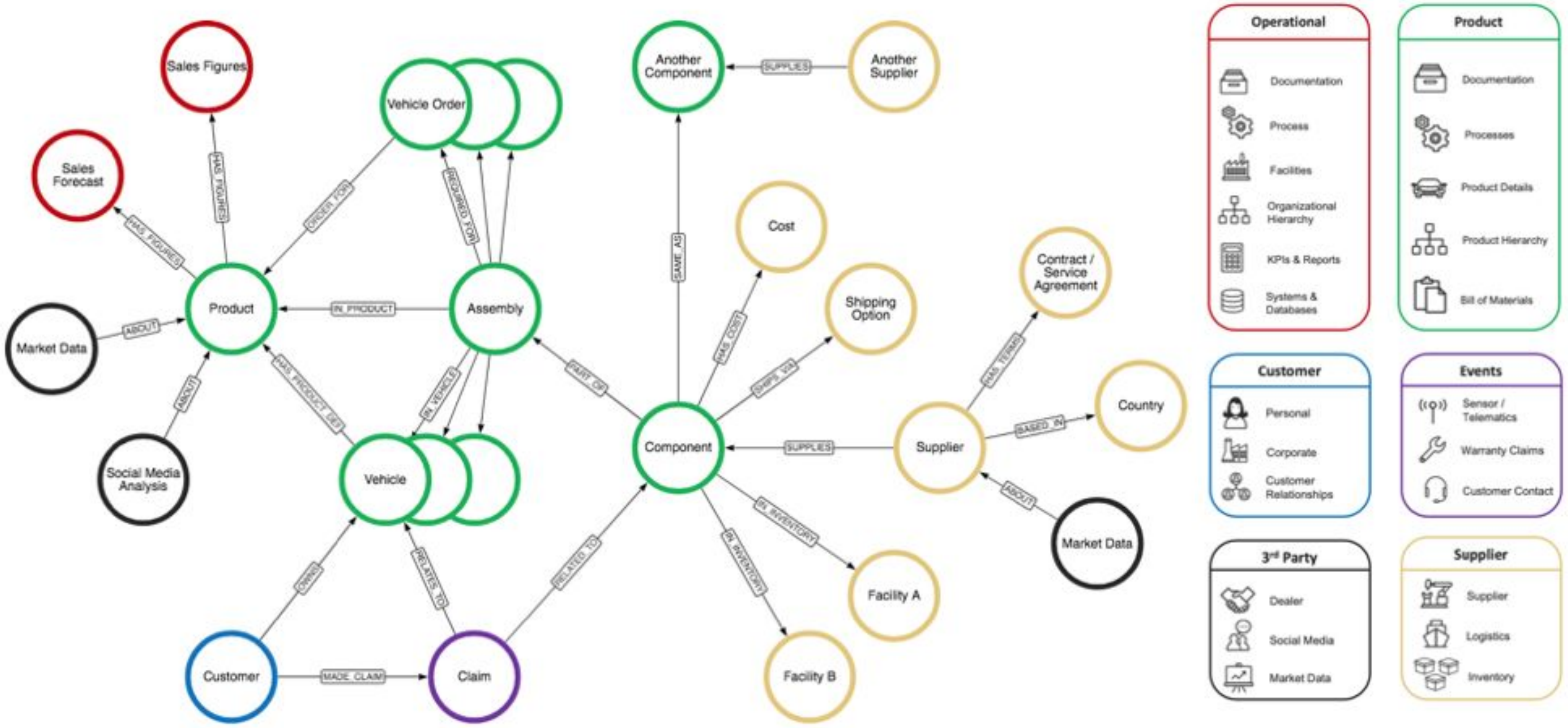
Distantly-supervised Supply Chain Extraction from News Data by combining Large Language Models and Data Programming

