ATLAS – The Multilingual Language Processing Platform*

ATLAS – La Plataforma de Procesamiento del Lenguaje Multilingüe

Maciej Ogrodniczuk
Institute of Computer Science
Polish Academy of Sciences
Warsaw, Poland
maciej.ogrodniczuk@ipipan.waw.pl

Diman Karagiozov
Tetracom Interactive Solutions Ltd.
Sofia, Bulgaria
diman@tetracom.com

Resumen: En este trabajo se presenta la plataforma ATLAS – marco multilingüe de procesamiento del lenguaje que integra el conjunto común de herramientas lingüísticas para un grupo de lenguas europeas (con menos recursos: búlgaro, croata, griego, polaco y rumano, junto con inglés y alemán como lenguas de referencia). La más avanzada funcionalidad PNL que ofrece la plataforma permite la anotación de textos multilingües en los niveles inferiores (segmentación, morfosintaxis) y a su vez soporta el procesamiento de más alto nivel como la categorización automática, extracción de información, la traducción automática o de resumen. Métodos de anotación más elaborados como la extracción de la entidad nombrada o lematización unitaria de varias palabras también están disponibles. La anotación multinivel de los textos se rige por las cadenas de procesamiento de lenguaje construidas con el estándar de la industria UIMA.
Para demostrar las capacidades del marco, se han construido en la parte superior del mismo tres servicios informados lingüísticamente: "i-Publisher" (plataforma de gestión de contenidos basada en la Web), "i-Librarian" (una biblioteca digital de trabajos científicos) y "EUDocLib" (página para la navegación y la búsqueda a través de documentos de EUR-LEX).
Palabras clave: herramientas lingüísticas, recursos lingüísticos, servicios Web, sistema de gestión de contenidos, servicios en línea, UIMA

Abstract: This paper presents the ATLAS platform – multilingual language processing framework integrating the common set of linguistic tools for a group of European languages (less-resourced: Bulgarian, Croatian, Greek, Polish and Romanian together with English and German as reference languages). State-of-the-art NLP functionality offered by the platform allows for multilingual annotation of texts on lower levels (segmentation, morphosyntax) which in turn supports higher-level processing such as automated categorization, information extraction, machine translation or summarization. More elaborate annotation methods such as named entity extraction or multiword unit lemmatization are also available. Multilevel annotation of texts is governed by language processing chains constructed with UIMA (*Unstructured Information Management Application*) industry standard.
To demonstrate capabilities of the framework, three linguistically-aware online services have been built on top of it: i-Publisher (Web-based content management platform), i-Librarian (a digital library of scientific works) and EUDocLib (site for browsing and searching through EUR-LEX documents).
Keywords: linguistic tools, language resources, Web services, content management system, online services, UIMA

* 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.

1 *Introduction*

The need for a common multilingual NLP framework, offering a coherent set of language tools for a number of European languages is getting more and more urgent in the integrating Europe. Recently, a number of pan-European initiatives fostering such development appeared, but they seem to be concentrating on gathering and standardizing resources rather than their practical interoperability. ATLAS project (Applied Technology for Language-Aided CMS) which started in 2010 responds to this need by offering a set of state-of-the-art NLP tools for several European languages wrapped in a common language processing framework which can be used and reused in various higher-level applications.

The most obvious of them is language-powered document management, made easy with seamless integration of NLP tools, hot-swap/hot add-on of LPC engines, scalability of NLP infrastructure, content- and context-based navigation and true multiliguality: language-independent information extraction, document categorization and clustering (similar documents in foreign languages), machine translation of abstracts (automated summarization).

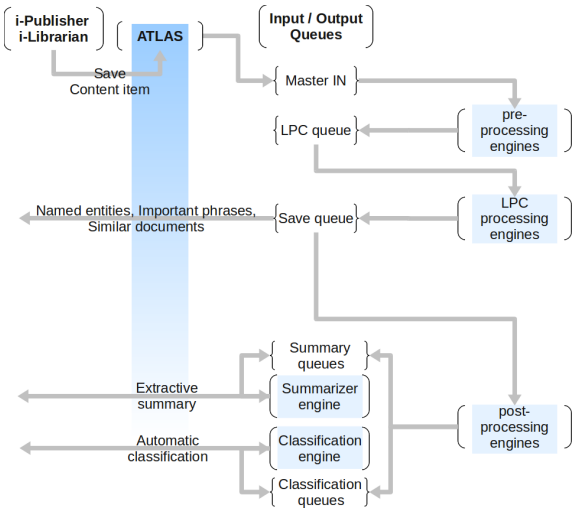
This article presents the capabilities of the platform, its architecture, components and formats together with sample interfaces to run the integrated linguistic machinery and performance results.

