E-law Module Supporting Lawyers in the Process of Knowledge Discovery from Legal Documents

Marek Kozłowski
National Information Processing Institute

Maciej Kowalski
National Information Processing Institute

Maciej Kazuła
National Information Processing Institute

{marek.kozlowski,maciej.kowalski,maciej.kazula}@opi.org.pl

Abstract

E-law module is the web application which works mainly as the set of information retrieval and extraction tools dedicated for the lawyers. E-law module consists of following tools: (1) document search engine; (2) context oriented search engine plugin; (3) legal phrase oriented machine translation; (4) document meta-tagger; (5) verdict finder. Machine translation, document meta-tagger and verdict finder tools are available for the general public. Other tools are restricted and are accessible after logging into the module.

1 Introduction

E-law module is being built by the CTI (Center of Information Technologies for the Social Science) consortium, which is granted with the European funds. The main goal of the consortium is to build innovative hardware and software infrastructure for lawyers, sociologists, psychologists and other humanists.

The consortium consists of three members: Cardinal Stefan Wyszyski University in Warsaw, Military Institute of Aviation Medicine, and National Information Processing Institute (OPI).

OPI as a member of the consortium is responsible for delivering software infrastructure, namely three modules: (1) E-law[1] – module supporting lawyers with text mining functionalities e.g., as classifiers, machine translation or search engines, (2) E-survey[2] – module responsible for creating questionnaires in a drag-and-drop wizard mode and sending them to respondents, (3) E-analytics[3] – module supporting social sciences researchers in performing the qualitative analysis for various data (statistical tools, predicative and simulation methods).

E-law module consists of following tools: (1) document search engine (2) context-oriented search engine plugin (3) legal phrase oriented machine translation (4) document meta-tagger (5) verdict finder.

2 Approach

2.1 Document Search Engine

During the project, 2 million legal documents have been downloaded from various, Polish and foreign European, open databases. Only the metadata about documents were collected: title, summary, depositors, date, keywords etc. The search engine retrieves results using Apache Lucene index[4] and well defined filters. For building the Lucene index, different analyzers depending on the language were used. Morfologik[5] analyzer was used for Polish, Standard analyzer was used for other languages.

2.2 Context Oriented Search Engine Plugin

We expanded our search engine with the plugin, which finds all relevant contexts for the query and cluster results (documents) according to the contexts. Most of currently used IR (Information Retrieval) approaches are based on lexico-syntactic analysis of text and they are mainly focused on words occurrences. Two main flaws of the approach are: inability to identify documents using different wordings and lack of context-awareness, which leads to retrieval of unwanted documents. Knowledge of an actual meaning of a polysemous word can improve the quality of the information retrieval process. However, the current generation

[1] http://eprawo-test.opi.org.pl/
[2] http://esurvey-test.opi.org.pl/
[3] http://eanalytics-test.opi.org.pl/
[4] https://lucene.apache.org
[5] It provides dictionary driven lemmatization filter and analyzer for the Polish Language, driven by the Morfologik library https://github.com/morfologik

Proceedings of the 5th Workshop on Balto-Slavic Natural Language Processing, pages 46–48,
Hissar, Bulgaria, 10–11 September 2015.

of search engines still lack an effective way to address the issue of lexical ambiguity. In a recent study (Sanderson, 2008) conducted using WordNet and Wikipedia as sources of ambiguous words it was reported that around 3% of Web queries and 23% of the most frequent queries are ambiguous. In the previous years, Web clustering engines (Carpineto et al., 2009) have been proposed as a solution to the issue of lexical ambiguity in IR. These systems group search results, by providing a cluster for each specific topic of the input query.

In the module presented, a novel result clustering method has been introduced, which exploits rule association mining in order to create coherent clusters of results concerning different subtopics. The core part is a frequent term sets mining method identifying closed frequent termsets using CHARM algorithm (Zaki and Hsiao, 2002). Discovered frequent termsets are hierarchized and used for building labeled trees of patterns.