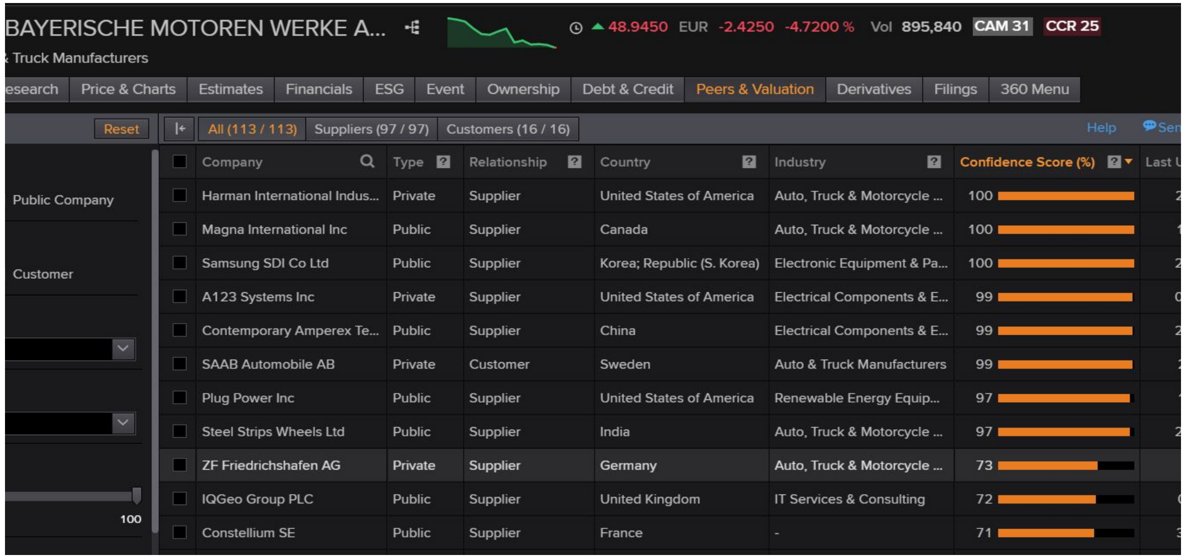


Source: https://neo4j.com/press-releases/neo4j-for-supply-chain-analytics/

Overview of the Idea

Use small existing KB to generate weak labels → Use weak label to train a classifier on news data → Use the trained classifier to identify new relationships in unseen news data

Knowledge Base of Relations



Company	Type	Relationship	Country	Industry	Confidence Score (%)
Harman International Indus...	Private	Supplier	United States of America	Auto, Truck & Motorcycle ...	100
Magna International Inc	Public	Supplier	Canada	Auto, Truck & Motorcycle ...	100
Samsung SDI Co Ltd	Public	Supplier	Korea; Republic (S. Korea)	Electronic Equipment & Pa...	100
A123 Systems Inc	Private	Supplier	United States of America	Electrical Components & E...	99
Contemporary Amperex Te...	Public	Supplier	China	Electrical Components & E...	99
SAAB Automobile AB	Private	Customer	Sweden	Auto & Truck Manufacturers	99
Plug Power Inc	Public	Supplier	United States of America	Renewable Energy Equip...	97
Steel Strips Wheels Ltd	Public	Supplier	India	Auto, Truck & Motorcycle ...	97
ZF Friedrichshafen AG	Private	Supplier	Germany	Auto, Truck & Motorcycle ...	73
IQGeo Group PLC	Public	Supplier	United Kingdom	IT Services & Consulting	72
Constellium SE	Public	Supplier	France	-	71

Source: Refinitiv Workstation

Corpus of News Data

TECHNOLOGY NEWS MARCH 10, 2021 / 2:55 AM / UPDATED A YEAR AGO

Exclusive: LG hopes to make new battery cells for Tesla in 2023 in U.S. or Europe - sources

By Hyunjoo Jin 4 MIN READ

SAN FRANCISCO (Reuters) - LG Energy Solution aims to build advanced battery cells for Tesla Inc electric vehicles in 2023 and is considering potential production sites in the United States and Europe, two people familiar with the matter told Reuters.

