Simultaneous Identification of Biomedical Named-Entity and Functional Relations Using Statistical Parsing Techniques *

Zhongmin Shi and **Anoop Sarkar** and **Fred Popowich**
School of Computing Science
Simon Fraser University
{zshi1,anoop,popowich}@cs.sfu.ca

Abstract

In this paper we propose a statistical parsing technique that simultaneously identifies biomedical named-entities (NEs) and extracts subcellular localization relations for bacterial proteins from the text in MEDLINE articles. We build a parser that derives both syntactic and domain-dependent semantic information and achieves an F-score of 48.4% for the relation extraction task. We then propose a semi-supervised approach that incorporates noisy automatically labeled data to improve the F-score of our parser to 83.2%. Our key contributions are: learning from noisy data, and building an annotated corpus that can benefit relation extraction research.

1 Introduction

Relation extraction from text is a step beyond Named-Entity Recognition (NER) and generally demands adequate domain knowledge to build relations among domain-specific concepts. A Biomedical Functional Relation (relation for short) states interactions among biomedical substances. In this paper we focus on one such relation: Bacterial Protein Localization (BPL), and introduce our approach for identifying BPLs from MEDLINE[1] articles.

BPL is a key functional characteristic of proteins. It is essential to the understanding of the function of different proteins and the discovery of suitable drugs, vaccines and diagnostic targets. We are collaborating with researchers in molecular biology with the goal of automatically extracting BPLs from text with BioNLP techniques, to expand their protein localization database, namely PSORTdb[2](Rey et al., 2005). Specifically, the task is to produce as output the relation tuple *BPL(BACTERIUM, PROTEIN, LOCATION)* along with source sentence and document references. The task is new to BioNLP in terms of the specific biomedical relation being sought. Therefore, we have to build annotated corpus from scratch and we are unable to use existing BioNLP shared task resources in our experiments. In this paper we extract from the text of biomedical articles a relation among: a LOCATION (one of the possible locations shown in Figure 1 for Gram+ and Gram- bacteria); a particular BACTERIUM, e.g. *E. Coli*, and a PROTEIN name, e.g. *OprF*.

(Nair and Rost, 2002) used the text taken from Swiss-Prot annotations of proteins, and trained a subcellular classifier on this data. (Hoglund et al., 2006) predicted subcellular localizations using an SVM trained on both text and protein sequence data, by assigning each protein name a vector based on terms co-occurring with the localization name for each organism. (Lu and Hunter, 2005) applied a hierarchical architecture of SVMs to predict subcellular localization by incorporating a semantic hierarchy of localization classes modeled with biological processing pathways. These approaches either ignore the actual location information in their predicted localization relations, or only focus on a small portion of eukaryotic proteins. The performance of these approaches are not comparable due to different tasks and datasets.

2 System Outline

During our system's preprocessing phase, sentences are automatically annotated with both syntactic information and domain-specific semantic information. Syntactic annotations are provided by a statistical parser (Charniak and Johnson, 2005). Domain-

*This research was partially supported by NSERC, Canada.

[1]MEDLINE is a bibliographic database of biomedical scientific articles at National Library of Medcine (NLM, http://www.nlm.nih.gov/).

[2]http://db.psort.org.

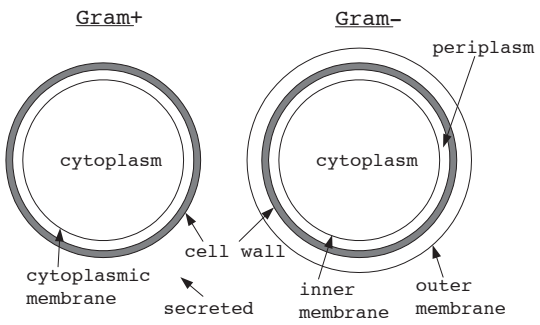

Figure 1: Illustration of possible locations of proteins with respect to the bacterial cell structure.

