Exploiting Rich Syntactic Information for Relation Extraction from Biomedical Articles*

Yudong Liu and **Zhongmin Shi** and **Anoop Sarkar**
School of Computing Science
Simon Fraser University
{yudongl,zshi1,anoop}@cs.sfu.ca

Abstract

This paper proposes a ternary relation extraction method primarily based on rich syntactic information. We identify PROTEIN-ORGANISM-LOCATION relations in the text of biomedical articles. Different kernel functions are used with an SVM learner to integrate two sources of information from syntactic parse trees: (i) a large number of syntactic features that have been shown useful for Semantic Role Labeling (SRL) and applied here to the relation extraction task, and (ii) features from the entire parse tree using a tree kernel. Our experiments show that the use of rich syntactic features significantly outperforms shallow word-based features. The best accuracy is obtained by combining SRL features with tree kernels.

1 Introduction

Biomedical functional relations (relations for short) state interactions among biomedical substances. For instance, the PROTEIN-ORGANISM-LOCATION (POL) relation that we study in this paper provides information about where a PROTEIN is located in an ORGANISM, giving a valuable clue to the biological function of the PROTEIN and helping to identify suitable drug, vaccine and diagnostic targets. Fig. 1 illustrates possible locations of proteins in Gram+ and Gram− bacteria. Previous work in biomedical relation extraction task (Sekimizu et al., 1998; Blaschke et al., 1999; Feldman et al., 2002) suggested the use of predicate-argument structure by taking verbs as the center of the relation – in contrast, in this paper we directly link protein named entities (NEs) to their locations; in other related work, (Claudio et al., 2006) proposed an approach that solely considers the shallow semantic features extracted from sentences.

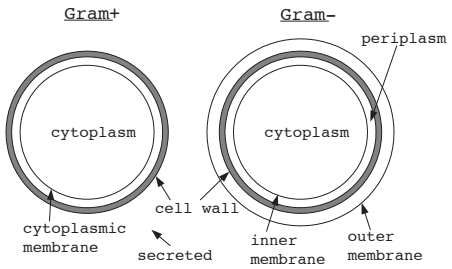

Figure 1: Illustration of bacterial locations

For relation extraction in the newswire domain, syntactic features have been used in a generative model (Miller et al., 2000) and in a discriminative log-linear model (Kambhatla, 2004). In comparison, we use a much larger set of syntactic features extracted from parse trees, many of which have been shown useful in SRL task. Kernel-based methods have also been used for relation extraction (Zelenko et al., 2003; Culotta and Sorensen, 2004; Bunescu and Mooney, 2005) on various syntactic representations, such as dependency trees or constituency-based parse trees. In contrast, we explore a much wider variety of syntactic features in this work. To benefit from both views, a composite kernel (Zhang et al., 2006) integrates the flat features from entities and structured features from parse trees. In our work, we also combine a linear kernel with a tree kernel for improved performance.

2 SRL Features for Information Extraction

Fig. 2 shows one example illustrating the ternary relation we are identifying. In this example, "Exoenzyme S" is a PROTEIN name, "extracellular" a LOCATION name and "Pseudomonas aeruginosa" an ORGANISM name. Our task is to identify if there exists a "PROTEIN-ORGANISM-LOCATION" relation among these three NEs.

To simplify the problem, we first reduce the POL

*This research was partially supported by NSERC, Canada.

Proceedings of NAACL HLT 2007, Companion Volume, pages 97–100,
Rochester, NY, April 2007.

