Mining Patents with tmChem, GNormPlus and an Ensemble of Open Systems

Robert Leaman[1], Chih-Hsuan Wei[1], Cherry Zou[1,2], Zhiyong Lu[1*]

[1]National Center for Biotechnology Information, Bethesda, Maryland, USA
[2]Poolesville High School, Poolesville, Maryland, USA

robert.leaman@nih.gov; chih-hsuan.wei@nih.gov;
chzou2009@gmail.com; *zhiyong.lu@nih.gov

Abstract. The significant amount of medicinal chemistry information contained in patents make them an attractive target for text mining. The CHEMDNER task at BioCreative V focused on information extraction from patents. This manuscript describes our submissions to the CEMP (chemical named entity recognition) and GPRO (gene and related object identification) subtasks. Our CEMP submission is an ensemble of five open systems, including both versions of tmChem, our previous work on chemical named entity recognition. Their output is combined using a machine learning classification approach. Our CEMP system obtained 0.8752 precision and 0.9129 recall, for 0.8937 f-score. Our submission to the GPRO task is an extension of GNormPlus, our previous work for gene and protein named entity recognition. Our GPRO system obtained a performance of 0.8143 precision and 0.8141 recall for 0.8137 f-score. Both submissions achieved the highest performance in their respective tasks.

Keywords: Chemicals; Patents; Information Extraction; Machine Learning; Ensemble Systems; Conditional Random Fields

1 Introduction

While publications such as those found in the biomedical literature contain a significant amount of useful chemical information [1], much of the useful information on medicinal chemistry is found in less formal documents, such as patents. The CHEMDNER task at BioCreative V, a major challenge events in BioNLP [2], addressed the extraction of chemical and biological entities from medicinal chemistry patents [3]. NCBI participated in both the CEMP (chemical named entity recognition) and GPRO (gene and protein related object identification) subtasks. Our CEMP submission consisted of an ensemble system combin-

ing the results 10 models from five open NER systems for chemical named entity recognition. Our submission to the GPRO subtask was based on the open-source GNormPlus system [4].

2 CEMP Task Methods

Our submission to the CEMP subtask was an ensemble system combining the results from five individual systems. Each of the individual systems included are machine learning systems, based on conditional random fields with a rich feature approach. These are tmChem Model 1 and tmChem Model 2 [5], becas[chemical] [6], the Wuhan university CHEMDNER tagger [7], and banner-chemdner [8]. All systems are retrainable, with the exception of becas[chemical] since it is provided as a software service.

We trained the constituent systems using combinations of two corpora. First, we used the training and development sets of annotated patents provided by the organizers. We pooled the training and development sets, and randomly split this into three sets: the training set, containing 12000 articles, and two evaluation sets containing 1000 articles each. We also used the full corpus of PubMed abstracts from the CHEMDNER task in BioCreative IV [9].

We combined the results of the constituent systems using a machine learning classification approach, representing each mention returned as an instance to be classified. We used one binary feature per constituent system; each feature representing whether the respective system returned the mention. Our implementation used Weka [10] and libsvm [11]. We found logistic regression and support vector machines to provide the highest performance.

We handle overlapping mentions by selecting the mention with the highest classification score. The result changes the precision / recall balance, which we overcome by determining the optimal classification score threshold for each classifier on the two evaluation sets and use their average for the final ensemble.

We submitted two runs with the intention of maximizing recall. The first ("high recall") simply omits the thresholding step, returning all mentions found after handling overlaps. The second ("higher recall") performs the same procedure but also adds new mentions whose text matches a mention found within the same abstract.

3 GPRO Task Methods

Our submission to the GPRO task was an adaptation of GNormPlus [4], our previous work on gene/protein name recognition and normalization. GNormPlus is a conditional random fields (CRF) [12] based method which can recognize gene/protein, family and domain mentions, and also determines their respective identifiers in NCBI Gene. By default, GNormPlus is trained using the refactored corpus of BioCreative II Gene Normalization task [13].

For the GPRO task, we used the BIEO (B: begin, I: inside, E: end and O: outside) labeling model and a CRF order of 2. More specifically, we created five individual models (M1-M5) based on different training data and features. We first separated all gene/protein-related annotations into two distinct types: mentions that can be normalized to a database record (type 1) and mentions that cannot (type 2). Next, our five models were designed as follows: In model 1, both types of mentions were used and were treated the same. In model 2, both types were used but treated as two separate classes. In model 3, the type 2 mentions were ignored and the model was trained with only the type 1 mentions. Models 4 and 5 resembled models l and 3 respectively, but also used the recognition result of the default GNormPlus system as an additional feature. The other features used in the five models were directly adapted from GNormPlus, including linguistic features, character calculation, semantic type and contextual words.

