Implicit Information Extraction from Clinical Notes

Sujan Perera, Advisor: Amit Sheth
Kno.e.sis Center, Wright State University, Dayton, OH, USA
Email: sujan@knoesis.org

***Abstract*—We address the problem of extracting implicit information from the unstructured clinical notes. Here we introduce the problem of 'implicit entity recognition in clinical notes', propose a knowledge driven approach to address this problem and demonstrate the results of our initial experiments.**

I. Introduction

Information extraction from the unstructured clinical documents is a well-studied research topic among the natural language processing community. This community has produced well-known information extraction tools like MedLEE, cTAKES, and MetaMap. These tools recognize the named entity recognition and linking as a primary task in clinical notes. Clinical notes consist of entities indicated in both explicit and implicit manner. For example, while the 'shortness of breath' mentioned explicitly in the sentence *'The patient has shortness of breath with re-accumulation of fluid in extremities'*, the entity 'edema' is indicated implicitly by the phrase *'re-accumulation of fluid in extremities'*. The tools mentioned above will recognize the entity 'shortness of breath' in this sentence but not the entity 'edema'.

We introduce the implicit entity recognition problem as: given input text that does not have explicit mentions of target entities, find which target entities are implied in the input text. The implicit entity recognition is particularly challenging since besides the fact that the sentences with implicit mentions lack the entity name, they can be embedded with negations. For example, the sentence *'The patients respiration become unlabored'* conveys that patient does not have 'shortness of breath'. The negation detection is particularly important task in clinical domain since it carries important information to understand the overall health status of the patients.

This work proposes a approach to recognize the implicit mentions of entities in the clinical notes by leveraging the domain knowledge present in Unified Medical Language System (UMLS). We have published the initial work on this topic [1].

II. Implicit Entity Recognition

Our proposed approach starts with finding the entity representative terms (ERT) for each entity from their definitions in the UMLS. The idea behind the ERT selection is to find the terms that may indicate the presence of an implicit entity mention. For example, implicit mentions of 'shortness of breath' are more likely to use term *'breathing'* or its synonyms, and implicit mentions of 'appendicitis' more likely to use term *'appendix'*. Hence, *'breathing'* and *'appendix'* selected as ERTs for 'shortness of breath' and 'appendicitis' respectively. The sentences that have ERTs of the entities but not their proper names become candidates to contain implicit entity mentions.

The ERT for each entity is selected by a measure that captures the dominance and the discrimination power of a term in the entity definitions. The term with highest dominance and discrimination power is selected as the ERT for the entity. Although ERT helps to find the candidate sentences, it is not sufficient to determine whether the sentence has an implicit entity mention. Hence the next step of the algorithm creates entity model for each entity. The entity model consists of bag of words that describe the characteristics of the entity. For example, an entity model for 'appendicitis' would be *'{acute, inflammation, appendix}'* and it is created by capturing the terms in the neighborhood of the ERT in the definition of the entity.

The next step of the algorithm is to determine which of the selected candidate sentences contain implicit entity mentions. This is performed by calculating the similarity between the candidate sentence and the entity model. We use WordNet as the linguistic knowledge base to determine the similarity between the entity model and the candidate sentence. This similarity calculation determines whether the sentence has similar meaning as the entity model or the opposite meaning, in which case it is recognized as a negated mention of the entity.

III. Dataset and Evaluation

The initial evaluation for this proposed approach is conducted using the dataset used by the SemEval-2014 task 7. We have re-annotated this dataset for implicit mentions of eight selected entities to create the ground truth for our experiments. The ground truth consists of 857 implicit entity mentions.

The experimental results showed that our algorithm is capable of identifying the implicit mentions of the entities and showed superior performance in identifying the negated mentions of the entities when compared with other applicable methods. Our approach showed F1-measure of 0.75 and 0.73 for positive and negative mentions of the entities respectively. Furthermore, the similarity value calculated by our algorithm proved to be highly informative feature for supervised approaches that designed to solve the implicit entity recognition problem. The SVM trained to solve this problem showed 0.77 and 0.67 F1 values for recognizing positive and negative mentions while it showed 0.81 and 0.73 F1 values when the similarity value calculated by our algorithm is incorporated as a feature.

References

[1] S. Perera, P. Mendes, A. Sheth, K. Thirunarayan, A. Alex, C. Heid, and G. Mott, "Implicit entity recognition in clinical documents," in *Proceedings of the Fourth Joint Conference on Lexical and Computational Semantics (*SEM)*, 2015, pp. 228–238.