A Semi-Supervised Generative Model Integrating Both Syntactic and Semantic Features for Bacterial Subcellular Localization Extraction

Zhongmin Shi¹, Li Zhong², Guishi Lin²
¹ The DGUT Innovation Center of Robotics and Intelligent Equipment, Dongguan, Guangdong, China
² Guangzhou Summba Information Technology Co., Ltd., Guangzhou, Guangdong, China
Email for correspondence: guishi@summba.com

Abstract. Our study on the Bacterial Subcellular Localizations (BPLs) extraction focuses on generative learning. We propose a generative model extracting BPLs from MEDLINE abstracts. The model integrates both syntactic and semantic features of a sentence, and capable of identifies biomedical named-entities and relations at the same time from a large set of noisy biomedical data. The overall performance of the model exhibits a significant improvement comparing to a supervised alternative.

1. Introduction
Subcellular Localization (SCL) is a typical biomedical relation. Bacterial SCL indicates where locations of proteins in bacteria. For example:

E. Coli produces [LOCATION membrane-bound] [PROTEIN lytic trans-glycosylase] and localizes it in [BACTERIUM murein sacculus].

The above example states a membrane-bound localization relation of above bacterium and protein. Since a protein has to be translocated to the correct subcellular compartments or stick to a membrane to function properly, the bacterial SCL is a core functional characteristic of proteins. This characteristic is the key to understanding the functions of different proteins and to discover suitable vaccines, drugs and diagnostic targets. Locations of bacterial proteins for Gram+ and Gram- bacteria are shown in Figure 1.

Figure 1. Locations of proteins and respecting bacterial structures.
Different from protein-protein, gene-gene and protein-disease relations that are associations of protein molecules from the perspective of biochemistry and signal transduction, the SCLs indicate functions of single proteins and therefore are more fundamental to the study of proteins.

To determine SCLs experimentally is however a time-consuming and laborious job. Studies with computational methods, for instance, Hidden Markov Models (HMMs) and Support Vector Machines (SVMs), have been carried out to predict SCLs automatically from protein sequences in the biomedical text, and exceed the accuracy of some high-throughput laboratory studies for to identify protein subcellular localization [1][2].

From the viewpoint of Natural Language Processing (NLP), identifying relations from the biomedical text is far more difficult than from other domains that have been widely used, for instance, newswire. The SCL extraction highly relies on protein name identification, which is much harder than the identification of location, person, organization names, etc. Moreover, the understanding of the biomedical text is highly relies on domain knowledge.

In this paper, we propose a generative model integrating both syntactic and domain-dependent semantic features from the parse tree of a sentence, and is able to simultaneously identify biomedical named-entities and relations.

2. Description of the Statistical Parser
As shown in Figure 2, our parser integrates both syntactic and semantic annotations into a single annotation, similar to the approach in [3] and [4]. We apply a lexicalized statistical parser [5], which is augmented by two kinds of semantic annotations:

1. Annotations on relevant PROTEIN, LOCATION and BACTERIUM NEs. Tags are PROTEIN-R, LOCATION-R and BACTERIUM-R respectively.

2. Annotations of paths between these relevant NEs. Nodes along the path to NEs are noted as _PTR, and lower-most nodes that span both NEs are noted as _LNK.

![Figure 2. An example of parsing results](image)

The BPL-relation is split into two binary relations: PROTEIN - LOCATION (PL) and BACTERIUM - PROTEIN PL (BP), which are more feasible to represent on the parse tree. The BPL-relation can be generated by integrating the extracted BP and PL. The integration of PL and BP relations is to generate the target BPL-relation from the extracted PL and BP relations. The relation between a BP and a PL will be fused if they are 1) in the same abstract and 2) both PROTEIN names referring to the same protein.