As in GNormPlus, we employed several post-processing steps: including enforcing tagging consistency and abbreviation resolution. In addition, we performed filtering, especially for the false positive predictions in two major types: “gene/protein family name” and “not a gene/protein mention.” We filtered these using a maximum entropy classifier trained with three types of features: the 5 tokens surrounding the span, whether the span can be found in NCBI Gene, UniProt or the list of type 1 mentions in the training and development sets, and morphological features: The number of uppercases, lowercases, digits, tokens, and binary features of common gene/protein (e.g., “alpha”) or family (e.g., “proteins”) suffixes. We also filtered composite mentions (“MULTIPLE” type) by applying our previous study SimConcept [14] to recognize these mentions rather than simplify them. Taken together, these post-processing steps improve the f-score by 3-5% on the development set.

For the final task submissions, we created two variants that used majority voting to aggregate the results of multiple individual models.

4 CEMP Task Results

We report the performance of both versions of tmChem on our two evaluation sets in Table 1. We found that applying the model trained on PubMed abstracts to the patent corpus reduced performance as much as 20% (data not shown). Performance improved considerably when the systems were retrained on the patent set, as would be expected. Less expected, however, is that training with a combination of the PubMed and patent sets consistently resulted in a slightly higher net f-score.

Table 1. Results for tmChem Model 1 and Model 2 in two training configurations. Each measure is averaged between the two evaluation sets. The highest value is shown in bold.

System	Training	Precision	Recall	F-score
tmChem.M1	Patent	**0.8819**	0.8088	0.8437
tmChem.M2	Patent	0.8721	0.7953	0.8319
tmChem.M1	Both	0.8741	**0.8232**	**0.8479**
tmChem.M2	Both	0.8711	0.8159	0.8426

Our task submissions were prepared with an ensemble of 10 models. This included 4 models trained on the patent training corpus (tmChem.M1, tmChem.M2, banner-chemdner and the Wuhan tagger), 4 models trained on both the patent training corpus and the full PubMed abstracts corpus (tmChem.M1, tmChem.M2, banner-chemdner and the Wuhan tagger), and also 2 models trained only on the PubMed abstracts corpus (becas[chemical] and the Wuhan tagger).

Table 2 shows the five versions of the ensemble we submitted. The first three runs used different classifiers: logistic regression, libsvm, and support vector machines using the modified Huber loss. The base classifier for both high recall configurations was logistic regression.

Table 2. Results for our ensemble systems as measured by precision (P), recall (R) and f-score (F). The internal evaluation values are averaged between the two evaluation sets. The highest value is shown in bold.

Runs	System	Internal Evaluation			Official Test		
		Precision	Recall	F-score	Precision	Recall	F-score
1	Logistic	0.8867	0.8979	**0.8923**	0.8752	0.9129	**0.8937**
2	Huber SVM	0.9091	0.8626	0.8853	0.8908	0.8918	0.8913
3	libsvm	**0.9255**	0.8753	0.8901	**0.8971**	0.8822	0.8896
4	High recall	0.6732	0.9562	0.7901	0.7967	0.9314	0.8588
5	Higher recall	0.5922	**0.9622**	0.7331	0.5202	**0.9762**	0.6787

5 GPRO Task Results

In this task the type 2 mentions, which cannot be mapped to a specific identifier, are not evaluated. Recognizing these mentions is therefore highly important, but we unfortunately found these mentions to be highly ambiguous with the type 1 mentions. We found that the CRF model could not differentiate between the two types well (models 2 and 3), but that combining the types and refining the result in post-processing provided better performance (model 1). Adding the recognition result of GNormPlus as an additional feature in the CRF models, increased recall about 4-6%, but significantly reduced precision. We produced two runs that aggregated the recognition results with a majority voting strategy. The last row in Table 3 aggregated the results of all 5 models, and obtained highest F-score (0.8137).

Table 3. Micro-averaged results for each model on the official test set, as measured by precision (P), recall (R) and f-score (F). The highest value is shown in bold.