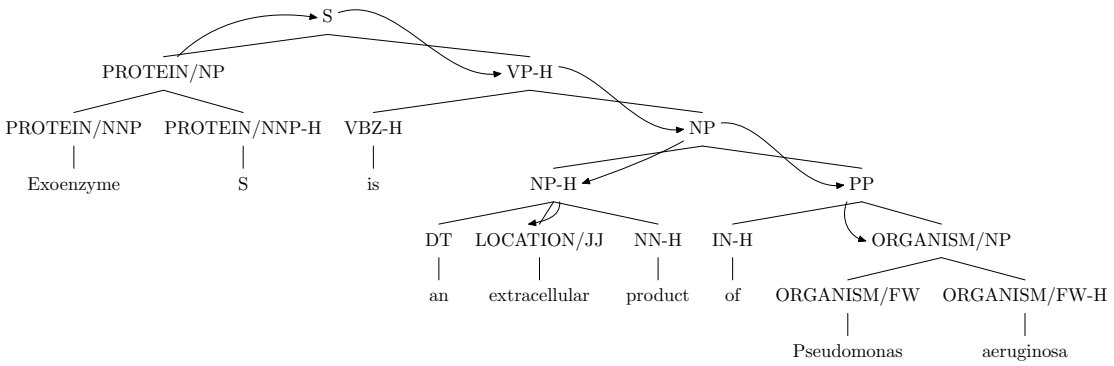

Figure 2: An example of POL ternary relation in a parse tree

ternary relation extraction problem into two binary relation extraction problems. Specifically, we split the POL ternary relation into binary relations as: (1) PO: PROTEIN and ORGANISM, and (2) PL: PROTEIN and LOCATION.

The ORGANISM-LOCATION relation is ignored because it does not consider the PROTEIN and is less meaningful than the PO and PL relations. Based on this simplification, and following the idea of SRL, we take the PROTEIN name in the role of the predicate (verb) and the ORGANISM/LOCATION name as its argument candidates in question. Then the problem of identifying the binary relations of PO and PL has been reduced to the problem of argument classification problem given the predicate and the argument candidates. The reason we pick PROTEIN names as predicates is that we assume PROTEIN names play a more central role in linking the binary relations to the final ternary relations.

Compared to a corpus for the standard SRL task, there are some differences in this task: first is the relative position of PROTEIN names and ORGANISM/LOCATION names. Unlike the case in SRL, where arguments locate either before or after the predicate, in this application it is possible that one NE is embedded in another. A second difference is that a predicate in SRL scenario typically consists of only one word; here a PROTEIN name can contain up to 8 words.

We do not use PropBank data in our model at all. All of our training data and test data is annotated by domain expert biologists and parsed by Charniak-Johnson's parser (released in 2006). When there is a misalignment between the NE and the constituent in the parse tree, we insert a new NP parent node for the NE.

3 System Description

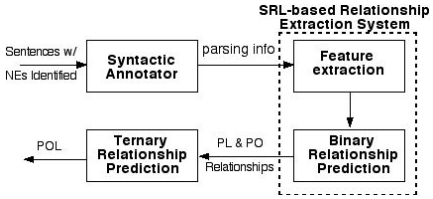

Figure 3: High-level system architecture

Fig. 3 shows the system overview. The input to our system consists of titles and abstracts that are extracted from MEDLINE records. These extracted sentences have been annotated with the NE information (PROTEIN, ORGANISM and LOCATION). The Syntactic Annotator parses the sentences and inserts the head information to the parse trees by using the Magerman/Collins head percolation rules. The main component of the system is our SRL-based relation extraction module, where we first manually extract features along the path from the PROTEIN name to the ORGANISM/LOCATION name and then train a binary SVM classifier for the binary relation extraction. Finally, we fuse the extracted binary relations into a ternary relation. In contrast with our discriminative model, a statistical parsing based generative model (Shi et al., 2007) has been proposed for a related task on this data set where the NEs and their relations are extracted together and used to identify which NEs are relevant in a particular sentence. Since our final goal is to facilitate the biologists to generate the annotated corpus, in future

