Leveraging Language Models for Company and Product Relation Extraction from News Articles

Abstract
Mapping the technological landscape entails a comprehensive analysis of important actors in technological development and deployment, such as companies and products. News articles serve as a rich repository of insights on these entities. In this paper, we propose two approaches for extracting company and product relations from news articles: an approach based on large language models (LLMs) and a pre-trained language model (PLM)-based approach. The LLM-based method utilizes GPT-3 in a few-shot setting, while the PLM-based approach incorporates named entity recognition and natural language inference models. We evaluate the performance of these approaches using a manually annotated dataset of more than 200 entities and 250 relations. We find that the PLM-based approach achieved the best results, while GPT-3 was better at detecting long-range implicit relationships.

Background
The fields of technology mining and forecasting play a crucial role in identifying emerging trends, understanding the dynamics of innovation, and making informed decisions in various industries. A valuable factor for these fields is the ability to analyze inter-organizational relations between companies and their products from a multitude of information sources, including news articles. These are a source of unstructured information about recent events between these types of entities such as acquisitions, investments, partnerships, etc., and are produced at a pace that is becoming increasingly hard to keep up with. A method of extracting structured data about companies from news articles would provide not only a clearer representation of past, current, and future relations but also give rise to data structures that allow for more meaningful and complex analysis, such as Social Network Analysis (Doehring, 2008). Such graph representations could also be used for grounding Language Models' generations (Pan et al. 2023).

Transformer models (Vaswani et al., 2017) have revolutionized Natural Language Processing in the last few years. Pre-trained Language Models (PLMs) achieved state-of-the-art results in all types of NLP tasks such as Machine Translation, Question Answering, and Sentiment Analysis (Devlin et al., 2018). A more recent application of these models is for the task of Information Extraction (Ding et al., 2021), which consists of extracting structured information, usually relations between entities, from unstructured data. The latest advancements in Generative Large Language Models such as GPT-3 (Brown et al., 2020) allow new possibilities for interacting with language models. However, factual hallucinations (McKenna et al., 2023) still plague textual generations when prompted freely. One way to infuse factual information in LLMs is through the use of Knowledge Graphs. Pan et al. (2023) described several ways KGs can be used to enhance LLMs and vice-versa, leading to more factually grounded and interpretable generations.

There have been several attempts at extracting company relations from text. Doehring (2008) planned an approach to extract company information from news articles, with the aim of applying Social Network Analysis, although the types of relations were not defined. Yamamoto et al. (2017) extracted company binary relations (collaborative/competitive) from news articles in the semiconductor industry. Khaldi et al. (2021) implemented a BERT-based business relation extraction system from individual sentences, specific to 6 types of company-company relations.

With our work, we aim to devise methods that allow the creation of Knowledge Graphs of entire news articles, by extracting company-company, company-product, and product-product relations from them.

Methods
We restrict the set of possible relations based on the entity types. Company-company relations can be Partners, Competitors, Owners or Investors. For company-product and product-product relations, only one relation type was considered for each, respectively Developer and Competitor.

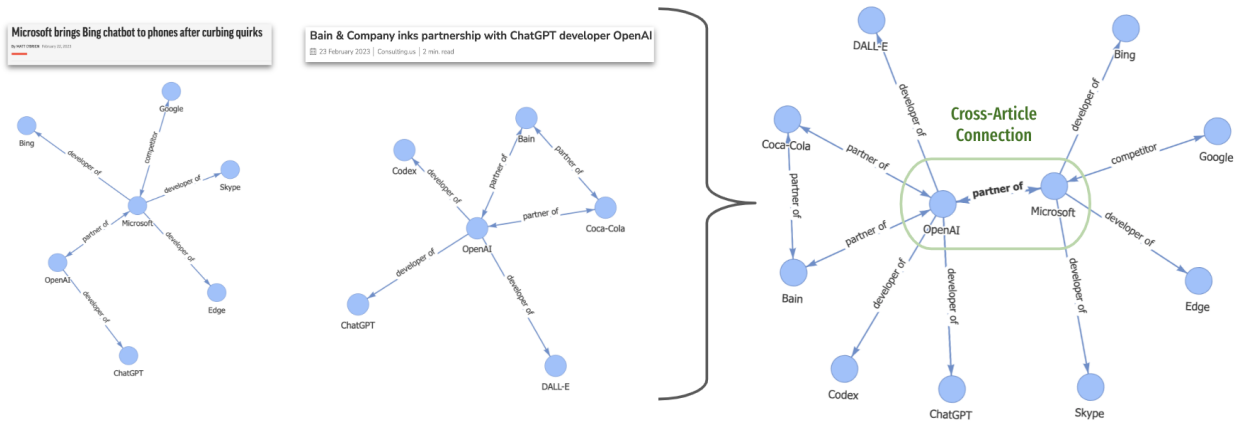
We explore two different approaches for this task. The first is an LLM-based method, that relies on prompting the GPT-3 model in a few-shot setting to extract entities and relations from each article. For each relationship, the model was asked to provide the corresponding passage from the article that justified it. Some postprocessing is applied to the output to filter out obvious hallucinations. The second approach is PLM-based. First, a Named Entity Recognition (NER) model is used to extract relevant entities from the article. Then, a Natural Language Inference (NLI) model is used to assign scores to every possible relationship between these entities, given the context of the article. The top-scoring relations are kept.

