ATLAS - Human Language Technologies integrated within a Multilingual Web Content Management System

Svetla Koeva
Department of Computational Linguistics, Institute for Bulgarian
Bulgarian Academy of Sciences
svetla@dcl.bas.bg

Abstract

The main purpose of the project ATLAS (Applied Technology for Language-Aided CMS) is to facilitate multilingual web content development and management. Its main innovation is the integration of language technologies within a web content management system. The language processing framework, integrated with web content management, provides automatic annotation of important words, phrases and named entities, suggestions for categorisation of documents, automatic summary generation, and machine translation of summaries of documents. A machine translation approach, as well as methods for obtaining and constructing training data for machine translation are under development.

1 Introduction

The main purpose of the European project ATLAS (Applied Technology for Language-Aided CMS)[1] is to facilitate multilingual web content development and management. Its main innovation is the integration of language technologies within a web content management system. ATLAS combines a language processing framework with a content management component (i-Publisher)[2] used for creating, running and managing dynamic content-driven websites. Examples of such sites are i-Librarian,[3] a free online library of digital documents that may be personalised according to the user's needs and requirements; and EUDocLib,[4] a free online library of European legal documents. The language processing framework of these websites provides automatic annotation of important words, phrases and named entities, suggestions for categorisation of documents, automatic summary generation, and machine translation of a summary of a document (Karagyozov et al. 2012). Six European Union languages – Bulgarian, German, Greek, English, Polish, and Romanian are supported.

2. Brief overview of existing content management systems

The most frequently used open-source multilingual web content management systems (WordPress, Joomla, Joom!Fish, TYPO3, Drupal)[5] offer a relatively low level of multilingual content management. None of the platforms supports multiple languages in their

[1] http://www.atlasproject.eu

[2] http://i-publisher.atlasproject.eu/

[3] http://www.i-librarian.eu/

[4] http://eudoclib.atlasproject.eu/

[5] http://wordpress.com/, http://www.joomla.org/, http://www.joomfish.net/, http://typo3.org/, http://drupal.org/

native states. Instead, they rely on plugins to handle this: WordPress uses the WordPress Multilingual Plugin, Drupal needs a module called Locale, and Joomla needs a module called Joomfish. There are modules, like those provided by ICanLocalize[6], than can facilitate selection within Drupal and WordPress of the material to be translated, but the actual translation is done by human translators. To the best of our knowledge, none of the existing content management systems exploits language technologies to provide more sophisticated text content management. This is proved by the data published at the CMS Critic[7] - an online media providing news, reviews, articles and interviews for about 60 content management systems. Taking into account that the online data are in many cases multilingual and documents stored in a content management system are usually related by means of sharing similar topics or domains it can be claimed that the web content management systems need the power of modern language technologies. In comparison ATLAS offers the advantage of integration of natural language processing in the multilingual content management.

3 Selection of "core" words

ATLAS suggests "core" words (plus phrases and named entities), i.e., the most essential words that capture the main topic of a given document. Currently the selection of core words is carried out in a two-stage process: identification of candidates and ranking. For the identification stage a language processing chain is applied that consists of the following tools: sentence splitter, tokenizer, PoS tagger, lemmatizer, word sense disambiguator (assigns a unique sense to a word), NP extractor (marks up noun phrases in the text) and NE extractor (marks up named entities in the text). After this stage, the target core words are ranked according to their importance scores, which are estimated by features such as frequency, linguistic correlation, phrase length, etc., combined by heuristics to obtain the final ranking strategy. The core words are displayed in several groups: named entities (locations, names, etc.) - both single words and phrases, and noun phrases - terms, multiword expressions or noun phrases with a hight frequency. For example among the "core" noun phrases extracted from Cocoa Fundamentals Guide[8] are the following phrases: *Object-Oriented Programming, Objective-C language, Cocoa application, Cocoa program*, etc. Even though the language processing chains that are applied differ from language to language, this approach offers a common ground for language processing and its results can be comfortably used by advanced language components such as document classification, clause-based summarisation, and statistical machine translation. Content navigation (such as lists of similar documents) based on interlinked text annotations is also provided.

4 Automatic categorisation

Automatic document classification (assigning a document to one or more domains or categories from a set of labels) is of great importance to a modern multilingual web content management system. ATLAS provides automatic multi-label categorisation of documents into one or more predefined categories. This starts with a training phase, in which a statistical model is created based on a set of features from already labelled documents. There are currently four classifiers, two of which exploit the Naïve Bayesian algorithm, the two others Relative entropy and Class-featured centroid, respectively. In the classifying phase, the model is used to assign one or more labels to unlabelled documents. The results from the different classifiers are combined and the final classification result is determined by a majority voting system. The automatic text categorisation is at the present stage able to handle documents in Bulgarian and English. For example, the Cocoa Fundamentals Guide is automatically categorised under the domain *Computer science*, and unter the Topics *Computer science, Graphics and Design, Database Management,* and *Programming*.

