Harnessing NLP Techniques in the Processes of Multilingual Content Management

Anelia Belogay
Tetracom IS Ltd.
anelia@tetracom.com

Svetla Koeva
Institute for Bulgarian Language
svetla@dcl.bass.bg

Adam Przepiórkowski
Instytut Podstaw Informatyki Polskiej Akademii Nauk
adamp@ipipan.waw.pl

Dan Cristea
Universitatea Alexandru Ioan Cuza
dcristea@info.uaic.ro

Diman Karagyozov
Tetracom IS Ltd.
diman@tetracom.com

Cristina Vertan
Universitaet Hamburg
cristina.vertan@uni-hamburg.de

Polivios Raxis
Atlantis Consulting SA
raxis@atlantisresearch.gr

Abstract

The emergence of the WWW as the main source of distributing content opened the floodgates of information. The sheer volume and diversity of this content necessitate an approach that will reinvent the way it is analysed. The quantitative route to processing information which relies on content management tools provides structural analysis. The challenge we address is to evolve from the process of streamlining data to a level of understanding that assigns value to content.

We present an open-source multilingual platform ATALS that incorporates human language technologies in the process of multilingual web content management. It complements a content management software-as-a-service component i-Publisher, used for creating, running and managing dynamic content-driven websites with a linguistic platform. The platform enriches the content of these websites with revealing details and reduces the manual work of classification editors by automatically categorising content. The platform ASSET supports six European languages.

We expect ASSET to serve as a basis for future development of deep analysis tools capable of generating abstractive summaries and training models for decision making systems.

Introduction

The advent of the Web revolutionized the way in which content is manipulated and delivered. As a result, digital content in various languages has become widely available on the Internet and its sheer volume and language diversity have presented an opportunity for embracing new methods and tools for content creation and distribution. Although significant improvements have been made in the field of web content management lately, there is still a growing demand for online content services that incorporate language-based technology.

Existing software solutions and services such as Google Docs, Slingshot and Amazon implement some of the linguistic mechanisms addressed in the platform. The most used open-source multilingual web content management

Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics, pages 6–10,
Avignon, France, April 23 - 27 2012.

systems (Joomla, Joom!Fish, TYPO3, Drupal)[1] offer low level of multilingual content management, providing abilities for building multilingual sites. However, the available services are narrowly focused on meeting the needs of very specific target groups, thus leaving unmet the rising demand for a comprehensive solution for multilingual content management addressing the issues posed by the growing family of languages spoken within the EU.

We are going to demonstrate the open-source content management platform ATLAS and as proof of concept, a multilingual library i-librarian, driven by the platform. The demonstration aims to prove that people reading websites powered by ATLAS can easily find documents, kept in order via the automatic classification, find context-sensitive content, find similar documents in a massive multilingual data collection, and get short summaries in different languages that help the users to discern essential information with unparalleled clarity.

The "Technologies behind the system" chapter describes the implementation and the integration approach of the core linguistic processing framework and its key sub-components – the categorisation, summarisation and machine-translation engines. The chapter "i-Librarian – a case study" outlines the functionalities of an intelligent web application built with our system and the benefits of using it. The chapter "Evaluation" briefly discusses the user evaluation of the new system. The last chapter "Conclusion and Future Work" summarises the main achievements of the system and suggests improvements and extensions.

Technologies behind the system

The linguistic framework ASSET employs diverse natural language processing (NLP) tools technologically and linguistically in a platform, based on UIMA [2] . The UIMA pluggable component architecture and software framework are designed to analyse content and to structure it. The ATLAS core annotation schema, as a uniform representation model, normalizes and harmonizes the heterogeneous nature of the NLP tools[3].

The processing of text in the system is split into three sequentially executed tasks.

Firstly, the text is extracted from the input source (text or binary documents) in the "pre-processing" phase.

Secondly, the text is annotated by several NLP tools, chained in a sequence in the "processing" phase. The language processing tools are integrated in a language processing chain (LPC), so that the output of a given NLP tool is used as an input for the next tool in the chain. The baseline LPC for each of the supported languages includes a sentence and paragraph splitter, tokenizer, part of speech tagger, lemmatizer, word sense disambiguation, noun phrase chunker and named entity extractor (Cristea and Pistiol, 2008). The annotations produced by each LPC along with additional statistical methods are subsequently used for detection of keywords and concepts, generation of summary of text, multi-label text categorisation and machine translation.

