MeDetect: Domain Entity Annotation in Biomedical References Using Linked Open Data

Li Tian[1], Weinan Zhang[1], Haofen Wang[1], Chenyang Wu[1]

Yuan Ni[2], Feng Cao[2], Yong Yu[1]

[1]Shanghai Jiao Tong University, Shanghai; [2]IBM China Research Laboratory, Beijing
[1]{tianli,wnzhang,whfcarter, wucy, yyu}@apex.sjtu.edu.cn; [2]{niyuan,caofeng}@cn.ibm.com

Abstract. Recently, with the ever-growing use of textual medicine records, annotating domain entities has been regarded as an important task in the biomedical field. On the other hand, the process of interlinking open data sources is being actively pursued within the Linking Open Data (LOD) project. The number of entities and the number of properties describing semantic relationships between entities within the linked data cloud are very large. In this paper, we propose a knowledge-incentive approach based on LOD for entity annotation in the biomedical field. With this approach, we implement MeDetect, a prototype system to solve the problems mentioned above. The experimental results verify the effectiveness and efficiency of our approach.

Keywords: Domain Entity Annotation, Linked Open Data

1 Introduction

Entity annotation aims at discovering entities in references automatically. It is quite useful for many tasks including information extraction, classification, text summarization, question answering, and literature-based knowledge discovery. On the other hand, the Web as a global information space is developing from a Web of documents to a Web of data. Currently, there are billions of triples publicly available in Web data sources of different domains. These data sources are becoming more tightly interrelated as the number of links in the form of mappings grows. Based on the two points, what have we done in this paper can be summarized as follows.

1. We have proposed a novel knowledge-incentive approach based on LOD for entity annotation in the biomedical field. This approach has data flexibility, language independence, and semantic relationship enrichment, which makes it more convenient and informative for further applications.
2. We have proposed to make use of collective annotation leveraged by LOD information to conduct entity filtering and disambiguation.

3. We have developed MeDetect to implement our proposal. The experimental results verify the effectiveness and efficiency of our approach.

2 Methods

The overall design of MeDetect is shown in Figure 1.

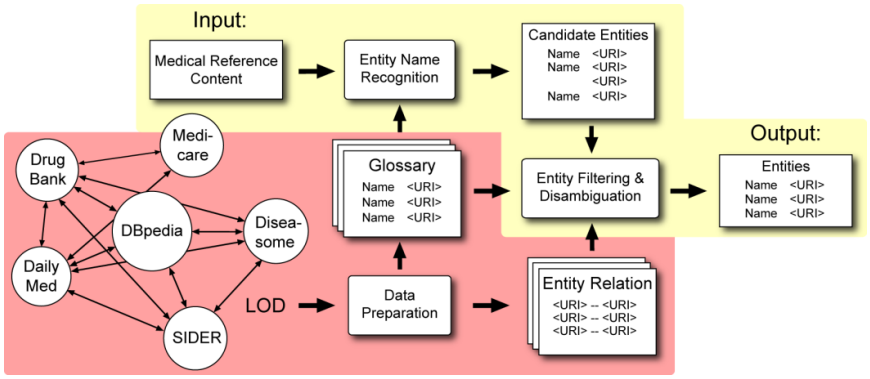

Fig.1.The architecture of MeDetect

Data Preparation. This step is conducted off-line. It generates two data structures for the on-line process of MeDetect. The first is an entity glossary, and the second is entity relation data. For Biomedical related ontology like DrugBank, we just use its entities and relations to build the entity glossary and relation data. For other general ontology like DBpedia which contains other non-biomedical entities, we propose a two-stage approach to handle it. In the first stage we use the link `owl:sameAs` from LODD to DBpedia to find the seed entities and expand the entity set from the original seed set using the sharing category information between entities. Secondly, we use a Support Vector Machine (SVM) [1] as the binary classification algorithm to pick biomedical entities from the candidate set.

Entity Name Recognition. With the biomedical entity glossary, the on-line entity name recognition step provides the syntactic match between the entity name in the glossary and the content of input biomedical references. This is based on syntactic matching, and the recognized entities with their URIs are passed to next step as candidate entities to be further filtered.

Entity Filtering and Disambiguation. The last but most important on-line step of MeDetect is entity filtering and disambiguation. We implement a system based on recent work on Web page annotation, Collective Annotation [2]. This approach not only detects the importance of each entity for the input text, but also, more importantly, filters out irrelevant entities based on the inter-entity relationship. There should be two functions in collective annotation: the single entity importance function and entity pair coherence function.

The single entity importance function estimates the relevance between an entity and the input text, based on their syntactic and semantic similarity, using logistic regression or category-based matching. Here, the entity description information in its URI can be utilized to match the input text.