5 Text summarization

Two different strategies for obtaining summaries are used in ATLAS. The strategy for short texts is based on identification of the discourse structure and produces a summary that can be classified as a type of excerpt, thus it is possible to indicate the length of the summary as a percentage of the original text. Summarisation of short texts in ATLAS draws on the whole language processing chain and also adds a couple of other modules to

[6] http://www.icanlocalize.com/

[7] http://www.cmscritic.com/

[8] https://developer.apple.com/library/mac/documentation/Cocoa/Conceptual/CocoaFundamentals/CocoaFundamentals.pdf

the chain: clause splitting, anaphora resolution, discourse parsing and summarization. The method used for short texts (Cristea et al. 2005) exploits cohesion and coherence properties of the text to build intermediate structures. Currently, the short text summarisation modules are implemented for English and Romanian.
The strategy for long texts assembles a template summary based on extraction of relevant information specific to different genres and is for the time being still under development.

6 Machine translation

For i-Publisher, machine translation serves as a translation aid for publishing multilingual content. The ability to display content in multiple languages is combined with a computer-aided localization of the templates. Text for a localization is submitted to the translation engine and the output is subject to human post-processing.
For i-Librarian and EuDocLib, and for any website developed with i-Publisher, the machine translation engine provides a translation of the document summary provided earlier in the chain. This will give the user rough clues about documents in different languages, and a basis to decide whether they are to be stored.

6.1 Obtaining training corpora

The development of a translation engine is particularly challenging, as the translation should be able to be used in different domains and within different text genres. In addition, most of the language pairs in question belong to the less resourced group for which bilingual training and test material is available in limited amounts (Gavrila and Vertan 2011). For instance, parallel corpora incorporating Bulgarian are relatively small and usually domain-specific, with mostly literary or administrative texts. ATLAS' administrative subcorpus contains texts from EU legislation created between the years 1958 and 2011, available as an online repositories, i.e., the EuroParl Corpus (Koehn 2005); the JRC-Acquis (Steinberger 2006), and includes all the accessible texts in the target languages. The scientific / administrative subcorpus consists of administrative texts published by the European Medicines Evaluation Agency (EMEA) in the years between 1978 and 2009. It is part of the OPUS collection (Tiedemann 2009). The mass media subcorpus contains news reports as well as some other journalistic texts published in nine Balkan languages and English from October 2002 until the present day on the East Europe information website[9]. The fiction subcorpus was compiled manually by harvesting freely available texts on the Internet, scanning, and from donations by authors. So far, it consists of texts in Bulgarian, English, and German. The subcorpus of informal texts consists of subtitles of films: feature films, documentaries, and animations, all part of the OPUS collection (Tiedemann 2009). Automatic collection of corpora is preferred to manual, and for that purpose a set of simple crawlers was designed. They are modified for each source to ensure efficiency. Figure 1 presents some statistical data for the Bulgarian-English parallel corpus, the largest in the collection (the vertical axis shows the number of words, while the horizontal - the domain distribution).

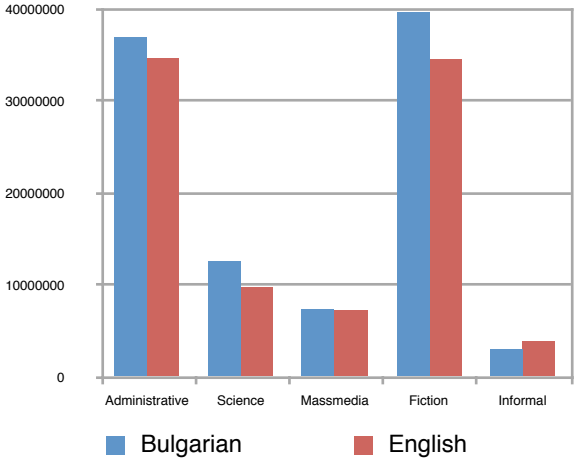

Figure 1 Bulgarian-English parallel corpus

Two basic methods are used to enlarge the existing parallel corpora. In the first, the available training data for statistical machine translation are extended by means of generating paraphrases (e.g. compound nouns are paraphrased into (semi-) equivalent phrases with a preposition, and vice versa). The paraphrases can be classified as morphological (where the difference is between the forms of the phrase constituents), lexical (based on semantic similarity between constituents) and phrasal (based on syntactic transformations). Paraphrase generation methods that operate both on a single monolingual corpus or on parallel corpus are discussed by Madnani and Dorr 2010. For instance, one of the methods for paraphrase generation from a monolingual corpus considers as paraphrases all words and phrases that are distributionally similar, that is, occurring with the

[9] http://setimes.com/

same sets of anchors (Paşca and Dienes 2005). An approach using phrase-based alignment techniques shows how paraphrases in one language can be identified using a phrase in a second language as a pivot (Bannard and Callison-Burch 2005).
The second method performs automatic generation of parallel corpora (Xu and Sun 2011) by means of automatic translation. This method can be applied for language pairs for which parallel corpora are still limited in quantity. If, say, a Bulgarian-English parallel corpus exists, a Bulgarian Polish parallel corpus can be constructed by means of automatic translation from English to Polish. To control the quality of the automatically generated data, multiple translation systems can be used, and the compatibility of the translated outputs can be calculated. Thus, both methods can fill gaps in the available data, the first method by extending existing parallel corpora and the second by automatic construction of parallel corpora.

