Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra

Federal Department of Defense Population Protection and Sport DDPS

armasuisse
Science and Technology S+T

Language Models for Company and Product Relation Extraction from News Articles

by André Charneca

Under the supervision of:
Dr. Ljiljana Dolamic

In collaboration with:
Angelika Romanou

CYD CYBER DEFENCE CAMPUS

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 into these entities. In this project, two approaches for extracting company and product relations from news articles were explored: one approach based on Large Language Models (LLMs) and another based on Pre-trained Language Models (PLM). The LLM-based method utilizes GPT-3 in a few-shot prompt 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 is able to detect longer-range implicit relationships.

Contents

Chapter 1

Background

1.1 Introduction

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 and "nowcast" 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 [4]. Such graph representations could also be used for grounding Language Models' generations [7].

1.2 Transformer-based Language Models

Transformer models [8] 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 [2]. A more recent application of these models is for the task of Information Extraction [3], 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 [1] allow new possibilities for interacting with language models. However, factual hallucinations still plague textual generations when prompted freely [6]. One way to infuse factual information in LLMs is through the use of Knowledge Graphs.

Pan et al. [7] described several ways KGs can be used to enhance LLMs and vice-versa, leading to more factually grounded and interpretable generations.

1.3 Related Work

There have been several attempts at extracting company relations from text. Extraconn [4] is a planned 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.[10] extracted binary business relations (collaborative/competitive) from news articles in the semi-conductor industry. A BERT-based business relation extraction system was implemented in [5], working at the sentence level, and specific to 6 types of company-company relations: Investment, Sale-Purchase, Competition, Cooperation, Legal Proceedings. However, none of these approaches considered relations involving products.

1.4 Problem Statement

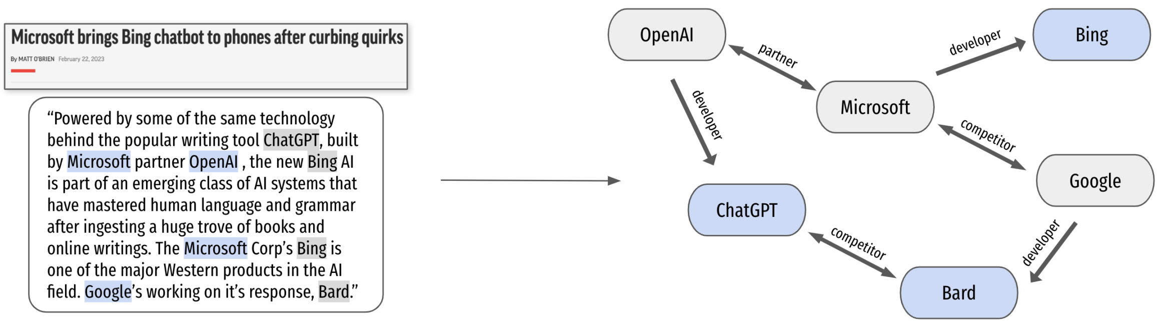

Figure 1.1: Illustration of the entity and relation extraction task from a news article passage. The passage has been slightly adapted.

Given a news article, we want a system that allows the creation of a Knowledge Graph of the entire article, by extracting company-company, company-product, and product-product relations. We consider three different types of relationships:

- Company-Company: [Partner, Competitor, Owner, Investor]
- Company-Product: [Developer]
- Product-Product: [Competitor]

The types [Partner, Competitor] are bidirectional, and the types [Owner, Investor, Developer] are unidirectional. As we don't have *a priori* access to the entities (companies and products) mentioned in each article, we must also devise a system to extract such entities from the text.

A Knowledge Graph including company and product relations could be included in LLM-based systems for Natural Language interactions with a tech expert user. For example, in Figure 1.1, the user could ask the question "Who is the competitor of OpenAI?", and a KG-enhanced LLM could reason: OpenAI → developer → ChatGPT ↔ competitor ↔ Bard ← developer ← Google. The output response would be "Google is a competitor of OpenAI, given that they develop competitor products." This graph representation improves factual grounding and allows for cross-article information to be infused in the model, without having to concatenate entire articles together.

Chapter 2

Methods

In this chapter, we will describe the two approaches considered for the task of business relation extraction between companies and products from news articles.