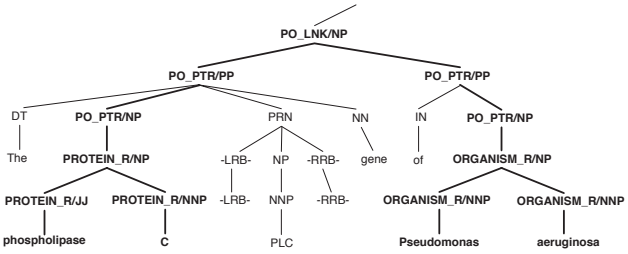

Figure 2: An example of parsing results

specific semantic information includes annotations on PROTEIN, BACTERIUM and LOCATION NEs by dictionary lookups from UMLS[3], NCBI Taxonomy[4] and SwissProt[5], and two automatic Bio-NE recognizers: MMTx[6] and Lingpipe[7].

We propose the use of a parser that simultaneously identifies NEs and extracts the BPL relations from each sentence. We define NEs to be **Relevant** to each other only if they are arguments of a BPL relation, otherwise they are defined to be **Irrelevant**. A sentence may contain multiple PROTEIN (LOCATION or ORGANISM) NEs, e.g., there are two PROTEIN NEs in the sentence below but only one, *OmpA*, is relevant. Our system aims to identify the correct BPL relation among all possible BPL tuples (candidate relations) in the sentence by only recognizing relevant NEs. Each input sentence is assumed to have at least one BPL relation.

Nine of 10 monoclonal antibodies mapped within the carboxy-terminal region of [$_{PROTEIN}$ *OprF*] *that is homologous to the* [$_{ORGANISM}$ *Escherichia coli*] [$_{LOCATION}$ *outer membrane*] *protein* [$_{PROTEIN}$ *OmpA*].

3 Statistical Syntactic and Semantic Parser

Similar to the approach in (Miller et al., 2000) and (Kulick et al., 2004), our parser integrates both syntactic and semantic annotations into a single annotation as shown in Figure 2. A lexicalized statistical parser (Bikel, 2004) is applied to the parsing task. The parse tree is decorated with two types of semantic annotations:

1) Annotations on relevant PROTEIN, BACTERIUM and LOCATION NEs. Tags are *PROTEIN_R*, *BACTERIUM_R* and *LOCATION_R* respectively.

2) Annotations on paths between relevant NEs. The lower-most node that spans both NEs is tagged as *_LNK* and all nodes along the path to the NEs are tagged as *_PTR*.

Binary relations are apparently much easier to represent on the parse tree, therefore we split the BPL ternary relation into two binary relations: BP (BACTERIUM and PROTEIN) and PL (PROTEIN and LOCATION). After capturing BP and PL relations, we will predict BPL as a fusion of BP and PL, see §4.1. In contrast to the global inference done using our generative model, heavily pipelined discriminative approaches usually have problems with error propagation. A more serious problem in a pipelined system when using syntactic parses for relation extraction is the alignment between the named entities produced by a separate system and the syntactic parses produced by the statistical parser. This alignment issue is non-trivial and we could not produce a pipelined system that dealt with this issue satisfactorily for our dataset. As a result, we did not directly compare our generative approach to a pipelined strategy.

4 Experiment Settings and Evaluations

The training and test sets are derived from a small expert-curated corpus. Table 1 lists numbers of sentences and relevant NEs in each BP/PL/BPL set.

Since the parsing results include both NE and path tags (note that we do not use any external NER system), there are two metrics to produce and evaluate PL or BP relations: **Name-only** and **Name-path** metrics. The **name-only** metric only measures **Rel-**

[3]http://www.nlm.nih.gov/research/umls/
[4]http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=Taxonomy
[5]http://www.ebi.ac.uk/swissprot/
[6]MetaMap Transfer, http://mmtx.nlm.nih.gov/
[7]http://www.alias-i.com/

	PL	BP	BPL
Training set	289 / 605	258 / 595	352 / 852
Test set	44 / 134	28 / 127	62 / 182

Table 1: Sizes of training and test sets (number of sentences / number of relevant NEs)

evant PROTEIN, BACTERIUM and LOCATION NEs (see Section 2). It does not take path annotations into account. The name-only metric is measured in terms of Precision, Recall and F-score, in which True Positive (TP) is the number of correctly identified NEs, False Positive (FP) is the number of incorrectly identified NEs and False Negative (FN) is the number of correct NEs that are not identified.

