Enhancing Entity Annotation using Web Service and Ontology Hierarchy in Biomedical Domains

Yue Shang, Yuan An, Xiaohua Tony Hu, Mi Zhang, Xia Lin
College of Computing and Informatics
Drexel University
Philadelphia, PA, USA
{ys439, ya45, xh29, mz349, linx}@drexel.edu

Abstract— Entity annotation is a fundamental step for entity relation extraction, information visualization and semantic web creation. Using domain specific ontologies to annotate a dataset for exploring, analyzing and integrating the entities and relations within the corpus is an important step for further study of the data. In this paper, we present an ontology based entity annotation system to annotate entities in neuroscience documents. To improve the annotation performance, we propose an ontology entity expansion method based on web service and ontology structure. We evaluate the proposed entity annotation method on real data obtained from Elsevier’s BrainNavigator. The results show that using web service and ontology structure to expand ontology entities can improve the annotation result.

Keywords—Entity Annotation; Ontology Expansion; Information Extraction

I. INTRODUCTION

The study of human brain has received a tremendous boost in recent years. With the exponential explosion of neuroscience journals, articles and data, it is necessary to develop tools to explore, compare, combine and integrate the extensive and growing neuroscience data. Information visualization is a good way to quickly acquire knowledge in a specific domain by showing the hot topics and their relationships. However, accurately annotating the entities within the corpus is a crucial step for achieving a good performance. In this paper, we propose an ontology based entity annotation method. The key idea to improve annotation performance is populating ontology entities from unstructured text using both web service and ontology hierarchical information.

In biomedical domain, ontologies are often used to support semantic queries, name entities recognition and data integration [1] [2] [3]. However, a large number of new entities and concepts continue to emerge due to the exponentially increment of biomedical literatures. Keeping ontologies up-to-date with extensive coverage is important for ontology-based applications. Manually building and expanding ontologies is a time-consuming task, which requires considerable human efforts. Thus, developing an automatic way to expanding the ontology instance is an important research topic. If candidate entities could be automatically recommended for ontology expansion, we can save manpower and time of ontology maintenance. Another consideration for ontology expansion is that the entities within the ontology used for annotation are professional and domain specific that they are not often used in writings. For example, “grey matter” is an important component for the central nervous system, including regions of the brain. But this entity does not involve in NIF (Neuroscience Information Framework) gross anatomy OWL [4] data. However, “grey matter” is often shown in the article discussing human memory and speech function of brain [5]. Without entity population, it is hard to annotate “grey matter” in documents. Thus, it is necessary to expand the ontologies in order to recall more entities to improve the performance of name entity recognition.

Moreover, because of the massive amount of documents and the rapid growth rate of knowledge, it is difficult to access, process and analyze each document directly. Web service like online repository and search engine provides an effective and efficient interface to acquire and utilize existing documents. Thus, web service is a feasible way to facilitate the ontology expansion process by bridge the semantic gap between ontology entities and candidates. In this paper, we present an entity annotation method on neuroscience data using web service and ontology hierarchical information to expand the ontology entity set in order to find more useful entities from unstructured documents.

The rest of the paper is organized as follows. Section 2 presents the related works of ontology expansion and entity based annotation. Section 3 describes the method we use for ontology expansion and entity annotation. Section 4 shows the evaluation and result analysis for annotation and ontology expansion respectively. The last section presents the conclusion of this work.

II. RELATED WORKS

With the evolution of semantic web, many studies have been carried out to solve the annotation problem of Web raw text in an ontology-based manner. Most of them use word-net and Wikipedia as key mean for entities annotation task [6]. Many tools are developed to extract entities and relations from web pages, like KnowItAll [7], TextRunner [8] [9], SEAL and Text2Onto. KnowItAll and TextRunner adopt an open information extraction method, which use redundant information from web documents to perform a bootstrapping information extraction process.

Using named entities to populate ontology has attracted a lot of attentions recently, and there are two main kinds of approaches: pattern-based and distributional [10]. The pattern
based method leverages lexical or syntactic patterns to extract patterns like “is -a”. Weakly supervision is often applied in form of a small set of human made seed patterns or seed instances [11] [12]. The distributional method uses context as evidence to find features for ontology population. A small seed set of entities are also required for new entities exploration. Our method aims to deal with entity annotation problem within biomedical domain. To obtain a better annotation performance, ontology is expanded using web services and ontology structure information.