2.1 Prompting Large Language Models

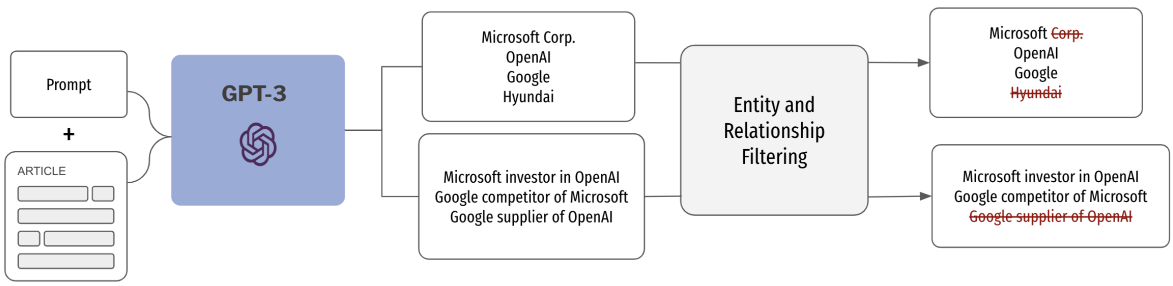

Figure 2.1: GPT-3 prompting and filtering pipeline.

The first approach consisted of few-shot prompting the GPT-3 model, in order to extract the mentioned entities and the existing relationships between them. The model outputs an entity list and a relationship table. For each relationship found, we also ask the model to provide the relevant text passage that confirms it. The full prompt can be found in Appendix A.

Entity Filtering Some filtering is applied to the entities extracted by GPT-3. Entities that are not mentioned in the article text or title are removed, as they are likely to be hallucinations. The company base name (that is, without legal suffixes) is kept, using the Python package `cleanco`[1].

Relation Filtering Relations whose type is not in the predefined relation list are removed. If one or more entities of a relation can't be found in the text or title, the relation is removed.

GPT-3 Model Parameters The model used in this project was `text-davinci-003`, which is a Generative model, trained for prompt continuation. The model behind ChatGPT (`gpt-3.5-turbo-0301`) was also experimented with, but was discarded, since it was much harder to make the model adhere to the strict format and relation types specified. Further exploration of an approach with this Conversational model is, however, encouraged, since the token cost is ten times cheaper than the used model. The parameters used when querying the model are presented in Table 2.1.

Table 2.1: Parameters for the GPT-3 model

Argument	**Value**
Model	text-davinci-003
Temperature	0
Max Tokens	512
Top-p	1
Frequency Penalty	0
Presence Penalty	0
Stop	None

2.2 Chaining Pre-trained Language Models

This approach consisted in chaining, in a pipeline, two types of language models: Named Entity Recognition (NER) and Natural Language Inference (NLI).

Named Entity Recognition These types of models are trained to classify each token of a given text as belonging to one of multiple classes, usually people, locations, organizations, etc. The NER model used was `xlm-roberta-large-finetuned-conll03-english`[2], and it has about 100 million parameters.

Natural Language Inference NLI models are trained to detect if a piece of text (hypothesis) entails or contradicts another (premise). These models are highly versatile, as they can also be used for

[1]`https://github.com/psolin/cleanco`
[2]`https://huggingface.co/xlm-roberta-large-finetuned-conll03-english`

zero-shot classification of text, by using an hypothesis such as "This text is about [topic]." The model used in our pipeline was `DeBERTa-v3-large-mnli-fever-anli-ling-wanli`[3], with about 100 million parameters.

2.2.1 Paragraph Concatenation

The input size for both NER and NLI models is smaller than the usual article, at 512 tokens. To satisfy this constraint while still keeping the ability to detect relations spanning multiple paragraphs, a paragraph concatenation strategy was devised. Articles are broken down into paragraphs, including the title, by `"\n"`. Then, paragraphs are concatenated until a maximum number of tokens has been reached. In our experiments, we used 128 tokens for this parameter. From now on, the term "paragraph" refers to these concatenations.

2.2.2 Entity Extraction

The pipeline for entity extraction is shown in Figure 2.2.