2 *Language Processing Architecture*

The architecture of the platform allows for asynchronous, queue-based processing of requests (see figure 1). When content processing is initiated (e.g. when a user saves a content item in the user interface) the item is stored in the master NLP queue which is a primary method of interaction between the integrated subsystems: subsequent processing engines are using the content of the queue a processing source and return its result to the queue.

All texts are first pre-processed with MIME type detectors, text converters (from MS Office, Open Office, PDF, HTML, RTF, ePub etc. to plain text), language detectors and cleaners – and finally sent back to the queue. The core processing is governed by language processing chains (see section 3.1)

Figure 1: Linguistic processing in ATLAS



which annotate text with properties relevant for further processing – most often tokens, paragraph and sentence boundaries, named entities and nominal phrases. As the last step, post-processing engines are producing higher-level annotation (summaries, classification).

2.1 The UIMA Framework

UIMA (*Unstructured Information Management Application*, see `http://uima.apache.org/`) is a pluggable component architecture and software framework designed especially for the analysis of unstructured content and its transformation into structured information. Currently UIMA is the only industry standard (OASIS standard) for content analytics.

Apart from offering common components (e.g. the type system for document and text annotations) UIMA 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.

2.2 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

Table 1: Summary of text annotations

Annotation type	Parameters[1]
Paragraph	–
Sentence	–
Token	POS tag, MSD, lemma, gender, number, case, *word sense*
Noun Phrase	head, *base form*
Named Entity	type (one of: Date, Location, Money, Organization, Percentage, Person, Time), *normalized value*
Markable	type, *reference*

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 (see table 1; optional parameters are shown in italics).

To facilitate introduction of further levels of annotation a general *markable* type has been introduced, carrying type and an optional reference to another markable object (which allows both for marking text fragments and creating mention chains). This way new annotation concepts can be tested and later included into the core model.

3 Language-related Functionality

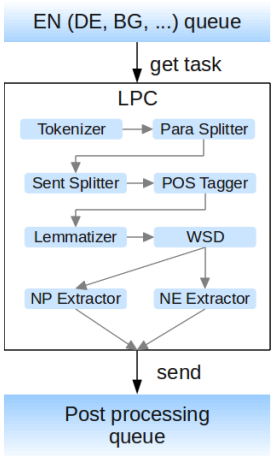
The annotations are used as input of the higher-level functionality of the platform. Individual functions are available as components which can be wrapped up in interfaces (see section 4).

3.1 Language Processing

Language processing chains (LPCs, see figure 2) provide the core 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

[1]Extending the standard set of parameters: begin offset, end offset and text value.

Figure 2: Language Processing Chain



does not require development of new software modules but rather combining existing tools.

The minimal set of annotation tools available across all integrated languages includes:

- tokeniser,
- sentence boundary detector,
- paragraph boundary detector,
- lemmatizer,
- POS tagger,
- NP (noun phrase) chunker,
- NE (named entity) extractor

(when quality of the tool is questionable for a given language, tools from the basic set are improved by project partners).

Currently, in the first phase of the project, the set of integrated tools (see table 2) is limited to English and is heavily OpenNLP-related. Tools for other project languages are currently being integrated and will be available in the next released of the platform.

Table 3 presents the current average performance of the linguistic chain.

3.2 Information Retrieval

The retrieval component can automatically compiles a summary of each document, featuring all relevant information – contextually important words and phrases, capitalized phrases, URLs, similar documents, extractive summary etc. With these extracts users can create personal keywords for each file and categorize their collections according to their personal liking.

