MEDTYPE: Improving Medical Entity Linking with Semantic Type Prediction

Shikhar Vashishth, PhD[1], Rishabh Joshi, BS[1], Denis Newman-Griffis, PhD[2],
Ritam Dutt, MS[1], Carolyn Rose, PhD[1]
[1]Carnegie Mellon University, Pittsburgh, PA; [2]University of Pittsburgh, Pittsburgh, PA

Introduction

Identifying the standardized concepts referred to in an unstructured text is a critical component of biomedical natural language processing, enabling harmonization across different documents for search and semantic analysis. Medical entity linking, also referred to as medical concept normalization, is the task of harmonizing different surface forms for the same concepts: for example, identifying that documents mentioning *Amyotrophic lateral sclerosis* and *Lou Gehrig's Disease* are referring to the same disease. A key step in this process is candidate generation, the identification of candidate medical concepts a text could be referring to. This process is prone to overgeneration, producing too many candidates that make normalization difficult and lead to erroneous predictions. In this work, we propose an intermediate step between candidate generation and final normalization decisions that alleviates the overgeneration. For this, we present MEDTYPE, a fully modular system that prunes out irrelevant candidate concepts based on the predicted semantic type of an entity mention. MEDTYPE utilizes a Transformer-based encoder for modeling the context of a given mention. To address the dearth of annotated training data for medical entity linking, we also present two novel large-scale datasets: WIKIMED, biomedical subset of Wikipedia corpus for entity linking, and PUBMEDDS, a distantly-supervised dataset of medical entity mentions, which help improve the performance on the task.

Method

We experimented with incorporating MedType into standard medical entity linking tools and evaluated its impact on entity linking performance in four datasets: NCBI, Bio CDR, ShARe, MedMentions for medical entity linking, as well as our novel WIKIMED dataset. The chosen datasets span across different domains such as biomedical research articles, Electronic Health Records (EHR), and general domain text. Thus, they allow us to evaluate the generality of MEDTYPE across diverse domains. MEDTYPE is evaluated for pruning candidate concepts generated from five entity linking models: MetaMap, cTAKES, MetaMapLite, QuickUMLS, and ScispaCy. We run four sets of experiments for demonstrating the effectiveness of our approach. Experiment 1 evaluates how incorporating MEDTYPE in existing entity linking systems helps improve medical entity linking. Experiment 2 demonstrates that the proposed datasets WIKIMED and PUBMEDDS are effective for semantic type prediction. Experiment 3 analyzes the gains obtained using the proposed datasets. Experiment 4 quantifies how type-based filtering helps prune irrelevant candidates.

Results

Experiment 1. The results show that the type-based filtering of candidates using MEDTYPE enhances entity linking systems across all 25 settings evaluated. MEDTYPE obtains statistically significant improvement of up to 4.2% F1-score over default entity linking without using semantic type prediction. Experiment 2 shows a substantial gain in performance on using the proposed datasets. Overall, we get an average absolute increase of 9.8, 12.7, and 13.4 AUC on type prediction from WIKIMED, PUBMEDDS, and the combined corpora respectively. Experiment 3. The results show that for semantic types such as *Sign or Symptoms*, utilizing WIKIMED and PUBMEDDS gives an absolute increase in F1-score of 24, 28 on Bio CDR and 12, 14 on ShARe respectively. Experiment 4. The type-based filtering removes all incorrect candidates in 36.5% cases while in 25.6% cases it helps to reduce the candidate set size. This clearly demonstrates the practical benefits of our proposed approach.

Conclusion

We found that filtering out irrelevant candidate concepts based on the predicted semantic type improves entity linking performance for a variety of popular medical entity extraction toolkits across several benchmark datasets. We further present two novel large-scale datasets: WIKIMED and PUBMEDDS. Pre-training on these datasets substantively improves MEDTYPE performance, and we share these datasets with the community as a resource for medical entity linking research at http://github.com/svjan5/medtype.