UTH-CCB@BioCreative V Track 2: Recognizing Chemical Entities in Patents

Yaoyun Zhang[1], Jun Xu[1], Jingqi Wang[1], Yonghui Wu[1], Manu Parkasam[2], Hua Xu[1]*

[1]School of Biomedical Informatics, The University of Texas Health Science Center at Houston, Houston, Texas, USA
[2]Mira Loma High School, 4000 Edison Avenue Sacramento, California, USA

Yaoyun.Zhang@uth.tmc.edu;Jun.Xu@uth.tmc.edu;
Jingqi.Wang@uth.tmc.edu;Yonghui.Wu@uth.tmc.edu;
prakasam.manu@gmail.com;Hua.Xu@uth.tmc.edu

Abstract. This paper describes the participation of UTH-CCB in the BioCreative V Track 2, a challenge of recognizing chemical entities in medicinal chemistry patents. We participated in the subtask 1 of chemical entity mention recognition in patents (CEMP) and subtask 2 of chemical passage detection (CPD). Our team ranked second in CEMP with a F-measure of 88.94% and ranked first in CPD with an accuracy of 94.75%.

Keywords. Chemical Entity Recognition, Text Categorization, Information Extraction, Medicinal Chemistry Patent

1 Introduction

Although biomedical entities of interest are contained in medicinal chemistry patents, such as chemical compounds, genes and proteins, the identification and further integration of such information for databases and life science research remains a tough challenge. The Spanish National Cancer Research Center (CNIO) and University of Navarra took the initiative to organize a challenge of chemical entity recognition in patents (CHEMDNER-patents), as Track 2 of the BioCreative V challenge. This task addressed the automatic extraction of chemical and biological data from medicinal chemistry patents.

2 System Description

The UTH-CCB team participated in two sub-tasks of this challenge:

1) **Chemical entity mention recognition in patents (CEMP)**: Machine learning-based systems were built for chemical entity mention recognition from patents. As the first step, a rule-based module was used for sentence boundary detection and tokenization. Machine learning classifiers were then built using the algorithms of conditional random fields (CRF) and structured support vector machines (SSVM) on the combined dataset of training and development sets.

Typical features for named entity recognition were employed in the system, including morphological features, bag-of-word, part-of-speech and n-grams. Moreover, the output of ChemSpot [1], which was proved to be effective in [2], lexicons and patterns to recognize various types of chemicals, as well as word representation related features were also examined and used.

The systems were then further improved by multiple post-processing steps, to recover missing chemicals and to handle ill-formed chemical mentions.

2) **Chemical passage detection (CPD)**: The system output for CPD was determined based on the patent titles and abstracts with recognized chemicals in the first subtask.

Overall, our team achieved the second rank in CEMP with a F-measure of 88.94% and the first rank in CPD with an accuracy of 94.75%.

3 Acknowledgment

This study is supported in part by grants from NLM 2R01LM010681-05, NIGMS 1R01GM103859 and 1R01GM102282, and CPRIT R1307.

REFERENCES

1. Rocktaschel, T., Weidlich, M., Leser, U.: ChemSpot: A Hybrid System for Chemical Named Entity Recognition. Bioinformatics 28, 1633-1640 (2012).
2. Leaman R., Wei C-H., Lu Z.: tmChem: a high performance tool for chemical named entity recognition and normalization, Journal of Cheminformatics, 7(Suppl 1): S3 (2015).