Source: Reuters.co.uk

Motivation / Research Question

Motivation & Relevant Work

- Recent events, such as the global **COVID-19** pandemic and **sanctions** against the Russian Federation, as well as the relevant literature have identified the **importance of mapping supply chain** relationship between companies
  - *Goh et. al. (2009), Dai et. al. (2021), Coqueret and Tran (2022), Wichmann et. al. (2018)*
- However, only little research exists on the **creation** and **maintenance/extension** of supply chain data sets
  - One key study (Wichmann et al. (2020)) uses **manually labelled training data**: Not scalable
- We gather a **novel supply chain data** and train a NLP model to **extract supply chains** using a **distant supervision** approach

Research Question

- Is supply chain mapping from news data using distant supervision (i.e. noisy labels) possible?
- Can model performance be improved by injecting knowledge into the model using data programming (i.e. systematic labelling based on heuristics)?

Contribution to the Literature

Provide a novel data set that can be used to study supply chain mapping using distant supervision

Validate that distant supervision with transformer-based large language models can be effective for supply chain mapping

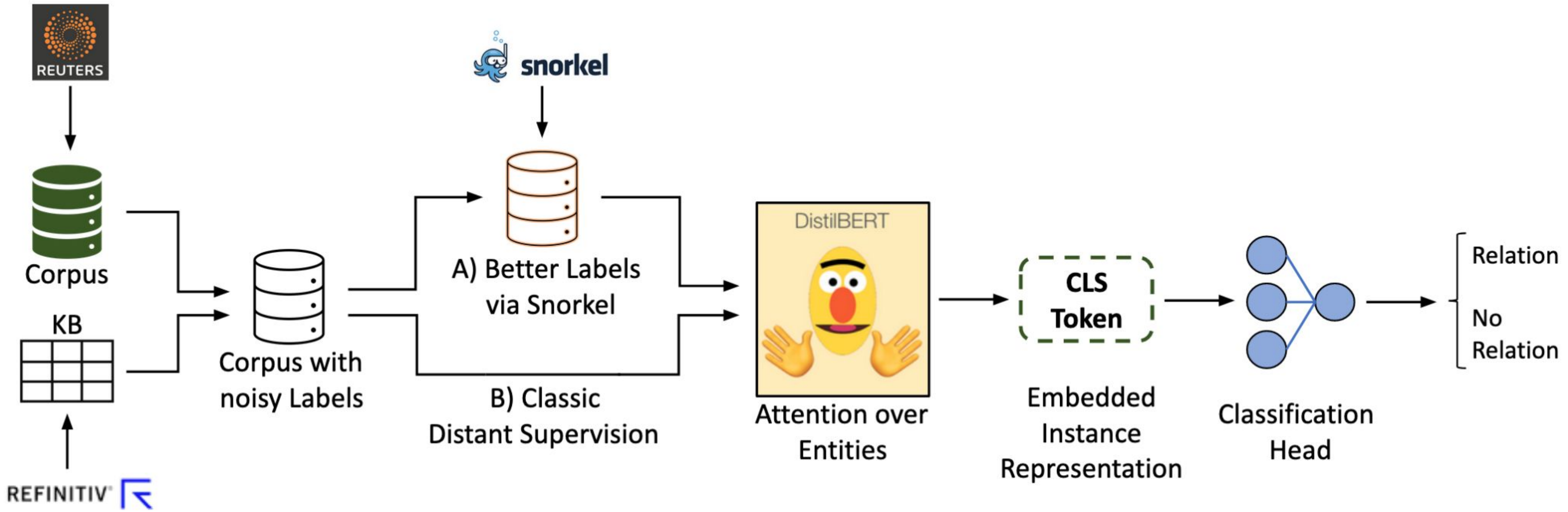
Investigate data programming can be an effective method to tackle the issue of noisy labels in the Distant Supervision paradigm to add domain-specific knowledge.