Table 2: NLP tools in ATLAS

Tool type	Tool name / Source
Paragraph splitter	Regexp-based solution by Tetracom
Sentence splitter	OpenNLP
Tokenizer	OpenNLP
Lemmatizer	RASP
POS tagger	OpenNLP
WSD	LESK-based[2]
NP extractor	14 English-specific rules by Tetracom
NE extractor	OpenNLP
Summarizer	LexRank[3] and Open Text Summarizer[4]
Categorization	MULAN[5] with a k-Nearest Neighbor-based algorithm

Table 3: Current performance of English LPC

Tool type	s/doc	% of total processing time
Paragraph splitter	0,01	0,2%
Sentence splitter	0,13	3,7%
Tokenizer	0,17	4,9%
Lemmatizer	0,10	3,0%
POS tagger	0,13	3,7%
WSD	1,67	49,4%
NP extractor	0,07	2,1%
NE extractor	1,11	32,9%
	3,37	100,0%

3.3 Automatic Categorization

Unlike most other repository approaches, no manual document categorization is necessary – after upload, the categorization component can automatically catalogue documents using a comprehensively trained model. The component can also make suggestions about other topics that are relevant to the document, therefore reducing the time spent on categorization to a minimum.

3.4 Full-text Search and Similarity Search

The documents are indexed by a powerful high-speed full-text search engine, based on Lucene. Using a simple, Google-like search form, users can quickly find words or phrases in all their documents. The search results provide up to three excerpts from the text, best matching the search terms. Additionally, similarity search is available basing on extracted essence of the documents.

3.5 Machine Translation

Currently integrated machine translation approach is based on the Serverland API (REST or XMLRPC) by DFKI – a middleware that provides a common interface to several machine translation services such as Google, Bing, Lucy, and Moses. The component is used to translate document summaries, noun phrases and the "long text" metadata fields (such as content item descriptions).

3.6 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.

4 Interfaces

For demonstration purposes three interfaces has been developed to illustrate how the linguistic building blocks can be interconnected into a useful application (Belogay et al., 2011). All three operate in a multilingual setting. Although in the first year of the project the implemented functionality is offered only for English (a reference language),

[2]See (Banerjee, 2002). The initial tool available in Perl was rewritten in C++ to improve performance (30 times).

[3]See (Erkan and Radev, 2004).

[4]See `http://libots.sourceforge.net/`.

[5]See (Tsoumakas et al., 2000; Tsoumakas, Katakis, and Vlahavas, 2010) and `http://mulan.sourceforge.net/`.

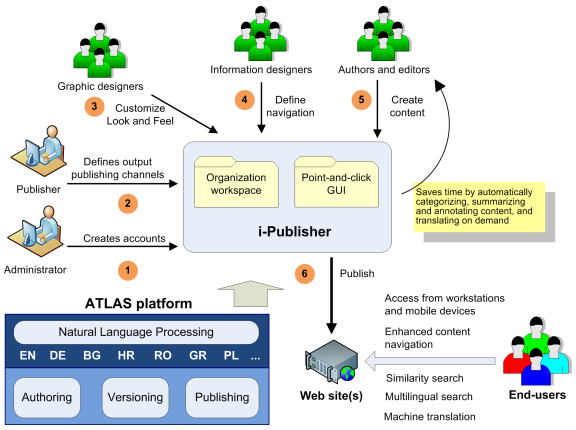


Figure 3: i-Publisher architecture

it will be soon available for other project languages – Bulgarian, Croatian, German, German, Greek, Polish and Romanian; hopefully for more due to the flexible architecture of the system.

The first of the currently available interfaces is i-Publisher – the online Web Content Management System integrating the language-based technology to make its processing output available on the managed Web sites.

Two other are thematic content-driven Web sites, i-Librarian and EUDocLib, built on top of ATLAS platform, using i-Publisher as content management layer. i-Librarian is intended to be a user-oriented personal workspace for storing, sharing and publishing various types of documents automatically assigned to appropriate subject categories, summarized and annotated with important words, phrases and names. EUDocLib is a publicly accessible repository of EU legal documents from the EUR-LEX collection with enhanced navigation and multilingual access.

4.1 i-Publisher: Web Content Management System

The i-Publisher service (see figures 3 and 4) available at `http://www.i-publisher.atlasproject.eu` (restricted access) offers a flexible WebCMS configuration interface with dynamic data model, content versioning, filtering and multi-level grouping, Web site layout and design editor, granular user access rights and library of predefined themes. The easy point-and-click graphical Multi-Document user interface supports drag-and-drop actions.