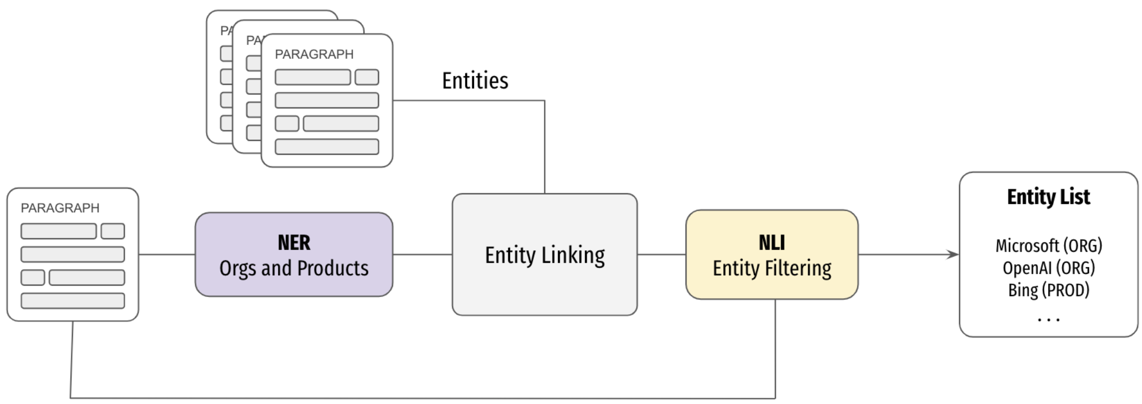

Figure 2.2: Entity extraction with the Pre-trained Language Model approach.

NER Output Filtering Candidate entities that surpass a score threshold (we used 0.995) are extracted from the paragraph using the NER model. Only entities of types `ORG` for companies, and `MISC` for products are kept. The entity's base name is extracted with `cleanco` and words with less than 2 characters are discarded.

[3] `https://huggingface.co/MoritzLaurer/DeBERTa-v3-large-mnli-fever-anli-ling-wanli`

Entity Linking Entities from all paragraphs of an article are linked together based on the shortest intersection of their set of words. As an example, if there's an entity "Nvidia Geforce" (ORG) in paragraph 1 of an article, and another entity "Nvidia AI" (ORG) in paragraph 2, both of these mentions are condensed to "Nvidia" (ORG). This allows us to narrow down the set of entities, which will speed up the relation extraction process later.

NLI Entity Filtering The last entity filtering step consists in using the NLI model to confirm the entities extracted by the NER model. For each entity, the paragraph's text is used as the premise, and the sentence "According to this example, [entity] is a [entity type]" is the hypothesis. The entity type is "company" or "product". The NLI score threshold for this task is 0.99. This leverages the superior textual understanding of the NLI model to filter entities that might have been mistakenly flagged by the NER model, e.g. "Western" or "Chinese".

At the end of the entity extraction process, each paragraph is associated to a set of entities, and each entity has an associated name and entity group.

2.2.3 Relation Extraction

The pipeline for relation extraction is shown in Figure 2.3.

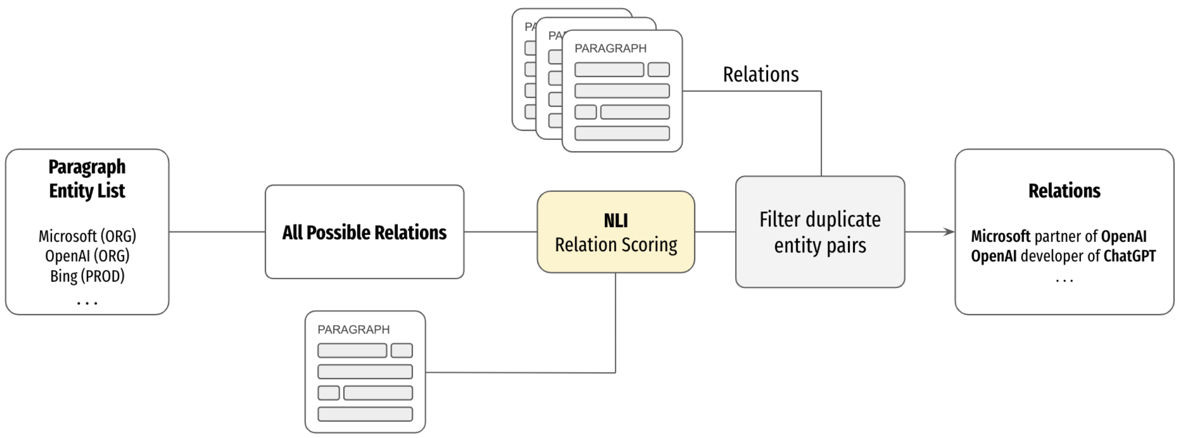

Figure 2.3: Relation extraction with the Pre-trained Language Model approach.

NLI Relation Scoring From the set of entities associated with a paragraph, we create a list of all the possible relations between them, respecting the entity types and directionality of each relation type. Each relation is translated to a natural language sentence, in order to be used as a hypothesis in the NLI task, as in Table 2.2.3. The best scoring relationships are kept (the threshold used was 0.995).

