Matching Consumer Health Vocabulary with Professional Medical Terms Through Concept Embedding

Yue Wang, MS[1], Jian Tang, PhD[2], V.G.Vinod Vydiswaran, PhD[3,2], Kai Zheng, PhD[4], Hua Xu, PhD[5], Qiaozhu Mei, PhD[2,1]
[1]Department of Electrical Engineering and Computer Science, University of Michigan, Ann Arbor, MI; [2]School of Information, University of Michigan, Ann Arbor, MI; [3]Department of Learning Health Sciences, University of Michigan, Ann Arbor, MI; [4]Department of Informatics, University of California, Irvine, CA; [5]School of Biomedical Informatics, The University of Texas Health Science Center at Houston, Houston, TX

Introduction: Matching consumer health vocabulary with professional medical terms is an important task for medical information retrieval, knowledge sharing, and effective health communication. Previous approaches studied mining consumer health vocabularies (CHV) with hand-crafted heuristics on medical search logs[1] and community-generated content[2]. Though the professional-consumer concept pairs mined through these approaches are very accurate, many relevant pairs could be missing. In other words, the recall is low. In this pilot study, we propose to match professional-consumer concept pairs through text embedding approaches. Text embedding approaches have proven to be very effective in capturing the similarity between words and phrases[3], which can yield high recall of professional-consumer concept pairs. We first learn the representation of the medical concepts with a large amount of unlabeled text. Afterwards, the professional-consumer concept pairs are matched according to the similarities of their embeddings.

Methods: We first learn word embeddings from large-scale text corpus and then represent a medical concept as the average of word vectors it contains. Three corpora are used to train word embeddings: 2.2 million MedHelp posts (consumer-generated text), 7 million Medline abstracts (professional-generated text), and 4.7 million Wikipedia articles (general-domain text). We use the Skip-gram model[3] to learn word embeddings and calculate the similarities between concepts as the cosine similarity of their word embeddings. The results are evaluated using a ranking-based metric, the mean reciprocal rank (MRR). Given a professional (consumer) concept in a CHV pair, we rank candidate medical concepts by cosine similarity to the query. Ideally, the corresponding concept in the pair is ranked high.

We use a large collection of candidate medical concepts and ground truth professional-consumer concept pairs. We collect all English medical concepts in the UMLS as the concept collection and use two existing CHVs as the ground truth pairs: CHV-1 from Vydiswaran et al.[2] and CHV-2 from Zeng et al.[1] To ensure fair comparison between word embeddings learned on different corpus, we filter out concepts that contain out-of-vocabulary words for any corpus. This gives us 494K medical concepts from UMLS, 892 concept pairs from CHV-1 and 124K concept pairs from CHV-2. We randomly select 1000 pairs from CHV-2 to make the sizes of CHV-1 and CHV-2 comparable for evaluation.

Results and Discussion: Our initial results show that it is feasible to identify alternative medical concepts by using professional or consumer concepts as queries in the concept embedding space. The MRR values indicate that on average, a professional or consumer concept is about 14th closest to its counterpart in CHV-1 and about 7th closest in CHV-2. Note that this is only the result on a small subset of the large medical concept collection where we have ground truth. We will generate an extended list of related concept pairs and pseudo-label them as professional- or consumer-oriented using the propensity measure[2]. This list can then be checked by medical experts and expand current CHVs. This will open up the opportunity to build more extensive CHVs with minimal effort from practitioners.

To make the concept embedding *specialize* in professional-consumer vocabulary translation and continuously improve itself, we are developing a semi-supervised representation learning algorithm. The algorithm uses a small set of known CHV pairs as the supervision signal, and learns word embeddings (and the way to compose word embeddings into a concept embedding) such that consumer concepts and professional counterparts are close to each other. With more and more supervision pairs, the recommended alternative concepts will become more precise.

References

1. Zeng QT, Tse T. Exploring and developing consumer health vocabularies. Journal of the American Medical Informatics Association. 2006;13(1):24–29.

2. Vydiswaran VV, Mei Q, Hanauer DA, Zheng K. Mining consumer health vocabulary from community-generated text. In: AMIA. vol. 2014. American Medical Informatics Association; 2014. p. 1150.

3. Mikolov T, Sutskever I, Chen K, Corrado GS, Dean J. Distributed representations of words and phrases and their compositionality. In: NIPS; 2013. p. 3111–3119.