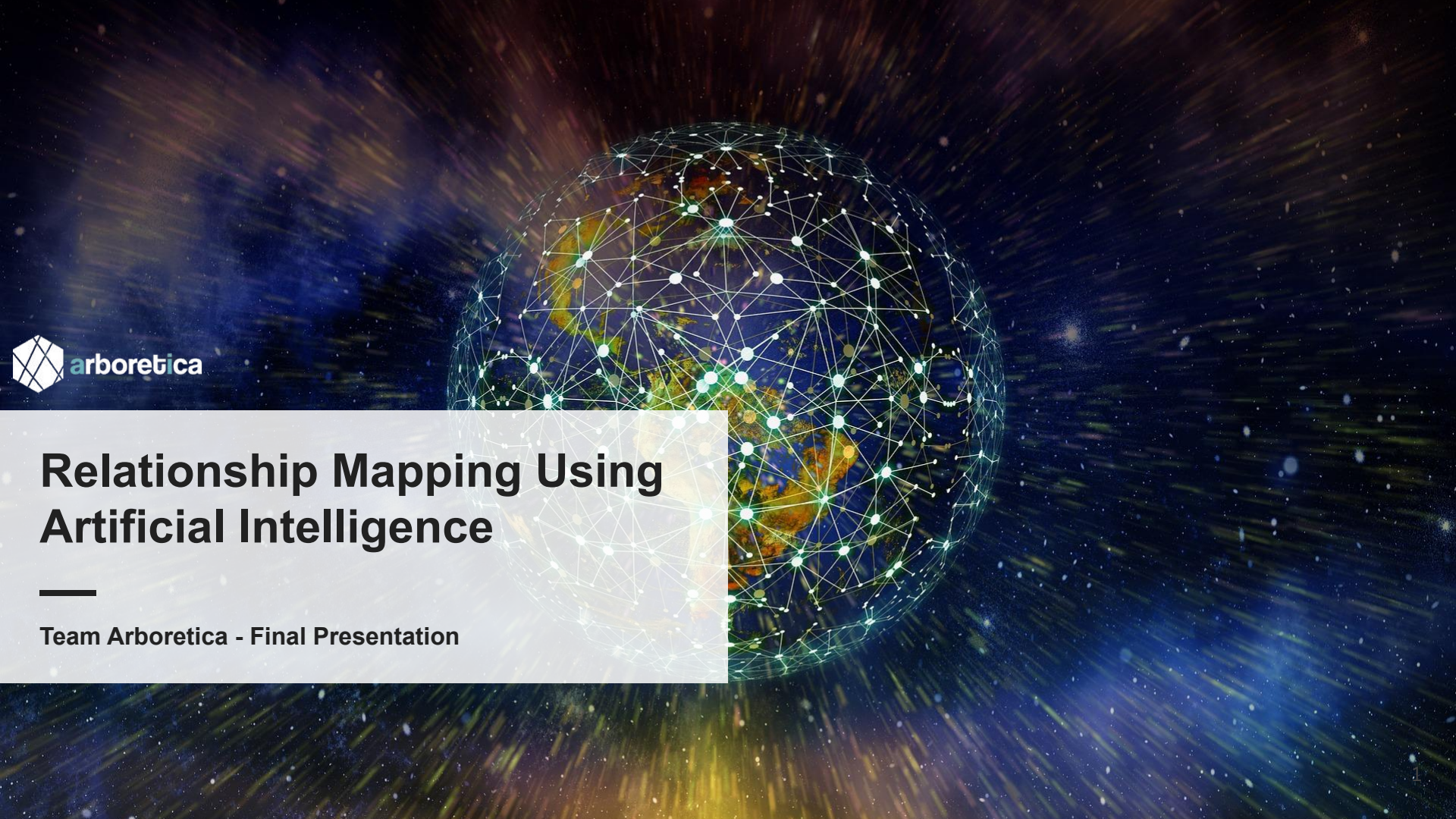


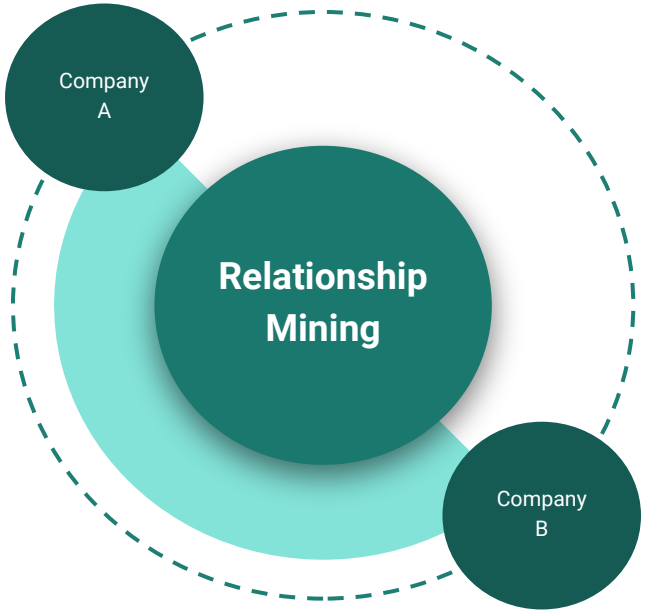
Relationship Mapping Using Artificial Intelligence

Team Arboretica - Final Presentation

Executive Summary

Our project goal is to develop an algorithm solution for our client, Arboretica, to extract financial relationships from texts more dynamically. To complete this project, we conducted literature research, explored different natural language processing techniques, expanded our dataset, and fine-tuned an advanced algorithm tailored to our client's needs. Our algorithm is expected to significantly reduce human efforts.

Agenda



- Team Introduction
- Project Background
- Problem Breakdown
- Project Methodology
- Data Analysis
- Model Training
- Key Insight & Future Work

Team Introduction

Team Arboretica

Clytze Sun | Project Manager

MISM - ‘22
Aspiring Data Scientist

Familiar with network graphing, distributed system, and machine learning algorithms

puxins@andrew.cmu.edu

Koko Wang

MISM - ‘22
Aspiring Data Analyst

Familiar with Machine Learning and NLP algorithms

kehanw@andrew.cmu.edu

Yusu Wang

MISM - ‘22
Aspiring Software Engineer

Familiar with Distributed System and NoSQL databases

yusuw@andrew.cmu.edu

Team Arboretica

Silvia Gu

MSPPM-DA ‘23
Aspiring Data Scientist

Familiar with NLP Research and Machine Learning

yunxing@andrew.cmu.edu

Florence Shen

MISM ‘22
Aspiring Machine Learning Engineer

Familiar with NLP Research and Machine Learning

jiayis2@andrew.cmu.edu

Kelly Zhang

MSPPM-DA ‘23
Aspiring Data Scientist

Familiar with NLP Research and Machine Learning

yizhang4@andrew.cmu.edu

Project Background

Background on Arboretica

| Arboretica is a Netherland-based company that develops rigorous technology to transform the environmental industry with **data science and automation**.

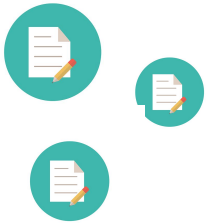
| A large amount of Arboretica's projects involve **collecting knowledge** from different industries to accelerate decision making for different organizations.

| Particularly, Arboretica uses **network analysis** to map out direct and hidden **relationship between entities** of interests from mass amount of public data as a key part of knowledge base building.

Project Goal

Design a Natural Language Processing (NLP) **algorithm solution** that can automatically identify entities from public data and extract relationships or financial flows between identified entities

NLP Algorithm



All public data (texts):
news articles, reports etc.

Relationship Network

Source: *Relationship Data Mining* (2022) *Arboretica*. Available at: https://www.arboretica.com/blog/relationship-data-mining-2/ (Accessed: October 22, 2022).