III. METHODS

To implement the ontology-based entity annotation, two main components consist in our method: ontology entity expansion and entity annotation. Professional and specialized entity terms in ontologies may limit the annotation performance. It is necessary to expand the ontology entities with the phrases with similar semantic meaning that are frequently used in writings as well. Thus, we proposed an ontology entity expansion method here using web services and ontology structure. And then, an entity annotation process is implemented with the expanded entity set. The whole process of our method is shown in Fig. 1.

![General Flowchart for Neuroscience Entities Annotation](image)

**A. Ontology Entity Expansion**

The goal of ontology entity expansion is to recognize more named entities that are semantically related to the existing ontology entities in order to improve the performance of entity annotation for unstructured corpus. To fulfill this goal, new entities need to be detected from unstructured text. Ontology entity expansion can be regarded as two sub-tasks: candidate entities detection and candidate ranking.

1. Candidate Entity Detection

To obtain candidates, we use the POS tagger in OpenNLP toolkit\(^1\), whose English POS model uses the Penn Treebank tag set. After the text has been tagged, expressions with patterns shown in Table 1 are used to detect the noun phrases, which may contain the full name of an ontology entity. Three patterns for candidate entity detection are used in this paper as shown in Table 1. The first pattern is about sequential noun phrase (“Brain Regions”). The second pattern is composed by a central component of one or several Proper Nouns, with no or several Nouns followed or leaded (“insula cortex”). The third pattern is composed by noun phrase or a Proper noun phrase leaded by one or more adjective (“White Matter”, “Premotor Cortex”).

<table>
<thead>
<tr>
<th>TABLE I. CANDIDATE ENTITY DETECTION</th>
</tr>
</thead>
<tbody>
<tr>
<td>Name</td>
</tr>
<tr>
<td>------------------------</td>
</tr>
<tr>
<td>Noun Phrase</td>
</tr>
<tr>
<td>Proper Noun</td>
</tr>
<tr>
<td>Phrase</td>
</tr>
<tr>
<td>Adjective Noun</td>
</tr>
</tbody>
</table>

2. Web service based Similarity

In this paper, we focus on dealing with neuroscience entity annotation problem so most of the entities are domain specific. Thus, we use PubMed as the online repository to get the co-occurrence hit count for calculating semantic similarity. Our goal is to extract biomedical entities from unstructured text. PubMed is a web repository for biomedical literature. Firstly, the candidate terms obtained from the previous section will be scored according to semantic similarity measures with the existing ontology terms. Four measurement approaches are used here: PMI [13], Dice [14], Jaccard, and NGD [15]. Since we have received a great number of candidate terms extracted by the patterns defined in the previous section, it would be time consuming if we run the similarity comparison with ontology entities with all of them. Here we make an assumption that if a phrase is not a real world entity, web service will not return or only return few hits exactly match the phrase. Then, we used the candidates remained after filtering as reasonable entities and calculate the similarities using the following formulas.

\[
PMI(M_1, M_2) = \log_2 \left( \frac{P(M_1 \cap M_2)}{P(M_1) \cdot P(M_2)} \right) = \log_2 \frac{C(M_1 \cap M_2)}{N \cdot C(M_1) \cdot C(M_2)}
\]

\[
\text{Dice}(M_1, M_2) = \frac{2 \cdot C(M_1 \cap M_2)}{C(M_1) + C(M_2)}
\]

\[
\text{Jaccard}(M_1, M_2) = \frac{C(M_1 \cap M_2)}{C(M_1) + C(M_2) - C(M_1 \cap M_2)}
\]

\[
\text{NGD}(M_1, M_2) = \frac{\max(\log C(M_1), \log C(M_2)) - \log C(M_1 \cap M_2)}{\log N - \min(\log C(M_1), \log C(M_2))}
\]

\(M_1, M_2\) represent the query entities. \(C(M)\) represents the number of hit count from search engine. \(C(M_1 \cap M_2)\) represents the hit count for the query “\(M_1 AND M_2\)”. The count of hits for “\(M_1 AND M_2\)” returned by PubMed can be regarded as an estimation of the semantic relations of two terms [15].

