MEDINFO 2017: Precision Healthcare through Informatics
A.V. Gundlapalli et al. (Eds.)

doi:10.3233/978-1-61499-830-3-1260

Concept Embedding for Relevance Detection of Search Queries Regarding CHOP

Yihan Deng[a], Lukas Faulstich[b], Kerstin Denecke[a]

[a] *Bern University of Applied Science, Bern, Switzerland,*
[b] *ID Information and Documentation GmbH, Berlin, Germany*

Abstract

Automatic encoding of diagnosis and procedures can increase the interoperability and efficacy of the clinical cooperation. The concept, rule-based and machine learning classification methods for automatic code generation can easily reach their limit due to the handcrafted rules and a limited coverage of the vocabulary in a concept library. As the first step to apply deep learning methods in automatic encoding in the clinical domain, a suitable semantic representation should be generated. In this work, we will focus on the embedding mechanism and dimensional reduction method for text representation, which mitigate the sparseness of the data input in the clinical domain. Different methods such as word embedding and random projection will be evaluated based on logs of query-document matching.

Keywords:

Automatic Encoding, Classification, Machine Learning

Introduction

In order to claim costs to the health insurance and for clinical documentation purposes, it is necessary and even legally required to encode diagnoses and procedures by classification codes from relevant classification systems. In Switzerland, these are ICD-10-GM for diagnoses and Schweizerische Operationsklassifikation (CHOP) for medical procedures. In order to facilitate the subsequent query matching process based on deep learning, we investigate the possibilities of direct embedding of concept through word2vec principle (Skip-gram, continuous bag of words, negative sampling) as well as the dimension reduction with the input of sparse concept vector. We would like to figure out whether the embedding method can be applied on the concept vector directly and which type of embedding (Skip-gram and CBOW) is more suitable for the concept embedding from domain specific data.

Methods

For test and development, we are using semantic representations in vector form generated from search entries and, on the other hand, target catalog texts that have been generated from the CHOP classification texts. As is illustrated in Figure 1, for each query – classification code text pair it has been assigned whether the classification text matched the user query or not. For generating the vectors, search entries and classification texts have been mapped to concepts of a medical terminology by the terminology server ID MACS®. As pre-processing for the query matching system, the input layer is in charge of the representation generation and dimension reduction. This layer is trying to represent the word in a corpus by a special instantiation of a set of hidden variables. The embedding process learns the representation of each word by maximizing the log likelihood of each word given its context (context window). Our evaluation platform will test the embedding result based on CBOW, Skip-gram, and random projection. The embedding is implemented with Tensorflow [1].

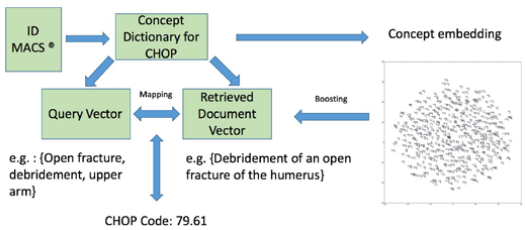

Figure 1- Concept embedding for a better automatic CHOP encoding.

Results

20067 logs are used to do the embedding and evaluation. Based on the annotated relevance metrics [1], the average match rate of Skip-gram embedding based on concept only vector representation has achieved the best relevance match (0.63), while the CBOW has achieved clearly less match rate (0.43). The random project has only reached the least match rate around 10%.

Discussion and Conclusion

Concept embedding can largely reduce the sparseness to make a suitable input for the deep neural network. The skip-gram is most suitable method for encoding short text queries referring to clinical or surgical procedure, since keywords are short and relatively independent. The portability and scalability of the method will be tested in other medical natural language text.

Acknowledgements

This work is supported by ID Information and Documentation GmbH, Germany

References

[1] Schnabel T, Labutov I, Mimno D, Joachims T. Evaluation methods for unsupervised word embeddings 2015, EMNLP, 298-307.

Correspondence:
Yihan Deng
yihan.deng@bfh.ch

[1] https://www.tensorflow.org