Keywords Clarification

- **Entity**: company names

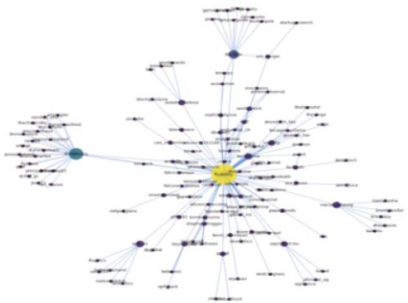
- **Different Types of Relationship**

  - **Partner**: if company A and B are working together, not necessarily in a direct contractual relationship
  - **People**: if company A and B are tied because of a person's relationship. e.g. if a person who use to work in company A is now hired by company B
  - **Financial**: if company A and B has any type of financial relationships, such as investment, acquisition, funding, grant, etc.
  - **Technical**: if company A uses company B's technology or vice versa.

Business Impact

Our algorithm solution can potentially help Arboretica serve its industry and academia clients, including leading environmental NGOs and policy research institutions.

Predictive Content Marketing



Automated Industry Intelligence Collection

Accelerated policy decisions

Source: *AI industry intelligence* (2022) *Arboretica*. Available at: https://www.arboretica.com/blog/ai-industry-intelligence/(Accessed: October 22, 2022).

Problem Breakdown

Relevant Industry and Academic solutions to NLP Problem

Main components

Entity Recognition

Implement algorithm to extract relevant entities from input texts

Relationship Extraction

Extract nature of relationship among the identified entities based on the whole context

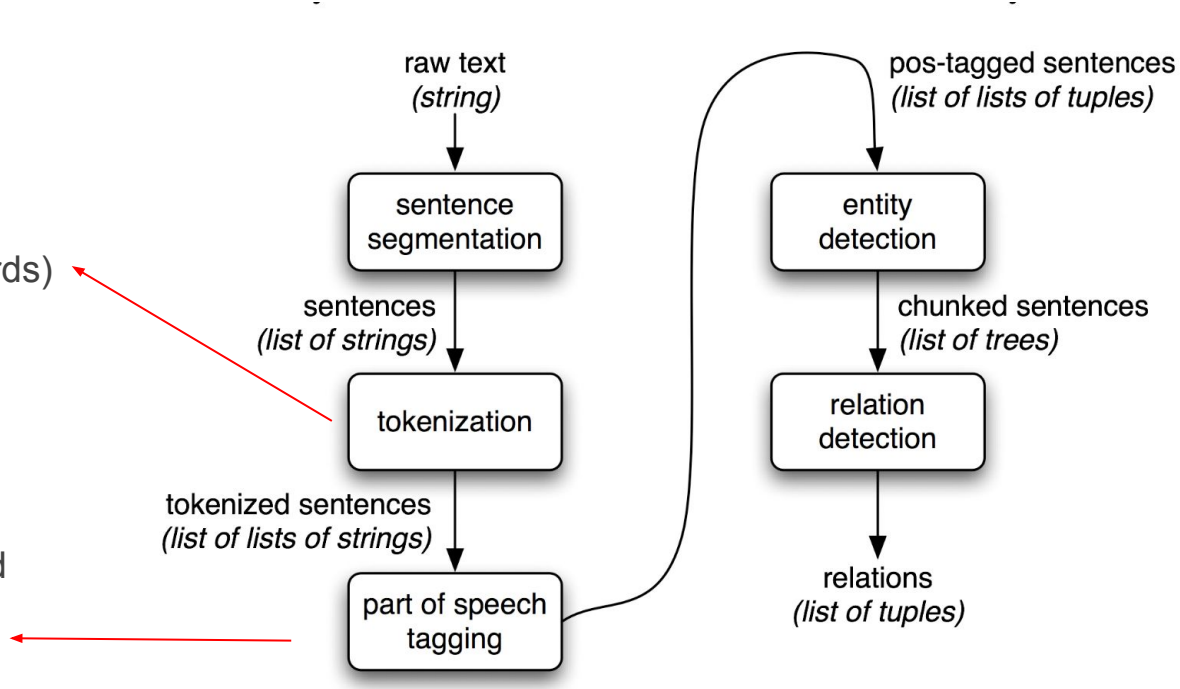
Process Breakdown

Tokenization:

break up a sentence into understandable units (mostly words)

Part of Speech tagging:

label each unit's grammatical and semantic functions



Project Methodology

spaCy, BERT, RoBERTa, LUKE

Methodology Evolution

Regular Expression

Word and Sentence Structure, only explicit facts

Word Vectorizers + Machine Learning

Topic Modeling

RNN-LSTM Deep Learning

Most are rule-based learning

Transformer-based Models

Advanced, with ability to extract implicit facts

spaCy name recognition algorithm

an open-source software library for advanced natural language processing. The library is published under the MIT license.

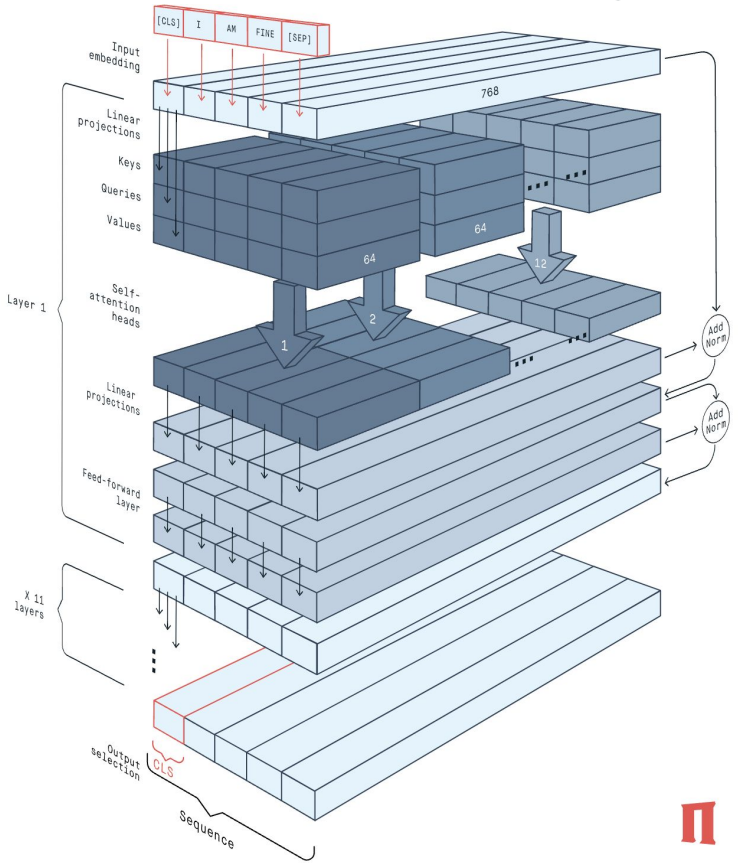
- Statistical models for 19 languages
- Multi-task learning with pre-trained transformers like BERT
- Support for custom models in PyTorch, TensorFlow and other frameworks
- State-of-the-art speed and accuracy

TEXT	START	END	LABEL	DESCRIPTION
Apple	0	5	ORG	Companies, agencies, institutions.
U.K.	27	31	GPE	Geopolitical entity, i.e. countries, cities, states.
$1 billion	44	54	MONEY	Monetary values, including unit.

