Towards a Platform for Curation Technologies: Enriching Text Collections with a Semantic-Web Layer

Peter Bourgonje, Julian Moreno-Schneider, Jan Nehring, Georg Rehm(✉),
Felix Sasaki, and Ankit Srivastava

Language Technology Lab, DFKI GmbH, Alt-Moabit 91c, 10559 Berlin, Germany
georg.rehm@dfki.de

Abstract. In an attempt to put a Semantic Web-layer that provides linguistic analysis and discourse information on top of digital content, we develop a platform for digital curation technologies. The platform offers language-, knowledge- and data-aware services as a flexible set of workflows and pipelines for the efficient processing of various types of digital content. The platform is intended to enable human experts (knowledge workers) to get a grasp and understand the contents of large document collections in an efficient way so that they can curate, process and further analyse the collection according to their sector-specific needs.

Keywords: Digital Curation · Linguistic Linked Data · NLP

1 Introduction

The target audience of our platform are knowledge workers who conduct research in specific domains with the goal of, for example, preparing museum exhibitions or writing news articles. Typically they only have limited time available to accomplish their tasks, ranging from several hours to one or two weeks at most. Owing to the diversity of tasks, the domains are often new to them. The output of their work is typically used in online or traditional media (e.g., newspapers, agencies, tv stations) or by a museum for an exhibition. In the project Digitale Kuratierungstechnologien[1] (DKT, Digital Curation Technologies, [7]), we aim at automating specific parts of the workflows, which consist of looking for information related to and relevant for the domain, learning the key concepts, selecting the most relevant parts and preparing the information to be used. We build a platform that integrates various Natural Language Processing (NLP), Information Retrieval (IR) and Machine Translation (MT) components to retrieve information and recombine them to produce output that improves the curation processes and makes them more efficient. We work with heterogeneous data sets from public and non-public sources. The output will be in the form of a semantically enriched hypertext graph, stored and accessed using linked data

[1] http://digitale-kuratierung.de.

H. Sack et al. (Eds.): ESWC 2016 Satellite Events, LNCS 9989, pp. 65–68, 2016.
DOI: 10.1007/978-3-319-47602-5_14

technologies. Our goal is to enable knowledge workers to explore and curate document collections easier and more efficiently [6].

2 Related Work

Our digital curation platform integrates individual components by linking content with metadata [2]. The components include open-source tools for NLP tasks such as Named Entity Recognition (NER), Information Extraction (IE), IR and MT. The platform uses an architecture developed in the project FREME [8]. A related platform focusing on the localisation industry as the main use case is described by [1]. Our platform targets multiple other industry sectors, has different digital content formats in focus and deploys a different approach to representing curated information.

3 Semantic Layer

For the purpose of this paper we focus on written text documents. Eventually we will also be able to handle the conversion of non-textual data into text (e.g., transcripts for audio, subtitles for video, etc.). On top of text data we generate a Semantic Layer (SL) that contains semantic annotations. The SL creates an interlinked representation connected to external information sources. It is produced by a set of tools that communicate using the NLP Interchange Format (NIF).[2] It operates in a pipelined workflow where the output of each service is used as input for the next one. The SL can be used for exploratory search. The user query is sent through the same pipeline used to generate the SL over the whole document collection. This allows us to search the index for the plain words in the query, but also any entities or temporal expressions that were recognized. The components of the pipelined workflow are:

- **NLP:** This component consists of NER combining a model approach and a dictionary approach. It works with three types of entities: persons, locations and organizations. Any entities found in the input are annotated using NIF. After NER, we also perform Entity Linking using DBPedia Spotlight [4] to retrieve the relevant (DBPedia) URI for entities recognized with the model and on the URI directly taken from the dictionary (the dictionary specifies a key – the entity – and a value, i.e., a URI in some ontology) for entities recognized with the dictionary. Subsequently the NLP workflow performs a temporal expression analysis. This module consists of a language-specific regular expression grammar and currently supports German and English. The expressions are normalized to a machine readable format and added to the NIF model.

[2] http://persistence.uni-leipzig.org/nlp2rdf/.

- **Information extraction:** We use Lucene[3] to create an index for our document collection that enables text-based IR. In addition to indexing the text content, entities and temporal expressions have their own specific fields in order to allow search in the SL as well. Indexing entities also allows us to disambiguate based on entity clustering (planned for the next phase in this two-year project).
- **Semantic Storage:** The semantic information generated during the NLP processes is stored in the triple store Sesame.[4] We use an ontology relating the semantic information extracted from the documents. It relies on Schema.org to describe entities and contains documents and concepts, where the concepts are divided into locations, organizations, persons and temporal expressions.
- **Multilingual component:** This component is based on Moses[5] enhanced with pre-/post-processing modules to leverage the information obtained from preceding steps (e.g., NER, temporal analysis). The MT system is capable of translating both segments (sentences, subtitles) and documents enabling knowledge workers to retrieve information and to present the semantically-enriched output in several languages (English, German, Spanish, Arabic). Preliminary experiments show as much as a 5 % improvement in the overall MT system performance for multiple language pairs and domains.