The entity pair coherence function judges the topic similarity or consistency of pairs of entity URIs so as to filter out noise and cope with ambiguities. For example, if a candidate entity has no relationship or common topic with others, it is quite possible that this candidate is noise. Also if a candidate entity name has more than one URI, the entity pair coherence function will calculate the coherence of each of these URI with the ones of other entities and choose the most coherent one as the final URI of this entity name. Thus the problem of ambiguity is handled. In MeDetect, we use a LOD neighborhood overlap calculation [3] to implement the entity pair coherence function, as is shown in Figure 2.

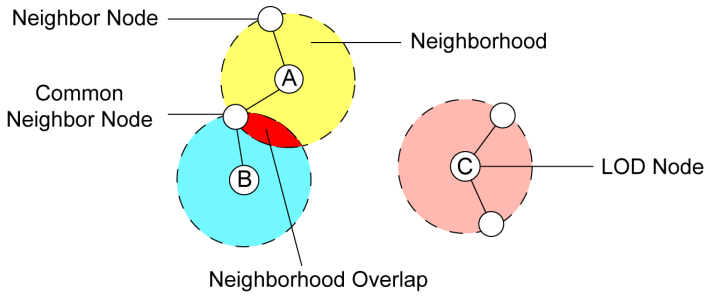

Fig.2. LOD neighborhood overlap calculation in collective annotation of MeDetect

Finally, we show a case of MeDetect entity annotation for a piece of biomedical reference in Figure 3. With the URI of each extracted entity, further information (such as the description of each entity and its links to related entities) can be directly provided in the annotation service.

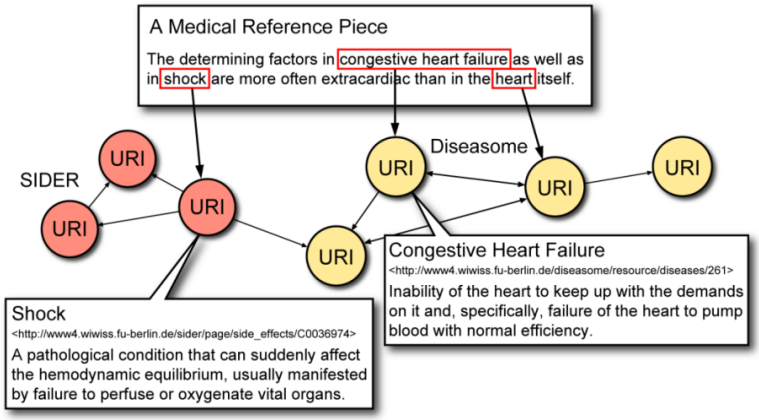

Fig.3. An Example of MeDetect Entity Annotation

3 Results

The effectiveness and efficiency of MeDetect is evaluated by an experimental study. In the experiment, 120 paper abstracts with different biomedical topics are randomly selected from PubMed and three human judgers with biomedical or computer background score the output entities for each paper abstract. To have a comparative evaluation of the quality of MeDetect, we import MetaMap and LingPipe. Here MeDetect\FD means MeDetect without filtering and disambiguation.

Compared systems	#Test references	#Output entities	#Corrected entities	Average accuracy	Average running time
MeDetect	120	598	455	**76.1%**	20.2ms
MeDetect\FD	120	683	468	68.5%	**11.9ms**
MetaMap	120	1,062	412	35.4%	601.4ms
LingPipe	120	782	510	65.2%	69.3ms

Table 1. Performance comparison among MeDetect, MeDetect\FD, MetaMap, and LingPipe

In Table 1 the average accuracy of MeDetect is much higher than MetaMap and LingPipe. Without entity filtering and disambiguation, MeDetect\FD provides a lower accuracy, despite its higher efficiency. In sum, MeDetect provides the most satisfactory performance.

4 Conclusion

This paper describes a novel knowledge-incentive approach based on LOD for entity annotation in the biomedical field. This approach has data flexibility, language independence, and semantic relationship enrichment, which makes it more adaptive and informative for further applications. We implement a prototype system MeDetect to demonstrate our approach for domain entity annotation for biomedical references. Its three key components are data preparation, entity name recognition, and entity filtering and disambiguation. Our system demonstrates its high annotation accuracy and data flexibility for adding more LOD sources. In future work, we will enrich the entity glossary of MeDetect by adding more LOD sources. More importantly, MeDetect will be further utilized for triple extraction from biomedical references.

References

1. Suykens J.A.K. and Vandewalle J. Least Squares Support Vector Machine Classifiers. Neural Processing Letters 1999.
2. Kulkarni S., Singh A., Ramakrishnan G. and Chakrabarti S. Collective Annotation of Wikipedia Entities in Web Text. SigKDD Proc. 2010.
3. Zhou W., Wang H., Chao J., Zhang W. and Yu Y. LODDO: Using Linked Open Data Description Overlap to Measure Semantic Relatedness Between Named Entities. JIST Proc. 2011.