text engineering platform "Ellogon" offers functionalities that facilitated the development of core CROSSMARC components as well as their porting into new domains and languages.

1 Introduction

The extraction of information from Web sites is a complex task. Most of the information on the Web today is in the form of HTML documents, which are designed for presentation purposes and not for automatic extraction systems. The extraction task becomes even harder in a multilingual context, where descriptions in web pages are written in different languages.

A number of systems have been developed to extract structured data from web pages. Such systems mainly 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 these systems use delimiter-based approaches. Texts processed by them are assumed to convey information in a rigidly structured manner, with entities and features mentioned in a fixed order (e.g. in a job offer description, job title always followed by job requirements and contact details), and fixed strings or mark-up acting as delimiters. Though the techniques of delimiter-based approaches have proven to be very efficient with rigidly structured pages, they are not applicable to descriptions written in free linguistic form. In the context of the IST project CROSSMARC[1], our aim was to implement techniques that can operate on pages without a standardised format (structured, semi-structured of free-text pages), as well as on pages from web sites that have not been represented in the training corpus.

CROSSMARC approach covers all the way from the identification of Web sites of interest (i.e. that contain Web pages relevant to a specific domain) in various languages, to the location of the domain-specific Web pages, to the extraction of specific information from the Web pages and its presentation to the end-user. The development of some of the core CROSSMARC components was facilitated by the exploitation of the general-purpose text engineering platform "Ellogon"[2]. Ellogon tools and functionalities were also exploited for the porting of some of the CROSSMARC components into new domains and languages.

The paper outlines first the CROSSMARC project and the Ellogon platform. It then presents the role of Ellogon in the various processing stages and procedures of the CROSSMARC project. Finally, it concludes summarising the current status of our work.

2 Related Work

The identification and retrieval of Web pages that are relevant to a particular domain or task is a complex process that has been studied by researchers in Artificial Intelligence, Web technologies and databases (e.g. Craven et al., 2000; Chakrabarti et al., 1999).

The term 'focused crawling' was introduced by (Chakrabarti et al., 1999). The system described there, starts with a set of representative pages and a topic hierarchy and tries to find more instances of interesting topics in the hierarchy by following the links in the seed pages. Pages are classified into top-

[1] http://www.iit.demokritos.gr/skel/crossmarc
[2] http://www.iit.demokritos.gr/skel/Ellogon

ics, using a probabilistic text classifier. Efficient porting to new domains and languages is a critical issue in focused crawling and spidering of web sites.

Apart from the process of identification and retrieval of Web pages that are relevant to a particular domain or task, the information management over the web requires also techniques for extracting information from the retrieved web pages. (Kushmerick, 1997) first introduced the technique of wrapper induction for Information Extraction from HTML pages. The technique works extremely well for highly structured document collections as long as the structure is similar across all documents, however it is less successful for more heterogeneous collections where structure based clues do not hold for all documents (Soderland, 1997). Most of the existing information extraction (IE) systems deal with textual content. Emphasis is given on porting IE technology into new domains either using machine learning techniques (Miller et al., 1998) or providing support for the writing of IE rules (Yangarber & Grishman, 1997). The development of some of the existing IE systems is supported by text engineering platforms, which offer an environment that facilitates the development, evaluation, deployment and maintenance of NLP applications. One such example is the ANNIE system developed over the GATE text engineering platform (Cunningham et al., 2003).

CROSSMARC methodology covers all the way from focused crawling to site-specific spidering, to information extraction from web pages. Concerning the retrieval of web pages of interest, CROSSMARC exploits machine learning techniques for web pages classification and link scoring. Concerning extraction from web pages, it combines language-based IE and wrapper induction based IE in order to develop a site-independent IE technique.

3 Overview of the CROSSMARC Project

The system developed in the context of CROSSMARC:

- exploits language technology methods as well as machine learning methods in order to facilitate the technology porting to new domains,
- exploits domain-specific ontologies and the corresponding language-specific lexica in order to facilitate the technology porting to new languages and domains,
- employs localisation and user modelling techniques in order to provide the results of web pages extraction taking into account the user's personal preferences and constraints,
- implements a multi-agent architecture which ensures clear separation of responsibilities and provides the system with clear interfaces and robust and intelligent information processing capabilities.

