Statistical Principle-based Approach for Gene and Protein Related Object Recognition

Po-Ting Lai[*1], Ming-Siang Huang[2,3], Chu-Hsien Su[4], Richard Tzong-Han Tsai[5], Wen-Lian Hsu[1]

[1] Department of Computer Science, National Tsing-Hua University, Taiwan, R.O.C
[2] Taiwan International Graduate Program Bioinformatics, Institute of Information Science, Academia Sinica, Taiwan, R.O.C
[3] Institute of Biomedical Informatics, National Yang Ming University, Taiwan, R.O.C
[4] Institute of Information Science, Academia Sinica, Taiwan, R.O.C
[5] Department of Computer Science and Information Engineering, National Central University, Taiwan, R.O.C
potinglai@gmail.com
elephant52381@iis.sinica.edu.tw
jason@iis.sinica.edu.tw
thtsai@csie.ncu.edu.tw
hsu@iis.sinica.edu.tw
[*]Corresponding Author

Abstract. We introduce a Statistical Principle-based Approach (SPBA) for named entity recognition (NER). SPBA is a pattern-based approach. It uses patterns to represent protein names, and uses the semantic labels to map sentence into labeled sentence. NER is then formulated as aligning labeled sentence with patterns. The weights of insertion/deletion/match are learned through logistic regression model in our refactored JNLPBA corpus. We participated in BioCreative V.5 Gene and Protein Related Object (GPRO) task to evaluate the ability of SPBA in processing patent abstracts. Since the NE types and boundaries are slightly different in two corpora. We adjusted SPBA's NER results by using a linear chain Conditional Random Fields (CRFs) model. In BioCreative V.5 GPRO task, our best configuration achieved an F-score of 73.73% on GPRO type 1; an F-score of 78.66% on combining GPRO type 1 and 2.

Keywords. Named Entity Recognition

1 Introduction

The goal of BioCreative V.5 Gene and Protein Related Objec (GPRO) task [3] is that given a patent abstract, a text mining system should (1) identify the boundaries of GPRO (2) and determine whether GPROs can be normalized to database ID or not.

To tackle this task, we developed a pipeline named entity recognition (NER) system that cascaded two NER components, an SPBA-based NER and a CRF-based NER[4].
SPBA is a pattern-based approach for named entity recognition (NER), and it was developed on our refactored JNLPBA corpus [6]. First, our domain experts constructed an entity knowledge base (EKB) by collecting public available dictionaries, like MeSH and UniProt. EKB will be used to map word or phrase into one or more semantic label (called concept). Through EKB, we can generate NE patterns. For instance, *"GATA1 erythroid transcription factor"* can be labeled as *"**GATA1**$_{GeneSymbol}$ **erythroid**$_{BiologicalProcess}$ **transcription factor**$_{ProteinEnd}$"*, thus be represented as a pattern [GeneSymbol] [BiologicalProcess] [ProteinEnd]. Using more general concept allows a pattern to identify more NEs. We employed an alignment mechanism in our pattern matching to allow flexible matching, and scored an alignment properly through the logistic regression model to improve accuracy. Different annotation criteria between refactored JNLPBA and GPRO task caused a lower performance in the strict evaluation metric. Thus, we developed a CRFs-based component trained on GPRO training set to adjust the output NE type and boundary. Finally, a GPRO normalization component was developed to map NE into its database ID.

2 Method

Here we describe SPBA in detail, and how we adjust the NE's type and boundary. First, we construct an entity knowledge base (EKB) consisting of concepts and patterns. Then our pattern matching approach is illustrated for identifying NEs. Furthermore, a logistic regression-based approach is employed to learn the weights of patterns for scoring a matched NE. Moreover, a linear chained CRF-based approach is proposed to adjusting the boundaries of NEs according to the annotation criteria of GPRO task. Finally, a normalization component is introduced for mapping GPRO into database ID.