The **name-path** measures nodes being annotated as *_LNK*, *_PTR* or *_R* along the path between NEs on the parse tree, therefore it represents **confidence** of NEs being arguments of the relation. The name-path metric is a macro-average measure, which is the average performance of all sentences in data set. In measurement of the name-path metric, TP is the number of correctly annotated nodes on the path between relevant NEs. FP is the number of incorrectly annotated nodes on the path and FN is the number of correct nodes that are not identified.

4.1 Fusion of BP and PL

The BPL relation can be predicted by a fusion of BP and PL once they are extracted. Specifically, a BP and a PL that are extracted from the same sentence are merged into a BPL. The predicted BPL relations are then evaluated by the same name-only and name-path metrics as for binary relations. In the name-path metric, nodes on both PL and BP paths are counted. Note that we do not need a common protein NER to merge the BP and PL relations. E.g., for name-only evaluation, assume true BPL($B1$, $P1$, $L1$): if we predict BP($B1$,) and PL($P1$, $L2$), then *TP*=2 due to $B1$, $P1$; *FP*=1 due to $L2$; and *FN*=1 due to $P1$.

5 NER and BPL Extraction

Baseline: An intuitive method for relation extraction would assume that any sentence containing PROTEIN, ORGANISM and LOCATION NEs has the relation. We employ this method as a baseline system, in which NEs are identified by the automatic NE recognizers and dictionary lookups as introduced in §2. The system is evaluated against the test set in Table 1. Results in Table 2 show low precision for PROTEIN NER and the name-path metric.

Extraction using Supervised Parsing: We first experiment a fully supervised approach by training the parser on the BP/PL training set and evaluate on the test set (see Table 1). The name-only and name-path evaluation results in Table 2 show poor syntactic parsing annotation quality and low recall on PROTEIN NER. The major reason of these problems is the lack of training data.

Extraction using Semi-supervised Parsing: Experiments with purely supervised learning show that our generative model requires a large curated set to minimize the sparse data problem, but domain-specific annotated corpora are always rare and expensive. However, there is a huge source of unlabeled MEDLINE articles available that may meet our needs, by assuming that any sentence containing BACTERIUM, PROTEIN and LOCATION NEs has the BPL relation. We then choose such sentences from a subset of the MEDLINE database as the training data. These sentences, after being parsed and BPL relations inserted, are in fact the very noisy data when used to train the parser, since the assumed relations do not necessarily exist. The reason this noisy data works at all is probably because we can learn a preference for structural relations between entities that are close to each other in the sentence, and thus distinguish between competing relations in the same sentence. In future work, we hope to explore explicit bootstrapping from the labeled data to improve the quality of the noisy data.

Two experiments were carried out corresponding to choices of the training set: 1) noisy data only, 2) noisy data and curated training data. Evaluation results given in Table 2.

Evaluation results on the name-only metric show that, compared to supervised parsing, our semi-supervised method dramatically improves recall for NER. For instance, recall for PROTEIN NER increases from 25.0% to 81.3%; recall on BACTERIUM and LOCATION NERs increases about 30%. As for the name-path metric, the overall F-score is much higher than our fully supervised method increasing from 39.9% to 74.5%. It shows that the inclusion of curated data in the semi-