Overview of Research

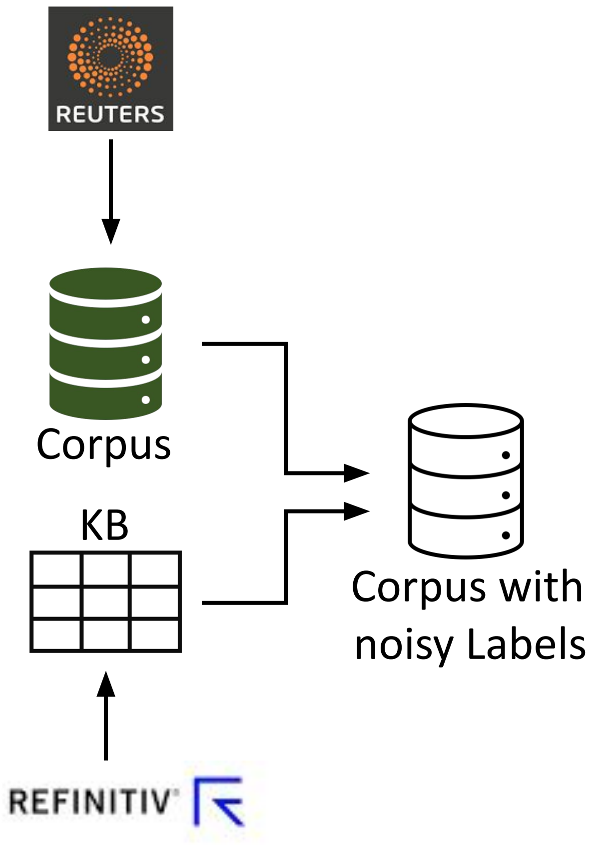
Creation of Novel Data Set of Supply Chain relations

Snorkel: Data Programming to improve weak labels produced by Knowledge Base

Predicting supplier relations using DistilBERT CLS Token



Supply Chain Relation Data Set



Data Retrieval

- ➔ Using ***Refinitiv***, create a Knowledge Base (KB) of supplier relations between pairs of companies:

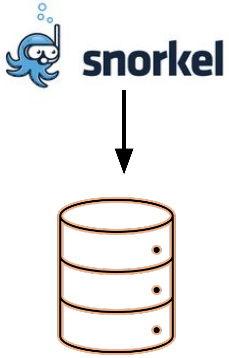
TSMC (supplier of) Apple Inc

- ➔ From ***Reuters,*** obtain 40,000 articles to create a corpus of news data containing entities in our corpus
- ➔ Extensive data processing to prepare corpus for NLP task

Label Creation

- ➔ If an instance in the corpus contains a pair of entities in the KB, then this instance is assumed to express the supplier relation
- ➔ Induces the *wrong labelling problem*
  - ◆ **Positive Rate: 57.7%**

Data Programming to Improve Noisy Labels



Labeling Functions

- Many functions to be applied each instance
- Each returns an **abstain**, **not specified** or **supplier** label
  - All our functions **abstain** if certain criteria not satisfied
- Combine output of all labelling functions into single probabilistic label (via Snorkel generative label model)
  - Follows a theoretically grounded methodology first proposed by Ratner et al (2016)

24 Functions

- 14 search words
  - "supplies", “supplier”, "customer", “client”, "buys”, … **supplier** label if word(s) found
- 7 count occurrence of certain characters
  - *, %, $, -, Q, … **not specified** if count of characters above a threshold
- 2 use KB
  - Entity pair not in KB: **not specified**. Confidence score (provided by KB) above 99.5%: **supplier**
- 1 attribute of instance
  - More than 5 companies from KB mentioned: **not specified**

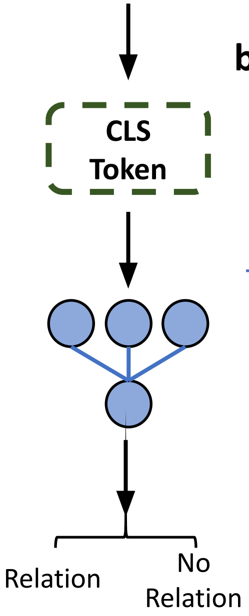
Synthetic Data to Improve Noisy Labels

Data Augmentation

- Using the probabilistic labels we generate 2,537 augmented positive instances
  - **Used rules:** Replace a random noun, verb or adjective with a synonym
- Example I:
  - **Original:** ‘Huawei suppliers Intel *rose* 0.1%, while Micron gained 4%.’
  - **Augmented:** 'Huawei suppliers Intel *rise* 0.1%, while Micron gained 4%.’
- Example II:
  - **Original:** ‘Nokia's *phone* is sold at AT&T stores.’
  - **Augmented:** ‘Nokia's *telephone* is sold at AT&T stores.’