3. Entity Ranking

For each term, we use the average co-occurrence score with all the ontology terms as its score and rank the terms for all these four kinds of scores respectively. We define \(\text{Score}(M)\) as the measurement of relatedness for candidate term \(M\) with the ontology. \(N\) is the number of total entities within the ontology and \(O_i\) is an ontology entity. \(S(M, O_i)\) represents one of the co-occurrence similarity measures in the previous section. \(\text{height}(O_i)\) represents the path length from the root of ontology to \(O_i\). The ontology is created in a tree

---

\(^1\) OpenNLP: http://opennlp.apache.org/
The nodes near the root of ontology tree are more general entities while the more specific nodes are at the bottom. As a result, entities on the top of ontology are more likely to return more hits from the search engine. For example, "retrosplenial cortex" is a term we found has high similarities with NIF brain structure ontology terms and it has a high score using the following formula.

\[ \text{Score}(M) = \frac{\sum_{i=1}^{N} S(M, O_i) \cdot \log(1 + \text{height}(O_i))}{N} \]

Then we rank the candidate terms according to \( \text{Score}(M) \) using PMI, Dice, Jaccard, and NGD respectively and four lists of terms are obtained. Finally, candidate terms are ranked by the average rank in four list and the top N candidate terms are chosen as new entities to expend the ontology we have to annotate the text.

### B. Entity Annotation

We design the ontology based entities annotation system, called "SemIntegrator", which is an open toolkit that allow users to explore, compare, and annotate documents using ontologies. Figure 2 show the interface for our system. This system is a Protégé (http://protege.stanford.edu) plugin that can create, modify, combine and compare ontologies and use the user input ontologies to annotate documents.

![User Interface of SemIntegrator for the Entity Annotation System](image)

**Fig. 2.** User Interface of SemIntegrator for the Entity Annotation System

### IV. EXPERIMENTAL RESULTS

We conduct two sets of experiments to evaluate the method proposed in this paper. In the first part of this section, we evaluate the overall performance of entity annotation by comparing with an online application from Elsevier for brain structure annotation about Neuroscience publication. Next, to evaluate the performance of ontology expansion, we also conduct an experiment on brain structure related ontology.

#### A. Experiment for Entity Annotation

As far as we known, there’s no dataset for brain related concept annotation. To evaluate entity annotation on real world data, we apply our method on Elsevier for brain structure annotation about Neuroscience publication. The goal of our experiment is to annotate the brain structure entities from the Neuroscience publications with brain structure ontology input by user. The gold standard we use is BrainNavigator, which is an application built and maintained by Elsevier to recognize brain structure in publications. We use the results from BrainNavigator as golden standard for our experiments. Precision, recall and F-score are used to evaluate the performance of our annotation performance.

To annotate the Elsevier articles with our method, firstly we need to select ontology with knowledge of brain structure. In our experiment, we use the brain structure part of NIF ontology, which are the sub-classes of “Regional part of Brain” in NIF gross anatomy OWL. The unstructured text we used to learn candidate entities are journals of Elsevier NeuroImaging journals. There are 15096 documents from the journal of Neuroimage in the dataset we got from Elsevier and we use these documents to learn new entities.

Six benchmark methods are used to compare the result with our method. The first benchmark only uses the ontology entities to annotate documents. PMI, Jaccard, Dice and NGD are methods we use to calculate the similarity of two entities. PMI, for example, indicates using PMI method to calculate the semantic similarity of ontology entities with candidate entities. And without combining the ranking results of four methods, we only use PMI and get the top N entities as the new entities. PJDN (PMI, Jaccard, Dice and NGD) indicates using the four similarity calculation methods, but without adding ontology hierarchical information.