• each word and its Part-of-Speech (POS) tag of PRO name
• head word (hw) and its POS of PRO name
• subcategorization that records the immediate structure that expands from PRO name. Non-PRO daughters will be eliminated
• POS of parent node of PRO name
• hw and its POS of the parent node of PRO name
• each word and its POS of ORG name (in the case of "PO " relation extraction).
• hw and its POS of ORG name
• POS of parent node of ORG name
• hw and its POS of the parent node of ORG name
• POS of the word immediately before/after ORG name
• punctuation immediately before/after ORG name
• feature combinations: hw of PRO name_hw of ORG name, hw of PRO name_POS of hw of ORG name, POS of hw of PRO name_POS of hw of ORG name
• path from PRO name to ORG name and the length of the path
• trigrams of the path. We consider up to 9 trigrams
• lowest common ancestor node of PRO name and ORG name along the path
• LCA (Lowest Common Ancestor) path that is from ORG name to its lowest common ancestor with PRO name
• relative position of PRO name and ORG name. In parse trees, we consider 4 types of positions that ORGs are relative to PROs: before, after, inside, other

Table 1: Features adopted from the SRL task. PRO: PROTEIN; ORG: ORGANISM

• subcategorization that records the immediate structure that expands from ORG name. Non-ORG daughters will be eliminated
• if there is an VP node along the path as ancestor of ORG name
• if there is an VP node as sibling of ORG name
• path from PRO name to LCA and the path length (L1)
• path from ORG name to LCA and the path length (L2)
• combination of L1 and L2
• sibling relation of PRO and ORG
• distance between PRO name and ORG name in the sentence. (3 valued: 0 if nw (number of words) = 0; 1 if $0 < nw <= 5$; 2 if $nw > 5$)
• combination of distance and sibling relation

Table 2: New features used in the SRL-based relation extraction system.

work we plan to take the relevant labeled NEs from the generative model as our input.

Table 1 and Table 2 list the features that are used in the system.

4 Experiments and Evaluation

4.1 Data set

Our experimental data set is derived from a small expert-curated corpus, where the POL relations and relevant PROTEIN, ORGANISM and LOCATION NEs are labeled. It contains ∼150k words, 565 relation instances for POL, 371 for PO and 431 for PL.

4.2 Systems and Experimental Results

We built several models to compare the relative utility of various types of rich syntactic features that we can exploit for this task. For various representations, such as feature vectors, trees and their combinations, we applied different kernels in a Support Vector Machine (SVM) learner. We use Joachims' SVM_light[1] with default linear kernel to feature vectors and Moschetti's SVM-light-TK-1.2[2] with the default tree kernel. The models are:

Baseline1 is a purely word-based system, where the features consist of the unigrams and bigrams between the PROTEIN name and the ORGANISM/LOCATION names inclusively, where the stopwords are selectively eliminated.

Baseline2 is a naive approach that assumes that any example containing PROTEIN, LOCATION names has the PL relation. The same assumption is made for PO and POL relations.

PAK system uses predicate-argument structure kernel (PAK) based method. PAK was defined in (Moschitti, 2004) and only considers the path from the *predicate* to the *target argument*, which in our setting is the path from the PROTEIN to the ORGANISM or LOCATION names.

SRL is an SRL system which is adapted to use our new feature set. A default linear kernel is applied with SVM learning.

TRK system is similar to PAK system except that the input is an entire parse tree instead of a PAK path.

TRK+SRL combines full parse trees and manually extracted features and uses the kernel combination.

[1] *http://svmlight.joachims.org/*

[2] *http://ai-nlp.info.uniroma2.it/moschitti/TK1.2-software/Tree-Kernel.htm*

Method	PL				PO				POL			
Measure	**Prec**	**Rec**	**F**	**Acc**	**Prec**	**Rec**	**F**	**Acc**	**Prec**	**Rec**	**F**	**Acc**
Baseline1	98.1	61.0	75.3	60.6	88.4	59.7	71.3	58.5	57.1	90.9	70.1	56.3
Baseline2	61.9	100.0	76.5	61.9	48.8	100.0	65.6	48.9	59.8	100.0	74.8	59.8
PAK	71.0	71.0	71.0	64.6	69.0	66.7	67.8	61.8	66.0	69.9	67.9	62.6
SRL	72.9	77.1	74.9	70.3	66.0	71.0	68.4	64.5	70.6	67.5	69.0	65.8
TRK	69.8	**81.6**	75.3	72.0	64.2	**84.1**	72.8	72.0	**79.6**	66.2	72.3	71.3
TRK+SRL	**74.9**	79.4	**77.1**	**72.8**	**73.9**	78.1	**75.9**	**72.6**	75.3	**74.5**	**74.9**	**71.8**

Table 3: Percent scores of Precision/Recall/F-score/Accuracy for identifying PL, PO and POL relations.