Source: spaCy. (2022, October 4). In Wikipedia. https://en.wikipedia.org/wiki/SpaCy

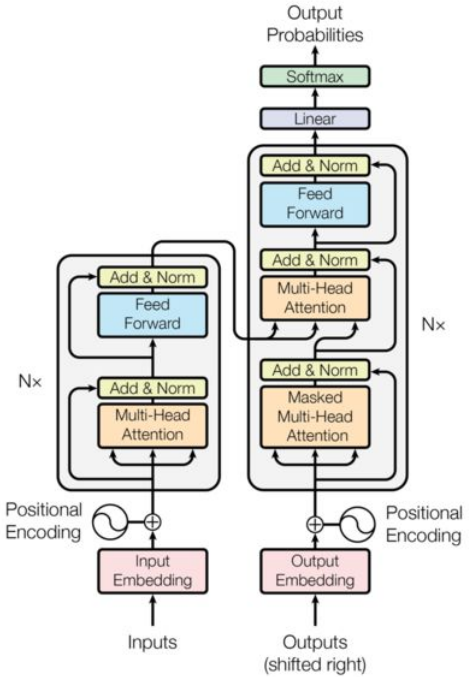
BERT Model

Bidirectional Encoder Representations from Transformers (BERT) is a transformer-based machine learning technique for natural language processing (NLP) pre-training developed by Google in 2018.



Source: BERT (language model). (2022, August 31). In Wikipedia. https://en.wikipedia.org/wiki/BERT_(language_model)

BERT unique benefits can solve our problem



Pretrained on a lot of data

- 2500 million words
- Tuning can be effective even with small dataset

Accounts for word's context

- Takes into account word's position and context
  He trusts you
  He has a trust fund
  I can't trust you

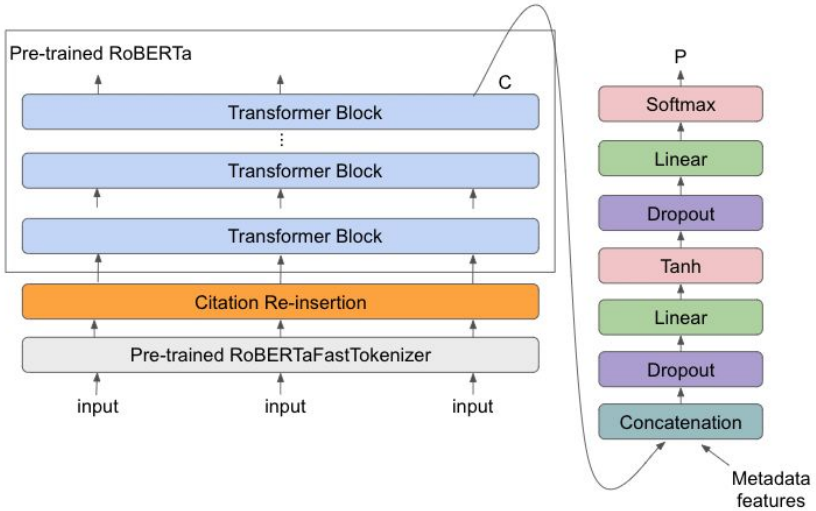
Open-source

- We can fine-tune according to our needs

RoBERTa Model

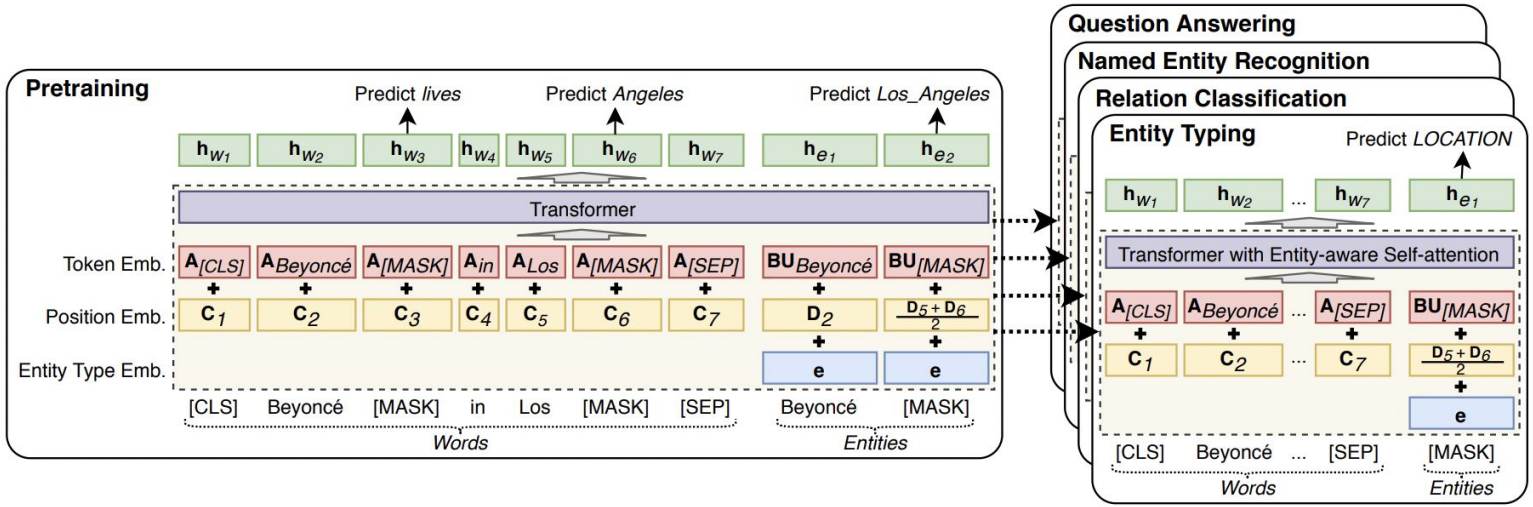
Similar architecture as compare to BERT, with design changes that improve results:

- Removing the Next Sentence Prediction (NSP) objective
- Training with bigger batch sizes & longer sequences
- Dynamically changing the masking pattern



"Overview of Roberta Model." GeeksforGeeks, June 23, 2022. https://www.geeksforgeeks.org/overview-of-roberta-model/.

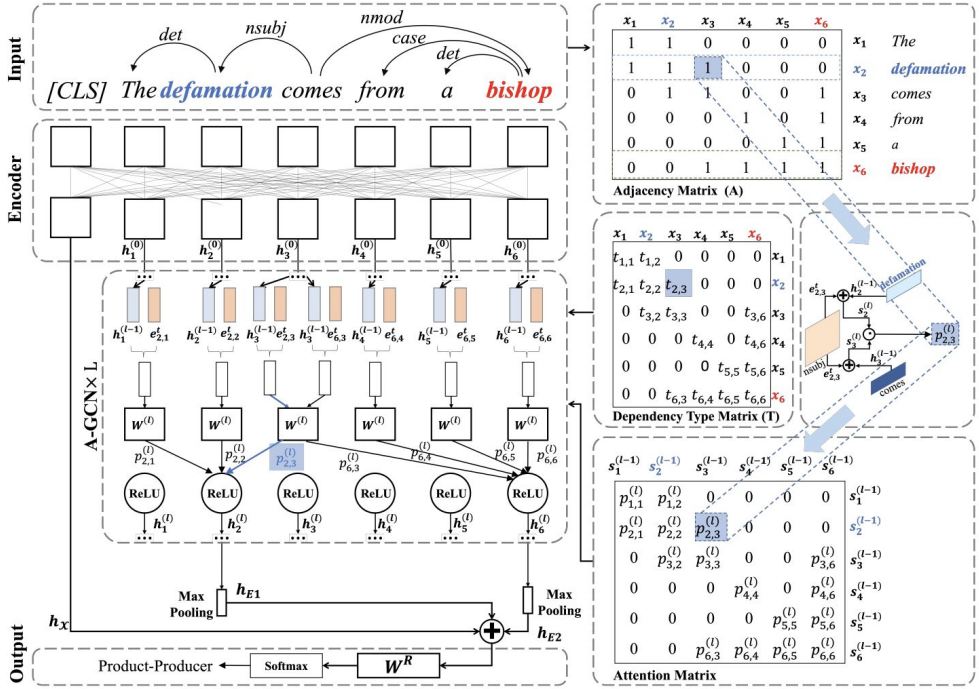
LUKE (Language Understanding with Knowledge-based Embeddings)



"Luke." LUKE. Accessed October 22, 2022. https://huggingface.co/docs/transformers/model_doc/luke.

