DKPro Core – Infrastructure for TDM/NLP applications and research

Commercial use (Y/N): unknown (open source)
Related project (*if any – please specify funder and project name*):
Author: Richard Eckart de Castilho, Iryna Gurevych
Organization: UKP Lab, Technische Universität Darmstadt
Date: Dec 12, 2014

Use case description – *Describe the problem you are trying to address, the research area it affects, and the TDM application used. Please state all participants, whether and how the results were subsequently used.*

DKPro Core is a software infrastructure that offers a wide range of state-of-the-art TDM tools through a common application programming interface. In this way, DKPro Core allows users to:

- **assemble TDM applications** by integrating a range of state-of-the-art TDM tools into a common framework;
- **create reproducible TDM research** based on stable releases of the portable DKPro Core software infrastructure;
- **deploy text analytics close to the text data to be processed**, allowing the protection of potentially sensitive data, and avoiding transferring potentially large amounts of data over the network unnecessarily.

Today, TDM relies significantly on language models that describe language on an abstract level, often through absolute and relative frequencies and weights learned from a text corpus. Such models are required at every stage of language analysis, from relatively simple text segmentation up to entity extraction, fact extraction, sentiment analysis, and so on.

As a consequence, an essential part of the DKPro Core infrastructure involves packaging these models, making them available as part of the infrastructure, and deploying them along with the tools. In order to serve as a stable point of reference for reproducible TDM research, DKPro Core and all associated artifacts such as software but also models are published and distributed through a resilient federated public online repository infrastructure.

Content sources – *List the targeted content sources and types of content used.*

Language models are offered by NLP tool authors, third-parties, or created by the end-user.

Models based on text corpora that are created from arbitrary sources depending on the task or research question at hand. They cover any kind of content including but not limited to news-wire data, literature, social network content.

Targeted users – *Describe end users, their number and expertise.*

All users of natural language processing technology, including researchers and commercial users. DKPro Core is being recognized and used by researchers across the globe. A prominent example is the Excitement Open Platform for Recognizing Textual Entailment[1], the main product of the project EXCITEMENT (Exploring Customer Interactions through Textual EntailMENT) which is funded by the European Commission under the European Union's Seventh Framework Programme (FP7).

Impact – *Describe all possible impacts of the use case. Specify any cross-border, societal and economic effects (if possible state any monetary benefits and market advantages) .*

- **Increasing TDM research productivity.** TDM generally involves obtaining and combining many different tools into TDM pipelines. However, these tools are generally not interoperable and wrappers need to be created to perform data transformation steps between each tool used in a pipeline. DKPro Core integrates many of such third-party tools with a generic TDM interoperability framework, providing researchers with choices and removing the need to wrap these tools over and over again.
- **Enabling reproducibility of TDM research.** The releases of DKPro Core offer a stable point of reference for TDM experiment setups. Its TDM components and third-party libraries are publicly accessible and distributed via repository system and remain unchanged after their release. This is in contrast to service-based TDM components that have been found to decay, either changing, or becoming intermittently or permanently unavailable.

Constraints – *Describe any legal, technical, economic, societal, organizational, cultural, multi-lingual or other limitations and you have encountered.*

- **Model creators generally unable to declare a license** – Those creating models from corpora are generally unable to tell what license the model has. Still, it is a general practice to make such models publicly available for download. While this is convenient for researchers, it is a problem for builders of research infrastructures, because In order to distribute language models through a federation of online repositories, it is important to know their license status. This significantly constrains our ability to make such models available as part of the infrastructure. The legal status of language models is also a problem for commercial users. Questions about language model licenses come up often and typically remain without a clear answer or even with no answer at all. A prominent example of models with an unclear license status are those offered for use with the OpenNLP software[2].
- **Scope of license of data being subject to aggregation and abstraction is unknown –** It is unclear in how far the license status of a corpus can restrict the potential license status of a language model created from the corpus. E.g. if a corpus is not redistributable and requires a (paid) license, can an owner of such a license create a model from data and redistribute it? If so, under which conditions, e.g. under that condition that the original corpus data must not be reconstructable from the language model (e.g. generally the case for abstract statistical models). Can abstract data models be protected by copyright or a similar concept at all, since two persons can easily and independently arrive at the same results by applying

[1] http://hltfbk.github.io/Excitement-Open-Platform/

[2] http://opennlp.sourceforge.net/models-1.5/

the same tool to the same data?

Comments – *Other comments and any recommendations you may have.*

We would welcome the definition of a clear boundary between text corpus data and language models abstracted from this data, covering the following aspects:

1) **When is the boundary between original data and abstract data crossed?** (e.g. when the original data cannot (reliably) be reconstructed)
2) **If the boundary is crossed, under which circumstances can the abstract data be protected by a data license?** (e.g. only if significant and non-trivial additional effort was involved to enrich the data)