Table 2.2: Relation to sentence translation for the PLM approach.

Relation(E1, E2)	**Sentence**
Owner	E1 is the owner of E2
Developer	E1 is the developer of E2
Investor	E1 is an investor in E2
Partner	E1 and E2 are business partners
Competitor	E1 and E2 are competitors

Filtering Duplicate Entity Pairs For a given article, only the highest scored relation involving a pair of entities (E1, E2) is kept, for each pair (E1, E2). That is, if in paragraph 1 the relation Investor(Microsoft, OpenAI) was extracted with a score 0.998, and in paragraph 3 there is Partner(Microsoft, OpenAI) with a score 0.995, the kept relation is Investor(Microsoft, OpenAI).

The output of this pipeline is a set of relations between entities Relation(E1, E2). In an article, for each pair (E1, E2), there is at most one Relation.

Chapter 3

Results

3.1 Evaluation Method

Dataset The specific task tackled by these approaches is, to our knowledge, a new one. Therefore, it was imperative to create a dataset of annotated news articles. A total of 54 news articles were manually annotated with entity and relation lists, and each relation is accompanied by the relevant passage from the text. They were chosen from reputable news sources, mainly: Reuters, CNN, Business Wire, CNBC, and more. The themes of the articles focus mostly on Technology, but other topics were also considered, such as Car Manufacturers, Social Media, Gaming or Health, among others. The articles were extracted using the Python package `newspaper3k`[1].

Metrics For each article, the Precision (Eq. (3.1)), Recall (Eq. (3.2)) and F1 (Eq. (3.3)) scores are computed between the set of relations (or entities) predicted by the models and the manually annotated dataset. These values are then averaged across the total number of articles in the dataset.

$$\text{Precision} = \frac{TP}{TP + FP} \tag{3.1}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{3.2}$$

$$\text{F1} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{3.3}$$

In the equations above, TP refers to True Positives, FP to False Positives and FN to False Negatives. On the task of Entity Extraction, when comparing the results from the models with the golden dataset,

[1] `https://github.com/codelucas/newspaper`

only the entities' names are taken into account, and not the entity type. This is because, in some cases, the distinction between company and product can be hard to define, for example, "Google" can refer to the search engine, which is a product, or the company.

3.2 Performance of both approaches

Table 3.1 shows the performance of both approaches on the tasks of Entity and Relation Extraction on the manually annotated dataset. In this table, metrics are calculated as specified in Section 3.1. In Table 3.2, the performance for each relation type is presented. The computation of these values is different from the ones in Table 3.1, as in this case, each metric is not averaged across each article, but computed for the dataset as a whole.

Table 3.1: Performance comparison between the GPT-3 prompting and Pre-trained Language Models approaches. Values were computed for each article and then averaged across articles.

Task	Model	Precision	Recall	F1
Entity Extraction	PLM	72.2	75.5	**72.5**
	GPT-3	**72.8**	**75.8**	71.2
Relation Extraction	PLM	**65.8**	**66.9**	**64.1**
	GPT-3	57.2	47.6	48.0

Table 3.2: Performance for each Relation Type. Values were computed using the total TP, FP, and FN values for the full dataset.

Relation (% of rels. in dataset)	Model	Precision	Recall	F1
Developer (33.3%)	PLM	**66.2**	**60.2**	**63.1**
	GPT-3	50.5	53.4	51.9
Competitor (30.7%)	PLM	35.0	**61.7**	**44.6**
	GPT-3	**53.5**	28.4	37.1
Partner (20.5%)	PLM	45.1	**59.3**	51.2
	GPT-3	**82.1**	42.6	**56.1**
Owner (11.0%)	PLM	**68.8**	**37.9**	**48.9**
	GPT-3	36.4	41.4	38.7
Investor (4.5%)	PLM	66.7	**66.7**	66.7
	GPT-3	**100.0**	50.0	66.7

3.3 PLM Pipeline Ablation Study

To investigate the impact of the additional filtering steps applied in the PLM pipeline, more specifically during the entity extraction phase, an ablation study was conducted, with the results presented in Table 3.3.

Table 3.3: Impact of different parts of the PLM pipeline on performance. EL stands for Entity Linking and NLIEF stands for NLI Entity Filtering.

