Concept Mentions within KDD-2009 Abstracts (kdd09cma1) Linked to a KDD Ontology (kddo1)

Gabor Melli
Simon Fraser University
Burnaby, BC, Canada
E-mail: lrec10@gabormelli.com

Abstract

We introduce the kddo1 ontology and semantically annotated kdd09cma1 corpus from the field of knowledge discovery in database (KDD) research. The corpus is based on the abstracts for the papers accepted into the KDD-2009 conference. Each abstract has its concept mentions identified and, where possible, linked to the appropriate concept in the ontology. The ontology is based on a human generated and readable semantic wiki focused on concepts and relationships for the domain along with other related topics, papers and researchers from information sciences. To our knowledge this is the first ontology and interlinked corpus for a subdiscipline within computing science. The dataset enables the evaluation of supervised approaches to semantic annotation of documents that contain a large number of high-level concepts relative the number of named entity mentions. We plan to continue to evolve the ontology based on the discovered relations within the corpus and to extend the corpus to cover other research paper abstracts from the domain. Both resources are publicly available at http://www.gabormelli.com/Projects/kdd/data/.

1. Introduction

We introduce a dataset composed of the kddo1 ontology and the kdd09cma1 corpus. The corpus is based on the 139 abstracts for the papers accepted into the proceedings of ACM's SIGKDD 15th annual conference on data mining and knowledge discovery (KDD-2009[1]). Each of the abstracts has been annotated to identify the domain-specific concepts mentioned. Further, where feasible, all mentions have been linked to the appropriate concept record in the ontology. To this end, the kddo1 ontology is focused on concepts and relationships for the domain of knowledge discovery in databases (which in turn inherits concepts from machine learning theory, optimization theory, numerical analysis, statistics, and computational linguistics). Both the corpus and ontology are publicly available[2] **Figure 1** graphically illustrates the structure of the dataset. To our knowledge this is the first ontology and interlinked corpus for a subdiscipline within computing science.

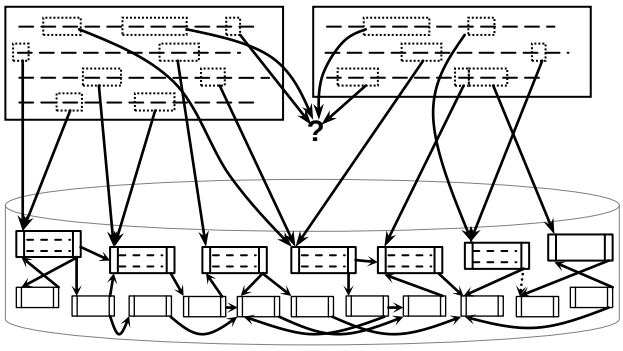

Figure 1 - Illustration of a pair of documents from the corpus with their concept mentions identified and linked to their corresponding ontology record (if it exists).

The kdd09cma1 corpus bears similarities to corpus from the natural science of biomedicine such as the GENIA (Kim & al, 2003), BioCreAtIvE (Hirschman & al, 2005), and PPLRE corpus (Melli & al, 2007). Typically the member documents of such biomedicine corpora are MEDLINE abstracts, and the annotation of the selected documents focuses on mentions of basic named entities such as named molecules, organisms, and locations. The kdd09cma1 corpus on the other hand contains very few named entities. Being from the formal science of computing, its text instead contain abstract concepts such as "*mining*", "*text data*", "*controlled experiment*", "*rounding integer linear program*", and "*minimal biclique set cover problem*". Further, in cases where named entities are mentioned they often are embedded within an abstract concept mention, as in "*Gibbs sampling method*" and "*hidden Markov model*".

Next, the kddo1 ontology bears similarities to ontologies from the bio-medical domain such as: he Gene Ontology[3] or even the MeSH controlled vocabulary[4]. Again, because this resource covers a formal science (rather than a natural science such as biomedicine) the concepts in the ontology tend to involve abstract specifications and relationships.