The main components of CROSSMARC architecture are the following (see Figure 1):

- Domain-specific Web crawling, which is managed by the Crawling Agent. The Crawling Agent consults Web information sources such as search engines and Web directories to discover Web sites containing information about a specific domain (e.g. laptops offers, job offers).
- Domain-specific spidering, which is managed by the Spidering Agent. The Spidering Agent identifies domain-specific Web pages grouped under the sites discovered by the Crawling Agent and feeds them to the Information Extraction Agent.
- Information Extraction, which is managed by the Information Extraction Agent. The Information Extraction Agent manages communication with remote information extraction systems. These systems process Web pages collected by the Spidering Agent and extract domain facts from them (Grover et al., 2002). The facts are normalised using a common domain ontology and stored in the system's database.
- Information Storage and Retrieval, which is managed by the Data Storage Agent. Its tasks consist of maintaining a database of facts for each domain, adding new facts, updating already stored facts and performing queries on the database.
- Information Presentation. The information presented to the end user can be adapted to his/her preferences. This user management is taken over by the Personalization Agent.

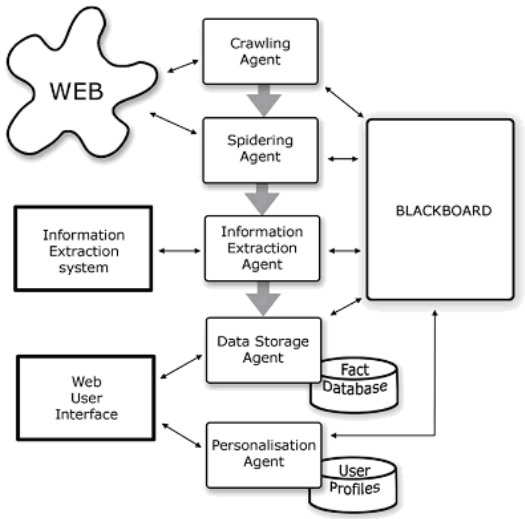

Figure 1: CROSSMARC's agent based architecture.

Our goal in CROSSMARC was to cover a wide area of possible knowledge domains and a wide range of conceivable facts in each domain, hence we imple-

mented a shallow representation of domain knowledge, the ontology of the domain (Pazienza et al., 2003). Cross-linguality is achieved through the onrology's lexica for the languages supported. During the information extraction stage, the Web pages found are matched against domain's ontology and an abstract representation of this information is generated.

4 The Text Engineering Platform "Ellogon"

Ellogon is a multi-lingual, cross-platform, general-purpose text engineering environment, developed in order to aid both researchers in the natural language field as well as companies that produce and deliver language engineering systems. Ellogon consists of mainly three subsystems (Petasis et al., 2002):

- A highly efficient core developed in C++, which implements an extended version of the TIPSTER data model. Its main responsibility is to manage the storage of the textual data and the associated linguistic information and to provide a well-defined programming interface that can be used in order to retrieve/modify the stored information.
- A powerful and easy to use graphical user interface (GUI). This interface can be easily tailored to the needs of the end user.
- A modular pluggable component system. All linguistic processing within the platform is performed with the help of external, loaded at runtime, components. These components can be implemented in a wide range of programming languages, including C, C++, Java, Tcl, Perl and Python.

Ellogon as a text engineering platform offers an extensive set of facilities, including tools for visualising textual/HTML/XML data and associated linguistic information, support for lexical resources (like creating and embedding lexicons), tools for creating annotated corpora, accessing databases, comparing annotated data, or transforming linguistic information into vectors for use with various machine learning algorithms. Additionally, Ellogon offers some unique features, like the ability to freely modify annotated textual data (with Ellogon automatically applying the required transformations on the associated linguistic information) and the ability to create stand-alone applications with customised user interfaces that perform specific tasks.