Finally, the annotations are stored in a fusion data store, comprising of relational database and high-performance Lucene[4] indexes.

The architecture of the language processing framework is depicted in Figure 1.

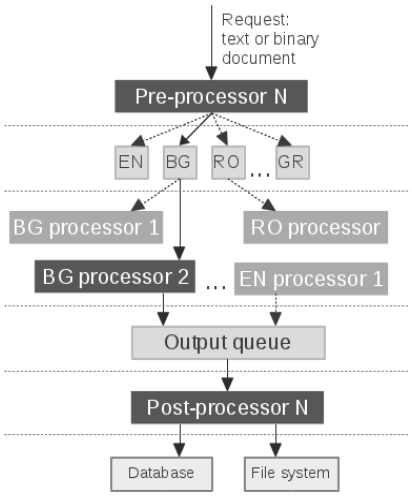

Figure 1. Architecture and communication channels in our language processing framework.

The system architecture, shown in Figure 2, is based on asynchronous message processing

[1] http://www.joomla.org/, http://www.joomfish.net/, http://typo3.org/, http://drupal.org/

[2] http://uima.apache.org/

[3] The system exploits heterogeneous NLP tools, for the supported natural languages, implemented in Java, C++ and Perl. Examples are: OpenNLP (http://incubator.apache.org/opennlp/), RASP (http://ilexir.co.uk/applications/rasp/), Morfeusz (http://sgjp.pl/morfeusz/), Panterra (http://code.google.com/p/pantera-tagger/), ParsEst (http://dcl.bas.bg/), TnT Tagger (http://www.coli.uni-saarland.de/~thorsten/tnt/).

[4] http://lucene.apache.org/

patterns (Hohpe and Woolf, 2004) and thus allows the processing framework to be easily scaled horizontally.

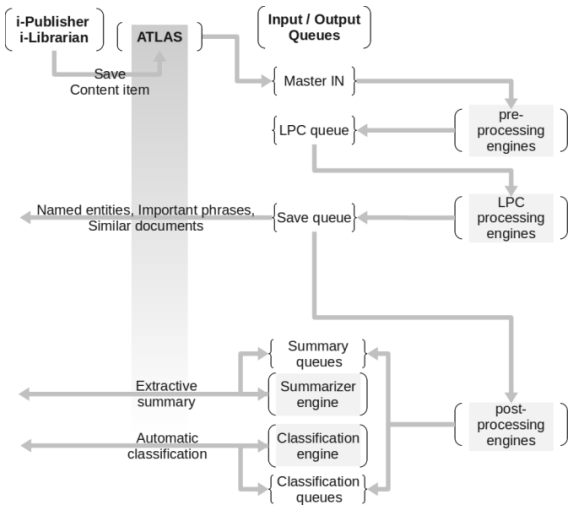

Figure 2. Top-level architecture of our CMS and its major components.

Text Categorisation

We implemented a language independent text categorisation tool, which works for user-defined and controlled classification hierarchies. The NLP framework converts the texts to a series of natural numbers, prior sending the texts to the categorisation engine. This conversion allows high level compression of the feature space. The categorisation engine employs different algorithms, such as Naïve Bayesian, relative entropy, Class-Feature Centroid (CFC) (Guan et. al., 2009), and SVM. New algorithms can be easily integrated because of the chosen OSGi-based architecture (OSGi Alliance, 2009). A tailored voting system for multi-label multi-class tasks consolidates the results of each of the categorisation algorithms.

Summarisation (prototype phase)

The chosen implementation approach for coherent text summarisation combines the well-known LexRank algorithm (Erkan and Radev, 2004) and semantic graphs and word-sense disambiguation techniques (Plaza and Diaz, 2011). Furthermore, we have automatically built thesauri for the top-level domains in order to produce domain-focused extractive summaries. Finally, we apply clause-boundaries splitting in order to truncate the irrelevant or subordinating clauses in the sentences in the summary.