3. Confidence of Relation Prediction
For each parse T, Bikel’s parser [5] generates a log probabilistic confidence score c_T, which however does not consider the confidence of a relation prediction that only covers a sub-tree of the entire parse
In the paper we do not optimize Bikel's parser to calculate the log probability of a sub-tree, instead we approximately assign a confidence score for the sub-tree $t$ from $c_T$:

$$c_i = c_T + \log \frac{l(t)}{l(T)}$$  \hspace{1cm} (1)

Where $l(T)$ and $l(t)$ indicate the entire tree $T$ and the number of words covered by the sub-tree $t$ respectively. Moreover, we apply a penalty to $c_i$ of recovered relations. A penalty coefficient $p_t$ is calculated as follows:

$$p_t = \frac{n_{\text{tags}}}{n_{\text{path}}}$$  \hspace{1cm} (2)

Where $n_{\text{path}}$ indicates the number of nodes on the path, and $n_{\text{tags}}$ denotes the number of nodes annotated with relation or NE tags by the parser along the path, before any recovery is applied. Therefore, the new definition of $c_i$ is:

$$c_i = c_T + \log \frac{L(t)\times n_{\text{tags}}}{L(T)\times n_{\text{paths}}}$$  \hspace{1cm} (3)

When merging two binary relations (from the same sentence and the same PROTEIN NE) into a BPL-relation, the confidence score of BPL is assigned as the addition of confidence scores of BP and PL. The confidence score of predicted BPL-relations is between (-3.0, -1.0) as shown by our experiments, which also indicates that any two binary relations with the same PROTEIN NEs in the same sentence have great chance to compose a valid BPL-relation. We therefore assign a threshold of -2.5 to the confidence score of BPL predictions, such that any BPL predictions with confidence score not greater than the threshold are ignored.

4. Extraction with Supervised Parsing

At the first place, we apply a fully supervised approach to train the parser with the BP/PL training set and to evaluate the parser with the test set. Table 1 that summarizes the training and test sets.

**Table 1.** Training and test sets: BPL instances, numbers of sentences, and relevant NEs

| Data set             | Sentences | BPL Instances | Relevant NEs |
|----------------------|-----------|---------------|--------------|
| Positive training set| 505       | 333           | 768          |
| Positive test set    | 100       | 65            | 160          |
| Negative training set| 653       | 649           | 0            |
| Negative test set    | 146       | 145           | 0            |

Table 2 shows Evaluation results with low precision and recall score of binary predictions. However, with the combination of binary relations to generate ternary predictions, the precision score raises but recall decreases.

**Table 2.** Training and test sets: numbers of sentences, BPL instances and relevant NEs

| Method                      | Performance (Precision/Recall/F-score) |
|-----------------------------|----------------------------------------|
|                            | BP          | PL          | BPL         |
| Baseline 1                  | 27.1/56.5/36.6 | 338.9/69.0/49.8 | 14.1/61.5/23.0 |
| Supervised (curated data only) | 15.8/10.3/12.5 | 37.9/25.0/30.1 | 50.0/10.0/16.7 |
| Semi-Supervised (curated data + newswire) | 8.7/7.1/7.8 | 28.0/15.5/20.0 | 16.7/3.3/5.5 |
We examine results of the parser and notice that the quality of the syntactic annotations is low. Figure 3 shows a parse tree generated by the supervised parser, in which highlighted dash regions are incorrect syntactic constituent dependencies, while the dashed arrows indicate correct dependencies. In addition, recall of the PROTEIN NER is as low as 14.3% due to two problem. The one is that the quantity of PROTEIN NEs in the training set is low, the other is that protein name sources are not available.

Figure 3. An example of parsing results from the supervised parser. The parse tree includes both syntactic and semantic (NEs and relations) annotations.

Figure 3 also indicates that the most of paths between PROTEIN and BACTERIUM/LOCATION NEs are not generated, because of missed and incorrectly identified NEs. A major reason that results in incomplete paths is the lack of syntactic and semantic annotations in the training set.

5. Extraction with Semi-Supervised Learning

5.1. Training Set Expansion with Newswire Data
To optimize the syntactic parsing performance, we introduce the Penn Treebank that contains nearly 1 million syntactically parsed sentences corpus into our training data. Evaluation results indicate that the accuracy of syntactic annotations is highly improved, as shown in Table 2. For instance, the sentence in Figure 4 is correctly parsed in terms of syntactic dependencies among constituents. The overall performance however is significantly worse comparing to the supervised system, due to the decrease of feature distributions of PROTEIN, ORGANISM and LOCATION NEs with the addition of the Penn Treebank, which contains a large number of non-biomedical articles.
Figure 4. An example of the supervised parsing results w/ additional Penn Treebank examples.

