Financial Documents Processing with NLP

Markus Leippold, Qian Wang

Nov, 2020

- Too much unstructured data
  - From clients: requests, contracts, etc.
  - From media: news, blog, etc.
  - From firms: financial filings, announcement, etc.
  - ...
- Requirements for decision making:
  - Efficiency
  - Speed
  - Consistency
  - Accuracy
  - Depth

Top Use Cases

- Chatbot: intent parsing, question answering
  - NLU or semantic parsing to know the query and locate the relevant data
  - NLG or reading comprehensive task

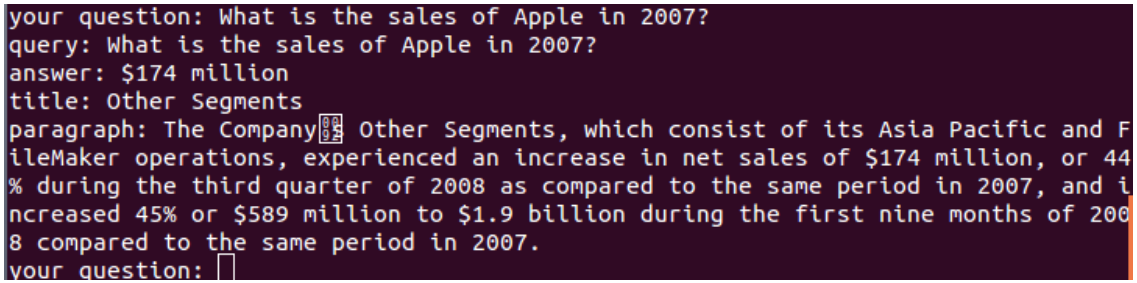

Figure: Question Answer Based on Financial Filings

- Chatbot: intent parsing, question answering
- Contract Analytics: entity recognition
- Sentiment Analysis: text classification

9-layer Deep Model				
Company	Text	Predict	Sentiment	Error
Glencore	Glencore shares in record crash as profit fears grow	-0.548	-0.971	-0.423
Barclays	London open: Taylor Wimpey and Ashtead drive markets higher, Barclays falls	0.348	-0.657	-1.004
Weir Group	Slump in Weir leads FTSE down from record high	-0.269	-0.827	0.558
AstraZeneca	News FeedFTSE 100 movers: Standard Chartered lifted while AstraZeneca sinks	0.0673	-0.666	-0.733
Lloyds Banking Group PLC	Lloyds Banking Group reports 7% dip in annual profits	-0.175	-0.696	-0.521

Table: Examples of Sentiment Labels and Prediction

- Document Retrieval and Semantic Search: text similarity, summarization, relationship extraction

