Fine-Tuned Relationship Extraction for Consumer Goods Concepts

Data Science Institute
COLUMBIA UNIVERSITY

Ruilin Liu, Zhifeng Zhang, Zhiqing Yang,
Jessie Wang, Zhucheng Zhan

Industry Mentors: John Labarga
Faculty Mentors: Sining Chen

Data Science Capstone Project
with Unilever

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.

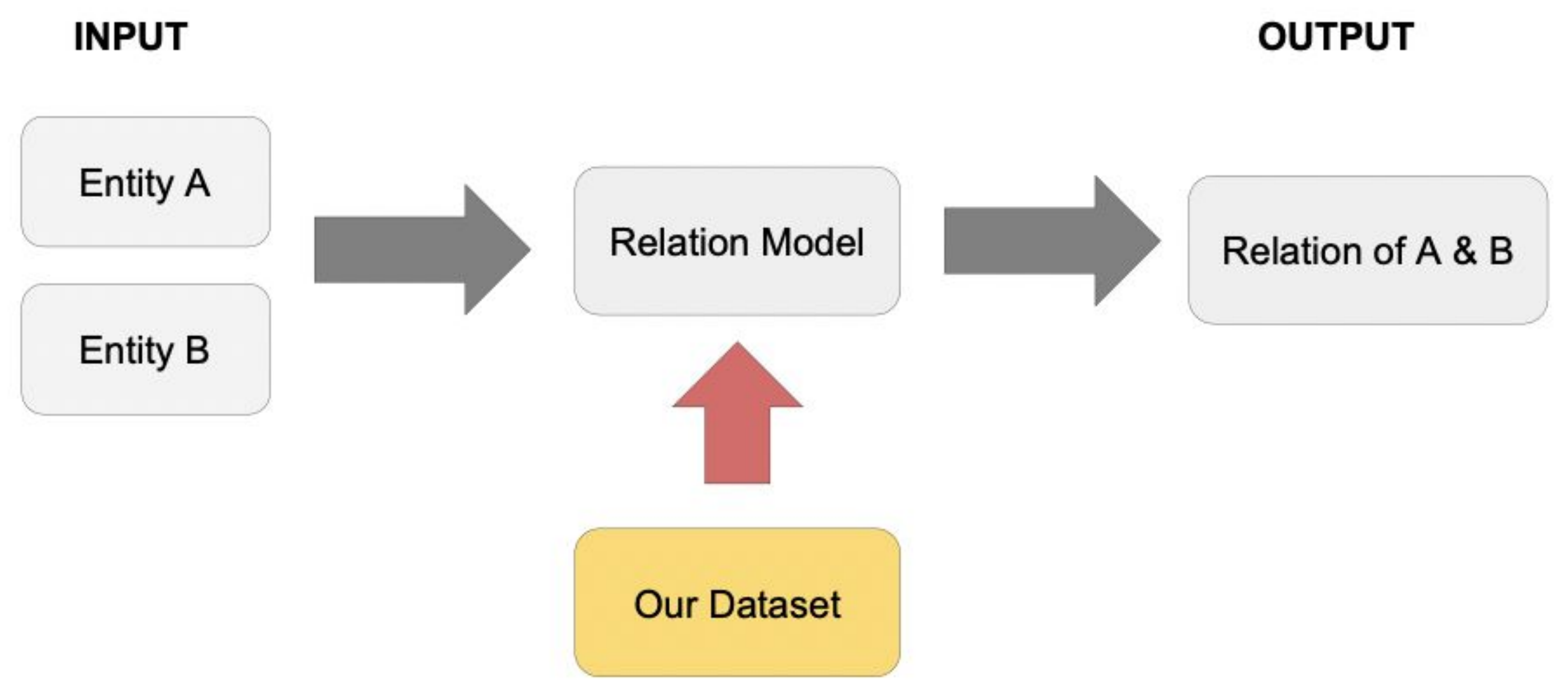
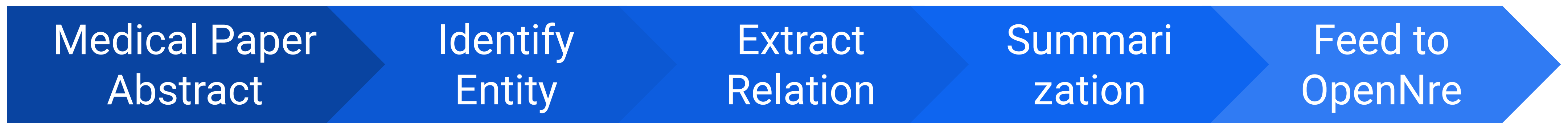


Figure 1. Input and output of relation model.