2.1 Statistical Principle-based Approach

Table 1. The resources of the initial concepts

Concept	Resource
BiologicalProcess	MeSH term
CellLineSymbol	CLDB [1]
CellTypeSymbol	ExPAsy [2]
Chemical	ChEBI [5]
Chromosome	Regular expression
Disease	MeSH term
DNASymbol	Entrez
Morphology	Manual
OrganTissue	SWISS-PROT
ProteinSymbol	Entrez and UniProt
RNASymbol	Entrez
Taxonomy	MeSH term
Structure	ExPAsy

Entity Knowledge Base: A NE is composed of one or more words. Some of these words could be generalized to concepts. For example, *"liver cancer"* could be generalized to the *"Cancer"* concept. If we express a NE as a set of sequence of concepts (called pattern), these patterns are likely to match unseen instances of that NE type. Therefore, EKB is constructed by collecting the concept set from publicly available biological databases shown in Table 1.

Pattern Generation: To generate pattern, we first employ prefix-tree matching to label all NEs in the training data by using the EKB. Then, unlabeled words are removed, and the remaining label sequence is called a pattern. For example, given a NE *"USF-related transcription factor"* and two EKB concepts, GeneSymbol and ProteinEnd. It will be labeled as: *"**USF**$_{GeneSymbol}$ -related **transcription factor**$_{ProteinEnd}$"*. Then [GeneSymbol] [ProteinEnd] will be the pattern. Since a NE may be labeled in more than one way, generating more than one pattern, we only keep the pattern with the highest ratio of labeled words to total words.

Pattern Matching: After pattern generation, the patterns will be used to recognize candidate NEs by an local alignment algorithm.

Logistic Regression: We use the logistic regression (LR) model [7] to learn weights for insertion, match and deletion. The score of an alignment of a pattern p and a labeled sentence l, $Similarity(p,l)$ is

calculated by the following formula:

$$imilarity(p,l) = \sum_i \lambda_M(\text{the } i\text{th matched word}) + \sum_j \lambda_D(\text{the } j\text{th deleted word}) + \sum_k \lambda_I(\text{the } k\text{th inserted word})$$

where λ_M (w) is the weight for the feature in which the matched word is w;

λ_D (w) is the weight for the feature in which the deleted word is w;

λ_I (w) is the weight for the feature in which the insertion word is w;

Then, *Similarity*(p,l) is transformed into a real number ranging from 0 to 1 by the sigmoid function:

$$h(p,l) = \frac{1}{1 + e^{-Similarity(p,l)}}$$

h is considered as the alignment score of p and l .

Word	ALL CAPS	Shape	Prefix [2]	Suffix [2]	POS	SPBA	GPRO
A	1	A	NULL	NULL	DT	O	O
promoter	0	a	pr	er	NN	O	O
sequence	0	a	se	ce	NN	O	O
of	0	a	of	of	IN	O	O
the	0	a	th	he	DT	O	O
human	0	a	hu	an	JJ	B	B-GPRO_TYPE_1
p1	0	a1	p1	p1	NN	I	E-GPRO_TYPE_1
TNF	1	A	TN	NF	NN	I	O
-	0	-	NULL	NULL	HYPH	I	O
R	1	A	NULL	NULL	NN	I	O
gene	0	a	ge	ne	NN	E	O
is	0	a	is	is	VBZ	O	O
provided	0	a	pr	ed	VBN	O	O
.	0	.	NULL	NULL	.	O	O

Fig. 1. An example of CRF features.

LR is applied as follows. Initially, all training sentences are labeled by using the EKB. The set of labeled training sentences is referred to as L. The size of L is greater or equal to the original set of training sentences because one sentence may have more than one ways of labeling. Then, patterns of each NE type are used to identify candidate NEs for each training sentence through the alignment. Moreover we collect the sets of true positive (TP) and false positive (FP) labeling results, referred to as E_{TP} and E_{FP}, respectively. Finally, we employ logistic regression model to learn feature weights.

2.2 Adjusting and Normalizing NE

NE Adjustment: There are two differences between SPBA's and GPRO's annotations. The first one is NE types. SPBA was trained on the refactored JNLPBA whose NE types are cell line, cell type, DNA, protein and RNA. However, GPRO task used two NE types, GPRO type 1 and GPRO type 2. GPRO type 1 denotes that NE can be normalized into database ID; GPRO type 2 denotes that NE cannot be normalized. The second one is NE boundaries. The curators of the refactored JNLPBA preferred to annotate longer phrase/chunk as NEs, but GPRO seems prefer to annotate the phrase/chunk which can exactly match the database's official name. Thus, we found that GPRO NEs were usually substrings of SPBA's NEs. Therefore, we used a linear chained CRF model to tackle this problem. The adjustment of NE types and boundaries are formulated as a NER problem. Given a sentence and SPBA's NE as feature, to predict the GPRO's NE. A subset of NERBio's features [8] including word, POS, affix, orthographical, word shape and POS features are used. Fig. 1 shows an example of our features.