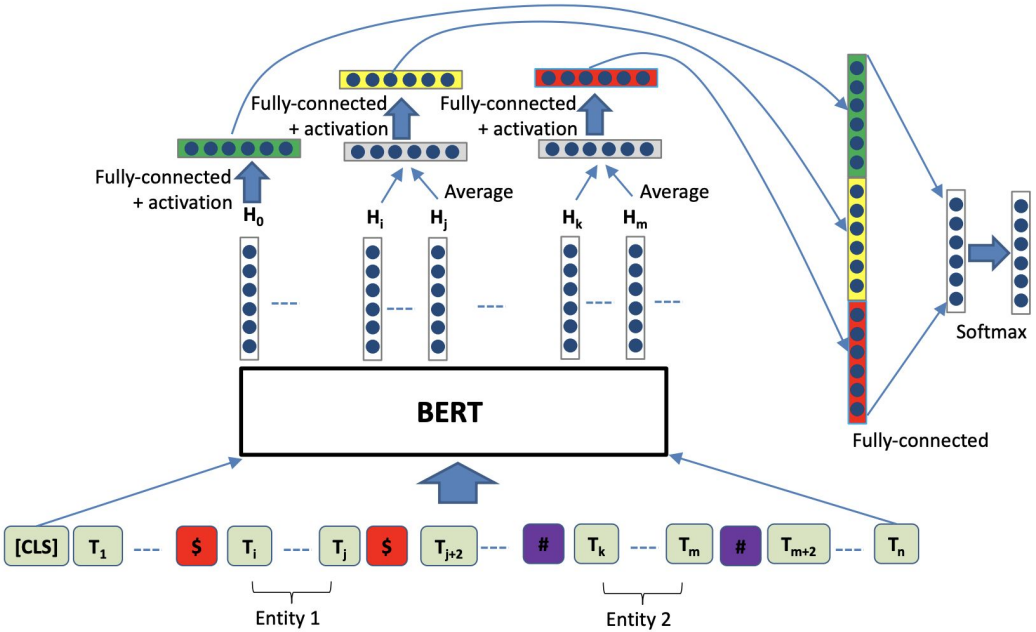
Deal with multiple-entity in a single sentence

A-GCN: leverage dependency information for relation extraction



Source: Tian, Yuanhe, Guimin Chen, Yan Song, and Xiang Wan. "Dependency-Driven Relation Extraction with Attentive Graph Convolutional Networks." ACL Anthology, 2021.

Incorporates information from the target entities



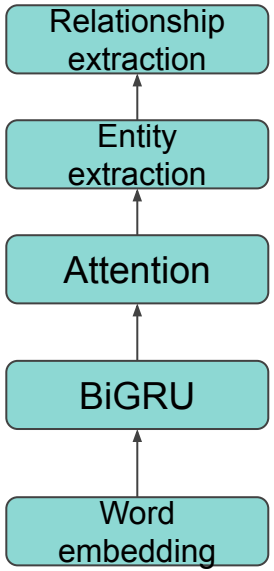
Source: Wu, Shanchan, and Yifan He. “Enriching Pre-Trained Language Model with Entity Information for Relation Classification.” arXiv.org, May 20, 2019. https://arxiv.org/abs/1905.08284.

Reference models built specifically for financial field

Research on Entity Relationship Extraction in Financial and Economic Field Based on Deep Learning

The authors propose BGAJM (BiGRU Attention Joint Model) to extract entity relationships in the financial field.

The model achieved strong performance compared to several competitive models at the time.



Z. Zhou and H. Zhang, "Research on Entity Relationship Extraction in Financial and Economic Field Based on Deep Learning," *2018 IEEE 4th International Conference on Computer and Communications (ICCC)*, 2018, pp. 2430-2435, doi: 10.1109/CompComm.2018.8780966.

Data Analysis

Data Augmentation

Source of data

We chose the most common data sources used by Arboretica

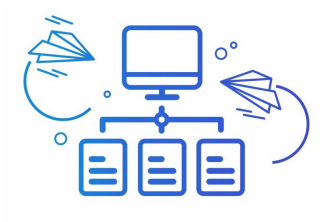
FINSMES

FinSMEs is the financial news site dedicated to covering venture capital, private equity, and merger and acquisition deals in real time.

ST RZ SHORT TERM RENTALZ

ShortTermRentalz provides the news and intelligence for the fast-growing and rapidly-evolving short term rental industry

Process of data augmentation



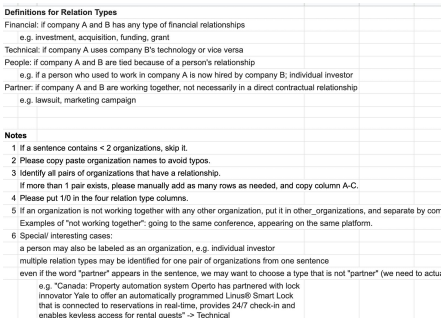
Web scraping

500 articles

Thousands of sentences

Annotate data points from almost 7000 sentences manually

Design well-rounded annotation standards



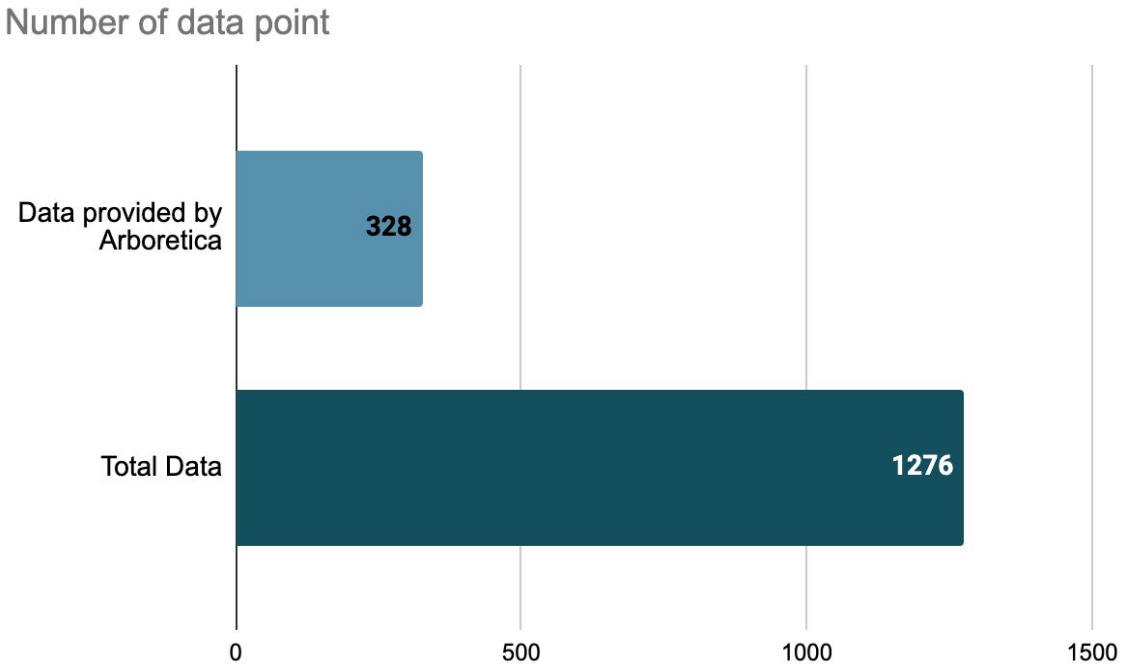
Definitions for Relation Types

Financial: if company A and B has any type of financial relationships
e.g. investment, acquisition, funding, grant
Technical: if company A uses company B's technology or vice versa
People: if company A and B are tied because of a person's relationship
e.g. if a person who used to work in company A is now hired by company B; individual investor
Partner: if company A and B are working together, not necessarily in a direct contractual relationship
e.g. lawsuit, marketing campaign