A large number of the functionalities provided by Ellogon have been exploited in the context of the CROSSMARC project. Supporting Java as a component development language has enabled the integration of the crawling and spidering agents as Ellogon components, thus easing their development as well as their deployment. The ability of Ellogon to automatically extract systems of components as stand-alone applications that run unmodified under different operating systems (Windows, Linux, Solaris, etc.) has been used extensively in CROSSMARC. In this way, various subsystems developed under Ellogon were integrated into the monolingual IE systems of the CROSSMARC partners. Furthermore, the Ellogon tools for creating vectors in various formats (e.g. WEKA ARFF, C4.5 vector format, etc.) for use with machine learning algorithms have significantly facilitated the process of training of some of the CROSSMARC components based on machine learning. Additionally, the extensive set of annotation tools offered by Ellogon (either for plain text or HTML documents) has played a central role in corpora annotation in CROSSMARC, as these tools have been used for annotating part-of-speech, named-entity, noun phrase and syntactic information on the collected web pages. Finally, functionalities like the processing and display of HTML documents, XML, DOM and XSLT support as well as the various viewers created a "comfortable" environment for CROSSMARC developers.

5 The role of "Ellogon" in CROSSMARC

The Ellogon text engineering platform offers a rich set of tools and facilities that facilitate the development of NLP systems based on machine learning. In the following sub-sections we present the contribution of Ellogon in the CROSSMARC components.

5.1 Web Pages Collection

The CROSSMARC Web Pages Collection involves two components:

- focused crawler: identifies web sites that are of relevance to the particular domain (e.g. retailers of electronic products).
- domain-specific spider: identifies web pages of interest (e.g. laptop product descriptions) within the retrieved web sites.

The focused crawler component is implemented as a meta-search engine, which exploits the topic-based website hierarchies used by various search engines and submits domain-specific queries to various search engines so as to collect web sites containing relevant to the domain information (Stamatakis et al., 2003). The returned list of web sites is filtered using a light version of the domain-specific spidering tool (NEAC). This light version of NEAC navigates the site until it finds an interesting web page. If it finds one, it considers the site as *fit* and stops navi-

navigating. If no such page is found, the whole site is navigated and if no fit page is found, the site is characterized as *unfit*.

CROSSMARC spidering tool comprises of three components:

a) Site navigation: It traverses a Web site, collecting information from each page visited, and forwarding part of the collected information to the "Page Filtering" module and another part to the "Link Scoring" module.
b) Page filtering: It is responsible for deciding whether a page is an interesting one (e.g. contains laptops offers) and therefore should be stored or not.
c) Link scoring: It validates the links to be followed, in order to accelerate site navigation (only links with score above a certain threshold are followed).

The crawler and spidering components were developed in Java as autonomous units. Both components were incorporated into Ellogon exploiting its Java support. Being an Ellogon component increases the deployment abilities of the two components, as they can be part of Ellogon generated applications, like for example an application that incorporates the crawler and the spidering tool configured for the Greek language along with the Greek IE system.

5.2 Information Extraction

Information Extraction (IE) from the domain-specific web pages collected by the crawling and spidering agents, involves two main sub-stages:

- *named entity recognition* (NERC) to identify named entities (e.g. product manufacturer name, company name) in descriptions inside the web page written in any of the project's four languages (Grover et al., 2002).
- *fact extraction* (FE) to identify those named entities that fill the slots of the template specifying the information to be extracted from each web page. For this task wrapper-induction approaches for fact extraction are combined with language-based information extraction in order to develop site-independent wrappers for the domain.

The architecture of the integrated 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 (see Figure 2).

The IE systems are not offered as Web services, therefore a proxy mechanism was 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 to a Web service. For this purpose, we developed a module named Information Extraction Remote Invocation (IERI) which takes the XHTML pages as input and routes them to the corresponding monolingual IE system according to the language they are written in. Again for deployment reasons, as it was the case for the focused crawler and the spider, this tool was embedded as an Ellogon component, offering to applications that use it the ability to remotely invoke any of the monolingual IE systems.