Method	Measure	Name-only Evaluation (%)					Name-Path Evaluation (%)		
		PL		BP		BPL	PL	BP	BPL
		PROT	LOC	PROT	BACT				
Baseline	**P**	42.3	78.6	41.9	81.3	40.7	27.1	38.9	31.0
	R	92.5	97.3	87.8	97.4	90.9	56.5	69.0	60.7
	F	58.0	87.0	56.7	88.6	56.2	36.6	49.8	41.0
Supervised (training data only)	**P**	66.7	87.5	66.7	72.7	76.9	45.9	41.2	43.9
	R	25.0	56.0	10.5	47.1	35.3	36.7	36.3	36.5
	F	36.4	68.3	18.2	57.1	48.4	40.8	38.6	39.9
Semi-supervised (noisy data only)	**P**	66.7	95.5	70.6	94.1	80.8	76.2	83.5	79.3
	R	84.2	80.8	80.0	84.2	81.8	67.8	72.4	67.0
	F	74.4	87.5	75.0	88.9	81.3	71.7	**77.5**	74.2
Semi-supervised (noisy data + training data)	**P**	73.9	95.5	76.5	94.1	84.8	77.0	81.1	78.7
	R	81.0	80.8	81.3	84.2	81.7	68.5	73.7	70.7
	F	**77.3**	**87.5**	**78.8**	**88.9**	**83.2**	**72.5**	77.2	**74.5**

Table 2: Name-only and name-path evaluation results. PROTEIN, LOCATION and BACTERIUM are PROT, LOC and BACT for short. The training data is the subset of curated data in Table 1.

supervised method does not improve performance much. Precision of PROTEIN NER increases 6.5% on average, while F-score of overall BPL extraction increases only slightly. We experimented with training the semi-supervised method using noisy data alone, and testing on the entire curated set, i.e., 333 and 286 sentences for BP and PL extractions respectively. Note that we do not directly train from the training set in this method, so it is still "unseen" data for this model. The F-scores of path-only and path-name metrics are 75.5% and 67.1% respectively.

6 Discussion and Future Work

In this paper we introduced a statistical parsing-based method to extract biomedical relations from MEDLINE articles. We made use of a large unlabeled data set to train our relation extraction model. Experiments show that the semi-supervised method significantly outperforms the fully supervised method with F-score increasing from 48.4% to 83.2%. We have implemented a discriminative model (Liu et al., 2007) which takes as input the examples with gold named entities and identifies BPL relations on them. In future work, we plan to let the discriminative model take the output of our parser and refine our current results further. We also plan to train a graphical model based on all extracted BP, PL and BPL relations to infer relations from multiple sentences and documents.

References

D. Bikel. 2004. A distributional analysis of a lexicalized statistical parsing model. In *Proc. of EMNLP '04*, pages 182–189.

E. Charniak and M. Johnson. 2005. Coarse-to-fine n-best parsing and maxent discriminative reranking. In *Proc. of ACL '05*, pages 173–180.

A. Hoglund, T. Blum, S. Brady, P. Donnes, J. Miguel, M. Rocheford, O. Kohlbacher, and H. Shatkay. 2006. Significantly improved prediction of subcellular localization by integrating text and protein sequence data. In *Proc. of PSB '06*, volume 11, pages 16–27.

S. Kulick, A. Bies, M. Libeman, M. Mandel, R. McDonald, M. Palmer, A. Schein, and L. Ungar. 2004. Integrated annotation for biomedical information extraction. In *Proc. of HLT/NAACL '04*, pages 61–68, Boston, May.

Y. Liu, Z. Shi, and A. Sarkar. 2007. Exploiting rich syntactic information for relation extraction from biomedical articles. In *NAACL-HLT '07, poster track*, Rochester, NY, April.

Z. Lu and L. Hunter. 2005. Go molecular function terms are predictive of subcellular localization. In *Proc. of PSB '05*, volume 10, pages 151–161.

S. Miller, H. Fox, L. Ramshaw, and R. Weischedel. 2000. A novel use of statistical parsing to extract information from text. In *Proc. of NAACL '06*, pages 226–233.

R. Nair and B. Rost. 2002. Inferring subcellular localization through automated lexical analysis. In *Bioinformatics*, volume 18, pages 78–86.

S. Rey, M. Acab, J. Gardy, M. Laird, K. deFays, C. Lambert, and F. Brinkman. 2005. Psortdb: A database of subcellular localizations for bacteria. *Nucleic Acids Research*, 33(D):164–168.