GPRO Normalization: Another way to determine GPRO type of NE can be done by checking whether NE can be mapped into database ID. Therefore, in another configuration, we used some normalization rules [9] to retrieve the ID of a NE. For examples, expanding both dictionary names and NEs like converting to lower cases; removing the symbols; removing the named entity suffix *"s"*.

3 Experiment Results and Future Work

We participated in BioCreative V.5 GPRO Task [3]. The official performances of our submissions on GPRO test set are shown in Table 2. The task uses a strict F1-measure evaluation metric. The column of *"Use GPRO dict"* means that GPROs of the training set are used to extend our normalization dictionary. The column of *"Use normalization"* means that we use the normalization component. Our best configuration achieves an F-score of 73.73% on GPRO type 1; an F-score of 78.66% on combining GPRO type 1 and 2. To our surprise, the results shows that the Config. 2 slightly outperform the Config. 1 which uses the normalization component. In the future, we would like to release a restful web service of our pipeline system.

Table 2. The performances of our submissions

Config.	Use GPRO dict	Use normalization	GPRO Type 1			Merge GPRO Type 1 and 2		
			Precision	Recall	F-score	Precision	Recall	F-score
1	O	O	**68.69%**	78.24%	73.15%	**81.44%**	74.67%	77.91%
2	X	X	66.53%	**82.68%**	**73.73%**	78.63%	**78.70%**	**78.66%**

REFERENCES

1. Romano, P., Manniello, A., Aresu, O., Armento, M., Cesaro, M., Parodi, B.: Cell Line Data Base: structure and recent improvements towards molecular authentication of human cell lines. Nucleic Acids Research 37, D925-D932 (2008)
2. Artimo, P., Jonnalagedda, M., Arnold, K., Baratin, D., Csardi, G., de Castro, E., Duvaud, S., Flegel, V., Fortier, A., Gasteiger, E., Grosdidier, A., Hernandez, C., Ioannidis, V., Kuznetsov, D., Liechti, R., Moretti, S., Mostaguir, K., Redaschi, N., Rossier, G., Xenarios, I., Stockinger, H.: ExPASy: SIB bioinformatics resource portal. Nucleic Acids Research 40, W597-W603 (2012)
3. Pérez, M.P., Rabal, O., Rodríguez, G.P., Vazquez, M., Fdez-Riverola, F., Oyarzabal, J., Valencia, A., Lourenco, A., Krallinger, M.: Evaluation of chemical and gene/protein entity recognition systems at BioCreative V.5: the CEMP and GPRO patents tracks. In: Proceedings

of the BioCreative V.5 Challenge Evaluation Workshop, pp. 3-11. (2017)
4. https://taku910.github.io/crfpp/
5. Hastings, J., Owen, G., Dekker, A., Ennis, M., Kale, N., Muthukrishnan, V., Turner, S., Swainston, N., Mendes, P., Steinbeck, C.: ChEBI in 2016: Improved services and an expanding collection of metabolites. Nucleic Acids Research 44, D1214-D1219 (2015)
6. Kim, J.-D., Ohta, T., Tsuruoka, Y., Tateisi, Y., Collier, N.: Introduction to the bio-entity recognition task at JNLPBA. Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and its Applications, pp. 70-75. Association for Computational Linguistics, Geneva, Switzerland (2004)
7. Ng, A.Y.: Feature selection, L_1 vs. L_2 regularization, and rotational invariance. Proceedings of the twenty-first international conference on Machine learning, pp. 78. ACM, Banff, Alberta, Canada (2004)
8. Dai, H.-J., Lai, P.-T., Chang, Y.-C., Tsai, R.: Enhancing of chemical compound and drug name recognition using representative tag scheme and fine-grained tokenization. J Cheminform 7, S14 (2015)
9. Tsai, R.T.-H., Lai, P.-T.: Multistage gene normalization for full-text articles with context-based species filtering for dynamic dictionary entry selection. BMC Bioinformatics 12 Suppl 8, (2011)