Relation Extraction Model

Overview of Models



DistilBERT base-uncased

Pre-trained language model can help in recognizing a broader set of relations

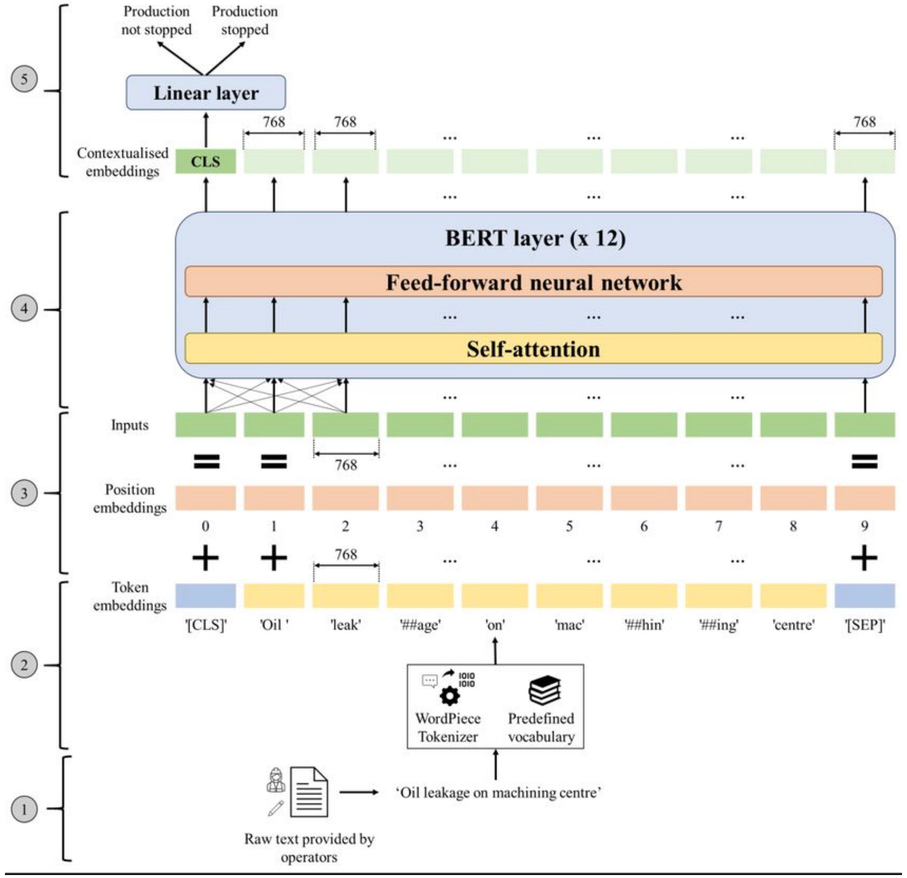
Bi-directional nature of model allows for token representations conditioned on contextual words

Bag of words

Treated as a benchmark model, since used in prior work

Cannot understand sentence structure or context like a transformer model

DistilBERT for Classification



Experiments

Experiment Overview

Standard Set Up: KBL

Train on KB labels	No Data Augmentation

Probabilistic labels from Label Model: PL

Train on Probabilistic labels	No Data Augmentation

Synthetic Positive Examples: AUG

Train on Probabilistic labels	Add 2,537 Augmented Positive Examples

Evaluation Strategy

Manual Evaluation

Based on 1,113 manually labelled "gold labels" randomly chosen from the Test Set

Evaluates the model's ability to predict if a supply chain relationship is mentioned