	Entity Extraction			Relation Extraction		
Pipeline	**Precision**	**Recall**	**F1**	**Precision**	**Recall**	**F1**
PLM_{base}	51.3	80.6	60.6	46.4	63.6	51.4
PLM_{EL}	+7.2	-2.5	+4.6	+9.8	+1.0	+6.8
PLM_{NLIEF}	+16.7	**+1.5**	+11.9	+3.6	-0.7	+1.9
$PLM_{EL+NLIEF}$	**+20.9**	-5.1	**+11.9**	**+19.4**	**+3.3**	**+12.7**

3.4 Error Analysis

In this section, we'll investigate some common error types committed by the Information Extraction pipelines developed.

Wrong entities Sometimes our models classified entities that are not companies or products, as such. This type of error was more present in the GPT-3 approach. For example, in an article from Reuters with the title *"Exclusive: EU drafts plan to allow e-fuel combustion engine cars"*[2], GPT-3 finds that (1) indicates the relation Developer(European Commission, CO2). In the same article, it also marks Germany as a company. In the article *"NASA, SpaceX postpone launch of next space station crew at 11th hour"*[3], it detects names of some mentioned individuals as companies, as in (2), where it detected the relation Owner(Elon Musk, SpaceX). In another article *"As chatbots boom, Nvidia sales outlook beats Wall Street expectations"*[4], the PLM system detects the relation Competitor(AMD.O, Nvidia), likely due to (3), instead of Competitor(Advanced Micro Devices, Nvidia).

(1) "The draft proposal, seen by Reuters on Tuesday, suggests creating a new type of vehicle category in the European Union for cars that can only run on carbon neutral fuels"
(2) "The private rocket company founded by billionaire **Elon Musk**"
(3) "**Nvidia**'s outlook also helped boost the share prices of competitors such as **Advanced Micro Devices** (**AMD.O**), whose stocks were up 3% after **Nvidia**'s results."

It's still noteworthy how the passages don't have to mention the entities by name for GPT-3 to

[2] https://www.reuters.com/business/autos-transportation/eu-proposes-exception-e-fuel-combustion-engines-2035-2023-03-21/
[3] https://www.reuters.com/lifestyle/science/nasa-spacex-postpone-launch-next-space-station-crew-2023-02-27/
[4] https://www.reuters.com/technology/nvidia-forecasts-first-quarter-revenue-above-expectations-2023-02-22/

extract relations from them, as it knows that "private rocket company" refers to SpaceX, mentioned previously in the article. This is an example of coreference resolution, performed by GPT-3.

Wrong relations As the natural language understanding of these models is not perfect, some extracted relations might not make sense given the text. Fully wrong relations were hard to find in the model's output, as most of the time it was possible to see how the article's wording could mislead the models. Likely due to (3), our PLM approach detects Competitor(AMD.O, Advanced Micro Devices), probably because it classified Advanced Micro Devices and AMD.O as two different companies, and they are present in a sentence that contains the word "competitors". In passage (4), from the article *"China tech companies are closely watching ChatGPT's A.I. skills. Here's what they're doing about it "*[5], GPT-3 finds the relation Developer(Kunlun Tech, ChatGPT), although the reality is that they are developing a separate open source alternative to the product.

> (4) "**Kunlun Tech** expects to release an open source Chinese version of **ChatGPT**, as early as the middle of this year"

Valid relations not in manual dataset Some relations can have overlapping meanings. For example, an Investment or Ownership could also be described as a Partnership. In the article *"Microsoft and NVIDIA announce expansive new gaming deal"*[6], the PLM pipeline finds the relation Partner(Microsoft, Activision), while the manual dataset has the relation Owner(Microsoft, Activision). This was probably due to passage (5), however, the acquisition of Activision by Microsoft is specifically mentioned later in the text. The correct relation was found by the NLI model, but it was then filtered out, since we only allow one type of relation per pair of entities. This type of error is found mostly in the Competitor relation. In (6) there are 6 companies and products mentioned, which could all be competitors of each other. However, the manual dataset does not have all the rival combinations, while the PLM pipeline usually finds all of them, resulting in a low Precision score (Table 3.2).

> (5) "Partnership will bring blockbuster lineup of **Xbox** games, including **Minecraft** and **Activision** titles like Call of Duty, to **NVIDIA GeForce NOW** cloud gaming service"
> (6) "Over the next two years, rivals including **General Motors**, **Ford**, **Mercedes-Benz**, **Hyundai** and **VW** will launch scores of new electric vehicles, from a **Chevrolet** priced below $30,000 to luxury sedans and SUVs that top $100,000."