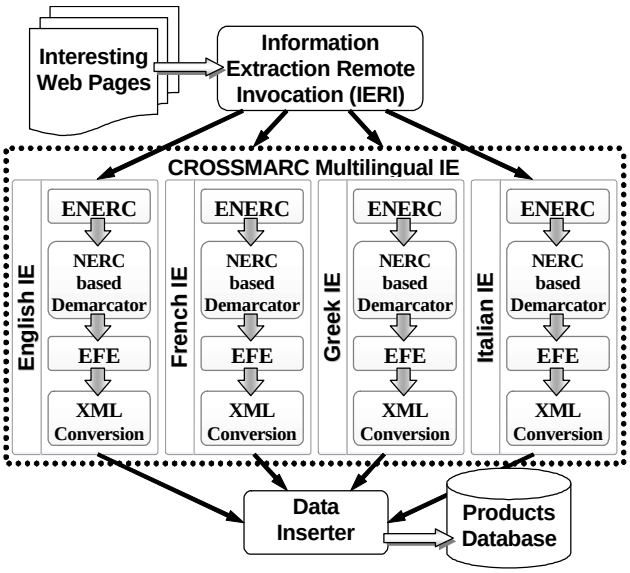

Figure 2: Architecture of the CROSSMARC IE system.

Named Entity Recognition

Although the individual NERC components differ in platforms and annotation methods, they have to produce a common output. According to CROSSMARC specifications, NERC components should add to each XHTML page they process, annotations for the named entities (NE), numeric expressions (NUMEX), time expressions (TIMEX) and terms they recognise according to a common DTD for all languages supported.

Two different NERC components have been developed for the Greek language. The first component is rule-based, targeting only the first CROSSMARC domain (laptop offers) while the second NERC component is based completely on machine learning, and can be adapted to various domains and languages. Ellogon has provided significant support in the development of both NERC components. The rule-based NERC has been mainly benefited by incorporating some ready-to-use Ellogon modules (HTokeniser, HBrill, HGazetteer) and the advanced annotation query facilities offered by Ellogon. The Hellenic NERC component that is based on machine learning (HNERC) has been built upon facilities provided by the platform in its entirety. The component uses many Ellogon modules but more importantly it takes advantage of Ellogon's ability to generate vectors to be used along with popular machine-

learning algorithms, its ability to instrument machine learning algorithms and the provided facilities for training with such algorithms.

The general architecture of both NERC components is similar: As a first step, the collected XHTML Web pages are converted into collections of Ellogon documents. Then, for each document the HNERC system applies a set of components that add Tipster-style annotations to the document. Finally, from these annotations an XML document is produced containing the output (by utilising the XML/XSLT support of Ellogon), which conforms to the NERC DTD. Additionally, the recognised named entities are marked using special tags in the original XHTML input pages. The HNERC system as appears inside Ellogon is shown in Figure 3.

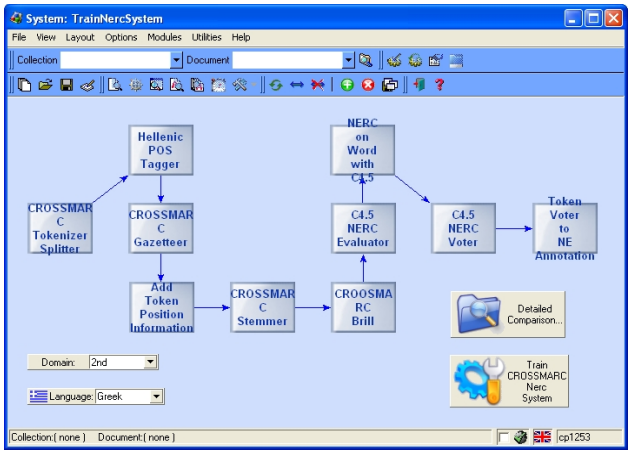

Figure 3: The HNERC system.

