Overview – Relation Dataset Without Human Labelling

Unilever uses a named entity recognition and relation extraction stack to build knowledge graphs and find relationships between concepts in text data. However, Unilever is tuning such models (e.g. OpenNRE) by having subject matter experts manually assign relationships to entities in candidate sentences, which is expensive.

This project aims to identify data sources and a less costly fine-tuning methodology to tune an open-source relation extraction model, OpenNRE, to chemistry and food science. The goal is to end up with a richer relationship set while still not requiring significant human labor.

Methodology and Technique

To generate data pairs (entityA, entityB, relation), one needs to first extract the correct entities from the sentences with meaningful relationships. To minimize false positives, we limited the results to be nouns, adjacent, and not common words. After experimenting with summarizer, paraphraser, and GPT-3, we applied a combined approach to summarize the text between entities to a couple of words, viewed as their relations. We further clustered the relations into centroids, which were used as the final labels.

Identified Entities and Clustered Relations

Through Entity Recognition, Relation Summarization and Hierarchical Clustering on Word2vec, a relation dataset that maps entity pairs with their relations is formed. The whole pipeline utilizes multiple NLP models so little human labor is needed.

Figure 3 illustrates how our pipeline converts raw article inputs to a relation dataset without the assistance of human. The graph on the top is an example of entity and relation extractions, and the dendrogram below shows the process of clustering the extracted relations into representative centroid terms. These centroid terms are then used to mark the corresponding entity pairs.

Conclusion

By utilizing models in multiple NLP tasks, such as entity recognition, POS tagging and autoregressive writing, we successfully built an automated pipeline that can take raw texts as input, and build entity-relation dataset from them. This dataset involves no human labelling, and representative relations come from hierarchical clustering.

Acknowledgments

Great thanks to Mr. John Labarga from Unilever and Professor Sining Chen for finding resources and providing valuable advice to this project.

References

Hierarchical Clustering: https://en.wikipedia.org/wiki/Hierarchical_clustering
OpenAI’s GPT-3: https://beta.openai.com/docs/quickstart