Notes

1 If a sentence contains < 2 organizations, skip it.
2 Please copy paste organization names to avoid typos.
3 Identify all pairs of organizations that have a relationship.
If more than 1 pair exists, please manually add as many rows as needed, and copy column A-C.
4 Please put 1/0 in the four relation type columns.
5 If an organization is not working together with any other organization, put it in other_organizations, and separate by com
Examples of "not working together": going to the same conference, appearing on the same platform.
6 Special/ interesting cases:
a person may also be labeled as an organization, e.g. individual investor
multiple relation types may be identified for one pair of organizations from one sentence
even if the word "partner" appears in the sentence, we may want to choose a type that is not "partner" (we need to actu
e.g. "Canada: Property automation system Operto has partnered with lock
innovator Yale to offer an automatically programmed Linus® Smart Lock
that is connected to reservations in real-time, provides 24/7 check-in and
enables keyless access for rental guests" -> Technical

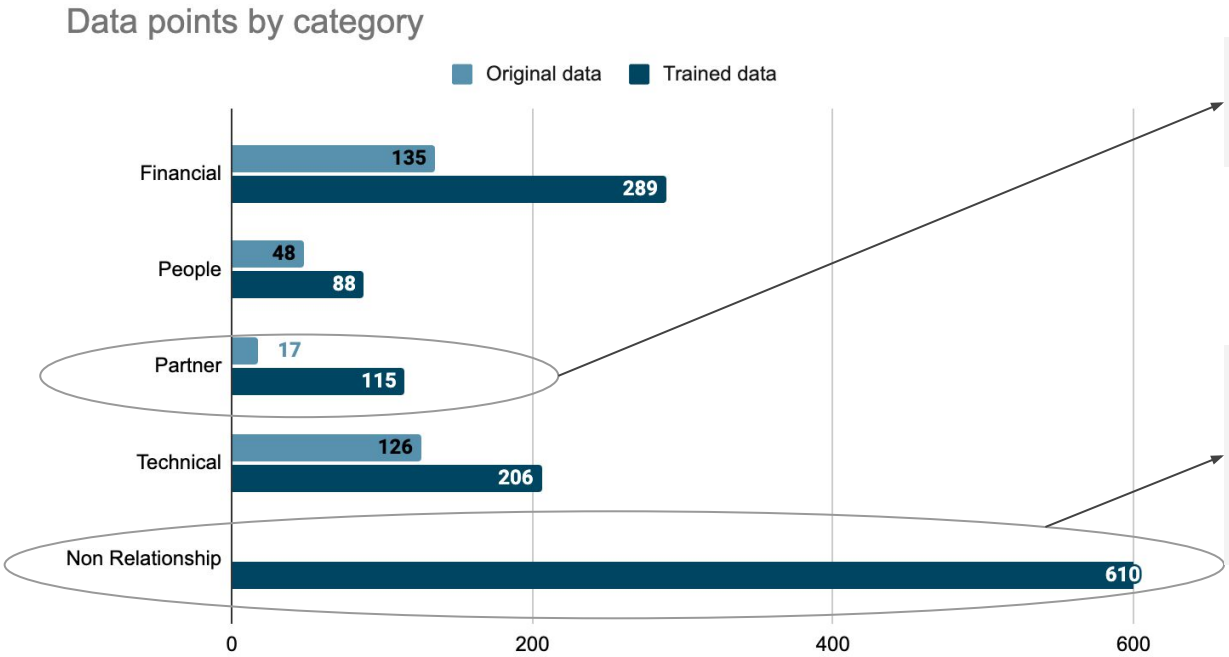
Result of data augmentation



To train a better model, we find more data and annotate the data manually.....

- > 900 data points added
- Total data points 4 times than original data points

More balanced trained data



Add many more data points with Partner relationship

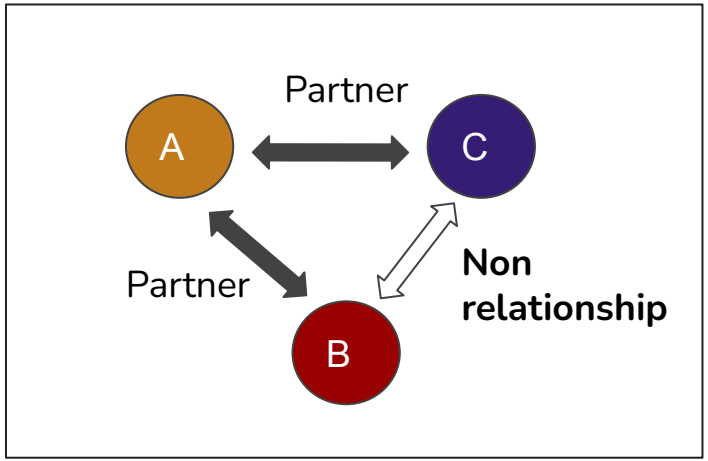
Add **non-relationship** data points to train model with the ability of identifying whether relationship exists

Definition of Non Relationship

Sample sentence from news

EzCare and BookingPal have both been highlighted as software platforms for their ability to link with Escapia.

Entities + Relationship



Training data provided by Arboretica

A	B	C	D	E	F
Company A	**Company B**	**Sentence**	**Url**	**Type**	**Degree**
Fortino Capital	Charles Souillard	As part of the transaction, Miguel Valdes and Charles Souillard, who founded Bo	https://finance.yahc	People	indirect
Fortino Capital	Miguel Valdes	As part of the transaction, Miguel Valdes and Charles Souillard, who founded Bo	https://finance.yahc	People	indirect
Fortino Capital	Autodesk	Belgium's Oqton scores $40 million to 'disrupt manufacturing' with a cloud-base	https://tech.eu/202	People	indirect
Fortino Capital	BE Semiconductor Industries	He currently serves as chairman of the Supervisory Board of BE Semiconductor I	https://www.ing.co	People	direct

Training data Columns

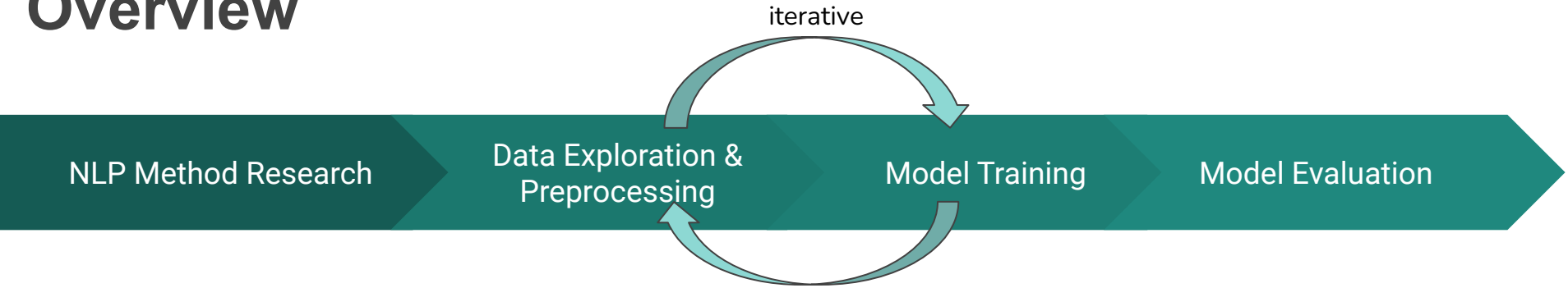
Scraped and annotated data

	B	C	D	E	F	G	H	I
	url	sentence	organization_a	organization_b	Financial	Technical	People	Partner
	https://shortterm	Canada: Property automation system Operto has partnered with lock innovator Yale to offer an automatically programmed Linus® Smart Lock that is connected to reservations in real-time, provides 24/7 check-in and enables keyless access for rental guests	Operto	Yale	0	1	0	0
	https://shortterm	The partnership has been forged to prevent time-consuming, and costly, key handovers, as well as reduce inconvenience for guests, especially in the times of a global pandemic						
	https://shortterm	The automated operations from Operto integrated with the Yale Linus® Smart Lock have been designed to free up guests' and property managers' time, offering them a secure and seamless rental experience	Operto	Yale	0	1	0	0

