AAAI Technical Report FS-12-05
Information Retrieval and Knowledge Discovery in Biomedical Text

Semantic Role Labeling for Biological Transport

He Tan
School of Engineering
Jönköping University, Sweden
he.tan@jth.hj.se

Srikanth Chowdari
Department of Biomedical Engineering
Linköping University, Sweden
srich816@student.liu.se

Abstract

Semantic role labeling (SRL) is a technique of semantic interpretation of text on the sentence level. In this paper, we present a corpus that is labeled with semantic roles for biological transport events. The corpus was built using domain knowledge provided by ontologies. We also report on a word-chunking approach for identifying semantic roles of biomedical predicates describing transport events. We trained a first-order Conditional Random Fields (CRF) for chunking applications with the traditional role labeling features and also domain-specific features. The results show that the system performance varies between different roles and the performance was not improved for all roles by introducing domain specific features.

Introduction

Semantic Role Labeling (SRL) is a process that, for each predicate in a sentence, indicates what semantic relations hold among the predicate and its associated sentence constituents. It plays a key role in many text mining applications such as Information Extraction and Question Answering. Recently, large corpora have been manually annotated with semantic roles in FrameNet (Fillmore, Wooters, and Baker 2001) and PropBank (Palmer, Gildea, and Kingsbury 2005) for general English. With the advent of resources, SRL has become a well-defined task with a substantial body of work and comparative evaluation.

Biomedical text considerably differs from the PropBank and FrameNet data, both in the style of the written text and the predicates involved. In this paper, we present a corpus that is labeled with semantic roles for biological transport events. The corpus was built based on the theory of frame semantics and using domain knowledge provided by ontologies. Then we report on a word-chunking approach for identifying semantic roles of biomedical predicates describing biological transport events.

Methods and Results

Corpus Construction

Frames Semantics (Fillmore 1985) begins with the assumption that in order to understand the meanings of the words in a language, we must first have knowledge of the background and motivation for their existence in the language and for their use in discourse. The knowledge is provided by the conceptual structures, or *semantic frames*. Ontology is a formal representation of knowledge of a domain of interest. They reflect the structure of the domain knowledge and constrain the potential interpretations of terms. Intuitively, ontological concepts, relations, rules and their associated textual definitions can be used as the frame-semantic descriptions imposed on a corpus.

For the details of ontology-driven construction of corpus with frame semantics annotations we refer to (Tan, Kaliyaperumal, and Benis 2012). Here, we outline the aspects of ontology driven frame-semantic descriptions: *1)* The structure and semantics of domain knowledge in ontologies constrain the frame semantics analysis, i.e. decide the coverage of semantic frames and the relations between them; *2)* Ontological terms can comprehensively describe the characteristics of events or scenarios in the domain, so domain-specific semantic roles can be determined based on terms; *3)* Ontological terms provide domain-specific predicates, so the semantic senses of the predicates in the domain are determined; *4)* The collection and selection of example sentences can be based on knowledge-based search engine for biomedical text.

Using the method, we have built a corpus for transport events strictly following the piece of domain knowledge provided by GO biological process ontology (The Gene Ontology Consortium 2000). In *Transport* frame and its sub-frame *Protein Transport*, we defined 10 possible frame elements (FEs), and identified the domain specific semantic types that indicate the typing of fillers of FEs, using the semantic types from the Unified Medical Language System (UMLS) Semantic Network. We collected 129 lexical units (LUs), and annotated minimally 10 sentences for each LU. Totally, we collected a set of 955 sentences from PubMed abstracts.

Machine Learning Model

Two main types of approaches to semantic role labeling are syntactic constituent approaches and word chunking approaches. The annotation effort that is required for a full syntactic parser is larger than that is required for taggers and chunkers. To void the expensive syntactic parsing process, we considered word chunking approaches (Hacioglu et

Copyright © 2012, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

Total	**TE**	**TO**	**TDS**	**TC**	TL	TP	TT	TDR	TA	TPL
955	786	200	320	501	112	92	46	4	69	31

Table 1: The number of sentences contain the FEs. Transport_Entity (TE), Transport_Origin (TO), Transport_Destination (TDS), Transport_Condition (TC) are core FEs. Others are Transport_Location (TL), Transport_Path (TP), Transport_Transporter (TT),Transport_Direction (TDR), Transport_Attribute (TA), Transport_Place (TPL). For the FE definitions we refer to (Tan, Kaliyaperumal, and Benis 2012)

	traditional features					traditional + domain specific features				
FE	All	TE	TD	TO	TC	All	TE	TD	TT	TC
R	0.56	0.70	0.75	0.61	0.18	0.56	0.74	0.67	1.0	0.21
P	0.59	0.66	0.93	0.61	0.20	0.57	0.73	0.75	0.60	0.19
A	0.85	0.97	0.77	0.73	0.89	0.84	0.92	0.75	0.50	0.90

