i-Publisher, i-Librarian and EUDocLib – linguistic services for the Web

Anelia Belogay, Diman Karagiozov
Tetracom Interactive Solutions
Damir Ćavar
University of Zadar
Dan Cristea
Alexandru Ioan Cuza University
Svetla Koeva
Institute for Bulgarian Language
Roumen Nikolov
Institute of Technologies and Development Foundation
Maciej Ogrodniczuk, Adam Przepiórkowski
Institute of Computer Science, Polish Academy of Sciences
Polivios Raxis
Atlantis Consulting
Cristina Vertan
Research Group "Computerphilology", Hamburg University

Abstract: This paper presents three linguistically-aware online services built on top of the multilingual framework prepared for the ICT PSP EU-co-financed project ATLAS (Applied Technology for Language-Aided CMS). The framework intends to use the state-of-the art text processing methods in order to extract information and cluster documents. These basic blocks provide the base for advanced CMS functions such as automatic categorization or text summarization.

The i-Publisher is a Web-based content management platform for visual website building which integrates linguistic features to improve content navigation e.g. by interlinking documents based on extracted phrases, words and names, providing short summaries and suggested classification concepts. Apart from comprehensive language analysis, the CMS supports data visualisation as well as large volume multilingual data storage and maintenance.

i-Librarian is a sample online service build with and on top of i-Publisher (as a content management layer) to illustrate the benefits of applying language technology to content administration. It allows visitors to maintain a personal workspace for storing, sharing and publishing various types of documents and have them automatically categorized, summarized and annotated with important words, phrases and names. It also allows similar documents in different languages to be found easily.

The EUDocLib service basing on the collection of EU legal documents illustrates how users can easily find similar documents, obtain the summaries of desired documents or extracted important phrases and words.

One of the innovative issues in project is the integration of linguistically and technologically heterogenous language tools within a common framework. Certain level of uniformity has been achieved by establishing the collective list of tools (tokeniser, sentence boundary detector, paragraph boundary detector, lemmatizer, part-of-speech tagger, noun phrase chunker, named entity extractor) to be used for all languages and sharing general annotation properties (minimal structure and features). To facilitate chaining, the language processing tools have been integrated into a common UIMA (Unstructured Information Management Application) framework.

Currently all three services and their underlying language processing chains are available for English (the reference language), but will be soon extended with integrated tools for Bulgarian, Croatian, German, Greek, Polish and Romanian, making the first framework offering uniform solutions for languages of Central and South-Eastern Europe.

Keywords: linguistic tools, language resources, Web services, content management system, online services, UIMA.

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 recently launched 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: iLibrarian, EUDocLib

and iPublisher 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.

The paper is organised as follows: Sec. 2 describes briefly the added-value of the NLP tools for Web applications. Sec. 3 is dedicated to the description of language processing chains in a multilingual setting, while Sec. 4 presents the Web services embedded in the ATLAS platform which make use of the described language processing chains. Finally Sec. 5 describes the more complex language processing tools to be built around the language chains within the ATLAS platform.

2. Linguistic Tools for the Web

A Web-based CMS for advanced language processing should support comprehensive language analysis, data visualisation, and information sharing for diverse languages (at the moment seven languages are envisaged, namely Bulgarian, Croatian, German, Greek, English as a reference language, Polish, and Romanian, but more languages should be expected) – thus combining automatic linguistic annotation that enable more sophisticated language analysis and advance technologies for large volume multilingual data storage, maintaining and processing. The more extensive, consistent and sophisticated the linguistic annotation is, the more useful the multilingual data will be for a whole range of Web services provided. The lower level annotation marks values of attributes such as part-of-speech, grammatical categories of the lemma and word forms, as well as lemma itself, allowing homographs to be distinguished, and word forms of the same lemma to be grouped together. The annotation process of large amount of data should be automated, thus the minimum set of language processing tools (either rule, or statistical, or mixed-approaches based)

integrated for each language are: tokenizer, tagger, and lemmatizer (which presupposes a morphological analyser as well). The benefit of having morphological and lexical information in a domain of humanities, i.e. cultural heritage digital libraries, has been demonstrated by the success of the Perseus Digital Library, oriented for wide user community.