4.3 Fusion of Binary relations

We predict the POL ternary relation by fusing PL and PO binary relations if they belong to the same sentence and have the same PROTEIN NE. The prediction is made by the sum of confidence scores (produced by the SVM) of the PL and PO relations. This is similar to the postprocessing step in SRL task in which the semantic roles assigned to the arguments have to realize a legal final semantic frame for the given predicate.

4.4 Discussion

Table 3 shows the results using 5-fold cross validation. We report figures on ternary relation extraction and extraction of the two binary relations. Comparison between the **PAK** model and **SRL** model shows that manually specified features are more discriminative for binary relation extraction; they boost precision and accuracy for ternary relation extraction. In contrast to the **SRL** model for binary relation extraction, the **TRK** model obtains lower recall but higher precision. The combination of **SRL** with the **TRK** system gives best overall accuracy of 71.8% outperforming shallow word based features.

5 Conclusion

In this paper we explored the use of rich syntactic features for the relation extraction task. In contrast with the previously used set of syntactic features for this task, we use a large number of features originally proposed for the Semantic Role Labeling task. We provide comprehensive experiments using many different models that use features from parse trees. Using rich syntactic features by combining SRL features with tree kernels over the entire tree obtains 71.8% accuracy which significantly outperforms shallow word-based features which obtains 56.3% accuracy.

References

C. Blaschke, M. Andrade, C. Ouzounis, and A. Valencia. 1999. Automatic extraction of biological information from scientific text: Protein-protein interactions. In *AAAI-ISMB 1999*.

R. C. Bunescu and R. J. Mooney. 2005. A shortest path dependency kernel for relation extraction. In *Proc. HLT/EMNLP-2005*.

G. Claudio, A. Lavelli, and L. Romano. 2006. Exploiting Shallow Linguistic Information for Relation Extraction from Biomedical Literature. In *Proc. EACL 2006*.

A. Culotta and J. Sorensen. 2004. Dependency tree kernels for relation extraction. In *Proc. ACL-2004*.

R. Feldman, Y. Regev, M. Finkelstein-Landau, E. Hurvitz, and B. Kogan. 2002. Mining biomedical literature using information extraction. *Current Drug Discovery*.

N. Kambhatla. 2004. Combining lexical, syntactic, and semantic features with maximum entropy models for information extraction. In *Proc. ACL-2004 (poster session)*.

S. Miller, H. Fox, L. Ramshaw, and R. Weischedel. 2000. A novel use of statistical parsing to extract information from text. *Proc. NAACL-2000*.

A. Moschitti. 2004. A study on convolution kernels for shallow semantic parsing. In *Proc. ACL-2004*.

T. Sekimizu, H.S. Park, and J. Tsujii. 1998. Identifying the interaction between genes and gene products based on frequently seen verbs in medline abstracts. In *Genome Informatics*. 62-71.

Z. Shi, A. Sarkar, and F. Popowich. 2007. Simultaneous Identification of Biomedical Named-Entity and Functional Relation UsingStatistical Parsing Techniques. In *NAACL-HLT 2007 (short paper)*.

D. Zelenko, C. Aone, and A. Richardella. 2003. Kernel methods for relation extraction. *Journal of Machine Learning Research*.

M. Zhang, J. Zhang, J. Su, and G.D. Zhou. 2006. A Composite Kernel to Extract Relations between Entities with Both Flat and Structured Features. In *Proc. ACL-2006*.