Drug Interaction Corpus

Shijun Zhang, MS1, Hengyi Wu, PhD1, Luis Rocha, PhD2, Hagit Shatkay, PhD3, Lang Li, PhD1
1The Ohio State University, Columbus, Ohio; 2Indiana University, Bloomington, Indiana; 3University of Delaware, Newark, Delaware

Introduction

Identifying drug-drug interactions has a significant impact on both clinical patient care and translational biomedical research1. The extraction and management of DDIs from the literature is a critical problem because of the overwhelming amount of DDI information. In order to develop and evaluate various text mining methods, a well annotated corpus is highly valuable. Therefore, in this study, a new drug-drug interaction (DDI) corpus is developed based on a novel annotation scheme.

Methods

The annotation guideline is developed based on the scheme recommended by Shatkay et al2. The unit of the annotation is a fragment. Each fragment is characterized along the following eight dimensions: focus, polarity, certainty, evidence, direction/trend, study type, interaction type, and mechanism. In selecting positive abstracts, DDI candidate abstracts are screened from a PubMed search with a keyword query [“drug interaction” AND (Type of Study)]. Negative abstracts are collected from the other abstracts identified from the same query. The DDI corpus is built up after the manual validation and classification process. Three annotators who develop the annotation guidelines form the Expert Group; while four other students form the Student Group. The sentence level annotation is processed through two rounds of annotation by all the annotators in two groups. Training is conducted at the beginning of each round of annotation. The expert annotator group validate and finalize all the annotations in the DDI corpus.

Results

DDI corpus has totally 1,650 DDI abstracts. It covers eight types of studies: in vivo DDI pharmacokinetics (PK) studies (n=300); in vitro DDI PK studies (n=300); clinical/in vivo DDI pharmacodynamics (PD) studies (n=300); in vitro DDI PD studies (n=200); drug-nutrition interaction studies (n=100); single drug studies (n=200); clinical case reports (n=50); and non-drug studies (n=200).

To examine the reliability of the annotations, inter-annotator agreement is used to assess the quality of corpus development in a random subset of our corpus, 102 abstracts and 938 sentences. It covers all the study types. In order to assess the quality of annotations for student annotators and expert annotators, we compared the annotation agreement between students vs experts (SE) and between experts vs experts (EE). The SE agreement has improved from the first round 82% to the second round 90%. While the EE agreement improves from 90% to 93%. This evidence suggests that after further training student annotators are equally accurate and consistent comparing to the expert annotators. The final inter-annotator agreements among eight annotation dimensions are within 90~98%.

Conclusion

The goal of this work is to construct a comprehensive DDI corpus allowing the training of machine learning algorithms for identifying relevant DDI abstracts and extract DDI drug pairs. To obtain the quality annotation, a two-round annotation process is conducted by seven well-trained annotators. After two rounds of annotation, the student annotators perform equally well as expert annotators. The overall annotation agreement among all dimensions reaches above 91%.

References