To evaluate the performance of these approaches, 54 articles from reputable news websites were manually annotated, creating a set of more than 200 entities and 250 relations and passages. To our knowledge, this is the only available dataset that maps news articles to company and product relations.

Findings
When compared to the manual dataset, the PLM-based approach achieves better performance in relation extraction than the LLM-based approach (more than 10% absolute increase in F1-score). In the task of entity extraction, both approaches perform equally well. In general, PLMs were better at extracting short-range relations (e.g. "Facebook parent Meta" → (*Meta, owner, Facebook*)), while GPT-3 performed better at long-range relations due to its bigger context window and better natural language understanding.

We also observed that some articles were mostly focused on one company and its products and partners. Finding the same relations or entities in different articles allows for the merging of the article-wise graphs into bigger graphs that contain cross-article information (Figure 1).

Figure 1: Output of PLM approach from two articles, and merged relationship graph.



The PLM approach was found to provide better overall results and faster inference speeds (<1s/article). This approach is also modular, as the models can be interchanged with any NER or NLI models, without requiring any fine-tuning. The LLM approach, while not achieving the highest results and being considerably slower (20s/article), allows more task flexibility by way of changing the prompt, and is able to provide a passage for every relation, which can be used to "fact-check" the relation.

Conclusion

Our findings highlight the power of leveraging the natural language understanding capabilities of different types of Language Models in the task of extracting structured data from news articles, without any fine-tuning or model training required. Our contributions can be summarized as follows: (i) created a LLM prompting strategy for extracting company and product relations from news articles, along with relevant passages for each relation, (ii) implemented a PLM-based pipeline for company and product entity and relation extraction from news articles that is fast, modular and doesn't require fine-tuning, and (iii) created a dataset of over 50 annotated news articles for company and product relation extraction. Several venues for future research could be explored, such as the linking of the product entities to their respective technologies, which would allow even richer Knowledge Graphs. It is also essential to emphasize the importance of creating larger and more diverse datasets for the evaluation of language models in this specific task.

References

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. Advances in neural information processing systems, 30.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

Ding, K., Liu, S., Zhang, Y., Zhang, H., Zhang, X., Wu, T., & Zhou, X. (2021). A Knowledge-Enriched and Span-Based Network for Joint Entity and Relation Extraction. Computers, Materials & Continua, 68(1).

Doehring, M. (2008, March). Extraconn: Extraction and analysis of company networks from news. In Proceedings of the 1st Internet of Services Doctoral Symposium 2008 at International Conference on Interoperability of Enterprise Systems and Applications (I-ESA'08), Berlin, Germany.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. Advances in neural information processing systems, 33, 1877-1901.

McKenna, N., Li, T., Cheng, L., Hosseini, M. J., Johnson, M., & Steedman, M. (2023). Sources of Hallucination by Large Language Models on Inference Tasks. arXiv preprint arXiv:2305.14552.

Pan, S., Luo, L., Wang, Y., Chen, C., Wang, J., & Wu, X. (2023). Unifying Large Language Models and Knowledge Graphs: A Roadmap. arXiv preprint arXiv:2306.08302.

Yamamoto, A., Miyamura, Y., Nakata, K., & Okamoto, M. (2017, January). Company relation extraction from web news articles for analyzing industry structure. In 2017 IEEE 11th International Conference on Semantic Computing (ICSC) (pp. 89-92). IEEE.

Khaldi, H., Benamara, F., Abdaoui, A., Aussenac-Gilles, N., & Kang, E. (2021, June). Multilevel entity-informed business relation extraction. In Natural Language Processing and Information Systems: 26th International Conference on Applications of Natural Language to Information Systems, NLDB 2021, Saarbrücken, Germany, June 23–25, 2021, Proceedings (pp. 105-118). Cham: Springer International Publishing.