2.3 Legal Phrase Oriented Machine Translation

One of the key features of E-law module was to aid law-related people with translating legal phrases from Polish to English and vice-versa.

The created translation system uses parallel bilingual data (Polish and English). The total amount of Polish-English data is approximately 42.000.000 pairs of words, phrases, sentences and whole documents (different granularity), incorporated from sources e.g., EUPARL, TED, CURIA, EURLEX.

The process starts from the data alignment based on the PoS oriented floating window of the correspondent block of text. Next there is processed final translation as follows: (1) Input phrase split by tokenizer into n-grams of the predefined maximum size; (2) Each n-gram is taken as a query to Lucene index and corresponding result text block is narrowed down using data alignment method. (3) Each result block is processed by the tokenizer (point 1) and stored in a sorted list. (4) Translation uses replacement by most frequent n-grams, starting from the longest n-grams.

Presented solution cannot compete against currently working SMT solutions like Joshua and Moses (up to 0.20 higher BLEU than the described solution) (Koehn, 2005; Machado and Hilario, 2014). Although the simplicity and little amount of RAM necessary makes this approach useful.

2.4 Document Meta-Tagger

Document Meta-Tagger is a tool, which assigns the high-level keywords to the text using the external knowledge resources i.e., BabelNet. BabelNet[6] is both a multilingual, encyclopedic dictionary with lexicographic and encyclopedic coverage of terms and a semantic network, connecting concepts and named entities in a very large network of semantic relations, called Babel synsets. Each BabelNet synset represents a given meaning and contains all the synonyms, which express that meaning in a range of different languages. BabelNet 3.0 covers and is obtained from the automatic integration of Wikipedia, WordNet, Wiktionary and Wikidata (Navigli and Ponzetto, 2012). The meta-tagger presented works on BabelNet synsets. It performs tokenization as the first step, removing stop-words, lower-casing, lemmatization and PoS tagging. We only persist the noun-phrases, because there are the most informative ones. Next we use the BabelNet API in order to disambiguate phrases. The result of the disambiguation step is the most probable synset. Each synset has it's categories (like Wikipedia categories describing articles). Within the text, all synsets are gathered and the most frequent categories of the synsets are retrieved as the meta-tags.

2.5 Verdict Finder

This tool refers to information extraction(IE). IE deals with unstructured or semi-structured machine-readable documents. The most popular tasks in IE are: named entity recognition, coreference and relationship identification, table extraction or the terminology extraction.

In the legal judgments we are interested in extracting article's legal numbers, which were used as the law references.

The IE is performed as follows: (1) Judgments processing using Apache Tika. (2) Article's legal numbers extraction using regular expressions, which come from the retrieved content files.

For each document the vector of legal article's numbers is build. Such vector representation is used in order to find similar verdicts. Similarity between vectors is measured by the Jaccard metric. The 10 most similar ones are returned as the potentially similar judgments.

[6] http://babelnet.org

References

Mark Sanderson. 2008. Ambiguous queries: test collections need more sense. *Proceedings of SIGIR*, pages 499–506. ACM, New York.

Claudio Carpineto, Stanislaw Osinski, Giovanni Romano and David Weiss. 2009. A survey of web clustering engines. *ACM Computing Surveys 41(3)*, pages 1–38. ACM, New York.

Mohammed Zaki and Ching Hsiao. 2002. CHARM: An efficient algorithm for closed itemset mining. *Proceedings 2002 SIAM Int. Conf. Data Mining*, pages 457–472. Arlington.

Roberto Navigli and Simone Ponzetto. 2012. The Automatic Construction, Evaluation and Application of a Wide-Coverage Multilingual Semantic Network. *Artificial Intelligence*, pages 217–250.

Maria Jose Machado and Hilario Leal Fontes. 2014. Moses for Mere Mortals. Tutorial. https://github.com/jladcr/Moses-for-Mere-Mortals/blob/master/Tutorial.pdf.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. *MT Summit*, pages 79–86.