Automatic Extraction of Location Relations from Text

Wu Zheng
University of Illinois at Urbana Champaign
wuzheng2@illinois.edu

Catherine Blake
University of Illinois at Urbana Champaign
clblake@illinois.edu

ABSTRACT

Automatically identifying semantic relationships from text plays an important role in knowledge discovery, for example to connect a researcher in one discipline to related research questions in a second discipline in which the researcher is not formally trained. This poster describes preliminary experiments in ongoing research project that explores the utility of semantics and syntax to identify relations from text automatically. We focus exclusively on the *location* relation, such as organization-location and gene-location. Location is an interesting case because it occurs in multiple text genres including news articles and scientific literature.

Keywords

Information extraction, relation mining, text mining, knowledge discovery

INTRODUCTION

The extraction of semantic relations plays an important role in knowledge discovery from text. Identifying relationships automatically can contribute to a variety of natural language processing applications, including information extraction, question answering, knowledge discovery, and information synthesis (Lapata, 2002; Morris & Hirst, 2004; Rosario & Hearst, 2005). The research presented in this poster is part of a broader project entitled Evidence-Based Discovery (EBD) where the goal is to enable a scientist in one domain to more easily identify findings from another domain in which the scientist has little or no formal training. Our goal in the broader project is to develop language technologies that automatically identify a range of semantic relations from text.

In this poster, we present preliminary results of experiments that explore the utility of semantics and syntactic constraints to identify a location relationship. The location relation is an interesting case because it spans multiple genre's for example identifying the geographical location of an organization's head office, the location of a gene within the body, or the location of a city within a country. Smith et al (Smith, et al., 2005) characterized location as a primitive instance-level relation in their research on relations in biomedical ontologies.

We define the location relation as a binary predicate where the arguments define the item of interest (in the previous examples the organization, the gene or the city) and the location of that item, specifically LOCATION(X, Y) indicates that X is located in Y. In sentence (1) below the system should instantiate the location relation as LOCATION(Slr0228, thylakoid membrane).

> Given the predicted orientation of these helices and assuming that Slr0228 is located in the thylakoid membrane, the conserved 81-amino acid feature would constitute a lumenal domain. (1)

METHOD

One of the fundamental tenants of computational linguistics is that there exists a relationship between the underlying form (the syntax) and the meaning conveyed (the semantics) in a sentence. Our goal is to explore the degree to which syntactic and semantic features of a sentence can enable us to identify new terms that are indicative of a location relation. We use a similar strategy to the development of the FrameNet database (see (Gildea & Jurafsky, 2001), pg 5).

(1) Identify seed terms for the binary predicate

(2) Sample sentences with the seed terms and identify arguments

(3) Characterize syntactic constructs

(4) Check the annotated sentences for consistency

(5) Use arguments to identify new seed terms for the binary predicate and repeat steps 1-4.

The first semantic constraint was the seed word "located" followed by the prepositions "in", "at", and "on". We drew a stratified random sample of 100 sentences that contained each of the three seed phrases from the 1.8 million sentences in the 2002 Journal of Biological Chemistry (JBC) articles from TREC (Hersh & Voorhees, 2009). Sentences in each category - located in, located at, and located on - were sampled at a rate proportional to their overall distribution in the journal (see Table 1).

	Sample	Journal
located in	58	3,528
located at	26	1,665
located on	16	834
Total	100	6,027

Table 1: Number of sample sentences in each category

With the semantic constraints in place, the next step is to identify syntactic patterns that characterize the location relations. The syntactic patterns explored in this paper are the typed dependencies produced by the Stanford Parser (Klein & Manning, 2003). Of the 1.8 million sentences in JBC, a parse was produced for 1.66 million (92%).

For step 3, we manually inspected each sample sentence to identify syntactic patterns that are indicative of location relations.

Figure 1 shows the dependency tree for sentence 1, where location (X,Y) should return X=Slr0228 and Y=thylakoid membrane. In this case, the passive nominal subject (nsubjpass) identified by the Stanford parser corresponds directly to first argument of the location relation and the object of the preposition (pobj) corresponds directly to the second argument of the location relation. Even though the subject and object returned by the Stanford parser do not necessarily reflect the subject and object of location verb we refer to X as the subject and Y as the object of the location relation to ease further discussion.

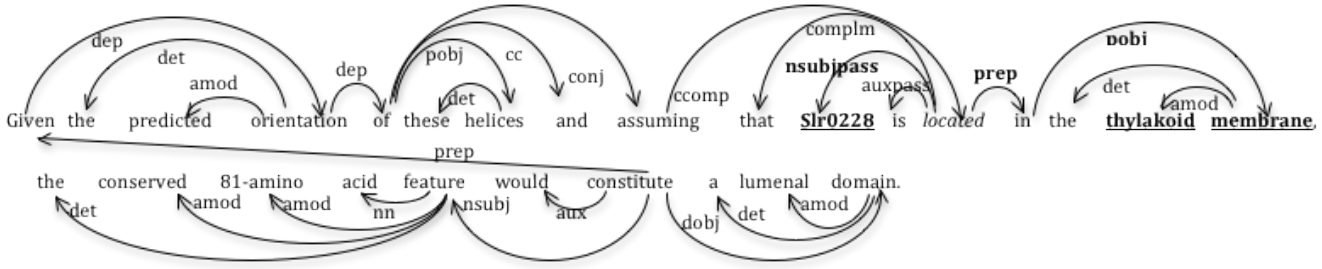
dep
det
amod
dep
pobj
cc
det
conj
ccomp
complm
nsubjpass
auxpass
prep
pobj
det
amod
Given the predicted orientation of these helices and assuming that Slr0228 is located in the thylakoid membrane,
prep
the conserved 81-amino acid feature would constitute a lumenal domain.
det
amod
amod
nn
nsubj
aux
dobj
det
amod

Figure 1: Dependency tree for sentence (1)

RESULTS AND DISCUSSION

Our analysis of the 100 sample sentences revealed two syntactic patterns that are highly indicative of location relations.