The Ellogon components comprising HNERC perform lexical pre-processing, gazetteer lookup and named-entity identification and classification for NE, NUMEX, TIMEX and terms. The components that depend on the domain and language (the part-of-speech tagger, the gazetteer lookup component and the NE identifier/classifier) automatically load the proper models (acquired during training), according to the documents that are currently processed. NE identification/classification is performed by 4 components, based on 2 machine learning algorithms. The first component uses Transformation-based Error-driven learning (Brill, 1995) in order to classify each word of a document into an NE category. The second and third component use decision trees (C4.5 – Quinlan, 1993) to perform the same task, with each component having different information as input. Finally, the fourth component operates as a voter, by utilising again C4.5 in order to decide upon the final classification of each word in the corpus. The implementation of these components has been extremely easy, as they heavily rely on facilities provided by Ellogon in order to interact with the two used learning algorithms.

Adapting HNERC to new domains and languages is an extremely easy process, if a training corpus is available. Ellogon infrastructure has facilitated the development of a graphical user interface through which HNERC modules can be trained (Figure 4). This interface provides abstraction over the specific training details of the various algorithms involved and resolves all dependencies among HNERC modules to ensure that all the needed information for training a module is available before its training starts.

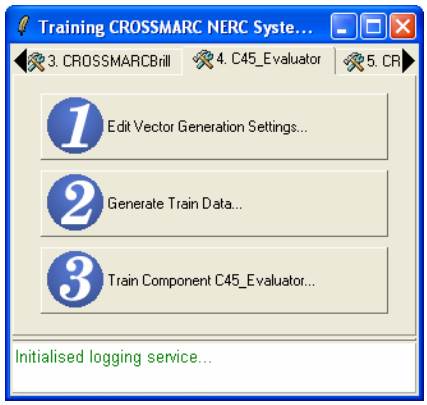

Figure 4: One of the provided facilities for adapting HNERC to new domains and languages.

HNERC technology has been tested in both CROSSMARC domains, not only for Greek, but also for the other 3 languages supported (English, French and Italian). It has also been tested in a 3rd domain (touristic packages offered by travel agencies) for Greek and English. Porting to new domains is a fairly easy process, as infrastructure is provided by Ellogon for annotating the needed training corpora, for training all the domain-specific components through a simple graphical interface (including POS tagger and gazetteers) and for evaluating the resulting system on new corpora.

Despite the fact that the HNECR system is build on top of generalised facilities provided by Ellogon, its processing performance is quite good, as Ellogon is a highly optimised platform. HNERC is able to process a collection of 50 XHTML documents in less than 2.5 minutes, requiring on average 3 seconds for each document. Finally its processing performance is the same on all domains and languages HNERC has been adapted to.

Demarcation

The Demarcation tool is responsible for locating different product descriptions inside a web page. It is quite common for web pages to contain more than one product descriptions and in many different ways. This specific tool uses heuristics in order to determine the number of descriptions and their boundaries.

The Demarcation tool is common for all languages and is developed under Ellogon in its entirety. As was also the case with the NERC component, two approaches were developed: a rule-based approach for the 1st CROSSMARC domain and an approach based on machine learning, adaptable to many domains. As the Demarcation tool must be integrated with each one of the monolingual NERC systems of CROSSMARC, a stand-alone application was created by using the relevant application generation wizards of Ellogon. The generated application has on optional user interface (Figure 5) and can run under both the Windows and Linux operating systems. As the graphical interface can be disabled when the application is executed (i.e. for incorporating the application in a command-line system), all the elements shown in the interface can be specified through special arguments.

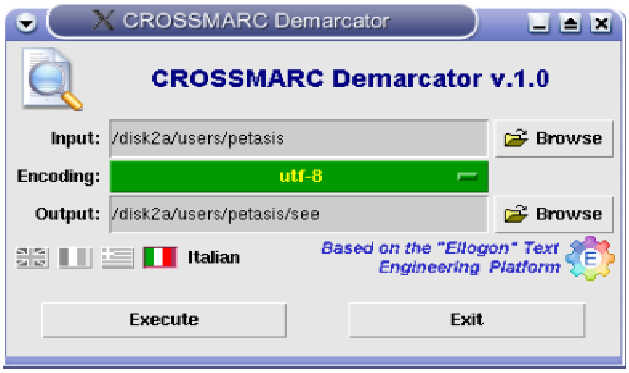

Figure 5: The user interface of the Demarcation tool.