Table 2: Unlabeled precision (P) and recall (R), and labeled accuracy (A).

al. 2004). The word chunking approaches convert the SRL problem into a word classification problem by selecting appropriate labels for each word in the phrase. These labels are a combination of a B(eginning), I(nside) or O(utside) prefix that indicates the location of the word within the role, and a role suffix that indicates the type of the role.

CRF for sequence labeling offer advantages over both generative models and classifiers applied at each sequence position (Sha and Pereira 2003). In this work we used the implementation of first-order chain CRF to chunking applications from LingPipe (http://alias-i.com/lingpipe). For the purposes of our experiments, 101 sentences were selected at random from our corpus and reserved as the test set. The remaining 854 sentences were used as the training data.

Two groups of features are used in training and testing. The first group include the traditional role labeling features: the text of words and predicates, the stem of predicates, the voice of predicates, the part-of-speech of words and predicates, the BIO tag for the phrases that includes the word (e.g. B-NP, I-NP and O) and the location of words relative to the predicate (i.e. ”before”, ”after” and ”-”). The output from the MetaMap tools (2011) (Aronson and Lang 2010) processing executed on corpus sentences are extracted into another group of features. The features provided from the MetaMap output include the domain specific noun phrases identified in the sentences, the head of the noun phrase, and the UMLS Semantic Types of the noun phrase.

The SRL task can be viewed as a two step process in which boundaries are first identified and then FEs are labeled. Therefore, we evaluated the system in labeled precision (P) and recall (R) and labeled accuracy (A). Tabel 2 presents P, R and A values when the model were trained using the traditional role labeling features, and the traditional features together with domain specific features. The results show that the system has performed better in identifying the type of the FEs, than for word chunking. The performance varies greatly between different FEs. Table 1 gives all the possible FEs we identified for transport events, and the number of sentences containing the different FEs. When the system was trained with the traditional role labeling features, it performed best on the TE, TD and TO. When the system was trained with the traditional features together with domain specific features, it performed best on the TE, TO and TT. In both cases the system performed worst on the TC. It also shows that the performance is not always improved for all FEs by introducing domain specific features. As a whole the system had similar performance when it was trained with and without domain specific features.

Conclusions

In this paper, we present a corpus that is labeled with semantic roles for biological transport events. It was built using ontological domain knowledge. We also report on a word-chunking approach for identifying semantic roles of biomedical predicates describing transport events. The results show that the system performance varies between different roles and the performance was not improved for all roles by introducing domain specific features. In the future work, we aim to extend the corpus to cover other biological events. We intend to develop a environment supporting corpus construction using ontological domain knowledge. We also intend to explore features based on full syntactic parses in machine learning models.

References

Aronson, A. R., and Lang, F.-M. 2010. An overview of metamap: historical perspective and recent advances. *JAMIA* 17(3):229–236.

Fillmore, C. J.; Wooters, C.; and Baker, C. F. 2001. Building a large lexical databank which provides deep semantics. In *Proceedings of the PACLIC*.

Fillmore, C. J. 1985. Frames and the semantics of understanding. *Quaderni di Semantica* 6(2):222–254.

Hacioglu, K.; Pradhan, S.; Ward, W.; Martin, J. H.; and Jurafsky, D. 2004. Semantic role labeling by tagging syntactic chunks. In *Proceedings of CoNLL 2004 Shared Task*, 110–113.

Palmer, M.; Gildea, D.; and Kingsbury, P. 2005. The proposition bank: an annotated corpus of semantic roles. *Computational Linguistics* 31:71–105.

Sha, F., and Pereira, F. 2003. Shallow parsing with conditional random fields. In *Proceedings of NAACL HLT 2003*, 134–141.

Tan, H.; Kaliyaperumal, R.; and Benis, N. 2012. Ontology-driven construction of corpus with frame semantics annotations. In *CICLing 2012, Part I, LNCS 7181*, 54–65.

The Gene Ontology Consortium. 2000. Gene ontology: tool for the unification of biology. *Nature Genetics* 25:25–29.