4 Experiments

Our goal is to reduce the time knowledge workers invest in their sector-specific curation processes. A proper evaluation would require us to measure the time it takes knowledge workers to get from input to output with and without utilization of our platform. This is rather difficult to measure and to quantify. We are in an early stage of the project and do not have access to suitable data for such an evaluation yet. As the project progresses, we will acquire real-world data (to be provided by the industrial partners involved in the project), annotate the data to construct a gold standard so that the platform can be evaluated as a whole. For now, we can offer isolated evaluations of individual components. For evaluating the German version of the temporal expression analyzer, we use the German WikiWars corpus [3]. This corpus is a collection of 22 documents sourced from Wikipedia pages about military conflicts and contains 2.240 temporal expressions. Evaluating against this corpus, we can report an f-score of *0.83*. However, we developed against this same corpus and since we are mainly interested in coverage of our regular expressions, the corpus was not divided into training and test sets. We consider our f-score an acceptable baseline and will continue to improve this during the project. For evaluating the German version of the NER module, we selected the German wikiNER corpus [5]. This corpus contains NER annotations in CoNLL format. For this we can report f-scores of *0.78*, *0.87* and *0.76* for locations, persons and organizations, respectively. These numbers will serve as baselines for future work as well.

[3] https://lucene.apache.org/core/.

[4] http://rdf4j.org/.

[5] http://www.statmt.org/moses/.

5 Conclusion and Future Work

This article addresses the issue of combining NLP, IR and MT procedures into a system that enables knowledge workers to explore a collection of documents in an intuitive and efficient way. Our focus is on combining the individual components and linking the output of the methods, rather than trying to improve upon the output of individual state-of-the-art procedures. In this early stage of the project, we can aggregate the information contained in multiple documents and present this in a way that allows the knowledge worker to see what is inside. For the future we plan to work on making our tools easily adaptable to new domains. This poses a challenge since we expect to deal with domains for which only limited amounts of training data are available. We also plan to exploit the linked open data framework more by plugging in new datasets. Future applications are related to the project goals: *text summarization* of documents will help the curation process and *semantic story-telling* will assist in text generation processes, relating individual document components at a semantic level.

Acknowledgments. "Digitale Kuratierungstechnologien" is supported by the German Federal Ministry of Education and Research, Unternehmen Region, WK-P (No. 03WKP45).

References

1. Lewis, D., Brennan, R., Finn, L., Jones, D., Meehan, A., O'sullivan, D., Hellmann, S., Sasaki, F.: Global intelligent content: active curation of language resources using linked data. In: Proceedings of LREC 2014, Reykjavik (2014)
2. Lewis, D., Gómez-Pérez, A., Hellman, S., Sasaki, F.: The role of linked data for content annotation and translation. In: Proceedings of the 2014 European Data Forum. EDF 2014 (2014). http://2014.data-forum.eu
3. Mazur, P., Dale, R.: WikiWars: a new corpus for research on temporal expressions. In: Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing. EMNLP 2010, pp. 913–922. Association for Computational Linguistics, Stroudsburg (2010)
4. Mendes, P.N., Jakob, M., García-Silva, A., Bizer, C.: DBpedia spotlight: shedding light on the web of documents. In: Proceedings of the 7th International Conference on Semantic Systems. I-Semantics 2011, pp. 1–8. ACM, New York (2011)
5. Nothman, J., Ringland, N., Radford, W., Murphy, T., Curran, J.R.: Learning multilingual named entity recognition from Wikipedia. Artif. Intell. **194**, 151–175 (2012)
6. Rehm, G.: Hypertextsorten: Definition - Struktur - Klassifikation. Ph.D. thesis, Institutfür Germanistik, Angewandte Sprachwissenschaft und Computerlinguistik, Justus-Liebig-Universität Giessen (2005)
7. Rehm, G., Sasaki, F.: Semantische Technologien und Standards für das mehrsprachige Europa. In: Ege, B., Humm, B., Reibold, A. (eds.) Corporate Semantic Web, pp. 247–257. Springer, Heidelberg (2015)
8. Sasaki, F., Gornostay, T., Dojchinovski, M., Osella, M., Mannens, E., Stoitsis, G., Richie, P., Declerck, T., Koidl, K.: Introducing freme: deploying linguistic linked data. In: Proceedings of the 4th Workshop of the Multilingual Semantic Web. MSW 2015 (2015)