Using SemRep and a medication indication resource to extract treatment relations from clinical notes

Cosmin A. Bejan, PhD[1], Wei-Qi Wei, MD, PhD[1], Joshua C. Denny, MD, MS[1,2]

[1]Department of Biomedical Informatics, Vanderbilt University, Nashville, TN;
[2]Department of Medicine, Vanderbilt University, Nashville, TN

Abstract: The goal of this study is to evaluate the contribution of SemRep and a medication indication (MEDI) resource to the task of extracting treatment relations from clinical notes. Although in many cases these relations link medications to diseases, there exist other types of treatment relations such as procedure-disease, procedure-patient, etc. Our preliminary results show that MEDI has a positive impact to this task when combined with SemRep.

Introduction and Background: Providers often record the reasons (i.e., the indications) for therapeutic interventions in their clinical notes. Our purpose was to investigate the impact of a medication indication resource on an existing relation extraction system for discovering treatment relations in clinical text. As a medication indication resource we selected MEDI,[1] a large database of medication-indication pairs, and as an extraction system we used SemRep,[2] a publicly-available and widely-used system successfully applied to literature data sets. Our ultimate goals are to create an automatic extraction system that will improve systems like SemRep for identifying treatment relations, and to expand MEDI with new medication-indication pairs. The ability of accurately extracting treatment relations could enable a more comprehensive understanding on a patient's treatment course, improve adverse reaction detection, discover off-label drug uses, and allow public health surveillance for common diseases.

Methods: In our study, we used a set 6864 discharge summaries from the Vanderbilt Synthetic Derivative, a de-identified version of the Vanderbilt electronic medical record. First, we processed the reports with SemRep (v1.5). For each sentence, the system extracted all corresponding UMLS concepts and the treatment relations between them. Next, we analyzed the concepts identified by SemRep (regardless of whether SemRep found a relationship between them) to identify possible medication-indication pairs from MEDI that co-occurred within one sentence. To evaluate how accurately SemRep and MEDI discovered treatment relations, two reviewers annotated the sentences in which both resources identified at least one relation. The annotation process consisted of manually linking pairs of UMLS concepts that represent treatment relations, limited to the identified UMLS concepts and blinded to the algorithms' results. The reviewers performed double annotations on 75% of the data. The inter-annotator agreement reached a Cohen's kappa value of 0.86 and the disagreements were adjudicated by an experienced clinical expert.

Results and Conclusion: After data processing, 943306 UMLS concepts and 3386 treatment relations were identified by SemRep, and 1590 UMLS concept pairs were matched to medication-indication concept pairs from MEDI (Figure 1). The small overlap of relations shown in Figure 1 is mainly due to the existence of non-medication relations (e.g., procedure-disease) that MEDI is not able to capture. In our evaluation, we compared the manual annotations (393 treatment relations and 4177 non-treatment relations) against the SemRep and MEDI relations. We also considered two simple ensemble methods which combine the predictions of the two resources – the union and intersection of SemRep and MEDI. As seen in Table 1, both MEDI and SemRep performed similarly with reasonable precision, despite having only 150 (3.1%) relations commonly identified over the entire dataset (Figure 1). The F-measure obtained by MEDI was slightly better than the SemRep, despite not using any linguistic information. The best results, in terms of F-measure, were achieved by the union of MEDI or SemRep results. The improvements of this method over both SemRep and MEDI stem primarily from significant gains in recall and comparatively smaller loss in precision. Not surprisingly, the more restrictive method (i.e., MEDI and SemRep) achieved high precision at the cost of large drops in recall. Further investigation is needed for research and clinical uses of such data.

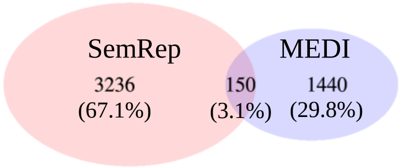

Figure 1 The connection between the relations identified by SemRep and MEDI.

Configuration	Precision	Recall	F-measure
MEDI	79.85	54.45	64.75
SemRep	76.70	54.45	63.69
MEDI and SemRep	93.66	33.84	49.72
MEDI or SemRep	72.84	75.06	73.93

Table 1 Results for treatment relation extraction.

References

1. Wei WQ, Cronin RM, Xu H, Lasko TA, Bastarache L, Denny JC. Development and evaluation of an ensemble resource linking medications to their indications. J Am Med Inform Assoc. 2013 Sep 1;20(5):954-61.
2. Rindflesch TC, Fiszman M. The interaction of domain knowledge and linguistic structure in natural language processing: interpreting hypernymic propositions in biomedical text. J Biomed Inform. 2003 Dec;36(6):462-77.