Fact Extraction

As mentioned above, CROSSMARC approach to IE relies on a pipeline of two components: the NERC component and the FE component. The NERC component identifies domain-specific named entities in pages from different sites. The FE component identifies domain-specific facts, i.e. assigns domain-specific roles to some of the entities identified by the NERC module. The FE component is based on wrapper induction algorithms that capitalise on the page-independent named entity information, rather than relying solely on the HTML tags, which vary among pages from multiple sites.

Using Ellogon as the development platform, we implemented a version of the STALKER wrapper induction algorithm (Muslea et al., 1998) that learns single-slot extraction rules by examples annotated by the user. Four separate sets of rules were created after the training of STALKER version in the four languages of the project. Similar to the Demarcation tool, in order to integrate the language-specific fact extraction modules with the corresponding monolingual NERC systems, a stand-alone application was generated by using the relevant wizards of Ellogon. The generated application also has on optional user interface and shares the same properties as the Demarcation tool regarding its execution.

5.3 Corpus Annotation

The need for porting NERC and FE technology into new domains requires the construction of a representative training and testing corpus for each domain. For this purpose, we specified a corpus collection and a corpus annotation methodology, which is used in the four languages of the project. The corpus collection methodology aims at finding a set of characteristics for a certain domain per language, determining how each characteristic must be represented in the training and testing corpora and establishing a set of rules to be followed for the formation of training and testing corpora in the project languages for a given domain. The aim of the annotation process has been the creation of good quality corpora annotated with the named entities and facts of a given domain. Good quality annotated corpora are corpora annotated consistently and according to specific Annotation Guidelines.

In order to annotate HTML corpora the annotation facilities provided by Ellogon have been extensively used. Ellogon provides a wide range of corpora annotation tools for annotating plain textual and HTML corpora as well as tools for annotating hierarchical related information (i.e. syntax trees). The tool for annotating HTML corpora has a simple and easy to use interface: The HTML rendering is presented to the user, along with a set of buttons, each of which is associated with a specific category. The user can select portions of the rendered HTML text and classify it into one of the available categories. Additional facilities are provided for correcting mistakes or by automatically annotating all occurrences of specific text with the same category within an HTML page. A screenshot of the HTML annotation tool is in Figure 6.

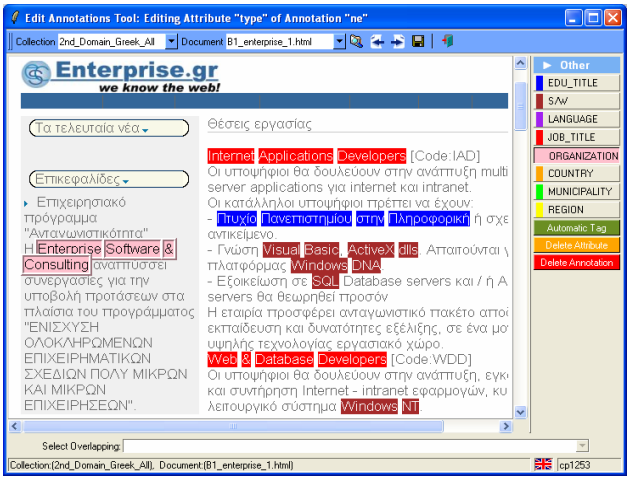

Figure 6: Annotating HTML corpora in "Ellogon".

5.4 Vector creation for machine learning algorithms

As machine learning plays an important role in NLP systems that are adaptable to more than one domains (or even languages), Ellogon provides extensive support for generating vectors from the linguistic information associated to documents, as well as for performing experiments from within Ellogon with widely used machine learning algorithms (i.e. C 4.5, etc.). Regarding vector generation, many common formats are supported (i.e. WEKA ARFF). The generated vectors are of fixed-length while many different data representations are supported (including vectors with features taken directly from the available linguistic information or vectors with features created from word occurrence frequencies based on a "bug of words" notation).

The vector creation facilities offer specialised graphical user interfaces (Figure 7) for generating vectors in an interactive mode. Additionally, these facilities can be accessed through a specialised set of functions (API) Ellogon components can use. This API allows the embedding of these facilities into components, thus enabling components to generate vectors in order to present them to a machine learning algorithm. The provision of such versatile vector generation facilities facilitated the exploitation of many widely used machine learning algorithms in the context of the CROSSMARC project.