Machine Translation (prototype phase)

The machine translation (MT) sub-component implements the hybrid MT paradigm, combining an example-based (EBMT) component and a Moses-based statistical approach (SMT). Firstly, the input is processed by the example-based MT engine and if the whole or important chunks of it are found in the translation database, then the translation equivalents are used and if necessary combined (Gavrila, 2011). In all other cases the input is processed by the categorisation sub-component in order to select the top-level domain and respectively, the most appropriate SMT domain- and POS-translation model (Niehues and Waibel, 2010).

The translation engine in the system, based on MT Server Land (Federmann and Eisele, 2010), is able to accommodate and use different third party translation engines, such as the Google, Bing, Lusy or Yahoo translators.

Case Study: Multilingual Library

i-Librarian[5] is a free online library that assists authors, students, young researchers, scholars, librarians and executives to easily create, organise and publish various types of documents in English, Bulgarian, German, Greek, Polish and Romanian. Currently, a sample of the publicly available library contains over 20 000 books in English.

On uploading a new document to i-Librarian, the system automatically provides the user with an extraction of the most relevant information (concepts and named entities, keywords). Later on, the retrieved information is used to generate suggestions for classification in the library catalogue, containing 86 categories, as well as a list of similar documents. Finally, the system compiles a summary and translates it in all supported languages. Among the supported formats are Microsoft Office documents, PDF, OpenOffice documents, books in various electronic formats, HTML pages and XML documents. Users have exclusive rights to manage content in the library at their discretion.

The current version of the system supports English and Bulgarian. In early 2012 the Polish, Greek, German and Romanian languages will be in use.

[5] i-Librarian web site is available at http://www.i-librarian.eu/. One can access the i-Librarian demo content using "**demo@i-librarian.eu**" for username and "**sandbox**" for password.

Evaluation

The technical quality and performance of the system is being evaluated as well as its appraisal by prospective users. The technical evaluation uses indicators that assess the following key technical elements:

- overall quality and performance attributes (MTBF6, uptime, response time);
- performance of specific functional elements (content management, machine translation, cross-lingual content retrieval, summarisation, text categorisation).

The user evaluation assesses the level of satisfaction with the system. We measure non functional elements such as:

- User friendliness and satisfaction, clarity in responses and ease of use;
- Adequacy and completeness of the provided data and functionality;
- Impact on certain user activities and the degree of fulfilment of common tasks.

We have planned for three rounds of user evaluation; all users are encouraged to try online the system, freely, or by following the provided base-line scenarios and accompanying exercises. The main instrument for collecting user feedback is an online interactive electronic questionnaire[7].

The second round of user evaluation is scheduled for Feb-March 2012, while the first round took place in Q1 2011, with the participation of 33 users. The overall user impression was positive and the Mean value of each indicator (in a 5-point Likert scale) was measured on AVERAGE or ABOVE AVERAGE.

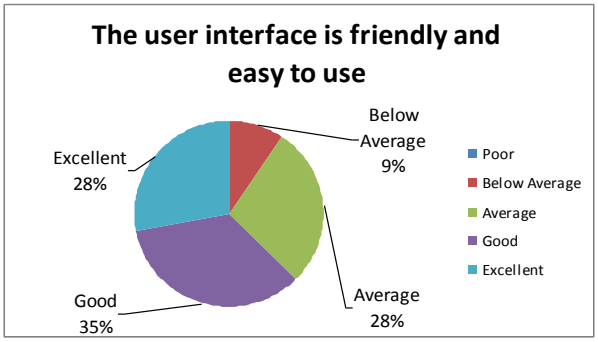

Figure 3. User evaluation – UI friendliness and ease of use.

[6] Mean Time Between Failures

[7] The electronic questionnaire is available at http://ue.atlasproject.eu

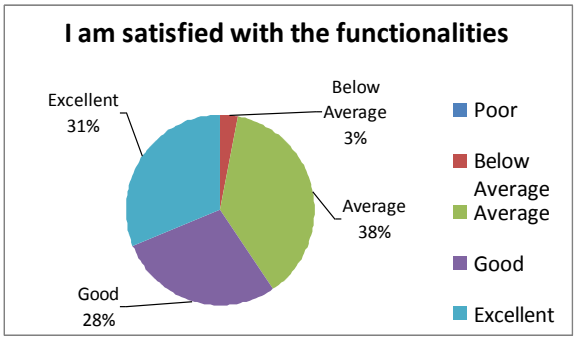

Figure 4. User evaluation – user satisfaction with the available functionalities in the system.