[5] https://www.cnbc.com/2023/02/23/china-tech-companies-are-closely-watching-chatgpts-ai-skills.html
[6] https://www.prnewswire.com/news-releases/microsoft-and-nvidia-announce-expansive-new-gaming-deal-301752099.html

3.5 Discussion

Overall results The PLM approach significantly surpasses the LLM approach in the Relation Extraction task, while the performance in Entity Extraction is very similar. For the Relation Extraction task, the classification tasks performed by the PLM pipeline are better suited. The NLI model only has to rate the possible relations given the context, and this is easier than generating formatted text expressing relations seen in the article. The PLM pipeline is also much more complex, involving several filtering steps, with threshold values that can be tuned to achieve the best results. While GPT-3 also has some inference parameters, we found that making the generation more variable (by increasing the temperature, for example) usually leads to more hallucinations and deviations from the fixed output format. We also saw that, in general, the PLM pipeline outputs many more relations than GPT-3. The lowest Precision scores for the PLM pipeline are achieved by Partner and Competitor, as these have a relatively broad meaning, and the pipeline outputs many relations that are not present in the golden dataset.

Relationship range We observed qualitatively that GPT-3 is better at capturing some nuanced and long-ranged relations. The article *"Warner Bros. Discovery's Zaslav Shows New Combative Streak, Tweaking Media Rivals"*[7] is about statements that the CEO of Warner Bros. made during a call with investors. From passage (7), GPT-3 detects that the text in bold indicates the relation Owner(Warner Bros., CNN). This requires the following multi-hop reasoning: "The executive" → Zaslav → CEO of Warner Bros. + talking about "strategy at CNN" = Warner Bros. owns CNN. While the PLM approach can also perform coreference resolution within its context length, this specific reasoning requires a longer context window to be achievable, since the executive's identity is mentioned in the text far before this passage. On the other hand, the PLM pipeline seems to be better at detecting more fine-grained, short-range relations. In *"Amazon cuts 9,000 more jobs, bringing 2023 total to 27,000"*[8], PLM correctly detects Owner(Meta, Facebook) and Owner(Alphabet, Google), likely due to passage (8), while GPT-3 finds, also validly, Competitor(Amazon, Alphabet) and Competitor(Amazon, Meta), but doesn't find the ownerships. We found that GPT-3 struggles with the Owner relation.

> (7) "The executive also showed a defiant streak, doubling down on a strategy at CNN that has generated lots of outside scrutiny. **In recent weeks, CNN has set in motion plans to overhaul the bulk of its daily schedule**"
> (8) "Like other tech companies, including **Facebook parent Meta and Google parent Alphabet**, Amazon ramped up hiring during the pandemic to meet the demand from homebound Americans"

[7] https://variety.com/2023/tv/news/warner-bros-discovery-zaslav-shows-combative-streak-1235534033/
[8] https://apnews.com/article/amazon-layoffs-jobs-cuts-jassy-0e857f39702de134c8f677c5b5731688

Impact of PLM pipeline steps In Table 3.3 we see that the Entity Linking and NLI Entity Filtering steps significantly improve F1 scores both in Entity and Relation Extraction. There is a decrease in Recall in Entity Extraction, which is expected since these steps decrease the set of entities detected by PLM$_{base}$. This shows the dependence of the Relation Extraction task on the quality of the entity set extracted.

Speed and Memory Usage In terms of speed, the PLM pipeline delivers the fastest entity and relation extraction, taking less than 1 second per article, and requiring about 4GB of GPU memory to run. While our LLM approach does not require a GPU, since it uses the OpenAI API, it is significantly slower, at about 20 seconds per article.

Chapter 4

Conclusion and Future Work

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. For the LLM approach, more complex prompting methods, such as chain-of-thought [9], could be explored. One could experiment with using an LLM as the NLI model that rates relations given to it in the prompt. 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.

Bibliography

[1] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. "Language models are few-shot learners". In: *Advances in neural information processing systems* 33 (2020), pp. 1877–1901.

[2] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. "Bert: Pre-training of deep bidirectional transformers for language understanding". In: *arXiv preprint arXiv:1810.04805* (2018).

[3] Kun Ding, Shanshan Liu, Yuhao Zhang, Hui Zhang, Xiaoxiong Zhang, Tongtong Wu, and Xiaolei Zhou. "A Knowledge-Enriched and Span-Based Network for Joint Entity and Relation Extraction." In: *Computers, Materials & Continua* 68.1 (2021).

[4] Markus Doehring. "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*. 2008.