Model Training

Entity Recognition & Relationship Classification

Overview



Relevant Industry and Academic advanced NLP theories, algorithms and their code bases:

- **spaCy**
- **BERT**
- **RoBERTa**
- **LUKE**

- **<u>Data analysis:</u>** entities & relationships distribution
- **<u>Data augmentation:</u>** data scraping; manual data annotation

- Baseline model training for entity recognition: spaCy (RoBERTa)
- LUKE-based models: NER & Relationship classification

- F-1 Score
- Precision
- Recall

Baseline Model Training: spaCy (RoBERTa)

```
Company A: fortino capital
Company B: newion
Url: https://www.eu-startups.com/2021/07/luxembourg-base
vations-software/
Recognized entities: ecoligo,fortino capital,tattoodo,wi
team advertising our newsletter,gravitee.io,tripadmit,fl
-startups   copenhagen,salonkee
```

In the first phase, we are able to identify **71%** of the required company entities from the sentences using spaCy. But precision is only **29%**.

The problems we face:

- Identified too much - Excessive companies entities identified than required (false positives)
- Identified none - Some required companies entities are not identified by the algorithm.

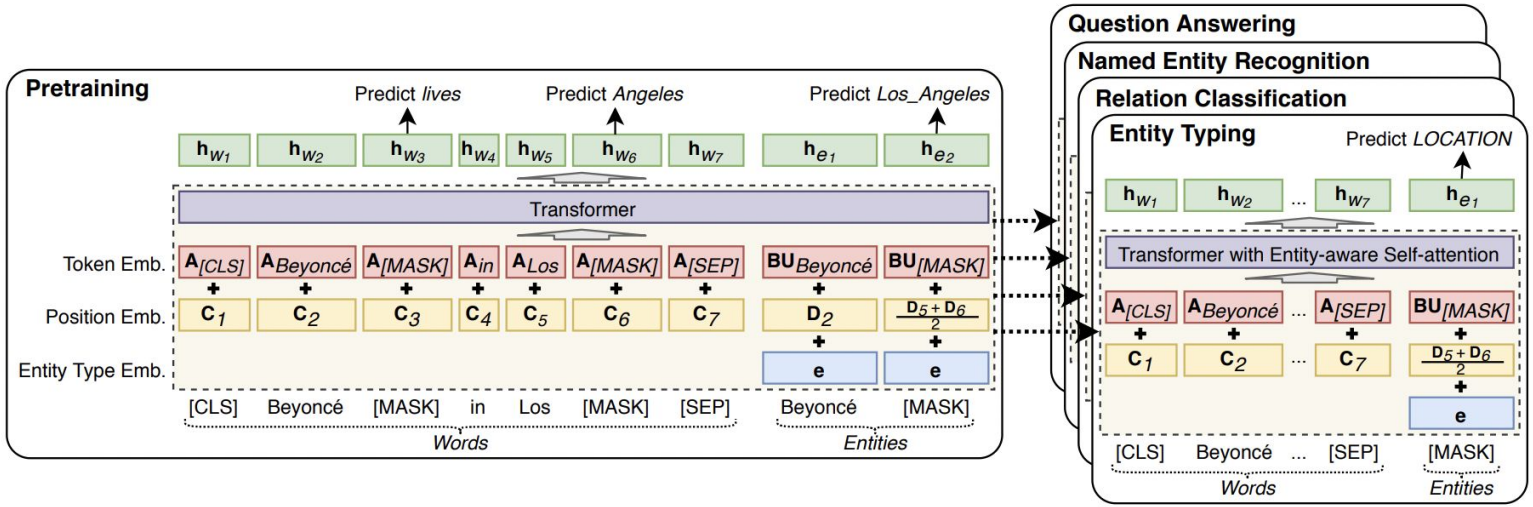
Source: *Our Github Repo: https://github.com/SilviaGu98/cmu_2022capstone_Arboretica*

Disadvantages of spaCy baseline model: not-customizable.

- spaCy **cannot fine-tune the model to customize for our need.**
  (We expect to always get many false positives and false negatives with spaCy)

- We want our training data to “teach” algorithms to **only find useful entities** and relationships and find it more **dynamically**.

LUKE (Language Understanding with Knowledge-based Embeddings)



"Luke." LUKE. Accessed October 22, 2022. https://huggingface.co/docs/transformers/model_doc/luke.

LUKE - a more advanced model

Named Entity Recognition

- Fine-tune pre-trained LUKE model with a span classification head on top
- LUKE solves the task by
  - enumerating all possible spans (or n-grams) in each sentence as entity name candidates
  - classifying them into the target entity type or non-entity type
- The model is trained using cross-entropy loss

LUKE - a more advanced model

Multi-label Relationship Classification

- Four types of relationships which can co-exist: **financial**, **technical**, **people** and **partner**
- LUKE Relation classification determines the correct relation between head and tail entities in a sentence
- A linear classifier based on a concatenated representation of the head and tail entities is used
- The model is trained using cross-entropy loss

Example of NER Prediction Result

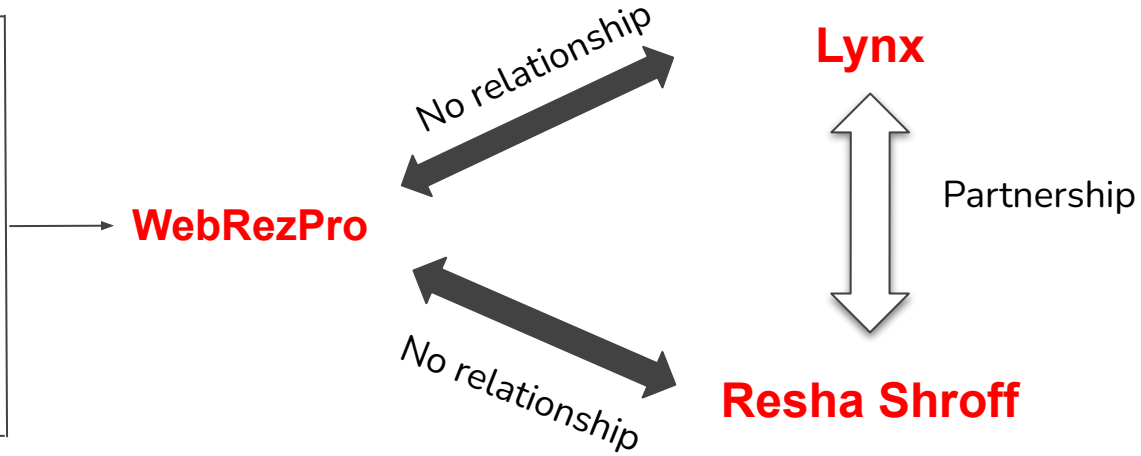
Lynx CEO **Resha Shroff** said:

"**WebRezPro** is a leading property management software and with this partnership, we are excited to extend the power of the Lynx platform to WebRezPro customers

In this sentence, three entities **Lynx**, **Resha Shroff** and **WebRezPro** were recognized by the algorithm

Example of Relationship Classification Prediction Result

Lynx CEO **Resha Shroff** said: "**WebRezPro** is a leading property management software and with this partnership, we are excited to extend the power of the Lynx platform to WebRezPro customers



Special case: only one entity recognized

Abdul Hadi Jameel, business manager at NZXT Middle East & Africa, says the partnership with **Rectron** will empower gaming enthusiasts in new markets to build computer systems which will help them pursue their passion for gaming.

