Towards Ontological Interpretations for Improved Text Mining

Robert Hoehndorf

Research Group *Ontologies in Medicine*, Institute for Medical Informatics, Statistics and Epidemiology and
Department of Computer Science, University of Leipzig and
Department of Evolutionary Genetics, Max-Planck-Institute for Evolutionary Anthropology
`hoehndorf@eva.mpg.de`

Axel-Cyrille Ngonga Ngomo
Department of Computer Science, University of Leipzig
`ngonga@informatik.uni-leipzig.de`

Michael Dannemann
Department of Evolutionary Genetics
Max-Planck-Institute for Evolutionary Anthropology
`michael_dannemann@eva.mpg.de`

1 Motivation

Text mining in biomedicine can be used for several tasks relating both to the extraction of domain-specific knowledge and the management of ontologies. These tasks include the identification of associations between biomedical entities, the extraction of relationships between biomedical entities, the alignment of ontologies and the generation of ontologies from text. Most of the methods used in text mining to perform these tasks are based on statistical measures, algorithms from natural language processing or machine learning. We believe that the overall performance of these methods remains limited as long as no semantic or *ontological* layer is added in the generation and analysis of text mining data. An ontological layer will allow to interpret the results of a text mining analysis with respect to formalized ontological background knowledge, and can be used to generate an *ontological interpretation* of the results of the analysis. In such an ontological interpretation, categories and individuals stand in well-defined ontological relations. The ontological interpretation of text mining results would present several advantages, of which the most important include consistency checks, automated belief revision (ontology curation) and ontologically founded data and information integration.

The generation and analysis of an ontological interpretation of text mining results are not straight forward, as it is necessary to deal both with inconsistent and incomplete knowledge. Classical logics will prove to be insufficient for such a task. Therefore, a non-classical, non-monotonic logic together with non-classical inferences such as abduction and induction is required.

2 Method

For our purpose, text mining identifies references to four kinds of ontological entities in text: categories C, individuals I, relations R and instances of relations T. A category is an intensional entity that can have instances. Instances of categories can be both individuals or other categories. Individuals cannot be instantiated (Herre et al., 2006). A relation such as *instance-of* or *part-of* is an ontological entity that specifies a kind of interaction between multiple entities. Relations have instances that are part of the world. The instances of relations are "the glue that holds things together, the primary constituents of the facts that go to make up reality" (Barwise, 1988). Without loss of generality, we restrict our discussion to binary relations and $R \subseteq (C \cup I) \times (C \cup I)$. We call the structure $\mathcal{TM} =< C, I, R, T >$ resulting from a text mining analysis a *text mining structure* (TMS).

The global aim of the research proposed herein is to provide an ontological interpretation of such a TMS. We apply this interpretation for the refinement of the TMS using the axioms of an ontology. In order to deal with inconsistent and incomplete knowledge, we use a non-monotonic form of logical deduction as a method to consistently generate explanations for facts resulting from this ontological interpretation.

In our work, an ontology is a structure $O =<$

$C', R', ::, isa, Ax >$ of categories C' and relations R' together with a set of axioms Ax.

Definition 1. *An ontological interpretation I of a TMS $\mathcal{TM} = < C, I, R, T >$ with respect to the ontology $O = < C', R', ::, isa, Ax >$ satisfies:*

- *for each $c \in C$, $c^I = c'$ such that $c' \in C'$ and either $c :: c'$ or $isa(c, c')$,*

- *for each $i \in I$, $i^I = i'$ such that there exists a $c' \in C'$ and $i :: c'$,*

- *for each $r \in R$, $r^I = r'$ such that $r' \in R'$ and $isa(r, r')$,*

- *for each $t \in T$, $t^I = t'$ such that there exists a $r' \in R'$ and $t' :: r'$.*

An ontological interpretation performs the following functions: for each category identified in the text, it identifies at least one category in the ontology O of which the category found in the text is either a sub-category or an instance; for each individual in the text, it identifies at least one category of which this individual is an instance; and similarly for relations and their instances.

Two major difficulties arise when trying to find an ontological interpretation of a TMS. First, it may occur that no ontological interpretation exists due to an inconsistency. In this case, we call the TMS $\mathcal{TM}$ classically inconsistent with the ontology O. Second, there may be many possible ontological interpretations for a TMS, and some measure of preference should be established to select the most appropriate ontological interpretation.

In order to deal with inconsistencies, we attempt to establish classical consistency by extending the ontological interpretation such that identified categories (or instances) are subclasses (or instances) of more general categories. For example, consider a TMS containing the following three relation instances:

$$IsA(Arsenic, Poison) \qquad (1)$$

$$PlaysRole(Arsenic, Poison) \qquad (2)$$

$$HasFunction(Arsenic, Poison) \qquad (3)$$

Here, poison is used in three mutually exclusive meanings: as a substance, a role and a function; any ontological interpretation interpreting *Poison*, *IsA*, *PlaysRole* and *HasFunction* in their usual understanding will be classically inconsistent. Interpreting *Poison* as a subclass of *Entity* avoids

the inconsistency, but does not permit inferences based on axioms pertaining to more specific categories. Abductive reasoning can be used to fill the gap: abduction is a non-classical form of inference that generates a *minimal* explanation for an observation. The general schema for abduction is: $B, A \rightarrow B \vdash A$. As an assumption, we use the following formula, where C_i ranges over all categories from O:

$$isa(Poison, C_1) \vee \ldots$$
$$\vee isa(Poison, C_n) \rightarrow isa(Poison, Entity) \qquad (4)$$

Abduction can then generate the desired and consistent minimal explanation for (4)

$$isa(Poison, Substance) \vee isa(Poison, Role) \vee$$
$$isa(Poison, Function) \qquad (5)$$

3 Conclusion

We suggest that ontological interpretations can improve text mining results by providing an additional semantic structuring layer. This layer can be used to disambiguate the kind of relations and categories identified through text mining, and to identify categories of which recognized named entities are instances. Formal ontologies play a crucial role in this step. The use of abductive reasoning can lead to rich and consistent ontological interpretations that contain explanations for the facts identified through text mining. These explanations can be used subsequently for the identification of novel hypotheses or the integration of knowledge. Ultimately, using ontological interpretations provides a starting point for elevating the results of text mining analyses from data to knowledge.

References

John Barwise. 1988. *The situation in logic*. CSLI Publications.

H. Herre, B. Heller, P. Burek, R. Hoehndorf, F. Loebe, and H. Michalek. 2006. General Formal Ontology (GFO) – A foundational ontology integrating objects and processes [Version 1.0]. Onto-Med Report 8, Research Group Ontologies in Medicine, Institute of Medical Informatics, Statistics and Epidemiology, University of Leipzig, Leipzig, Germany.