(1) Terms depicted as subjects of the predicate seed term. For example, in the following sentence, the phrase *the catalytic serine (Ser_153)* corresponds to the first argument in location.

> The catalytic serine (Ser_153) is in a characteristic epsilon conformation and is located in a tight turn with the G-H-S-Q-G sequence belonging to the usual consensus sequence of the alpha/beta hydrolase fold family. (2)

(2) The seed term is a participial modifier of the first location argument. For example in the sentence (3), *located* is the participial modifier of term *AU-rich elements (ARE).*

> Decay of these cytokine mRNAs is normally regulated in part by the presence of AU-rich elements (ARE) located in their 3'-untranslated regions. (3)

We also observed variations of these two rules. For example, in some sentences the seed term was connected to the location arguments via a conjunction. In these cases, the Stanford parser associated the subjects of the sentence with just the first verb and the subject of our seed term was not represented directly in the dependency tree. In these sentences, the first location argument can be identified by tracing back though the dependency tree to identify the subject. For example in sentence (4) *Slt2-GFP* is the first located argument. In the output of the parser, *Slt2-GFP* is the subject of the verb *concentrated*, but the dependency tree connects the second argument (cytoplasm) via the conjunction *and*.

> In accordance with the previous report localizing Slt2-HA, Slt2-GFP was concentrated in the nucleus and also located in the cytoplasm at 25 C. (4)

Another variation is that the seed term appeared in the open clausal complements of other verbs in a sentence. (An open clausal complement is a clausal complement without its own subject, whose reference is determined by an external subject.) In this case, the first argument of the location relation is the subject of the verb that the open clausal complement modifies. For example in sentence (5), the dependency parser comprises list the phrase *some of the ER retention signals* as the subject of the verb *shown* and the seed term *located* is the open clausal complement of *shown*.

> Since some of the ER retention signals have been shown to be located in the C-terminal end of polypeptides, we generated constructs with deletion of 15, 40, and 97 amino acids from the C-terminal tail of RyR. (5)

In contrast to the syntactic patterns that would accurately identify the first argument of the location relation, we found only one syntactic pattern that worked surprisingly well for the second argument. Specifically, the head noun identified by the Stanford parser for the seed term was typically the second argument in the located relation. For example (1), the phrase *the thylakoid membrane,* which is the object of *located in*.

We identified rules that would identify both the predicate and arguments for the location relation. Errors from seven of the 100 sentences were caused by incorrect parsing. The performance on the development collection of sentences is shown in Table 2. The rules were evaluated using two different criteria. Under the strict criterion, we considered an extracted term correct if and only if the term matched the whole phrase of each of the location relation arguments. Under the loose criterion, we considered an extracted term correct if it captured the head noun of location relation arguments. In the later case missing modifiers were not considered errors.

The evaluation of the development collection suggests that syntactic features are highly effective in capturing arguments of the location relation. We are currently working on steps 4 and 5, which will establish the consistency of these relations and use the arguments identified from the seed term located, to generate additional verbs that are indicative of location relation.

Our approach does depend on the performance of the parser. Sentences from scientific literature, which are quite different from the sentences in newspaper articles on which the Stanford parser was trained, in particular they tend to be longer and more complex. Training the parser on scientific literature may further improve the system results.

			Correct Counts		Precision		Recall	
	Number of Prediction	Actual Number	Strict	Loose	Strict	Loose	Strict	Loose
First argument	101	103	66	89	65.3%	88.1%	64.1%	86.4%
Second argument	122	103	71	100	58.2%	82.0%	68.9%	97.1%
Both arguments	223	206	137	189	61.4%	84.8%	66.5%	91.7%

Table 2: Performance of rules on development collection

REFERENCES

Gildea, D., & Jurafsky, D. (2001). Automatic labeling of semantic roles. *Computational Linguistics, 99*(9), 1-43.

Hersh, W. R., & Voorhees, E. M. (2009). TREC genomics special issue overview. *Information Retrieval, 12*(1), 1-15.

Klein, D., & Manning, C. D. (2003). *Fast Exact Inference with a Factored Model for Natural Language Parsing.* . Paper presented at the Advances in Neural Information Processing Systems.

Lapata, M. (2002). The disambiguation of nominalisations. *Computational Linguistics, 28*(3), 357–388.

Morris, J., & Hirst, G. (2004). *Non-classical lexical semantic relations*. Paper presented at the Proceedings of the 4th Human Language Technology Conference and 5th Meeting of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL 2004) - Workshop on Computational Lexical Semantics.

Rosario, B., & Hearst, M. (2005). *Multi-way relation classification: application to protein-protein interaction.* Paper presented at the Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing Vancouver, British Columbia, Canada

Smith, B., Ceusters, W., Klagges, B., Köhler, J., Kumar, A., Lomax, J., et al. (2005). Relations in biomedical ontologies. *Genome Biology, 6*(5), R46.