3. Language Processing Chains in a Multilingual Setting

As mentioned in Sec. 1, the ATLAS platform provides Web services involving advanced language technology components for seven languages. 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 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 (Vertan et al. 2009).

One of the innovative issues in project ATLAS is the integration of linguistically and technologically heterogenous 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 project 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 has 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 will be 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, for example `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.

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 (Koeva 2007). 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[1] 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 key words 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.

3.1. Integration of the Tools into UIMA Framework

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 and aggregate engines which define the sequence of execution of chained primitives.

3.2. The Annotation Model

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

[1] Unstructured Information Management Application. See http://uima.apache.org/.

object. This way new annotation concepts can be tested and later included into the core model.

4. WebCMS and its Online Applications

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* 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.

4.1. i-Publisher

The i-Publisher service (see Fig. 1):

- is mainly targeted at small enterprises and non-profit organizations,
- gives the ability to build via point-and-click user interface content-driven Web sites, which provides a wide set of predefined functionalities and whose textual content is automatically processed, i.e. categorized, summarized, annotated, etc.,
- enables publishers, information designers and graphic designers to easily collaborate,

- saves authors, editors and other contributors valuable time by automatically processing textual data and allows them to work together to produce high-quality content.

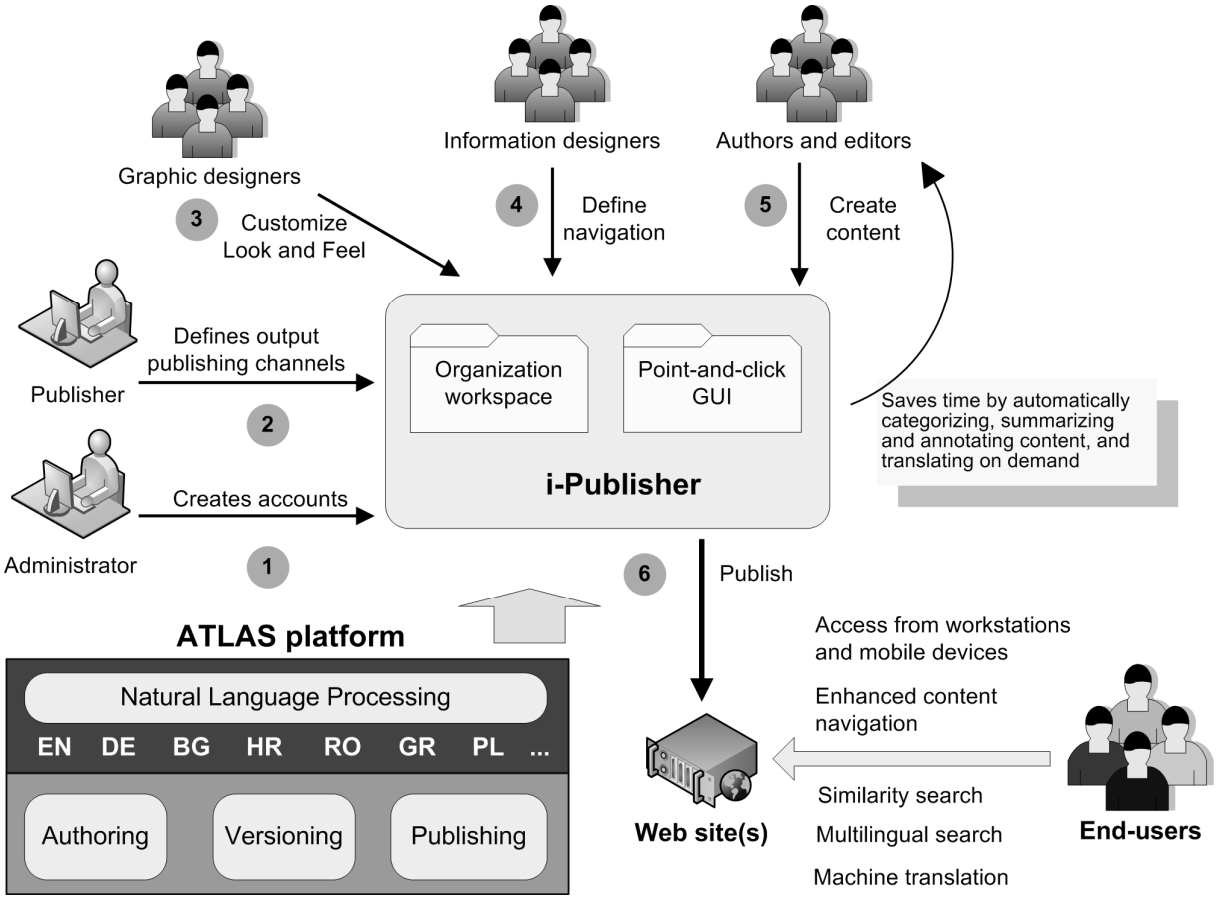

Figure 1. i-Publisher architecture

4.2. i-Librarian

The i-Librarian service (see Fig. 2):

- addresses the needs of authors, students, young researchers and readers,
- gives the ability to easily create, organize and publish various types of documents,
- allows users to find similar documents in different languages, to share personal works with other people, and to locate the most essential texts from large collections of unfamiliar documents.

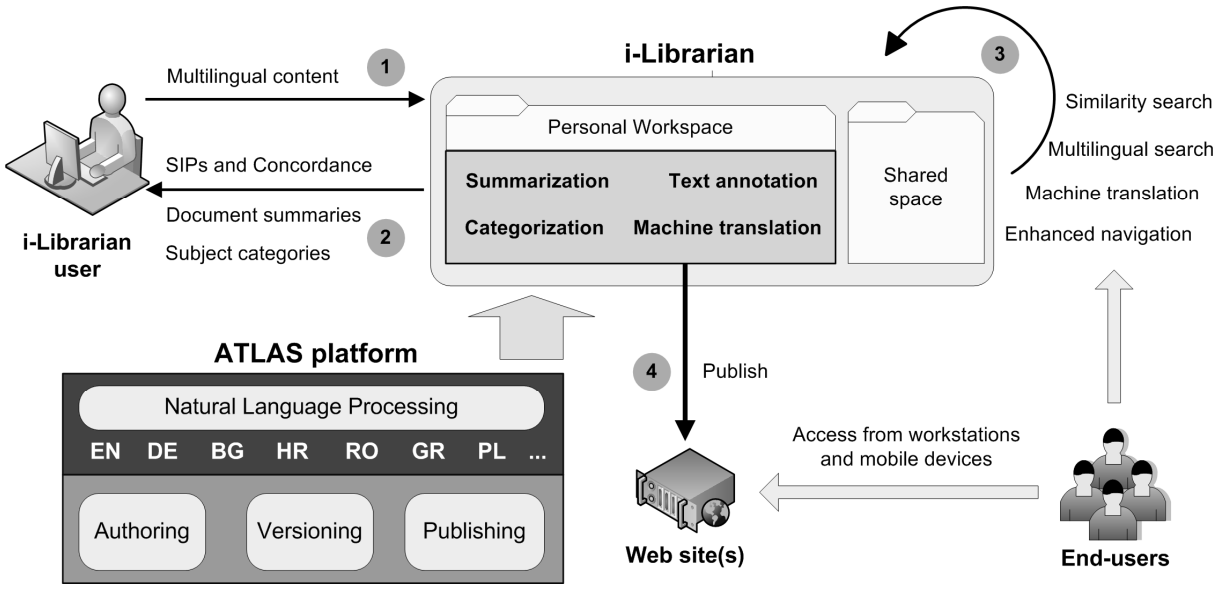

Figure 2. i-Librarian architecture

4.3. EUDocLib

The EUDocLib service:

- addresses the needs of people who require easier access to EU documents in their own language,
- enables users to easily find similar documents, read the summaries of desired documents, or read extracted important phrases and words.

5. Further Steps

The project outlines numerous directions to be followed, ranging from novel approach to advanced linguistic topics through building on user's opinions on to improve their usability up to sustainability issues. From the Semantic Web point of view the most innovative research areas to be covered in the project in the next 1.5 years are connected with integration of Machine Translation engine, Cross-language retrieval engine, Summarization engine and Categorization engine.

According to the project schedule, public versions of these online services will be launched in the beginning of 2012, after their evaluation is completed.

