Semantic Enrichment and Analysis of Legal Domain Documents

M. Besher Massri
Jožef Stefan Institute
Jamova 39, 1000 Ljubljana, Slovenia
besher.massri@ijs.si

Sara Brezec
Jožef Stefan Institute
Jamova 39, 1000 Ljubljana, Slovenia
sara.brezec@ijs.si

Erik Novak
Jožef Stefan Institute
Jožef Stefan International Postgraduate School
Jamova 39, 1000 Ljubljana, Slovenia
erik.novak@ijs.si

Klemen Kenda
Jožef Stefan Institute
Jožef Stefan International Postgraduate School
Jamova 39, 1000 Ljubljana, Slovenia
klemen.kenda@ijs.si

ABSTRACT

In text mining document enrichment processes are used to improve information retrieval. Document enrichment helps us extract metadata from the text which can then be used in document classification.

This paper presents the legal domain document enrichment process and analysis of the enriched data. The process of enriching the documents with multiple layers of annotations is described. The focus is on legal domain documents data set, but the proposed procedure can be generalized to any type of documents.

Keywords

document enrichment, semantic annotations, ontology, analysis, legal domain

1. INTRODUCTION

Document enrichment process helps to improve information retrieval. Nowadays, more and more data has to be processed which makes information retrieval systems extremely valuable. Using document enrichment, more information can be gained about the documents which can be optimized for retrieval.

In the legal domain, extracting meta data about the legal domain documents improves building search engines which are designed to help lawyers efficiently access documents related to a certain topic. In this paper, we present an enrichment process of the legal domain documents. Different types of annotations are used to enrich the data; word-level features which are associated with word information, Wikipedia concepts gained by the process of Wikification and InforMEA ontology terms that cover the field of Environmental Law and Governance. Next, preliminary analysis on the enriched documents is used to review the results. Throughout the paper the focus is on legal domain documents. This approach can be generalized to other document data sets. Our contribution is applying semantic annotation and mapping with ontology on environmental legal domain documents.

The remainder of the paper is structured as follows: Section 2 is related work. Next, the data set is described in section 3. Section 4 presents the methodology used for the document enrichment process. Analysis of the results is in section 5 and finally, we present future work and conclusion in section 6.

2. RELATED WORK

Much work has been done on semantic enrichment of text. Some tools provide a generic pipeline that can be applied and embedded into more complex pipelines. Such pipelines include word and sentence tokenization, part of speech tagging, dependency parsing, and named entity recognition. Examples of such tools are software packages or libraries for different languages, like Spacy [5], Scikit Learn [14], Stanford CoreNLP [11], and MITIE [4].

Semantic enrichment methods have been used to improve the features when building classification models of documents in different domains. An example of this can be found in [7], where two levels of semantic enrichment were used before and after training to classify medical domain documents. In [1], they used dependency parsing, ProbBank [9], and hypernyms from WordNet [13] among other syntactic and semantic features to build relation classification models for the SemEval-2010 Task 8. We also see in [10], the use of mapped cross-domain ontologies in improving information retrieval in the biomedical and chemical domain documents.

In this paper, some of the tools and techniques will be used plus others, mentioned above, and applied to the legal environmental domain documents, providing further analysis about information extracted from the corpus based on the enrichment process.

3. DESCRIPTION OF DATA

We used EUR-Lex, an online service that provides different documents regarding the European Union, as a source to extract our data [3]. For each document, a set of descriptors or keywords was provided among other metadata, in addition to the document title and text. Based on the descriptors and the language of the text, the environmental legal documents were filtered which were provided in the English language and used as the main source of data for document enrichment. The resulting data set, after filtering and cleaning, was around 72k documents.

After preliminary inspection of the data, the documents vary greatly in length. The longest document contains about 560k words whereas the shortest contains 27 words. Nevertheless, approximately 99% of the documents have less than 30k words, 90% of them have less than 5k words and 66.6% have under a 1000 words. Sometimes it can be noticed that classification models produce better results on sets of documents with similar length. Mentioned numbers indicate the potential of providing more precise classification on a set of documents where only few documents are removed from the initial data set.