<table>
<thead>
<tr>
<th>Method</th>
<th>Precision</th>
<th>Recall</th>
<th>F-score</th>
</tr>
</thead>
<tbody>
<tr>
<td>NIF only</td>
<td>0.7243</td>
<td>0.5750</td>
<td>0.6411</td>
</tr>
<tr>
<td>PMI</td>
<td>0.6382</td>
<td>0.6190</td>
<td>0.6285</td>
</tr>
<tr>
<td>Jaccard</td>
<td>0.7192</td>
<td>0.6683</td>
<td>0.6928</td>
</tr>
<tr>
<td>Dice</td>
<td>0.6842</td>
<td>0.6438</td>
<td>0.6634</td>
</tr>
<tr>
<td>NGD</td>
<td>0.7176</td>
<td>0.6679</td>
<td>0.6918</td>
</tr>
<tr>
<td>PJDN</td>
<td>0.7113</td>
<td>0.6524</td>
<td>0.6806</td>
</tr>
<tr>
<td>Proposed method</td>
<td>0.7164</td>
<td>0.6855</td>
<td>0.7006</td>
</tr>
</tbody>
</table>

The proposed method shows the best performance. Comparing with PJDN, the proposed method improves the recall by 2.6%, which indicates that using ontology hierarchy information can improve the performance of entity annotation. For the four kinds of similarity measures, NGD have a highest performance and Jaccard also have a relatively high experimental result.

#### B. Experiment for Ontology Entity Expansion

We also test the performance of ontology expansion using the previous brain structure ontology from NIF.

<table>
<thead>
<tr>
<th>Removed amount</th>
<th>#Entities</th>
<th>#NGD</th>
<th>#PJDN</th>
<th>#Proposed Method</th>
</tr>
</thead>
<tbody>
<tr>
<td>10%</td>
<td>79</td>
<td>43</td>
<td>41</td>
<td>48</td>
</tr>
<tr>
<td>20%</td>
<td>158</td>
<td>74</td>
<td>58</td>
<td>73</td>
</tr>
<tr>
<td>30%</td>
<td>237</td>
<td>88</td>
<td>78</td>
<td>96</td>
</tr>
<tr>
<td>40%</td>
<td>316</td>
<td>126</td>
<td>125</td>
<td>139</td>
</tr>
<tr>
<td>50%</td>
<td>395</td>
<td>171</td>
<td>158</td>
<td>182</td>
</tr>
</tbody>
</table>

Firstly, we randomly remove certain percentage of terms in ontology, in our experiment 10% to 50% respectively. Then, we run our method on the remaining ontology entities to see if our method can expand the ontology by learning from unstructured documents. Then we would like to evaluate the correctness of the new entities our method found. We return
the same number of terms as removed from ontology. For time consuming problem, we filter the NIF ontology by PubMed hit count. If a term receives the hit count less than 10, it is removed from the ontology entity set. And after filtered by PubMed, the size of ontology is reduced to 790.

The experimental results for our method as well as NGD and PJDN over the NIF brain structure ontology are shown in Table 3. We show the number of correctly learned entities for NGD, PJDN and our method, according to the original ontology entity. From the result shown in Table 3 and we can see that our method can find more correct entities than the other two methods. Both NGD and our method perform better than PJDN, indicating that the ontology structure information can improve the performance. One of the reasons is that if a candidate entity has a high similarity with the entities at the bottom of ontology it would be more related to the ontology. For example, “vivo studies” is a candidate entity we get and it has high similarity with many NIF ontology entities. “Grey Matter” is a central nervous system entity involved in many brain functions. Both two entities do not exist in NIF ontology. We calculate the average height of the related 46 entities that have a similarity with “vivo studies” is 1.36, while for the 125 entities related to “grey matter” the average height is 1.62. This, to some extent, indicates that “grey matter” has high similarity with more specific entity than “vivo studies” does.

C. Application for Entity Annotation

We apply the ontology based entity annotation method on Elsevier dataset and use the annotated entities to visualize the relationship between entities.

V. CONCLUSION

In this paper, an ontology based entity annotation method is presented using web service as well as ontology structure information. Our goal is to design an ontology based entity annotation tool for users to tag the entities within unstructured text using their own ontologies. The challenge of the work is that ontology entities are domain specific, professional and sometimes hard to cover the entities used by researchers in writings. As a result, synonyms, phrases that have similar semantic content with ontology entities are unable to be detected. To solve this problem, we make use of the web service as an external context to calculate the semantic similarity between entities. And ontology structure information aids the entities ranking process by adding weights for candidates that related to specific ontology entities. Furthermore, an annotation tool “SemIntegrator” is implemented for entity annotation using ontologies.

ACKNOWLEDGMENT

This work is supported by the NSF I/UCRC Center for Visual and Decision Informatics (CVDI, NSF IIP 1160960)

VI. REFERENCES