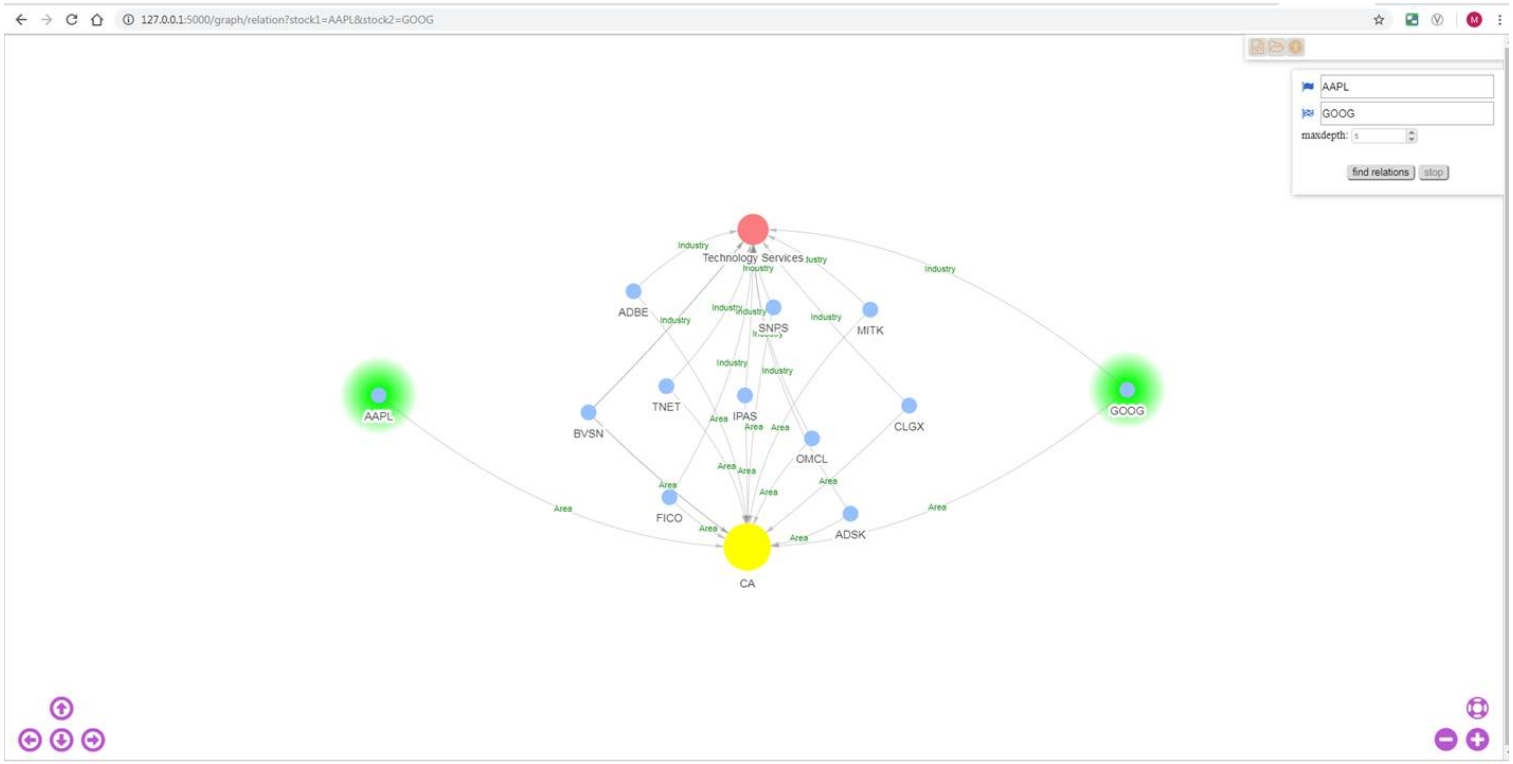

Figure: Knowledge Graph for Company Relationship

Available Resources

- Raw Text
  - Financial filings (e.g. EDGAR)
  - Financial news (e.g. WSJ, NYTimes)
  - Market Forums (e.g. Stockaholics)
  - Social Media (e.g. Twitters)
- Labeled Data
  - Sentiment Analysis (market sentiment): SemEval-2017 Task 5
  - Summarization (Financial Filings): FNS 2020, FinTOC 2020
  - Named Entity Recognition: Alvarado et al. (2015)

Scarcity in Labeled Data

- Utilize pretrained language model
  - First learn the language (English), then learn the financial tasks
  - BERT, XLNet, GPT-3
- Domain adaption
  - Learn the tasks as well as the analogy between domains

Utilize Pretrained Language Model

- Kölbel et al. (2020)
  - Extract all the sentences in a financial filing that are relevant to climate risks
  - Analyze the relation between the climate risks and the asset returns
- The pipeline:
  - Build a language model on financial texts: train BERT on a large corpus including financial filings, company conference transcripts, financial news, etc.
  - Manually label a small dataset and iteratively expand it
  - Fine-tuning the BERT model

- Sentence Classification

Figure: Neuron View

- Force the model to learn the tasks while find the analogy between domains at the same time

$$\min_{f \in \mathscr{F}} \epsilon_{\mathbb{P}}(f) + \hat{d}(\mathbb{P}, \mathbb{Q})$$

$$\max_{f' \in \mathscr{F}} \hat{d}(\mathbb{P}, \mathbb{Q})$$

- Lead to some adversarial learning algorithm

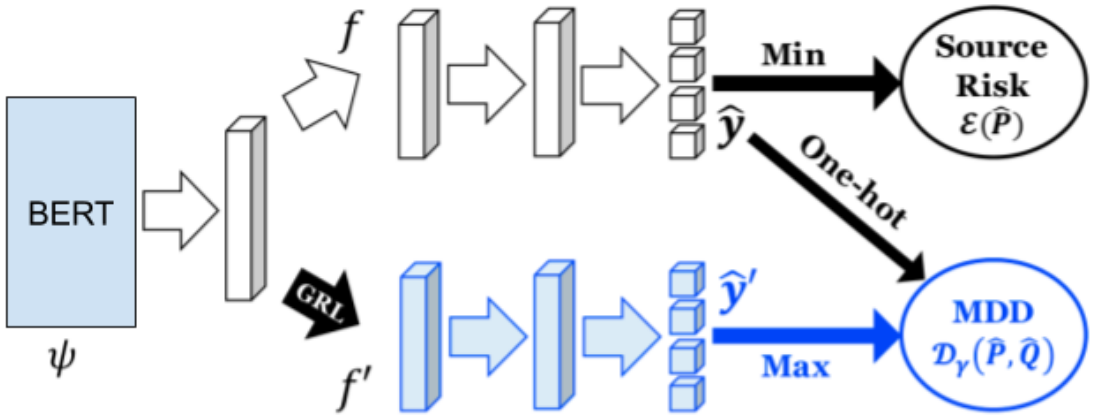

Figure: Adversarial Network

- Experiment Results (training on Amazon product reviews and test on financial news)

	Baseline	MDD-1	MDD-2	MDD-3
Acc	0.78	0.84	0.85	0.87
F1	0.81	0.88	0.89	0.91

- When training on Amazon product reviews and test on movie reviews, MDD shows similar performance with baseline (BERT generalizes quite well)

Reference I

Alvarado, J. C. S., Verspoor, K., & Baldwin, T. (2015). Domain adaption of named entity recognition to support credit risk assessment. In *Proceedings of the australasian language technology association workshop 2015* (pp. 84–90).

Kölbel, J. F., Leippold, M., Rillaerts, J., & Wang, Q. (2020). Ask bert: How regulatory disclosure of transition and physical climate risks affects the cds term structure.