ATLAS – A Robust Multilingual Platform for the Web[†]

Diman Karagiozov*, Svetla Koeva, Maciej Ogrodniczuk***, Cristina Vertan******

* Tetracom Interactive Solutions Ltd., diman@tetracom.com
** Bulgarian Academy of Sciences, svetla@dcl.bas.bg
*** Polish Academy of Sciences, maciej.ogrodniczuk@gmail.com
**** University of Hamburg, cristina.vertan@uni-hamburg.de

Abstract

This paper presents a novel multilingual framework integrating linguistic services around a Web-based content management system. The language tools provide semantic foundation for advanced CMS functions such as machine translation, automatic categorization or text summarization. The tools are integrated into processing chains on the basis of UIMA architecture and using uniform annotation model. The CMS is used to prepare two sample online services illustrating the advantages of applying language technology to content administration.

1. Introduction

During the last years, the number of applications which are entirely Web-based, or offer at least some Web front-end has grown dramatically. As a response to the need of managing all this data, a new type of system appeared: the Web-content management system. In this article we will refer to these type of system as WCMS.

Existent WCMS focus on storage of documents in databases and provide mostly full-text search functionality. These types of systems have limited applicability, due to two reasons:

- data available online is often multilingual, and
- documents within a CMS are semantically related (share some common knowledge, or belong to similar topics).

Shortly currently available CMS do not exploit modern techniques from information technology like text mining, semantic Web or machine translation.

The ICT PSP EU project ATLAS – Applied Technology for Language-Aided CMS aims to fill this gap by providing three innovative Web services within a WCMS. These three Web services: i-Librarian, EUDocLib and i-Publisher are not only thematically different but offer also different levels of intelligent information processing.

The ATLAS WCMS makes use of state-of-the art text technology methods in order to extract information and cluster documents according to a given hierarchy. A text summarization module and a machine translation engine are embedded as well as a cross-lingual semantic search engine.

The cross-lingual search engine implements Semantic Web technology: the document content is represented as RDF triples and the search index is built up from these triples. The RDF representation of documents collects not only metadata information about the whole file but also exploits linguistic analysis of the document and store as well the mapping of the file on some ontological concept.

This paper presents the architecture of the ATLAS system with particular focus on the language processing components to be embedded aiming to show how robust NLP (natural language processing) tools can be wrapped in a common framework.

2. Language resources in the ATLAS System

The linguistic diversity in the project is a challenge not to be neglected: the languages belong to four language families and involve three alphabets. To our knowledge it is the first WCMS which will offer solutions for documents written in languages from Central and South-Eastern Europe. Whilst the

[†] The work reported here was carried out within the Applied Technology for Language-Aided CMS project co-funded by the European Commission under the Information and Communications Technologies (ICT) Policy Support Programme (Grant Agreement No 250467). The authors would like to thank all representatives of project partners for their contribution.

standardised development of tools for wide-spread languages as English and German is more common, the situation is quite different when involving languages from Central and South Eastern Europe (see http://www.c-phil.uni-hamburg.de/view/Main/LrecWorkshop2010). Tools with different processing depth, different output formats and sometimes very particular approach are current state of the art in the language technology map of the above-mentioned area [2]. One of the innovative issues in project ATLAS is the integration of linguistically and technologically heterogeneous language tools within a common framework.

The following description presents the steps taken in order to provide such common representation.

- Starting from the fixed desiderata to include text summarisation, automatic document classification, machine translation and cross-lingual information retrieval the minimal list of tools required by such engines which can be provided by all languages involved in the pro ject has been collated and includes:
    - tokeniser,
    - sentence boundary detector,
    - paragraph boundary detector,
    - lemmatizer,
    - PoS Tagger,
    - NP (noun phrase) chunker,
    - NE (named entity) extractor.

  Some of these tools are not completely available for particular languages (e.g. NP chunker for Croatian) but can be developed within the project. Regarding the NE extractor the following entities have been agreed upon: persons, dates, time, location and currency.
- The annotation levels in the texts and the minimal features to be annotated have been defined: Paragraph, Sentence, Token, NP and NE. In order to provide a common representation all linguistic information regarding lemma, PoS etc. have been agreed to be provided at the token level. For a token following features are retained:
    - begin – an integer representing the offset of the first character of the token,
    - end – an integer representing the offset of the last character of the token,
    - pos – a string representing the morphosyntactic tag (PoS, gender, number) associated with the token,
    - lemma – a string containing the lemma of the token.
- For each of the above-mentioned tools the list of additional linguistic features to be represented (if necessary and available) have been defined, e.g. *antecedentBegin* and *antecedentEnd* representing the offset of the first and respectively the last character of the referent in an NP. This feature is necessary for processing German NPs and is therefore included as optional in the NP annotation frame.

A glossary of tagsets delivered by each tool is also maintained, ensuring cross-lingual processing.

Each of the language tools can be included as primitive engine, i.e. part of an UIMA aggregate engine, but also as an aggregate engine. In this way any language component can reuse results produced by a particular tool and exploit its full functionality if required.

3. Language Processing chains

One of the goals of the ATLAS WCMS is to offer documented language processing chains (LPCs) for text annotation. A processing chain for a given language includes a number of existing tools, adjusted and/or fine-tuned to ensure their interoperability. In most respects a language processing chain does not require development of new software modules but rather combining existing tools.