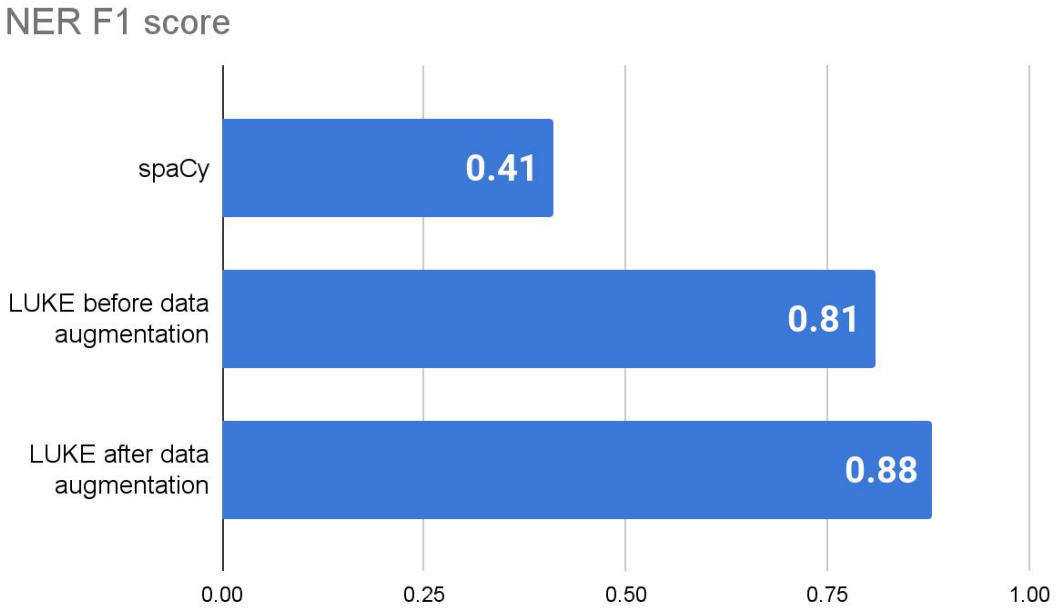
Only one entity **Rectron** was recognized. In this case, the sentence and entity will be **dropped**, since there are no paired companies.

Special case: indirect relationship

Macronix, Micron, and **SK Hynix** all recently applied to the U.S. Department of Commerce for licenses to sell electronic components to **Huawei**.

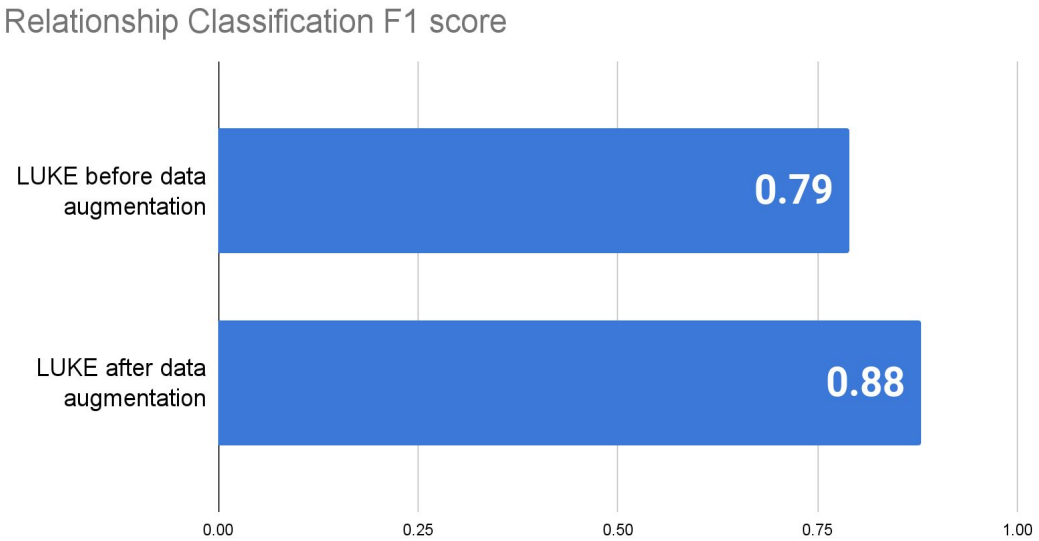
Macronix, Micron, and SK Hynix all have a direct relationship with Huawei. In this case, we would pair up **Macronix, Micron**, and **SK Hynix** to have **no direct relationships**

NER Evaluation Results



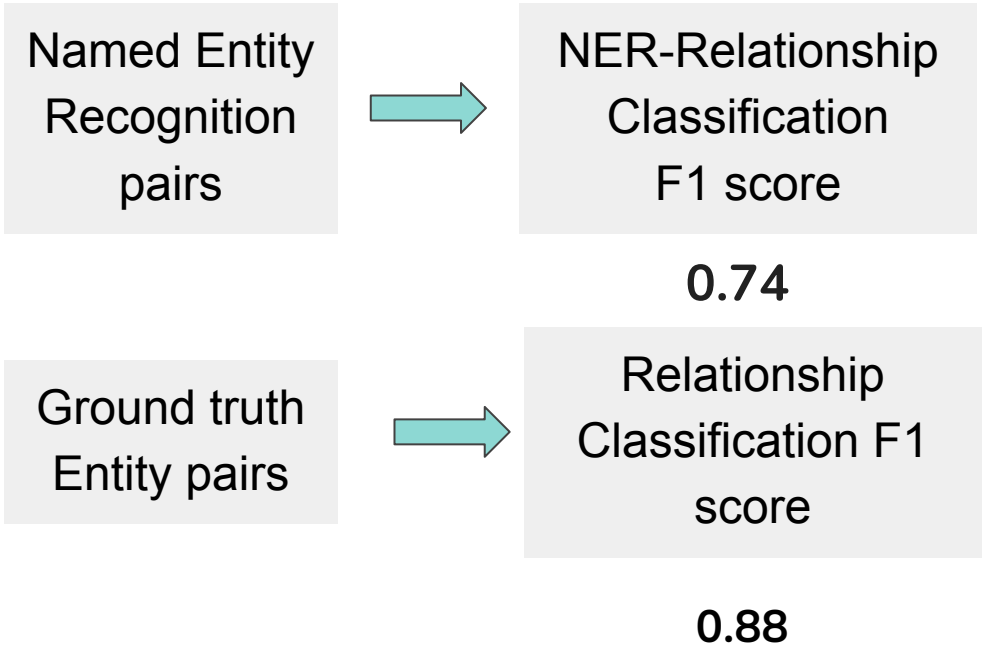
- NER F1 score nearly **doubled** when we updated the model from spaCy to LUKE.
- It also increased by **8%** after data augmentation.

Relationship Classification Evaluation Results



- Relationship classification F1 score increased by **11%** after data augmentation.