[5] Hadjer Khaldi, Farah Benamara, Amine Abdaoui, Nathalie Aussenac-Gilles, and Eunbee Kang. "Multilevel entity-informed business relation extraction". In: *International Conference on Applications of Natural Language to Information Systems*. Springer. 2021, pp. 105–118.

[6] Nick McKenna, Tianyi Li, Liang Cheng, Mohammad Javad Hosseini, Mark Johnson, and Mark Steedman. "Sources of Hallucination by Large Language Models on Inference Tasks". In: *arXiv preprint arXiv:2305.14552* (2023).

[7] Shirui Pan, Linhao Luo, Yufei Wang, Chen Chen, Jiapu Wang, and Xindong Wu. "Unifying Large Language Models and Knowledge Graphs: A Roadmap". In: *arXiv preprint arXiv:2306.08302* (2023).

[8] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. "Attention is all you need". In: *Advances in neural information processing systems* 30 (2017).

[9] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. "Chain-of-thought prompting elicits reasoning in large language models". In: *Advances in Neural Information Processing Systems* 35 (2022), pp. 24824–24837.

[10] Ayana Yamamoto, Yuichi Miyamura, Kouta Nakata, and Masayuki Okamoto. “Company relation extraction from web news articles for analyzing industry structure”. In: *2017 IEEE 11th International Conference on Semantic Computing (ICSC)*. IEEE. 2017, pp. 89–92.

Appendix A

GPT-3 Prompt

Below is the few-shot prompt used for entity and relation extraction with GPT-3.

Each article below has an associated list of entities and a table with relationships. The list has items of the form "entity (type)", where "entity" is the name (e.g. Amazon) and "type" is one of these types: ORG (for companies) or PROD (for products). The table links the mentioned companies to other mentioned companies. Here's the explanation of every column. "Entity" relates to a company or a product. "Relationship" is the connection between the entities, only using entities from the list. For example, in the sentence "KitKat, produced by Nestle, is working with a new cocoa farm", a valid relationship is |Nestle|developer of|KitKat|, but a non-valid relationship is |KitKat|partner of|new cocoa farm|, because "new cocoa farm" is not a valid entity. "Passage" is the text from the article relevant to that relationship. Some articles don't have entities (Entity list = []) or no relationships (the only table row is |end|).
Allowed relationship types:
Company-Company (ORG-ORG) = [owner of, partner of, investor in, competitor]
Company-Product (ORG-PROD) = [developer of]
Product-Product (PROD-PROD) = [competitor]
Here are some examples:
Title: JPMorgan restricts employee use of ChatGPT
Date: Wed February 22, 2023
Text: JPMorgan Chase is temporarily clamping down on the use of ChatGPT among its employees, as the buzzy AI chatbot explodes in popularity. The biggest US bank has restricted its use among global staff, according to a person familiar with the matter. The decision was taken not because of a particular issue, but to accord with limits on third-party software due to compliance concerns, the person said. JPMorgan Chase (JPM) declined to comment.
ChatGPT was released to the public in late November by artificial intelligence research company Open AI. Since then, the much-hyped tool has been used to turn written prompts into convincing academic essays and creative scripts as well as trip itineraries and computer code.

Adoption has skyrocketed. UBS estimated that ChatGPT reached 100 million monthly active users in January, two months after its launch. That would make it the fastest-growing online application in history, according to the Swiss bank's analysts.

The viral success of ChatGPT has kickstarted a frantic competition among tech companies to rush AI products to market. Google recently unveiled its ChatGPT competitor, which it's calling Bard, while Microsoft (MSFT), an investor in Open AI, debuted its Bing AI chatbot to a limited pool of testers.

But the releases have boosted concerns about the technology. Demos of both Google and Microsoft's tools have been called out for producing factual errors. Microsoft, meanwhile, is trying to rein in its Bing chatbot after users reported troubling responses, including confrontational remarks and dark fantasies. Another big player in AI innovation is Meta, with a higher focus on open source models.

Their recent push for the Metaverse and the Oculus glasses has put them behind the chatbot race.

Some businesses have encouraged workers to incorporate ChatGPT into their daily work. But others worry about the risks. The banking sector, which deals with sensitive client information and is closely watched by government regulators, has extra incentive to tread carefully. Schools are also restricting ChatGPT due to concerns it could be used to cheat on assignments. New York City public schools banned it in January.