4. DATA ENRICHMENT PROCESS

4.1 Standard NLP pipeline Annotations

As a first step in data enrichment process, the traditional natural language processing analysis methods were used. The Stanford CoreNLP library was chosen, which is a set of human-language technology tools developed at Stanford University [11]. Using the library, the documents were tokenized into words and then a set of basic syntactic and semantic information was extracted for each word:

- The tokenized word
- The lemma, or dictionary form of the word
- The part of speech of the word in the text.
- Set of synonyms for the word using WordNet lexical database [13], when applicable.

In addition, entity recognition methods were used to identify entities that were categorized into following 11 category classes:

- Named entity classes: PERSON, LOCATION, ORGANIZATION, and MISC
- Numerical entity classes: MONEY, NUMBER, ORDINAL, and PERCENT.
- Temporal entity classes: DATE, TIME, and DURATION.

The MISC category represents an entity mention that was not classified in any of the mentioned classes. An example of these entities are document types (‘Regulation’) and languages (‘English’). Other classes are self-explanatory.

4.2 Wikification

The second annotation step was wikification, which is extracting entities with a relevant Wikipedia concept from the text.The JSI Wikifier tool was used, which is a service developed in Jozef Stefan Institute, that annotates a given raw text with annotations each representing a Wikipedia concept [8].

For each document in our data set, we used Wikifier on the raw text provided and obtained a list of annotation objects; each contains the following information:

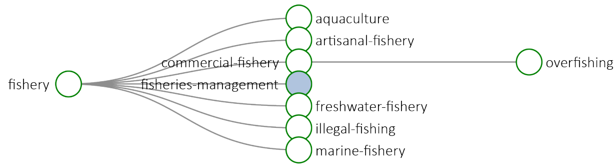

Figure 1: A snapshot that contains a subset of the InforMEA ontology tree.

- The annotation name representing the Wikipedia concept
- Wikipedia page URL of the annotation
- Wiki data classes: the set of classes from WikiData knowledge base [6] that this annotation belongs to.
- One of the DBPedia [1] identifiers that corresponds to the annotation.
- The page rank score of the annotation.
- The cosine similarity between the the document text and the Wikipedia page that the annotation represents.

4.3 InforMEA Ontology

Finally, to provide information about the potential environmental categories that the documents are categorized into, InforMEA ontology was used to map the document with relevant environmental ontology terms. The ontology has 532 unique terms that form a hierarchical structure based on the ‘broader’ relation between ontology concepts. A subset of the ontology tree visualization representing the branch ‘fishery’ is shown in figure 1. More detail, along with the ontology tree, is available on GitHub [12].

To annotate the documents with InforMEA Ontology terms, a simple string matching method was used between the ontology terms and the metadata provided. For each document, the following enrichment data was used to search through for words that matched with any ontology terms:

- The normalized words of the documents
- The synonyms of those words
- The wiki-data classes of the Wikipedia annotations extracted from the document

The reason for using the wiki-data classes instead of the Wikipedia concepts themselves is that the Wikipedia concepts are usually too specific to match with an ontology term, whereas the Wiki data classes represent the topic or the category that this concepts falls into. In fact, the Wikipedia concepts were included in the initial experiments, but had to be omitted later as they did not produce any matches.

5. ANALYSIS OF RESULTS

After annotation was done, extracted information was analysed to get an initial evaluation about the nature of the corpus.

Content analysis produced the most frequent words which are associated with document type or legal body, such as council, state, and member. After removing stop words and numbers as they were not relevant, the TF-IDF analysis produced similar outcomes to the normal word counting analysis. The TF-IDF analysis is presented as a word cloud in Figure 2.

Figure 2: Words with the highest TF-IDF value counted over all documents. The TF-IDF value measures the importance of the word to a document.