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 word, viewed as their relations. We further clustered the relations into centroids, which were used as the final labels



Medical Paper Abstract

- To identify data sources for fine tuning relationship models
- Focus on chemistry and food science

Identify Entity

- Started from NER libraries
- Refined by dictionary, POS, syntactic rules
- Kept most confident pairs

Extract Relation

- Summarization
- Paraphrasing
- Verb

Summarization

- Clustered similar relationship
- Generated labels

Feed to OpenNre

- Feed entities and relations to OpenNre

Figure 2. Project Workflow, Methodology, and Techniques

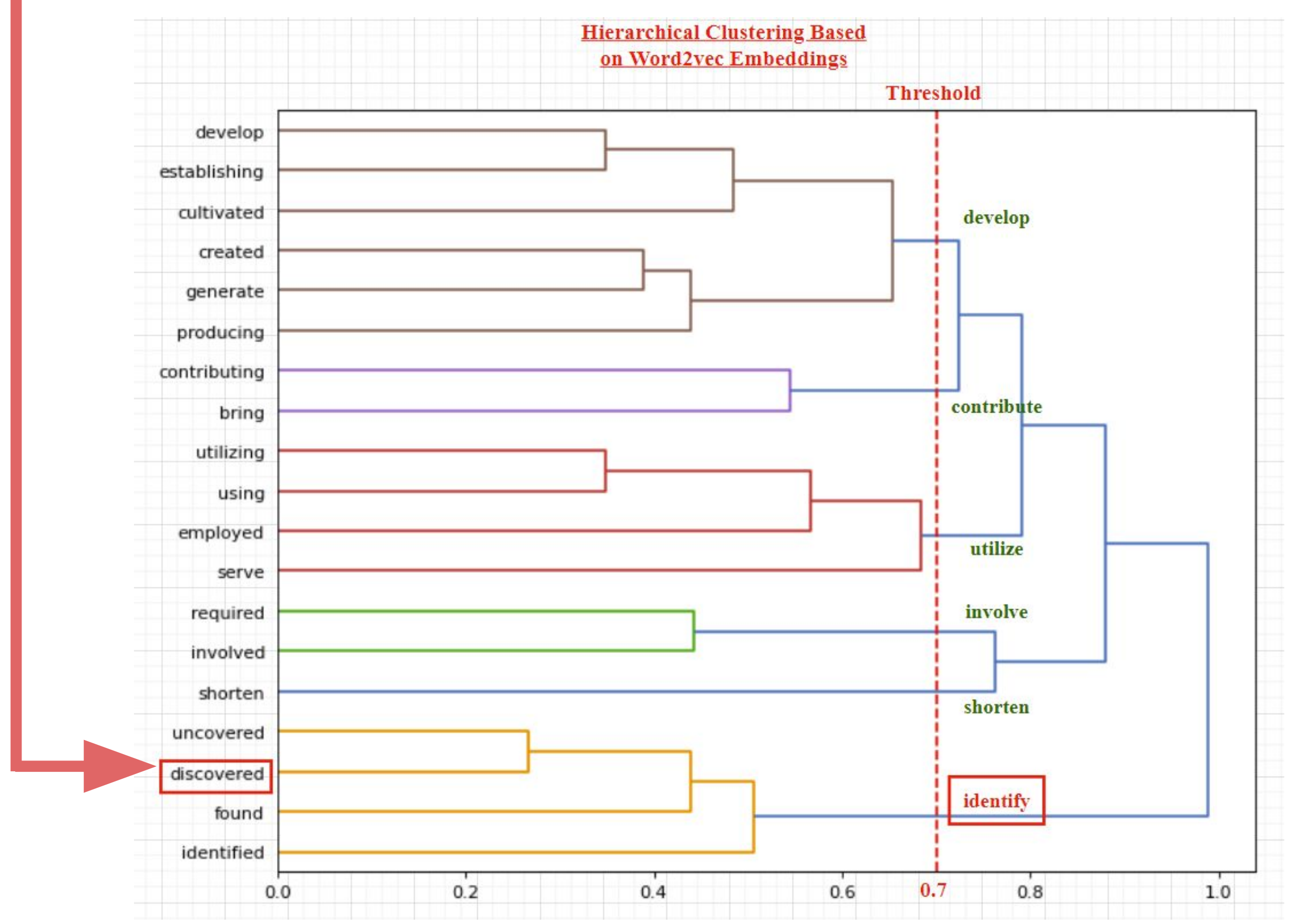
Identified Entities and Clustered Relations

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.

Here, we summarise the origins and types **Entity 1** of 375 compounds that have been discovered **Relation** from the symbiosis **Entity 2** over the past four decades and discuss the potential for synergistic actions between compounds within the complex chemical mixtures in which they exist.

The principle of treatment for advanced stage CRC is a multidisciplinary and comprehensive treatment based on chemotherapy **Entity1** , which always bring **Relation** significant toxic side effects **Entity2** .



Resulted Dataset

Entity 1	Entity 2	Relation
Origins and types	Symbiosis	identify
Chemotherapy	Toxic side effects	contribute
…	…	…

Figure 3. Process of automatically producing relation dataset from raw text inputs.

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 advices to this project.

References

OpenNRE: https://opennre-docs.readthedocs.io/en/latest/get_started/introduction.html

Hierarchical Clustering: https://en.wikipedia.org/wiki/Hierarchical_clustering

OpenAI’s GPT-3: https://beta.openai.com/docs/quickstart