5.1. Machine Translation and Cross-lingual Retrieval

Machine translation ensures integration of foreign-language content, achieved by means of using the EuroMatrix technology (Koehn et al. 2007), improving existing translation models and incorporating them in the proposed online

services. The users will be offered a computer-aided translation service based on the database with existing translations in the form of a translation memory tool. As a next step, the system to process user-submitted translations will be implemented to further improve the quality of existing models.

Cross-lingual retrieval is intended to give users the opportunity to rapidly access content made available by the online content services independently of its language. Existing concept-based cross-lingual search engine together with its underlying ontology, both developed during the LT4eL project (Monachesi et al. 2006, Degórski et al. 2008)[2], will be adapted and further improved. Basing on categories and keywords a conceptual space will be constructed and integrated into a conceptual search mechanism (Vertan et al. 2007).

5.2. Text Summarization

Summarization tools will be prepared by adjusting and fine-tuning existing software components for the project target languages basing on a discourse-parsing based method.

After the text is segmented into elementary discourse units (mainly clauses), a discourse tree is composed for each sentence based on cue-phrases recognized by the parser. The sequence of sentence trees is arranged into discourse tree by maximizing a score contributed from centering transitions and anaphoric links. The discourse tree is in turned used for computing the summaries, both general and focused, with the support of specialized resources and tools such as a collection of discourse markers and (optionally) an anaphora resolver.

5.3. Automatic Categorization

A language-independent text categorization tool fine-tuned to work with each project language will be prepared and tuned to heterogeneous domains. Its ultimate goal will be effective organization of content in the online services by using vector space models (VSMs) with lexical distribution patterns or alternative features selected from the documents. Initial VSMs will be generated on the basis of lexical distribution in documents, using various lexical windows, of 1 to n N-grams, as well as normalization methods for matrix reduction (i.e. by elimination of specific lexical classes of elements, or elimination of lexical covariation etc.). Finally, classification of documents will be performed by applying similarity measures over the vector space models of classes and

[2] Language Technology for eLearning FP6 Specific Targeted Research Project (Information Society Technologies), contract number 027391. See http://www.lt4el.eu/.

particular documents. For the particular classification algorithms, we will use k Nearest Neighbours, Support Vector Machines and Latent Semantic Analysis.

6. Conclusions

The ATLAS platform opens the doors to standardized multilingual online processing of language and it offers localized demonstration tools built on top of the linguistic modules. We intend it to be a contribution to the development of text processing chains for the Web, especially for underrepresented languages.

The framework is ready for integration of new types of tools and new languages to provide wider online coverage of the needful linguistic services in a standardized manner.

References

Banerjee, S. (2002). *Adapting the Lesk algorithm for word sense disambiguation to WordNet*. University of Minnesota, Duluth. Master's thesis.

Degórski, Ł., Marcińczuk, M. and A. Przepiórkowski. (2008). *Definition extraction using a sequential combination of baseline grammars and machine learning classifiers*. In Proceedings of the Sixth International Conference on Language Resources and Evaluation, LREC 2008, Marrakech. ELRA.

Erkan, G. and D. R. Radev. (2004). *LexRank: Graph-based Centrality as Salience in Text Summarization*. Journal of Artificial Intelligence Research 22.

Monachesi, P., Cristea D., Evans D., Killing A., Lemnitzer L., Simov K. and C. Vertan. (2006). *Integrating Language Technology and Semantic Web techniques in eLearning*. In Proceedings of The International Interactive Computer-Aided Learning Conference (ICL 2006), Villach, Austria, September.

Tsoumakas, G., Katakis I. and I. Vlahavas. (2010). *Mining Multilabel Data*. In Maimon O. and L. Rokach (eds.). *Data Mining and Knowledge Discovery Handbook*, Springer, 2nd edition.

Tsoumakas, G., Vilcek J., Spyromitros E. and I. Vlahavas. (2000). *Mulan: A Java Library for Multi-Label Learning*. Journal of Machine Learning Research 1.

Vertan, C., Monachesi P., Simov K., Osenova P., Lemnitzer L., Killing A. and D. Evans. (2007). *Crosslingual retrieval in an e-Learning environment*. In R. Basili and M. T. Pazienza, (eds.). Proceedings of The 10th Congress of the Italian Association for Artificial Intelligence (AIIA 2007), pages 839-847, Berlin, Heidelberg. Springer-Verlag.