Linguistic features are available through

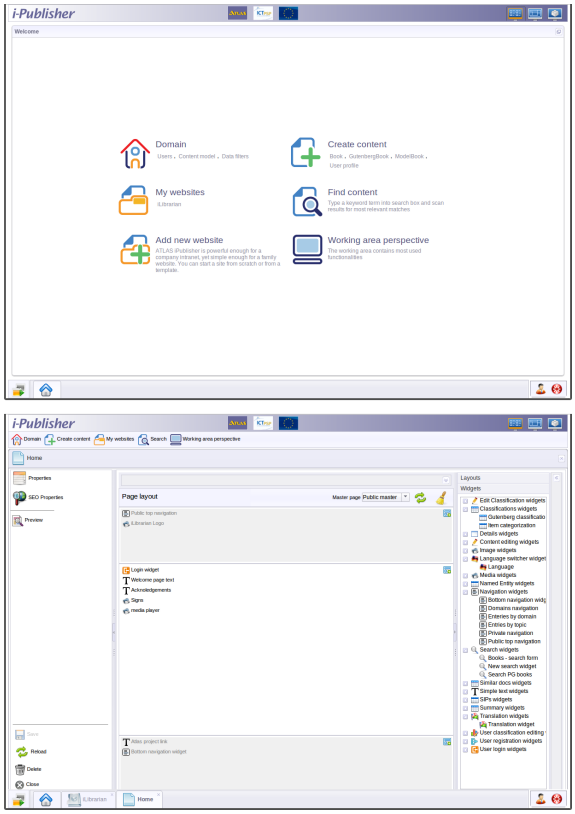


Figure 4: i-Publisher interface: welcome screen and working area

widgets placed on Web pages; currently a wide set of predefined functionalities for automated processing of textual content is integrated, with categorization, summarization, identification of named entities and noun phrases etc.

4.2 i-Librarian: A Personal Library

The i-Librarian service (see figures 5 and 6) available at `http://www.i-librarian.eu` (restricted access) is a demonstration of ATLAS NLP capabilities in the form of a digital library Web site created using i-Publisher. It can be used e.g. as a repository of scientific papers, therefore addressing the needs of authors, students and researchers by giving them the ability to easily create, organize and publish various types of documents and then searching for similar documents in different languages, sharing personal works with other people, and locating the most essential texts from large collections of unfamiliar documents.

The library offers the public and private workspace and employs language technology to extract important phrases and

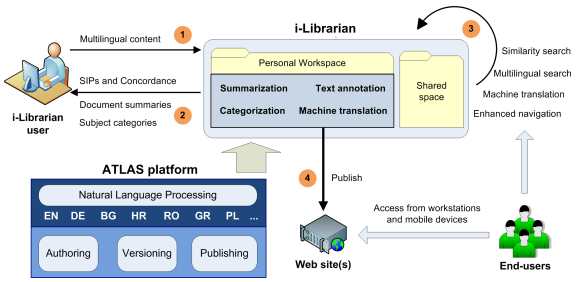


Figure 5: i-Librarian architecture

named entities from indexed documents; similar items are then displayed on demand, abstract translated and document summaries produced.

To evaluate performance and scalability of the site even before users started uploading their papers, 4.4 K documents (165 M tokens) from Project Gutenberg have been uploaded.

