The Impact of learning Unified Medical Language System Knowledge Embeddings in Relation Extraction from Biomedical Texts

Anuja Tayal
657111812
UIC,Chicago
atayal4@uic.edu

REVIEW

Objective

The objective of the paper[1] is to explore how knowledge embeddings learned from the UMLS Metathesaurus impact the quality of relation extraction on two diverse sets of biomedical texts. To accomplish this, they designed a system- Relation Extraction with Knowledge Embeddings(REKE) to incorporate the knowledge embeddings and quantify the impact of knowledge embeddings on the quality of relation extraction results. Authors learn 2 type of embeddings mainly Lexicalised Knowledge Embeddings (LKEs) and Unlexicalised Knowledge Embeddings (UnLKEs). In short, authors learn the knowledge embeddings from the UMLS, incorporate the embeddings learned in BlueBert and develop the REKE system to finally evaluate the impact of learned Knowledge Embeddings.

As biomedical concepts can be mentioned in various ways in biomedical texts, so the authors made use of lexical knowledge to account for different type of concept and relation type mentions by introducing Lexical Knowledge Embeddings (LKE). It can be addressed in unLKE by performing entity linking from a concept mention encoded in UMLS and retrieving knowledge embedding for the linked concept. Secondly, new relation types between biomedical concepts could be of interest which are not yet encoded in UMLS. To mitigate this, the authors designed Relation Extraction using Knowledge Embedding(REKE) system which incorporates the embeddings in the BlueBERT system by scoring the plausibility of the relation types of interest between pair of concept mentions.

Dataset

To evaluate their model, authors used 2 relation extraction dataset mainly 2010 i2b2/VA dataset and 2013 Drug-Drug Interaction(DDI) Extraction Challenge dataset. In i2b2/VA dataset, medical discharge summaries and progress notes were annotated for 3 medical concepts (problems, tests and treatments) and 8 relation types between them. Whereas, Drug-Drug Interaction (DDI) corpus consisted of 4 types of pharmacological substances (Drug, brand, group) and 4 types of relations(Mechanism, Effect, Advise and Int) between them. DDI corpus consisted of two types of datasets mainly DrugBank dataset (with 792 texts and 15756 substances) and MedLine Abstracts (with 233 abstracts and 2746 substances).

ISBN 978-1-4503-2138-9.
DOI: 10.1145/1235

REKE System

REKE system extends NCBI BlueBERT as it does not incorporates UMLS knowledge for which authors implemented Knowledge Embedding Encoder (KEE).KEE maps the biomedical concept mentioned in the sentence and the set of relation types known to their respective UMLS embedding which is then used to compute a vector of relation plausibility scores $s_{A,B}$. The authors in their previous work had shown that **TransD** scoring function gives the best results. Contextual embedding h is calculated using NCBI Bluebert model for each sentence of a biomedical text which is concatenated to the plausibility scores to calculate its softmax scores.

Knowledge Embedding Encoder (KEE)

Lexicalised and unLexicalised Knowledge Embeddings were learned through Knowledge Embedding Encoder using Generative Adversarial Network (GAN). In KEE, the discriminator tries to learn the relation plausibility scores between two UMLS concepts while generator tries to fool the discriminator by generating incorrect relations by computing the score with wrong concept. In KEE, the word piece tokenisation of the atom of UMLS concept or relation type is passed through BERT language model to capture lexicalised context and then through a span encoder which is a biLSTM or bidirectional long short term memory. A similar procedure happens for relation types and unlexicalised concept mentions.

Evaluation

To evaluate the REKE system, standard metrics of micro precision, recall and F1 scores were used. REKE system used the same hyperparameters as BlueBERT and surpassed it in all metrics. It was interesting to see that performance metrics for LKE model were more as compared to UnLKE.

Pros and Cons

I really liked that the authors had performed ablation study to support their model and exploited UMlS Metathesaurus to

incorporate concept mentions and relation types in BlueBERT and would also work for new relation types found in the future. But, the performance metric increased was just slightly which could be improved. Also, the authors could have provided more specific details about the model or an example as I had to go through the supplementary details to get through the nitty gritty of the paper. Secondly, the authors used TransD scoring function without any explanation which was a bit surprising. Lastly, some specifics about the GAN model used for knowledge embedding encoder were not clear to understand.

Conclusion

In conclusion, authors successfully learned the impact of adding UMLS embedding in relation extraction system.

REFERENCES

[1] Maxwell A Weinzierl, Ramon Maldonado, and Sanda M Harabagiu. 2020. The impact of learning Unified Medical Language System knowledge embeddings in relation extraction from biomedical texts. *J. Am. Med. Inform. Assoc.* 27, 10 (Oct. 2020), 1556–1567.