5.2. Training Set Expansion with Noisy Data
The experimental results of the supervised learning indicate that our generative model requires a large size of curated biomedical text to avoid the sparse data issue, but domain-specific annotated corpora are normally rare and expensive. However, a huge set of unlabelled MEDLINE articles is available and may be handful, with assumption that BACTERIUM, PROTEIN and LOCATION NEs has the BPL-relation in many of the articles. Therefore, we choose around 14 thousand sentences that contains that BACTERIUM, PROTEIN and LOCATION NEs from a subset of the MEDLINE database as the training data. These sentences are then parsed and annotated with BPL-relation tags, resulting in a very noisy data set, since the assumed relations may not exist. However, this noisy data works since entities close to each other prefer structural relations over the competing relations in the sentence.

We then choose two sets of training data in the upcoming experiments: 1) noisy data only, 2) noisy and annotated data. Table 2 indicates that our semi-supervised methods dramatically outperform the pure supervised parsing in terms of precision and recall for both binary and ternary predictions. For the semi-supervised method trained on the noisy and curated data, the precision of ternary predictions increases 37.9% (from 51.0% to 88.9%), and the recall increases 16.5% (from 10.1% to 27.6%). Evaluation results show that the semi-supervised learning with addition of curated data greatly improves the overall performance. We also experimented the semi-supervised method with the noisy set alone, and test by the curated set containing 286 and 333 sentences respectively for PL and BP extractions. The F-score of ternary predictions is 25.1%.

6. Conclusion and Future Work
In this paper we introduce a statistical parsing-based method that integrates both syntactic and semantic features to extract specific biomedical relations from biomedical articles. We utilize a large unlabelled data set to train the relation extraction model. Experiments indicate that the semi-supervised model significantly increases F-score from 16.7% to 40.5% against the fully supervised method.

This study suggests that the BPL-relation extraction task is difficult. The system should be able to not only identify NEs, but also identify networks that jointly complete some biomedical functions. Moreover, the inadequate of annotated data makes it tougher to build an effective model. Our wish is that, with increasing the size of curated data set and utilizing various machine learning methods, the relation extraction studies would be more valuable to the Biomedical and NLP research and other relevant community [6][7].
ACKNOWLEDGMENT
This is a joint work with the Natural Language Processing Lab, Simon Fraser University, and supported by KEY Laboratory of Robotics and Intelligent Equipment of Guangdong Regular Institutions of Higher Education (2017KSYS009), The DGUT Innovation Center of Robotics and Intelligent Equipment (KCYCXPT2017006), and National Key Research and Development Program of China (2017YFB1300200, 2017YFB1300203).

References
[1] Ray, S., & Craven, M. (2001, August). Representing sentence structure in hidden Markov models for information extraction. In International Joint Conference on Artificial Intelligence (Vol. 17, No. 1, pp. 1273-1279). LAWRENCE ERLBAUM ASSOCIATES LTD.
[2] Li, L., Yu, S., Xiao, W., Li, Y., Li, M., Huang, L., Zheng, X., Zhou, S., & Yang, H. (2014, September). Prediction of bacterial protein subcellular localization by incorporating various features into Chou's PseAAC and a backward feature selection approach. Biochimie, Vol. 104 (pp. 100-107).
[3] Miller, S., Fox, H., Ramshaw, L., & Weischedel, R. (2000). A novel use of statistical parsing to extract information from text. In 1st Meeting of the North American Chapter of the Association for Computational Linguistics.
[4] Kulick, S., Bies, A., Liberman, M., Mandel, M., McDonald, R., Palmer, M., ... & White, P. (2004). Integrated annotation for biomedical information extraction. In HLT-NAACL 2004 Workshop: Linking Biological Literature, Ontologies and Databases (pp. 61-68).
[5] Bikel, D. M. (2004, July). A distributional analysis of a lexicalized statistical parsing model. In Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (pp. 182-189).
[6] Wang, J., Brodmann, M., & Basler, M. (2019, June). Assembly and Subcellular Localization of Bacterial Type VI Secretion Systems. Annual Review of Microbiology, Vol. 73 (pp. 621-638).
[7] Rudner, D., & Losick, R. (2010, April). Protein Subcellular Localization in Bacteria. Cold Spring Harb Perspect Biol. Vol. 2(4).