# Accepted Manuscript

Chemical-induced Disease Relation Extraction with Dependency Information and Prior Knowledge

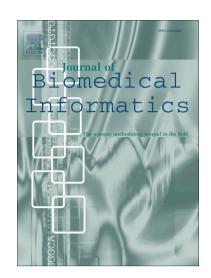Huiwei Zhou, Shixian Ning, Yunlong Yang, Zhuang Liu, Chengkun Lang, Yingyu Lin

Please cite this article as: Zhou, H., Ning, S., Yang, Y., Liu, Z., Lang, C., Lin, Y., Chemical-induced Disease Relation Extraction with Dependency Information and Prior Knowledge, *Journal of Biomedical Informatics* (2018), doi: https://doi.org/10.1016/j.jbi.2018.07.007

# Chemical-induced Disease Relation Extraction with Dependency Information and Prior Knowledge

Huiwei Zhou[*], Shixian Ning, Yunlong Yang, Zhuang Liu, Chengkun Lang, Yingyu Lin

School of Computer Science and Technology, Dalian University of Technology, Dalian 116024, Liaoning, China

[*] Corresponding author.

Email: zhouhuiwei@dlut.edu.cn, ningshixian@mail.dlut.edu.cn, SDyyl_1949@mail.dlut.edu.cn, zhuangliu1992@mail.dlut.edu.cn, kunkun@mail.dlut.edu.cn, lyydut@sina.com

**Abstract** Chemical-disease relation (CDR) extraction is significantly important to various areas of biomedical research and health care. Nowadays, many large-scale biomedical knowledge bases (KBs) containing triples about entity pairs and their relations have been built. KBs are important resources for biomedical relation extraction. However, previous research pays little attention to prior knowledge. In addition, the dependency tree contains important syntactic and semantic information, which helps to improve relation extraction. So how to effectively use it is also worth studying. In this paper, we propose a novel convolutional attention network (CAN) for CDR extraction. Firstly, we extract the shortest dependency path (SDP) between chemical and disease pairs in a sentence, which includes a sequence of words, dependency directions, and dependency relation tags. Then the convolution operations are performed on the SDP to produce deep semantic dependency features. After that, an attention mechanism is employed to learn the importance/weight of each semantic dependency vector related to knowledge representations learned from KBs. Finally, in order to combine dependency information and prior knowledge, the concatenation of weighted semantic dependency representations and knowledge representations is fed to the softmax layer for classification. Experiments on the BioCreative V CDR dataset show that our method achieves comparable performance with the state-of-the-art systems, and both dependency information and prior knowledge play important roles in CDR extraction task.

*Keywords*—CDR extraction, Dependency information, Prior knowledge, Attention mechanism.

## 1 Introduction

The extraction of chemical-disease relation (CDR) provides additional support to precision medicine efforts. It is of essential importance to the clinical disease diagnosis, treatment and drug development [1, 2]. However, manually extracting these relations from biomedical literature into structured knowledge bases, such as Comparative Toxicogenomics Database (CTD) [3], is expensive, time-consuming, and difficult to keep up-to-date. Automatically extracting CDR from the literature is becoming increasingly important for precision medicine as well as drug discovery and basic biomedical research.

To further promote the development of systems for extracting chemical-disease interactions, a challenging task of automatic extraction of CDR from biomedical literature is published on BioCreative V [4]. It consists of two specific subtasks: (i) disease named entity recognition and normalization (DNER) and (ii) chemical-induced diseases (CID) relation extraction. This paper focuses on the CID subtask at both intra- and inter-sentence levels. The intra-sentence level means a given pair of entity mentions is within the same sentence, while the inter-sentence level means a mention pair is in two different sentences.

Since then, much research has been investigated for CDR extraction, such as rule-based methods [5],