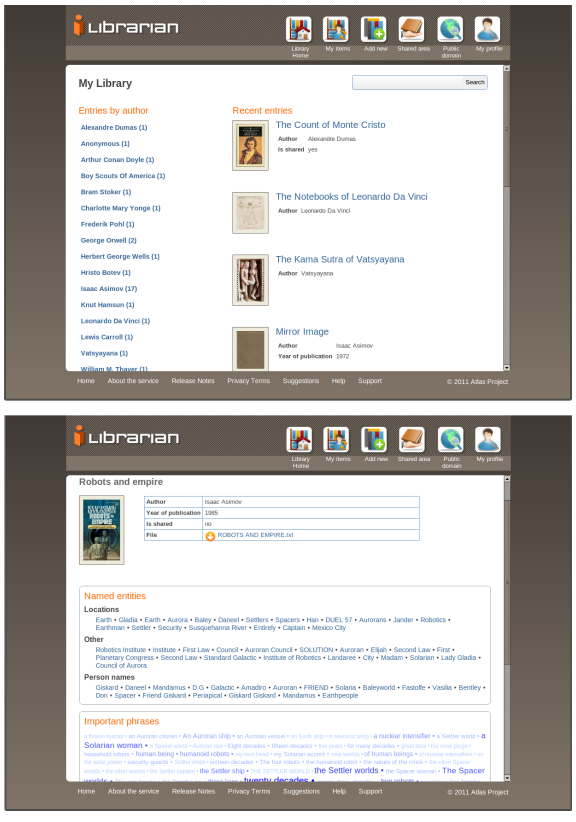


Figure 6: i-Librarian interface: library view and item view

4.3 EUDocLib: Electronic Library of Legal EU Documents

Another demonstration of ATLAS functionality is the EUDocLib Web site (freely available at `http://eudoclib.atlasproject.eu`, see figure 7), also created with i-Publisher. The library offers easy access to EU documents in all project languages with automatic categorization (by assigning documents to Eurovoc, the EU's multilingual thesaurus), extraction of important phrases, named entities, similar items and cross-lingual information retrieval.

Figure 7: EUDocLib interface: search results, browse perspective and individual item

Currently the site covers 140 K documents (182 M tokens).

5 *Further Steps*

The first obvious direction to be followed is internationalization of the platform and improvement of quality offered by currently integrated tools. The generic OpenNLP tools

will be supplemented with their local counterparts and statistical models are planned to be combined with rule-based approach (for example, combining the OpenNLP named entity extractor with a rule-based named entity recognition tool focusing mainly of exceptions of the general rules, covered by OpenNLP).

Another important general further step will be improvement of the overall performance of the chained tools in ATLAS, mainly post-processing engines and data stores. However, the most innovative research areas to be covered are related to three subjects: machine translation, summarization and categorization.

5.1 Machine Translation

Existing translation approach is planned to be amended with a system to process user-submitted translations. Concept-based cross-lingual search engine together with its underlying ontology, both developed during the LT4eL[6] project (Degórski, Marcińczuk, and Przepiórkowski, 2008; Monachesi et al., 2006), 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 Summarization

Summarization tools are planned to be adjusted and fine-tuned basing on a discourse parser developed by Iasi University. 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 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 particular documents.

Apart from the already mentioned ones, Support Vector Machines, Latent Semantic Analysis, Naïve Bayes, MaxEnt, Centroid and advanced feature space reduction algorithms will be used for classification.

6 Conclusions

The ATLAS platform opens the door 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.

New versions of the online services are planned to be launched in the beginning of 2012.

References

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

Belogay, Anelia, Damir Cavar, Dan Cristea, Diman Karagiozov, Svetla Koeva, Roumen Nikolov, Maciej Ogrodniczuk, Adam Przepiorkowski, Polivios Raxis, and Cristina Vertan. 2011. i-Publisher, i-Librarian and EUDocLib – linguistic services for the Web. To appear in the proceedings of PALC 2011.

Degórski, Łukasz, Michał Marcińczuk, and Adam Przepiórkowski. 2008. Defini-

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

tion 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, Gunes and Dragomir R. Radev. 2004. LexRank: Graph-based Centrality as Salience in Text Summarization. Journal of Artificial Intelligence Research 22.

Monachesi, Paola, Dan Cristea, Diane Evans, Alex Killing, Lothar Lemnitzer, Kiril Simov, and Cristina 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, Grigorios, Ioannis Katakis, and Ioannis Vlahavas. 2010. Mining Multilabel Data. In O. Maimon and L. Rokach *Data Mining and Knowledge Discovery Handbook*, Springer, 2nd edition.

Tsoumakas, Grigorios, Jozef Vilcek, Eleftherios Spyromitros, and Ioannis Vlahavas. 2000. Mulan: A Java Library for Multi-Label Learning. Journal of Machine Learning Research 1.

Vertan, Cristina, Paola Monachesi, Kiril Simov, Petya Osenova, Lothar Lemnitzer, Alex Killing, and Diane Evans. 2007. Crosslingual retrieval in an eLearning environment. In Roberto Basili and Maria Teresa Pazienza, editors, *Proceedings of The 10th Congress of the Italian Association for Artificial Intelligence (AIIA 2007)*, pages 839–847, Berlin, Heidelberg. Springer-Verlag.