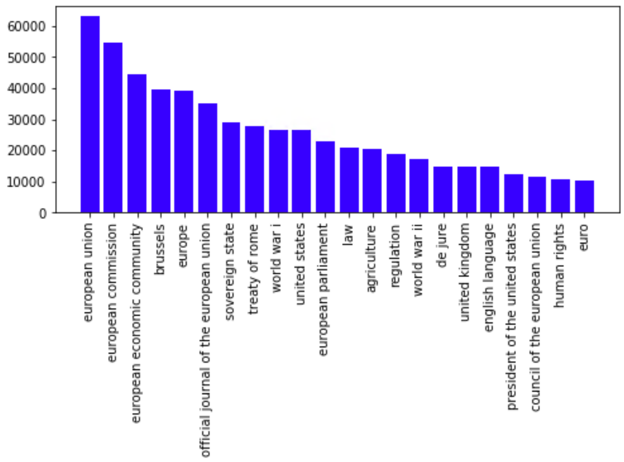

Figure 3: The most frequent Wikipedia concepts found in the enriched data set. The majority of the concepts are associated with law, environment and geo-spatial features.

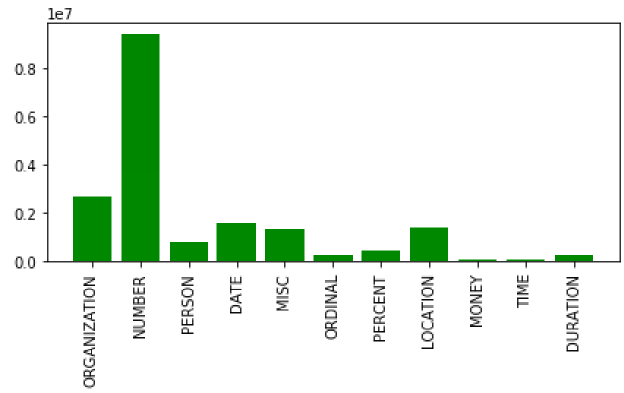

Figure 4: Distribution of 11 entity classes in the annotated data set. The class with by far the highest frequency is NUMBER; we can find them in page numbers, article numbers and other similar locations.

Out of the 72k documents in the corpus, 157k unique Wikipedia concepts were extracted, with only 22 of them having occurrences in over 10k documents. Furthermore, about 50k concepts appear in only one document and 100k concepts appear in up to three documents. This indicates that most of the concepts are unique to the documents. In regards to Wikipedia concepts, the most frequent Wikipedia concepts are shown in Figure 3. The majority of the concepts can be associated with the European union. From the same figure, some concepts can be associated with law and environment, such as "law", "agriculture" and "regulation". This indicates that the process of wikification is able to acquire relevant information. In addition, Geo-spatial concepts are extracted through the process. Their presence can be acknowledged in the country names which are also amongst the most frequently found Wikipedia concepts. Nonetheless, the wikification process was able to find concepts for which connection with the documents is not clear. This will be investigated in future work.

When inspecting the entities extracted, it was observed that 18M entities were obtained through the annotation process with 1.08M distinct entities. Entities were categorized into 11 classes mentioned in the section 4.1. Figure 4 shows the distribution of the classes across all documents. The most frequent class was NUMBER. Numbers appeared in page numbers, article numbers and other similar locations. After removing the NUMBER class, the number of unique entities became 483k. The classes which were the most interesting were LOCATION, ORGANIZATION, and PERSON, since these classes can help in identifying people and organizations that are mentioned in the legal documents, and locations enable the mapping of the legal documents with the geo-spatial information.

Most frequently occurred LOCATION named entities were country names. In the ORGANIZATION class legal bodies were mainly found; almost all of them were associated with the European Union. In almost every document at least one ORGANIZATION and one LOCATION entity appeared.

In comparison to Wikipedia concepts and entity results, a similar pattern of results was obtained from the analysis performed on the word-level features. Therefore, we omit the representation of the outcomes.

Finally, the analysis of the InforMEA ontology mapping is presented. The mapping was done between ontology terms and terms from Wikipedia data classes, normalized words and word synonyms. The distribution of most frequent ontology terms is presented in Figure 5.

The most frequent ontology term 'committee' can be found in other annotation classes as 'commission'. Additionally, ontology terms associated with organizations, logistics and the environment appear amongst the most frequent ontology terms.