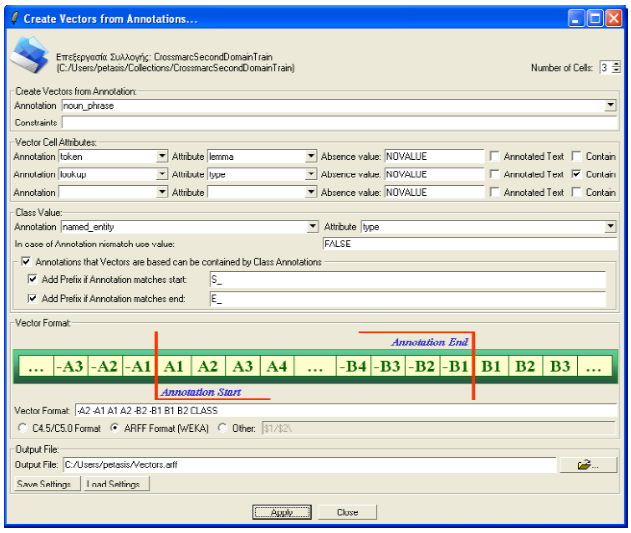

Figure 7: An "Ellogon" tool for creating vectors from Annotations.

5.5 Comparison functionalities

Ellogon provides significant infrastructure for comparing the linguistic information associated with the textual data. The Collection Comparison tool (Figure 8) can be used for comparing the linguistic information stored in a set (or collection) of documents. Various constraints regarding the information that will be compared can be specified through the graphical user interface of the comparison tool and the comparison results are presented by utilising standard figures, like recall, precision and F-measure. Additionally, the comparison tool can present a comparison log. This log is a graphical representation of the differences found during the comparison process and can provide valuable help to the user in order to locate and possibly correct the errors.

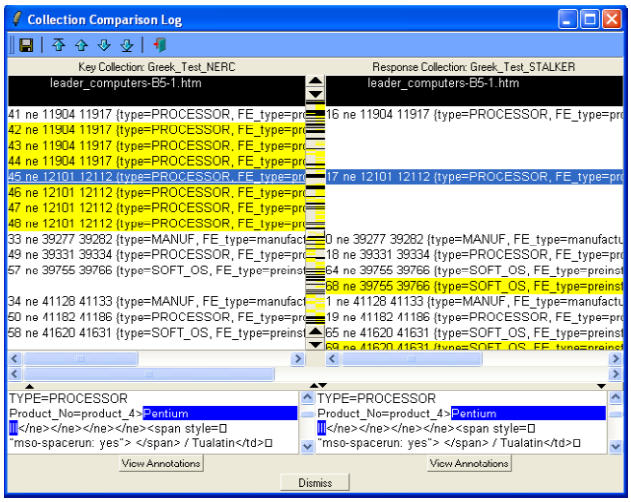

Figure 8: The "Ellogon" Corpora Comparison tool.

The comparison infrastructure has played a significant role in the evaluation of the CROSSMARC information extraction system, as the French and Hellenic information extraction systems are evaluated with the use of Ellogon and the performance of the Demarcation and Fact Extraction tools are evaluated with Ellogon for all languages. Finally, the comparison infrastructure has significantly contributed in the development of the Hellenic information extraction system, as the quick identification of errors through the comparison log has accelerated considerably the process of tuning the system into the CROSSMARC domains.

6 Conclusions

The CROSSMARC information extraction system can be perceived as an elaborate meta-search engine, which identifies domain-specific information from the Web, as it incorporates all the needed sybsystems, from the identification of Web sites of interest in various languages, to the location of the domain-specific Web pages, to the extraction of specific information from the Web pages and its presentation to the end-user.

However, the process of extracting information from Web pages poses some difficulties, as 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, which must be "ported" from raw text processing to process HTML texts. We investigated how the use of a suitable text engineering platform can ease this transition to the Web page processing. Our case study suggests that the use of such a platform can be valuable, as any language processing system can reuse and thus benefit from the provided infrastructure.