Automatic Evaluation

Based on 5,346 automatically labelled instances of the Test Set (KB induced Label)

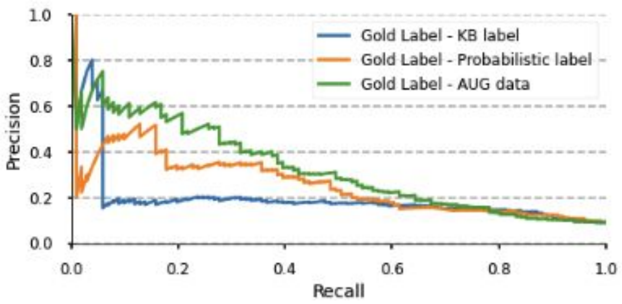
Evaluates the model's ability to predict if a supply chain relationship exists

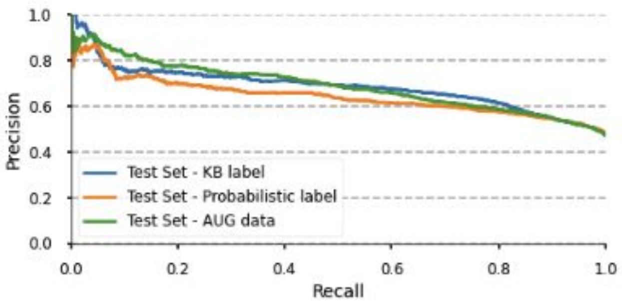
Results

Results

AUC	KBL	PL	AUG
Gold Labels	0.70	0.73	**0.74**
Test Set	**0.73**	0.69	0.72
P@100	**KBL**	**PL**	**AUG**
Gold Labels	0.18	0.34	**0.40**
Test Set	0.91	0.85	**0.91**

Table 3: Evaluation - BERT Language Model





Key Findings

Manual Evaluation:	Injecting knowledge into the model by using automatic labels created with heuristics improves the model's ability to predict if text *mentions* a supply chain relationship
	Adding synthetic positive examples to the training data further improves the model's precision on that task
Automatic Evaluation:	Using noisy labels, the model has relatively high precision in predicting if a (supply chain) relationship *exists* between two entities
	Using less noisy labels and additional positive examples does not further improve the model's ability to predict on this task

Case Study / Further Research

Case Study

(1) All labels are in line. Simple for the model to predict
(2) Two related entities are mentioned but **not** their relationship. The Label Model accounts for this.
(3) The Label Model is not able to account for complex multi entity instances.

Instance	KB	PL	GD
(1) **LG Chem**'s wholly owned battery subsidiary, LG Energy Solution, an EV battery supplier to **Tesla** and GM, praised the ruling on Wednesday.	1	1	1
(2) It also approved a huge new wholly-owned Shanghai factory for U.S. electric car maker **Tesla**, and a \$2.3 billion joint venture organic light-emitting diode plant to be built by South Korea's **LG Display**.	1	0	0
(3) Panasonic has been the exclusive battery cell supplier for **Tesla**, but the U.S. electric vehicle maker is in advanced talks with South Korea's **LG Chem** as it seeks to diversify sources of the key component.	1	1	0

* *KB = KB label, PL = probabilistic label, GD = gold label*

Further Research

Model multiple relationships between

- Currently multiple relationships (e.g. Owner, joint venture same Sector,) are absorbed into the KB Label
- Having a KB with several relationships would allow the model to distinguish between them
- For example negative examples could be generated using competitors that are unlikely to supply one another

Predict at the "bag level"

- Combining several instances as input reduces the risk of false positives
- This would bring this research closer in line to common approaches such as Riedel et. al (2010), Hoffmann et. al (2011) and Christou and Tsoumakas 2021)

Improved Label Model

- Further improving the label functions used in the Label Model could further reduce the noise in the labels and thus improve model performance
- For example, multi entity instances could be accounted for (see Case Study 3)

Thank you for your attention!