The corpus and ontology have several possible applications. Initially the annotation can be used directly to analyze the concepts that play an important role in the conference. Without the annotation this type of topic analysis would be restricted to surface level (n-gram) orthographic word combinations; a method which is susceptible to the merging of identical concepts with different surface representations, or the incorrect merging of concept mentions that refer to different concepts. Examples from the corpora of mentions that can refer to the same concept include: "*feature*", "*attribute*", and "*variable*"; and "*empirical test outcome*" and "*experimental result*". Common polysemous concept mentions include: "*model*", "*work*", and "*feature*".

Beyond topic modeling, the dataset may be of interest to fields such as terminology mining (Daille, 2002) to

[1] http://www.kdd.org/kdd/2009
[2] www.gabormelli.com/Projects/kdd/data/kddo/kddo1
www.gabormelli.com/Projects/kdd/data/sigkdd/kdd09cma1
[3] http://www.geneontology.org/
[4] http://www.nlm.nih.gov/mesh/

improve extraction of key phrases. The dataset could be also be used to benchmark future automated annotation systems for information science documents (Melli & Ester, 2010).

Finally, we believe that in the future all peer reviewed published papers in the sciences will require that their abstracts be annotated and reviewed by the authors. This requirement would support the suggestion by researchers from library information sciences that scientific literature will become ever more navigable at a semantic (Renear & Palmer, 2009). The dataset (both the corpus and ontology) could become the seed of a valuable and naturally expanding corpus.

2. The `kddo1` ontology

The `kddo1` ontology is composed of concepts and semantic relations from the domain of knowledge discovery from databases. Development of the ontology began in January 2009. It precedes the creation of the corpus (described in the following section) which took place in November 2009. Minimal enhancement to the ontology was performed once the annotation of the corpus was performed. The choice to limit changes to the ontology was made in part due to the realization that it would not be possible to create an entry for every concept mentioned in the corpus in the allocated amount of time. Further, it is realistic to assume that many similar annotation initiatives by others will also have only a partial ontology for a substantial period of time.

The ontology is based on a semantic wiki (Schaffert, 2006) - that is, the underlying wiki isstructured in a way to facilitate the conversion of its contents to a machine processable ontology. The semantic wiki was created by the author for the field of knowledge discovery from databases[5]. It makes use of structured English and follows the structure proposed in (Melli & McQuinn, 2008), where each concept record contains 1) A unique preferred name, 2) A one sentence definition in the form of "`X` *is a type of* `Y` *that* …", 3) words that are commonly synonymous with the concept, 4) A context that contains statements relating the concept to other concepts in the ontology, 5) Examples of instances of the concept. 6) Examples of related but differing instances, 7) a set of related concepts whose relationship has not been formally defined, and 8) Relevant external references for the concept. Table 1 summarizes some statistics of the ontology.

Concepts in ontology			5,067
Internal links			27,408
	Min	Median	Max
Links into a concept	0	3	157
Links out of a concept	2	3	444
Synonyms per concept	0	1	8

Table 1– Summary statistics of the `kddo1` ontology

[5] Sample records can be found at:
www.gabormelli.com/RKB/Information_Extraction
www.gabormelli.com/RKB/Text_Classification

3. The `kdd09cma1` corpus

The `kdd09cma1` corpus is based on the 139 abstracts of the papers accepted for ACM's SIGKDD annual conference in 2009 (KDD 2009) that are freely accessible from ACM's Digital Library[6]. KDD is a competitive peer-reviewed conference with acceptance rates in the range of 20% -25%. The conference topic is data mining and knowledge discovery from databases.

The abstracts were manually annotated by the author for concept mentions. We define a *concept mention* to be a sequence of tokens (orthographic words and punctuation) whose meaning is deemed by an expert to be used within their community of speakers, and whose meaning is not necessarily well understood by a member of the general public. Often concept mentions are words (*terminological units*), but not always. The mentions can also be phrases. For example the phrase "*problem of web classification*" could be identified as a mention of the `Web_Object Classification_Task` concept.