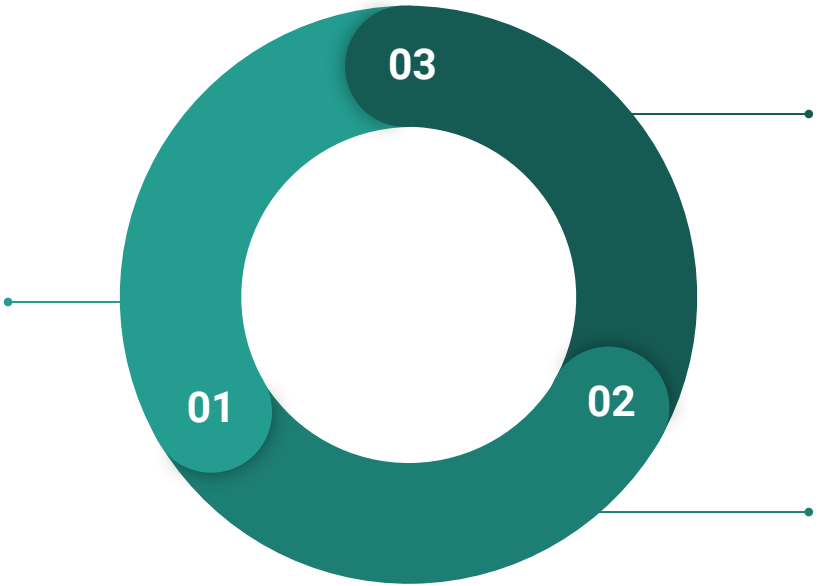
Relationship Classification Results



Evaluation Results :

Best Evaluation Result on Validation Dataset			
Task	F1	Precision	Recall
Named Entity Recognition (Spacy) [1]	0.4148	0.2938	0.7051
Named Entity Recognition	0.8076	0.7949	0.8213
Named Entity Recognition(+)[2]	0.8797	0.8468	0.9198
Relationship classification	0.7890	0.7989	0.7807
Relationship classification(+)	0.8807	0.8859	0.8757
NER-Relationship classification(+)	0.7361	0.7969	0.6945

Insights & Limitations



Limitations of Sentence-level Relation Extraction Models

Dataset Size and Coverage

Inefficiency of Span-level NER Models

Dataset Size and Coverage

- Adding our own annotated data to the client's dataset significantly boosts performance of both the NER model and the relationship classification model.
- A larger and more diverse dataset can contribute significantly to the model.
- Although we doubled the dataset size by collecting and annotating additional data, the augmented dataset is still very limited in size and scope.

Inefficiency of Span-level NER Models

- As a span-level model, LUKE considers all possible n-grams in each sentence as entity candidates (n does not exceed 16 in the original implementation; we cap n at 1 because the entities of interest in this project are never longer than 10 tokens).

- This approach is feasible but not computationally efficient. Ideally, most n-grams should not be considered entity candidates because they are not meaningful phrases.

Challenges of Sentence-level Relation Extraction Models

- In many cases, it is difficult to conclude a particular type of relation between two entities from one sentence alone. Often, we need more context from previous or following sentences to better understand if a relation exists between two entities and what type of relation it is.

- Sometimes one sentence refers to an entity by pronoun, and its full name appears in a separate sentence, where the other entity of interest is not mentioned.

Future Work

Data Augmentation

A large dataset that is representative of the real-world data is crucial to building a strong model.

- Build industry-specific dataset for intended use case.
- Apart from human annotation, consider using methods such as GPT-3 to increase data size. GPT-3 can be leveraged to paraphrase existing sentences.

arboretica

Sentence-level Model with Context

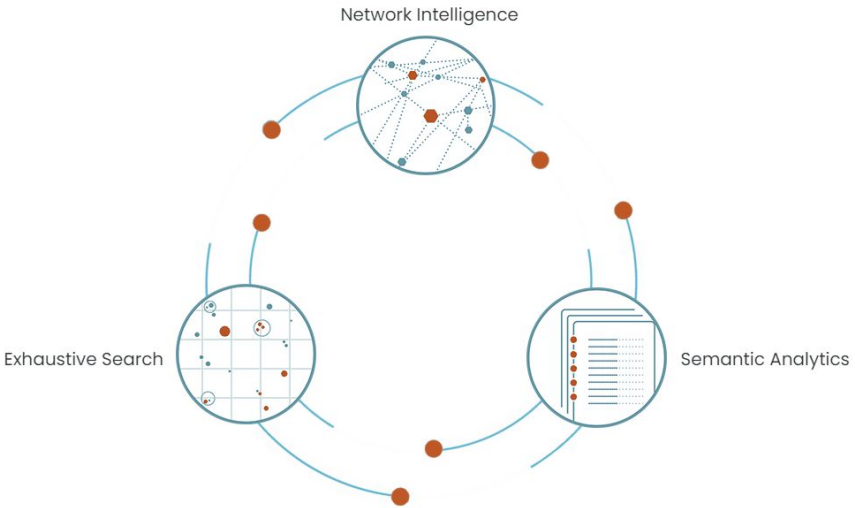
For future experiments, each sample can include the sentence of interest as well as a few sentences preceding and following it for context to utilize 512 sequence max-length in BERT-based models .

Span-level Models

Generate a set of entity candidates to have the model predict on, instead of passing in all n-grams. Semantic parsing can help determine which spans correspond to meaningful phrases.

Conclusion

Efficiency and Precision Gain



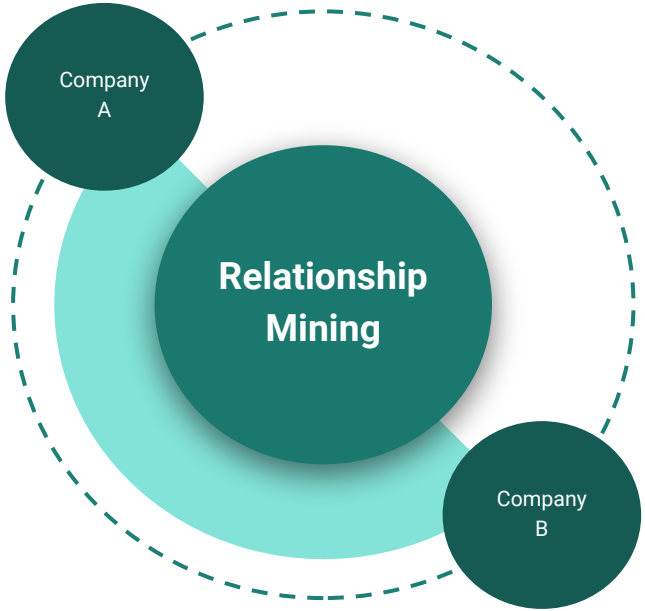
Increase Data Coverage

Decrease Processing Time

Enhance Relationship Extraction Accuracy

Source: *Advantage* (2022) *Arboretica*. Available at: https://www.arboretica.com/technology/#advantage.(Accessed: October 22, 2022).

Conclusion



- Model has great performance, significantly increased relationship extraction accuracy and decreased human post-processing.

- Please refer to our code base to see instructions and implementation details of our model pipeline.

Questions?

Contact our Project Manager at:
puxins@andrew.cmu.edu

Reference

Relationship Data Mining (2022) *Arboretica*. Available at: https://www.arboretica.com/blog/relationship-data-mining-2/ (Accessed: October 22, 2022).

AI industry intelligence (2022) *Arboretica*. Available at: https://www.arboretica.com/blog/ai-industry-intelligence/(Accessed: October 22, 2022).

Advantage (2022) *Arboretica*. Available at: https://www.arboretica.com/technology/#advantage.(Accessed: October 22, 2022).

spaCy. (2022, October 4). In Wikipedia. https://en.wikipedia.org/wiki/SpaCy

BERT (language model). (2022, August 31). In Wikipedia. https://en.wikipedia.org/wiki/BERT_(language_model)

"Overview of Roberta Model." GeeksforGeeks, June 23, 2022. https://www.geeksforgeeks.org/overview-of-roberta-model/.

"Luke." LUKE. Accessed October 22, 2022. https://huggingface.co/docs/transformers/model_doc/luke.

Tian, Yuanhe, Guimin Chen, Yan Song, and Xiang Wan. "Dependency-Driven Relation Extraction with Attentive Graph Convolutional Networks." ACL Anthology, 2021.

Wu, Shanchan, and Yifan He. "Enriching Pre-Trained Language Model with Entity Information for Relation Classification." arXiv.org, May 20, 2019. https://arxiv.org/abs/1905.08284.

Z. Zhou and H. Zhang, "Research on Entity Relationship Extraction in Financial and Economic Field Based on Deep Learning," *2018 IEEE 4th International Conference on Computer and Communications (ICCC)*, 2018, pp. 2430-2435, doi: 10.1109/CompComm.2018.8780966.