6.2 Accepted approach

Given that the ATLAS platform deals with languages from different language families and that the engine should support several domains, an interlingua approach is not suitable. Building transfer systems for all language pairs is also time-consuming and does not make the platform easily portable to other languages. When all requirements and limitations are taken into account, corpus-based machine translation paradigms are the best option that can be considered (Karagyozov et al. 2012). For the ATLAS translation engine it was decided to use a hybrid architecture combining example-based and statistical machine translation at the word-based level (i.e., no syntactic trees will be used). The ATLAS translation engine interacts with other modules of the system. For example, the document categorisation module assigns one or more domains to each document, and if no specific trained translation model for the respective domain exists, the user gets a warning that the translation may be inadequate with respect to lexical coverage. Each input item to the translation engine is then processed by the example-based machine translation component.
If the input as a whole or important chunks of it are found in the translation database, the translation equivalents are used and, if necessary, combined (Gavrila 2011). In all other cases the input is sent further to the Moses-based machine translation component which uses a part-of-speech and domain-factored model (Niehues and Waibel 2010).
Like the architecture of the categorization engine, the translation system in ATLAS is able to accommodate and use different third-party translations engines, such as those of Google, Bing, and Yahoo.
The ATLAS machine translation module is still under development. Some experiments in translation between English, German, and Romanian have been performed in order to define: what parameter settings are suitable for language pairs with a rich morphology, what tuning steps lead to significant improvements, wheather the PoS-factored models improve significantly the quality of results (Karagyozov et al. 2012).

7 Conclusion

To conclude, ATLAS enables users to create, organise and publish various types of multilingual documents. ATLAS reduces the manual work by using automatic classification of documents and helps users to decide about a document by providing summaries of documents and their translations. Moreover, the user can easily find the most relevant texts within large document collections and get a brief overview of their content. A modern web content management systems should help users come to grips with the growing complexity of today's multilingual websites. ATLAS answers to this task.

Acknowledgments

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

References

Bannard and Callison-Burch 2005: Bannard, Colin and Chris Callison-Burch. Paraphrasing with bilingual parallel corpora. In *Proceedings of ACL,* pages 597–604, Ann Arbor, MI.

Cristea et al. 2005: Cristea, D., Postolache, O., Pistol, I. (2005). Summarisation through Discourse Structure. *Computational Linguistics and Intelligent Text Processing, 6th International Conference CICLing 2005* (pp. 632-644). Mexico City, Mexico: Springer LNSC, vol. 3406.

Gavrila 2011: Gavrila, M. Constrained recombination in an example-based machine translation system. In M. L. Vincent Vondeghinste (Ed.), *15th Annual Conference of the European Association for Machine Translation*, Leuven, Belgium, pp. 193-200.

Gavrila and Vertan 2011: Gavrila Monica and Cristina Vertan. Training data in statistical machine translation – the more, the better? In *Proceedings of the RANLP-2011 Conference*, September 2011, Hissar, Bulgaria, pp. 551-556.

Karagyozov et al. 2012: Diman Karagiozov, Anelia Belogay, Dan Cristea, Svetla Koeva, Maciej Ogrodniczuk, Polivios Raxis, Emil Stoyanov and Cristina Vertan. i-Librarian – Free online library for European citizens, In *Infotheca*, Belgrade, to appear.

Koehn 2005: Koehn, Ph. Europarl: A Parallel Corpus for Statistical Machine Translation*, Proceedings of MT Summit*, pp. 79–86.

Madnani and Dorr 2010: Nitin Madnani and Bonnie Dorr. 2010. Generat- ing phrasal and sentential paraphrases: A survey of data-driven methods. *Computational Linguistics,* 36(3), pp. 341–388.

Niehues and Waibel 2010: Niehues Jan and Alex Waibel, *Domain Adaptation in Statistical Machine Translation using Factored Translation Models*, Proceedings of EAMT 2010 Saint-Raphael.

Paşca and Dienes 2005: Paşca, Marius and Péter Dienes. 2005. Aligning needles in a haystack: Paraphrase acquisition across the Web. In *Proceedings of IJCNLP*, Jeju Island, pp. 119-130.

Steinberger et al. 2006: Steinberger, R., Pouliquen, B., Widiger, A., Ignat, C., Erjavec, T., Tufiş, D., Varga, D. (2006). The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages. *Proceedings of LREC 2006*. Genoa, Italy.

Tiedemann 2009: Tiedemann, J. News from OPUS - A Collection of Multilingual Parallel Corpora with Tools and Interfaces. In: N. Nicolov, K. Bontcheva, G. Angelova, R. Mitkov (eds.) *Recent Advances in Natural Language Processing* (vol. V), John Benjamins, Amsterdam/Philadelphia, pp. 237–248.

Xu and Sun 2011: Jia Xu and Weiwei Sun. Generating virtual parallel corpus: A compatibility centric method. In Proceedings of the Machine Translation Summit XIII.