The identification and linking of concept mentions was mostly performed as two separate phases. We first identified mentions of concepts that would be understood and/or often used within the data mining community. This phase was performed without consideration for what concepts existed in the ontology. Next, an attempt was made to link the mentions to the concept in the ontology (described in the next section) that stood for the intended concept in the mention. On average, the identification task took approximately 6 minutes per abstract. The linking task on the other hand took approximately 17 minutes per abstract. During linking however we occasionally divided long mentions into component spans that would be found in the ontology. For example:

- "... *cascading non-homogeneous Poisson process*" → "… [[`Cascading Stochastic Process`|*cascading*]] [[`Non-Homogeneous Stochastic Process`|*non-homogeneous*]] [[`Poisson Stochastic Process`| *Poisson process*]]
- "… *training multi-label text classifiers*." → "… [[`Training Phase`|*training*]] [[`Multi-Label Classifier`|*multi-label*]] [[`Text Object`|*text*]] [[`Classifier`| *classifiers*]]."

To evaluate the quality of the annotation, sixteen abstracts were randomly selected and the paper's author was asked to review the annotation. Fourteen authors responded and simply accepted the annotation as is.

The text was tokenized and assigned a part-of-speech role by using Charniak's parser (Charniak, 2000). Table 2 summarizes some key statistics about the corpus. Of the 7,580 concept mentions approximately two thirds are single token mentions (e.g. "*data*", "*algorithm*", and "*f-measure*"), and the remaining third are multi-token mentions (e.g.: "*experimental results*", "*real-valued data set*", and "*minimal biclique set cover problem*"). Table 3 summarizes some additional key statistics of the linking

[6] http://portal.acm.org/toc.cfm?id=1557019

(*external links*) between the corpus and ontology.

Documents		139	PER DOCMNT. min|med|max
Sentences		1,186	3 | 8 | 17
Tokens		29,139	105 |220| 367
Concept Mentions	(100%)	7,580	26 | 52 | 96
Single Token	(~66%)	5,001	12 | 35 | 65
Multi Token	(~33%)	2,579	4 | 18 | 38

Table 2 – Summary statistics of the `kdd09cma1` corpus, including the minimum, median, and maximum per abstract.

Documents		139	PER DOCMNT. min|med.|max
Linked Mentions	51.7%	3,920	10 | 26 | 66
Unlinked Mentions	48.3%	3,660	3 | 25 | 49
Distinct Concepts linked to by corpus		820	9 | 19 | 50
Concepts uniquely linked to by a single document			0 | 2 | 17

Table 3 – Summary statistics of the external links from the `kdd09cma1` corpus to the `dmswo1` ontology, including the minimum, median, and maximum per abstract.

illustrates the type of annotation performed on each of the abstracts. Notice that:

1. Some of the demarcated mentions include a phrase followed by a 'bar' (|) character. These expressions are meant to represent ontology concept identifiers – such as a preferred name[7]. The format is intended to replicate the approach used by wiki-based systems such as Wikipedia to redirect a hyperlink. For example "[[*Collaborative Filtering Algorithm*|Collaborative filtering]] identifies the concept record in ontology O with the label "*Collaborative Filtering Algorithm*".

2. Not all concept mentions are mapped to an ontology concept. For example, "*cold start users*" is not associated with an ontology concept. The reason for this assumption is to simulate the real-world scenario were not all concepts are entered into the ontology. An example of this assumption at work is how Wikipedia pages can include links that do not yet lead to some other destination within Wikipedia.

3. Although these two sample sentences do not include named entities, concept mentions can include entity mentions. For example, the term "*Eschirichia Coli*" is a valid concept mention, which can be found in an ontology.

4. Future Work

Several directions for future work are envisioned. The four main directions are to extend the corpus to include other documents; to loosen the annotation requirements to enable focus on more interesting concepts; to expand and enrich the ontology based on the concepts and relations mentioned in the corpus; and, to annotate the relation mentions found in the corpus. Each of these directions is briefly explored.

Unannotated: "*Collaborative filtering is the most popular approach to build recommender systems and has been successfully employed in many applications . However , it cannot make recommendations for so-called cold start users that have rated only a very small number of items*."