Acknowledgements

We would like to express our thanks to our partners in the CROSSMARC project, who used the Ellogon-based components and functionalities, for their useful and constructive comments. We would also like to thank our colleagues from the University of Sheffield, developers of GATE. Our cooperation in the context of the R&D projects ECRAN and GIE, a few years ago, motivated and helped us to make our first efforts towards the development of Ellogon.

References

(Brill, 1995) E. Brill, "Transformation-Based Error-Driven Learning and Natural Language Processing: A Case Study in Part of Speech Tagging". *Computational Linguistics*, 21, 1995.

(Chakrabarti et al., 1999) S. Chakrabarti, M. H. van den Berg and B. E. Dom, "Focused Crawling: a new approach to topic-specific Web resource discovery". *In Proceedings of the Eighth International World Wide Web Conference*, Toronto, Canada, May 1999.

(Craven et al., 2000) M. Craven, D. DiPasquo, D. Freitag, A. McCallum, T. Mitchell, K. Nigam and S. Slattery, "Learning to construct knowledge bases from the World Wide Web". *Artificial Intelligence*, 118 , pp. 69–113, 2000.

(Cunningham et al., 2003) H. Cunningham, D. Maynard, K. Bontcheva, V. Tablan, C. Ursu and M. Dimitrov, "Developing Language Processing Components with GATE (a User Guide)". http://gate.ac.uk/sale/tao/index.html#annie

(Grover et al., 2002) C. Grover, S. McDonald, V. Karkaletsis, D. Farmakiotou, G. Samaritakis, G. Petasis, M. T. Pazienza, M. Vindigni, F. Vichot and F. Wolinski, "Multilingual XML-Based Named Entity Recognition". *In Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC 2002)*, Las Palmas, Canary Islands, Spain, May 2002.

(Kushmerick, 1997) N. Kushmerick, "Wrapper Induction for Information Extraction". *Ph.D. thesis*, University of Washington, 1997.

(Miller et al., 1998) S. Miller, M. Crystal, H. Fox, L. Ramshaw, R. Schwartz, R. Stone, R. Weischedel and the Annotation Group, "Algorithms that learn to extract information BBN: Description of the SIFT System as used for MUC–7". *In Proceedings of the Seventh Message Understanding Conference (MUC–7)*, April, 1998.

(Muslea et al., 1998) I. Muslea, S. Minton, and C. Knoblock, "Stalker: Learning extraction rules for semistructured, web-based information sources". *In Proceedings of AAAI-98 Workshop on AI and Information Integration*, 1998.

(Pazienza et al., 2003) M. T. Pazienza, A. Stellato, M. Vindigni, A. Valarakos and V. Karkaletsis, "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"*, Heraklion, Crete, Greece, June 2003.

(Petasis et al., 2002) G. Petasis, V. Karkaletsis, G. Paliouras, I. Androutsopoulos and C. D. Spyropoulos, "Ellogon: A New Text Engineering Platform". *In Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC 2002)*, Las Palmas, Canary Islands, Spain, pp. 72-78, May 2002.

(Quinlan, 1993) Quinlan, J. R., "C4.5: Programs for machine learning". Morgan-Kaufmann, San Mateo, CA, 1993.

(Stamatakis et al., 2003) K. Stamatakis, V. Karkaletsis, G. Paliouras, J. Horlock, C. Grover, J.R. Curran, S. Dingare. "Domain-Specific Web Site Identification: The CROSSMARC Focused Web Crawler". In *Proceedings of the Second International Workshop on Web Document Analysis (WDA 2003)*, Edinburgh, UK, August 3, 2003.

(Soderland, 1997) Soderland S., "Learning to extract text-based information from the world wide web". *In Proceedings of 3rd International Conference in Knowledge Discovery and Data Mining (KDD-97)*, pp. 251–254, 1997.

(Yangarber & Grishman, 1997) Yangarber, R. and Grishman, R., "Customization of Information Extraction Systems". In *Proceedings of the International Workshop on Lexically Driven Information Extraction*, Frascati, Italy, July 16, 1997.