6. CONCLUSION AND FUTURE WORK

In conclusion, a semantic enrichment methodology consisting of three main processes; annotation, wikification and

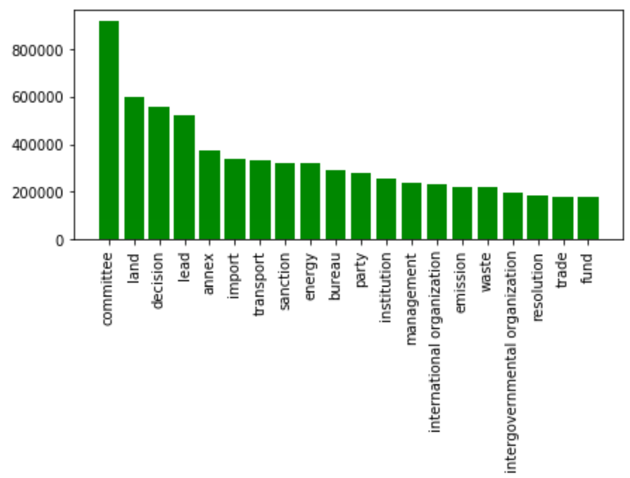

Figure 5: The 20 most frequent ontology terms. Terms are chosen from the aggregated set of normalized words, word synonyms, and wikidata classses of the extracted wikipedia annotations.

mapping to the InforMEA ontology, was performed on legal domain documents. In addition, the analysis on the extracted metadata was provided on the corpus scale to examine the nature of the dataset semantics.

Based on the analysis, some problems were observed with the wikification process as it produced a few unrelated matches. The plan is to address this problem in more detail, observe the reasons behind them, and if possible, try to partly solve the problem.

Regarding the named entities annotation, consideration of adding more finely-tuned annotations, like geo-spatial locations, would help in providing more accurate metadata about the documents. Furthermore, improvement could be made on the baseline string matching that was used to match documents with InforMEA ontology terms. By building classification models, the intention is to use the extracted annotations as features among others.

Finally, the enrichment was mainly done to provide additional metadata on the documents that will be used in later processes. Later plans for further work will be to use the annotations in query expansion to improve legal document retrieval.

7. ACKNOWLEDGMENTS

This work was supported by the Slovenian Research Agency and EnviroLens European Unions Horizon 2020 project under grant agreement No 821918 [2].

8. REFERENCES

[1] DBPedia knowledge graph. `https://wiki.dbpedia.org/`. Accessed in: August 2019.

[2] EnviroLens project. `https://envirolens.eu/`. Accessed in: August 2019.

[3] Eur-Lex. `https://eur-lex.europa.eu/homepage.html`. Accessed in: August 2019.

[4] MITIE: Mit information extraction. `https://github.com/mit-nlp/MITIE`. Accessed in: August 2019.

[5] spaCy industrial-strength natural language processing in python. `https://spacy.io/`. Accessed in: August 2019.

[6] WikiData the free knowledge base. `https://www.wikidata.org`. Accessed in: August 2019.

[7] Albitar, S., Espinasse, B., and Fournier, S. Semantic enrichments in text supervised classification: Application to medical domain. In *FLAIRS Conference* (2014).

[8] Brank, J., Leban, G., and Grobelnik, M. Annotating documents with relevant wikipedia concepts.

[9] Kingsbury, P., and Palmer, M. From TreeBank to PropBank. In *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC'02)* (Las Palmas, Canary Islands - Spain, May 2002), European Language Resources Association (ELRA).

[10] KÖHNCKE, B., and BALKE, W.-T. Enriching documents with context terms from cross-domain ontologies. *Information and Media Technologies 10*, 2 (2015), 294–304.

[11] Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S. J., and McClosky, D. The Stanford CoreNLP natural language processing toolkit. In *Association for Computational Linguistics (ACL) System Demonstrations* (2014), pp. 55–60.

[12] Massri, M. Ontology tree visualizer. `https://github.com/besher-massri/OntologyTreeVisualizer`. Accessed in: August 2019.

[13] Miller, G. A. Wordnet: A lexical database for english. *Commun. ACM 38*, 11 (Nov. 1995), 39–41.

[14] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research 12* (2011), 2825–2830.