Annotated: [[Collaborative Filtering Algorithm |*Collaborative filtering*]] *is the most popular* [[Algorithm|*approach*]] *to build* [[Recommender System|*recommender systems*]] *and has been successfully* [[Computing System Employment Act|*employed*]] *in many* [[Computing Application| *applications*]]. </S> *However, it cannot make* [[Item Recommendation Prediction|*recommendations*]] *for so-called* [[*cold start users*]] *that have* [[Item Rating Act|*rated*]] *only a very small* [[Number|*number*]] of [[Item|*items*]]. </S>

Figure 2 - Annotation sample of two sentences from (Jamali & Ester, 2009)

We plan to extend the corpus to other conferences within the domain of knowledge discovery in databases, including, for example, all past ACM SIGKDD and IEEE ICDM conferences. In order to expedite the process we have begun to develop a system that can automatically perform the annotation by training classifications models based on the dataset (Melli & Ester, 2010). The system will then be applied to the abstracts of the candidate documents to be added. The automation is hoped to accelerate the annotation process. More ambitiously we hope to introduce the annotation task into the paper submission process of future publication venues.

The assumption found in the `kdd09cma1` corpus that every concept mention is identified and linked is likely too restrictive. It is foreseeable that annotation may only exist for particularly interesting and descriptive mentions. For example, articles in Wikipedia only demarcate a small portion of their concept mentions. While extending the corpus we plan to include documents that have only been partially annotated. This extension will allow for the real-world simulation of scenarios where annotators chose only to annotate a passage within a document that contains a particularly clear definition of a concept.

We also plan to release extended versions of the ontology on a regular basis. The next release will intentionally cover more of the unlinked concept mentions and relations found in the `kdd09cma1` corpus. Sample terms to be enhanced or included are: *framework, inference, monitor, online advertising, Netflix Prize,* and *parameter-free*. We also hope to align the ontology to existing ones, such as `SUMO`[8].

Finally we hope to annotate the different relations in the ontology that are mentioned found in the corpus.

[7] Typically a word/phrase with capitalized first letters.

[8] http://www.ontologyportal.org/

5. Conclusion

We present a publicly available dataset composed of a corpus and ontology for the field of knowledge discovery in databases: `kdd09cma1` and `kddo1`. The dataset represents one of the first attempts to add semantic information to concepts from a computing discipline.

6. Acknowledgements

We thank the authors of KDD papers who participated in the review of the annotated versions of their papers' abstracts.

7. References

Eugene Charniak. (2000). "*A Maximum-Entropy-Inspired Parser*." In: Proceedings of NAACL Conference (NAACL 2000)

Béatrice Daille. (2002). "*Terminology Mining*." In: Proceedings of the Summer Convention on Information Extraction (SCIE 2002). [doi>10.1007/b11781]

Lynette Hirschman, Alexander Yeh, Christian Blaschke, and Alfonso Valencia. (2005) "*Overview of BioCreAtIvE: critical assessment of information extraction for biology*." In: BMC Bioinformatics 2005, 6(Suppl 1):S1

Mohsen Jamali, and Martin Ester. (2009). "*TrustWalker: A Random Walk Model for Combining Trust-based and Item-based Recommendation.*" In: Proceedings of ACM SIGKDD Conference (KDD 2009). [doi>10.1145/1557019.1557067].

Jin-Dong Kim, Tomoko Ohta, Yuka Teteisi, and Jun'ichi Tsujii. (2003). "*GENIA Corpus - a semantically annotated corpus for bio-textmining*." In: Bioinformatics. 19(suppl. 1).

Gabor Melli, and Martin Ester. (2010). "*Supervised Identification of Concept Mentions and their Linking to an Ontology*" submitted

Gabor Melli, and Jerre McQuinn. (2008). "*Requirements Specification Using Fact-Oriented Modeling: A Case Study and Generalization*." In: Proceedings of ORM-2008.

Gabor Melli, Martin Ester, and Anoop Sarkar. (2007). "*Recognition of Multi-sentence n-ary Subcellular Localization Mentions in Biomedical Abstracts*." In: Proceedings of LBM-2007.

Allen H. Renear, and Carole L. Palmer. (2009). "*Strategic Reading, Ontologies, and the Future of Scientific Publishing*." In: Science, 325(5942). [doi>10.1126/science.1157784].

Sebastian Schaffert. (2006). "*IkeWiki: A Semantic Wiki for Collaborative Knowledge Management.*" In: 1st International Workshop on Semantic Technologies in Collaborative Applications (STICA 2006).