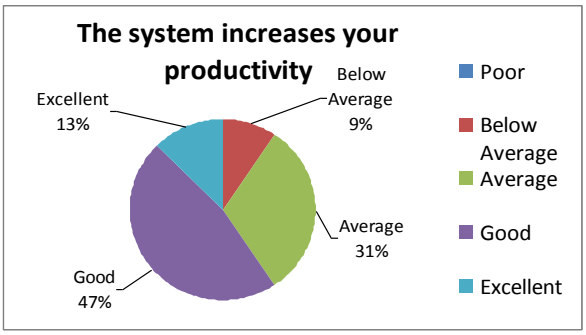

Figure 5. User evaluation – users productivity incensement.

Acknowledgments

ATLAS (Applied Technology for Language-Aided CMS) is a European project funded under the CIP ICT Policy Support Programme, Grant Agreement 250467.

Conclusion and Future Work

The abundance of knowledge allows us to widen the application of NLP tools, developed in a research environment. The tailor made voting system maximizes the use of the different categorisation algorithms. The novel summary approach adopts state of the art techniques and the automatic translation is provided by a cutting edge hybrid machine translation system.

The content management platform and the linguistic framework will be released as open-source software. The language processing chains for Greek, Romanian, Polish and German will be fully implemented by the end of 2011. The summarisation engine and machine translation tools will be fully integrated in mid 2012.

We expect this platform to serve as a basis for future development of tools that directly support decision making and situation awareness. We will use categorical and statistical analysis in order to recognise events and patterns, to detect opinions and predictions while processing

extremely large volumes of disparate data resources.

Demonstration websites

The multilingual content management platform is available for testing at http://i-publisher.atlasproject.eu/atlas/i-publisher/demo . One can access the CMS demo content using "demo" for username and "sandbox2" for password.

The multilingual library web site is available at http://www.i-librarian.eu/. One can access the i-Librarian demo content using "demo@i-librarian.eu" for username and "sandbox" for password.

References

Dan Cristea and Ionut C. Pistol, 2008. Managing Language Resources and Tools using a Hierarchy of Annotation Schemas. In the proceedings of workshop 'Sustainability of Language Resources and Tools for Natural Language Processing', LREC, 2008

Gregor Hohpe and Bobby Woolf. 2004. Enterprise Integration Patterns: Designing, Building, and Deploying Messaging Solutions. Addison-Wesley Professional.

Hu Guan, Jingyu Zhou and Minyi Guo. A Class-Feature-Centroid Classifier for Text Categorization. 2009. WWW 2009 Madrid, Track: Data Mining / Session: Learning, p201-210.

OSGi Alliance. 2009. OSGi Service Platform, Core Specification, Release 4, Version 4.2.

Gunes Erkan and Dragomir R. Radev. 2004. LexRank: Graph-based Centrality as Salience in Text Summarization. Journal of Artificial Intelligence Research 22 (2004), p457–479.

Laura Plaza and Alberto Diaz. 2011. Using Semantic Graphs and Word Sense Disambiguation Techniques to Improve Text Summarization. Procesamiento del Lenguaje Natural, Revista nº 47 septiembre de 2011 (SEPLN 2011), pp 97-105.

Monica Gavrila. 2011. Constrained Recombination in an Example-based Machine Translation System, In the Proceedings of the EAMT-2011: the 15th Annual Conference of the European Association for Machine Translation, 30-31 May 2011, Leuven, Belgium, p. 193-200

Jan Niehues and Alex Waibel. 2010. Domain adaptation in statistical machine translation using factored translation models. EAMT 2010: Proceedings of the 14th Annual conference of the European Association for Machine Translation, 27-28 May 2010, Saint-Raphaël, France.

Christian Federmann and Andreas Eisele. 2010. MT Server Land: An Open-Source MT Architecture. The Prague Bulletin of Mathematical Linguistics. NUMBER 94, 2010, p57–66