Runs	Methods	Precision	Recall	F-score
1	M1 results	0.7835	0.8302	0.8062
2	M2 results	**0.8224**	0.7852	0.8034
4	M4 results	0.7677	**0.8502**	0.8069
3	Majority voting based on M1 – M4 results	0.8059	0.7982	0.8020
5	Majority voting based on M1 – M5 results	0.8143	0.8141	**0.8137**

6 Acknowledgment

The authors thank the organizers of the BioCreative 5 CHEMDNER task. This research is funded by the National Institutes of Health Intramural Research Program, National Library of Medicine (RL, CH, ZL) C.Z. was a summer intern at the NCBI/NIH and supported by the NIH Intramural Research Training Award.

REFERENCES

1. Krallinger, M., Leitner, F., Rabal, O., Vazquez, M., Oyarzabal, J., Valencia, A.: CHEMDNER: The drugs and chemical names extraction challenge. Journal of cheminformatics 7, S1 (2015)
2. Huang, C.-C., Lu, Z.: Community challenges in biomedical text mining over 10 years: success, failure and the future. Briefings in Bioinformatics (2015)
3. Krallinger, M., et al.: Overview of the CHEMDNER patents task. Proceedings of the Fifth BioCreative Challenge Evaluation Workshop, (2015)
4. Wei, C.H., Kao, H.Y., Lu, Z.: GNormPlus: An Integrative Approach for Tagging Genes, Gene Families, and Protein Domains. BioMed Research International 2015, (2015)
5. Leaman, R., Wei, C.H., Lu, Z.: tmChem: a high performance approach for chemical named entity recognition and normalization. Journal of cheminformatics 7, S3 (2015)
6. Campos, D., Matos, S., Oliveira, J.L.: Chemical name recognition with harmonized feature-rich conditional random fields. Fourth BioCreative Challenge Evaluation Workshop, vol. 2, pp. 82-87 (2013)
7. Lu, Y., Ji, D., Yao, X., Wei, X., Liang, X.: CHEMDNER system with mixed conditional random fields and multi-scale word clustering. Journal of cheminformatics 7, S4 (2015)
8. Munkhdalai, T., Li, M., Batsuren, K., Ryu, H.: BANNER-CHEMDNER: Incorporating Domain Knowledge in Chemical and Drug Named Entity Recognition. Fourth BioCreative Challenge Evaluation Workshop, vol. 2, pp. 135-139 (2013)
9. Krallinger, M., Rabal, O., Leitner, F., Vazquez, M., Salgado, D., Lu, Z., Leaman, R., Lu, Y., Ji, D., Lowe, D.M., Sayle, R.A., Batista-Navarro, R.T., Rak, R., Huber, T., Rocktaschel, T., Matos, S., Campos, D., Tang, B., Xu, H., Munkhdalai, T., Ryu, K.H., Ramanan, S.V., Nathan, S., Zitnik, S., Bajec, M., Weber, L., Irmer, M., Akhondi, S.A., Kors, J.A., Xu, S., An, X., Sikdar, U.K., Ekbal, A., Yoshioka, M., Dieb, T.M., Choi, M., Verspoor, K., Khabsa, M., Giles, C.L., Liu, H., Ravikumar, K.E., Lamurias, A., Couto, F.M., Dai, H.J., Tsai, R.T., Ata, C., Can, T., Usie, A., Alves, R., Segura-Bedmar, I., Martinez, P., Oyarzabal, J., Valencia, A.: The CHEMDNER corpus of chemicals and drugs and its annotation principles. Journal of cheminformatics 7, S2 (2015)
10. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.: The WEKA Data Mining Software: An Update. SIGKDD Explorations 11, (2009)
11. Chang, C.-C., Lin, C.-J.: LIBSVM : a library for support vector machines. ACM Transactions on Intelligent Systems and Technology 2, 27:21--27:27 (2011)

12. Lafferty, J.D., McCallum, A., Pereira, F.: Conditional random fields: Probabilistic Models for Segmenting and Labeling Sequence Data. Proceedings of the International Conference on Machine Learning, pp. 282-289 (2001)
13. Morgan, A.A., Lu, Z., Wang, X., Cohen, A.M., Fluck, J., Ruch, P., Divoli, A., Fundel, K., Leaman, R., Hakenberg, J., Sun, C., Liu, H.-h., Torres, R., Krauthammer, M., Lau, W.W., Liu, H., Hsu, C.-N., Schuemie, M., Cohen, K.B., Hirschman, L.: Overview of BioCreative II gene normalization. Genome biology 9, S3 (2008)
14. Wei, C.H., Leaman, R., Lu, Z.: SimConcept: A Hybrid Approach for Simplifying Composite Named Entities in Biomedical Text. IEEE journal of biomedical and health informatics 19, 1385-1391 (2015)