Most of the basic linguistic tools (sentence splitters, stopword filters, tokenizers, lemmatizers, part-of-speech taggers) for languages in scope of our interest have already existed as standalone offline applications. The multilinguality of the system services requires high level of accuracy of each monolingual language chain – simple example is that a word with part-of-speech tag ambiguity in one language may correspond to an unambiguous word in the other language. The complexity grows at the level of structure and sense ambiguity differs among languages. Thus the high precision and performance of language specific chains predefines to the great extend the quality of the system as a whole. For example the Bulgarian PoS tagger has been developed as a modified version of the Brill tagger applying a rule-based approach and techniques for the optimization leading to the 98.3% precision [4]. The large Bulgarian grammar dictionary used for the lemmatization is implemented as acyclic and deterministic finite-state automata to ensure a very fast dictionary look-up.

The language processing chains have been fine-tuned and adjusted to facilitate integration into a common UIMA framework. Other tools (such as noun phrase extractors or named entity recognizers) had to be implemented or multilingually ported. The annotation produced by the chain along with additional tools (e.g. frequency counters) results in higher-level functions such as detection of keywords and phrases along with improbable phrases from the analyzed content, and utilisation of more sophisticated user functionality deserves complex linguistic functions as multilingual text summarisation and machine translation.

UIMA is a pluggable component architecture and software framework designed especially for the analysis of unstructured content and its transformation into structured information. Apart from offering common components (e.g. the type system for document and text annotations) it builds on the concept of analysis engines (in our case, language specific components) taking form of primitive engines which can wrap up NLP (natural language processing) tools adding annotations aggregate engines which define the sequence of execution of chained primitives.

Making the tools chainable requires ensuring their interoperability on various levels. Firstly, compatibility of formats of linguistic information is maintained within the defined scope of required annotation. The UIMA type system requires development of a uniform representation model which helps to normalize heterogeneous annotations of the component NLP tools. With ATLAS it covers properties vital for further processing of the annotated data, e.g. lemma, values for attributes such as gender, number and case for tokens necessary to run coreference module to be subsequently used for text summarisation, categorization and machine translation.

To facilitate introduction of further levels of annotation a general markable type has been introduced, carrying subtype and reference to another markable object. This way new annotation concepts can be tested and later included into the core model.

4. Integration of language processing chains in ATLAS

The language chains are used in order to extract relevant information such as named entities and keywords from the documents stored within the ATLAS WCMS. Additionally they provide the baseline for further engines: Text summarization, Clustering and Machine translation [3] and as such they are the foundation of the enhanced ATLAS platform.

The core online service of the ATLAS platform is i-Publisher, a powerful Web-based instrument for creating, running and managing content-driven Web sites. It integrates the language-based technology to improve content navigation e.g. by interlinking documents based on extracted phrases, words and names, providing short summaries and suggested categorization concepts.

Currently two different thematic content-driven Web sites, i-Librarian and EUDocLib, are being built on top of ATLAS platform, using i-Publisher as content management layer. i-Librarian is intended to be a user-oriented web site which allows visitors to maintain a personal workspace for storing, sharing and publishing various types of documents and have them automatically categorized into appropriate subject categories, summarized and annotated with important words, phrases and names. EUDocLib is planned as a publicly accessible repository of EU legal documents from the EUR-LEX collection with enhanced navigation and multilingual access.

An important aspect of ATLAS System is that all three services operate in a multilingual setting.

Similar functionality will be implemented within the project for Bulgarian, Croatian, German, English, German, Greek, Polish and Romanian. The architecture of the system is modular and allows anytime a new language extension.

5. References

[1] Belogay, A., Ćavar, D., Cristea, D., Karagiozov, D., Koeva, S., Nikolov, R., Ogrodniczuk, M., Przepiórkowski, A., Raxis, P., Vertan C.: i-Publisher, i-Librarian and EUDocLib – linguistic services for the Web. In: Proceedings of the 8th Practical Applications in Language and Computers (PALC 2011) conference. University of Łódź, Poland, 13-15 April 2011 (to appear)

[2] Degórski, Ł., Marcińczuk, M., Przepiórkowski, A.: Definition extraction using a sequential combination of baseline grammars and machine learning classifiers. In: Proceedings of the 6th International Conference on Language Resources and Evaluation, LREC 2008. ELRA, Marrakech (2008), http://nlp.ipipan.waw.pl/~adamp/Papers/2008-lrec-lt4el/213_paper.pdf

[3] Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bo jar, O., Constantin, A., Herbst, E.: Moses: Open Source Toolkit for Statistical Machine Translation. In: ACL (ed.) Annual Meeting of the Association for Computational Linguistics, (ACL), demonstration session. Prague (2007), http://acl.ldc.upenn.edu/P/P07/ P07- 2045.pdf

[4] Koeva, S.: Multi-word Term Extraction for Bulgarian. In: Piskorski, J., Pouliquen, B., Steinberger, R., Tanev, H. (eds.) Proceedings of the Workshop on Balto-Slavonic Natural Language Processing, pp. 59–66. Association for Computational Linguistics, Prague, Czech Republic, June 2007. http://www.aclweb.org/anthology/W/W07/W07- 1708

[5] Ogrodniczuk, M., Karagiozov, D.: ATLAS – The Multilingual Language Processing Platform. In: Proceedings of the 27th Conference of the Spanish Society for Natural Language Processing. University of Huelva, Spain, 5-7 September 2011 (to appear)