Entity list = [JPMorgan (ORG), ChatGPT (PROD), Open AI (ORG), Google (ORG), Meta (ORG), Metaverse (PROD), Oculus (PROD), UBS (ORG), Bard (PROD), Microsoft (ORG)]

|Entity|Relationship|Entity|Passage|

|Open AI|developer of|ChatGPT|"ChatGPT was released to the public in late November by artificial intelligence research company Open AI"|

|Google|competitor|Open AI|"Google recently unveiled its ChatGPT competitor"|

|Microsoft|investor in|Open AI|"Microsoft (MSFT), an investor in Open AI"|

|Bard|competitor|ChatGPT|"Google recently unveiled its ChatGPT competitor, which it's calling Bard"|

|Google|competitor|Microsoft|"Demos of both Google and Microsoft's tools"|

|Meta|competitor|Microsoft|"Another big player in AI innovation is Meta"|

|Meta|competitor|Google|"Another big player in AI innovation is Meta"|

|Meta|competitor|OpenAI|"Another big player in AI innovation is Meta"|

|Meta|developer of|Metaverse|"Their recent push for the Metaverse"|

|Meta|developer of|Oculus|"Their recent push for the Metaverse and the Oculus glasses"|

|end|

–

Title: Amplifon expects acquisitions to boost revenue after record full-year results

Date: March 1, 2023

Text: March 1 (Reuters) - Italy's Amplifon (AMPF.MI) expects to boost revenue through bolt-on acquisitions this year, the world's largest hearing aid retailer said on Wednesday after reporting record core profit for 2022.

The Milan-based company posted full-year recurring earnings before interest, taxes, depreciation, and amortisation (EBITDA) of 525.3 million euros ($560.28 million) in its best-ever results, compared

with 482.8 million euros a year earlier.
Amplifon, which proposed a divided of 29 cents per share, also reported record annual recurring revenue of 2.12 billion euros, compared with 1.95 billion euros a year earlier.
By 1241 GMT, the company's shares were up 1.7%, while Italy's blue-chip index FTSE MIB (.FTMIB) was up 0.65%.
Entity list = [Amplifon (ORG)]
|Entity|Relationship|Entity|Passage|
|end|
–
Title: Tensions between gaming companies soar to new highs
Date: Fri, June 20, 2021
Text: Microsoft's pending $68.7 billion acquisition of Activision-Blizzard has been a highly debated topic in recent times. The conversation around the deal only appears to be intensifying, particularly after Microsoft's recent meeting with EU regulators. During the meeting, Microsoft looked to make its case, mostly countering Sony's opposition to the deal. One of the biggest points about the Activision purchase pertains to the potential implications for the Call of Duty franchise.
A similar tension was observed back in 2016 with From Software, the studio behind the famous Dark Souls franchise. It was acquired by Bandai in a similar fashion, leading to high amounts of speculation. Bandai Namco, since last year, has partnered with Activision to publish their most important titles but has since decided to stop this venture, as a report from last week indicates. Some rumors have been circulating about Mojang being Activision's replacement for Bandai Namco's upcoming game releases, which had the markets going wild.
The latest player in the videogame industry, Meta (previously known as Facebook), is also looking to take advantage of its massive user base to enter the gaming industry, taking advantage of their VR technology.
Entity list = [Microsoft (ORG), Activision-Blizzard (ORG), Sony (ORG), Call of Duty (PROD), From Software (ORG), Dark Souls (PROD), Bandai Namco (ORG), Meta (ORG), Facebook (ORG)]
Entity	Relationship	Entity	Passage
Microsoft	owner of	Activision-Blizzard	"Microsoft's pending $68.7 billion acquisition of Activision-Blizzard"
Microsoft	competitor	Sony	"Microsoft looked to make its case, mostly countering Sony's opposition to the deal"
Bandai Namco	owner of	From Software	"From Software, the studio behind of the famous Dark Souls franchise. It was acquired by Bandai in a similar fashion"
Bandai Namco	partner of	Activision	"Bandai Namco, since last year, has partnered with Activision"
Bandai Namco	partner of	Mojang	"Some rumors have been circulating about Mojang being Activision's replacement for Bandai Namco's upcoming game releases"
Meta	competitor	Microsoft	"The latest player in the videogame industry, Meta (previously known as Facebook), is also looking to take advantage of its massive user base to enter the gaming industry"
From Software	developer of	Dark Souls	"From Software, the studio behind the famous Dark Souls

franchise"|
|Activision-Blizzard|developer of|Call of Duty|"Activision purchase pertains to the potential implications for the Call of Duty franchise"|
|end|
–
Title: **title**
Date: **date**
Text: **text**
Entity list =