Extracting Knowledge Graphs from Financial Filings

Extended Abstract

Jay Pujara
University of California
Santa Cruz, CA 95064
jay@cs.umd.edu

ABSTRACT

Textual corpora, such as financial documents, contain a wealth of knowledge. Recently, knowledge graphs have become a popular approach to capturing structured knowledge of entities and their interrelationships. In this paper, we evaluate open information extraction (IE) and knowledge graph construction techniques for assessing the relevance of textual segments in the Financial Entity Identification and Information Integration Challenge. Our approach is to extract several textual signals, including topics and open IE triples, and combine these in a probabilistic framework to predict the relevance of each potential relationship.

ACM Reference format:
Jay Pujara. 2017. Extracting Knowledge Graphs from Financial Filings. In *Proceedings of DSMM'17, Chicago, IL, USA, May 14, 2017,* 2 pages.
https://doi.org/http://dx.doi.org/10.1145/3077240.3077246

1 MOTIVATION

Financial filings can contain a wealth of knowledge, capturing relationships between financial entities and organizations [2], detailing product offerings [5], specifying policies and restrictions for monetary flows between entities, and providing details about the organizational and governing structure of corporations. Unfortunately, accessing this knowledge can be difficult, as financial filings are long, technical documents which contain a vast amount of text. One approach to improving the accessibility of the information in a large document collection is constructing a knowledge graph, a structured representation of entities and the relationships between them [4]. In this work, we present preliminary results describing an approach to knowledge graph construction on financial filings, and identifying several challenges and possible solutions.

2 INFORMATION EXTRACTION FROM FINANCIAL DOCUMENTS

In many knowledge graph construction pipelines, the initial step is information extraction (IE). IE translates unstructured text into candidate structural relationships between entities. In many cases, IE systems are built for a particular purpose or topic, and trained using a known ontology of entity attributes and relationships [8]. In contrast to such ontology-based IE systems, "open" IE does not have a fixed ontology and specifies the relationships between entities in terms of the raw textual structure of the document [6]. For example, given the sentence

```
For example, because Ally controls Ally Bank, Ally is an
affiliate of Ally Bank for purposes of the Affiliate
Transaction Restrictions.
```

an open IE system will generate structured extractions: *(Ally; controls; Ally Bank); (Ally; is an affiliate of; Ally Bank)* , both of which are useful additions to a knowledge graph that captures relationships between Ally and Ally Bank. However, for the more complex sentence

```
Furthermore, there is an "attribution rule" that provides
that a transaction between Ally Bank and a third party
must be treated as a transaction between Ally Bank and a
nonbank affiliate to the extent that the proceeds of the
transaction are used for the benefit of, or transferred
to, a nonbank affiliate of Ally Bank.
```

Open IE systems have difficulty extracting a meaningful set of relationships, and produce relationships such as *(the proceeds of the transaction; are used for; the benefit of, or transferred to, a nonbank affiliate of Ally Bank).* This extraction does not represent a crisp of meaningful representation of the associated textual content and is unlikely to be considered relevant by an end-user.

3 PROPOSED MODELING APPROACH

Our goal is to exploit the flexibility of open IE systems that can adapt to a wide variety of text, while avoiding the pitfall of generating noisy textual relationships that would impede understanding financial documents. To accomplish this goal, we propose to use corpus-level textual signals, such as topics, term counts, and term salience (via TFIDF statistics), to re-rank the extracted triples and capture the most meaningful extractions. Re-ranking extracted triples requires fusing statistical signals with semantic relationships, we plan to use the approach of Pujara et al. [7], which constructs a probabilistic model over possible knowledge graphs using evidence from textual signals and ontological relationships using the probabilistic soft logic (PSL) modeling framework [3]. In the subsequent sections, we illustrate the textual signals our system will exploit, provide an overview of the knowledge fusion approach, and describe a PSL model that can represent textual signals and implement the knowledge fusion across extracted triples.

DSMM'17, May 14, 2017, Chicago, IL, USA

ACM ISBN 978-1-4503-5031-0/17/05…$15.00
https://doi.org/http://dx.doi.org/10.1145/3077240.3077246

Topic 0	ge northern receivables comerica gecc energy corporations bancorporation suntrust noninterest discontinued nonaccrual trusts fte
Topic 1	card companys receivables discover cards american network express member student merchant percent merchants contents
Topic 2	bb noninterest contents leases companys shareholders thousands tier huntington fdic alll ts nonperforming nonaccrual
Topic 3	citi firm citigroup firms citis sachs goldman level3 contents condensed client vies gs america
Topic 4	card dollars percent ally nonperforming leases securitization pci receivables warranties representations securitizations basel america
Topic 5	family mae fannie freddie multifamily mac fhfa guaranty trusts conservatorship foreclosure sheets book servicers
Topic 6	companys clients mellon morgan basel client bny stanley tier schwab york var contents intangible
Topic 7	contents dollars condensed companys securitization hedges lien quarters vies reflecting junior 143 securitizations prime
Topic 8	bancorp pnc noninterest bancorps leases nonperforming contents fifth nonaccrual alll condensed card percent visa
Topic 9	firm firms jpmorgan chase card pages chases pci wholesale basel predominantly nonaccrual client securitization

Table 1: Topics learned from 648 10-K and 10-Q filings from financial corporations

4 TOPIC MODELS

One promising direction that we explore in our preliminary work is applying topic models. We first learn a topic model from a large corpus of financial filings, and subsequently apply the learned model to estimate the topic coherence of a particular extraction. Our hypothesis is that those triples featuring terms that have a high probability in one or more topics will be associated with entities or relationships that are important for constructing a knowledge graph. In Table 1 we show examples of learned topics, listing the top fifteen terms from each of ten topics[1].

5 KNOWLEDGE FUSION OVERIVEW

Topic models are one example of a useful textual signal for reranking triples. In order to integrate multiple textual signals, we plan to develop a model for knowledge fusion from complementary sources of statistical signals. Since extracted knowledge is formulated as subject, predicate, object triples, we introduce a separate set of features for subject and object entities and predicates specifying relationships between these entities. This approach allows our system to use diverse sources such as named entity recognition systems for subject entities. The fusion model will combine these sources for each triple and provide a score for the relevance of the extraction. Using training data, the value of each input feature can be learned.

6 PSL MODEL

Probabilistic soft logic is a modeling framework that allows probabilistic relationships between variables to be specified using a series of first-order logic rules. Using PSL, we can specify the knowledge fusion model described in the previous section. For each triple subject, predicate, and object, each source of statistical relevance can be associated with a corresponding rule in the PSL model. In addition, extractions from a named entity recognition system can provide a coarse-grained set of selectional preferences for each set of relationships, allowing rules to enforce stricter constraints for meaningful relationships. Overlaps in arguments between relationships can similarly introduce dependencies between pairs of relationships. We provide examples of PSL rules that capture these modeling ideas below.

[1]Topics learned on 648 financial filings provided by IBM

$$\textsc{SubjTopic}(E1, T) \wedge \textsc{OIETriple}(E1, R, E2) \rightarrow \textsc{RelevantTriple}(E1, R, E2) \quad (1)$$

$$\textsc{SalientVerb}(R) \wedge \textsc{OIETriple}(E1, R, E2) \rightarrow \textsc{RelevantTriple}(E1, R, E2) \quad (2)$$

$$\textsc{NERType}(E1, T) \wedge \textsc{RelDomain}(R, T) \wedge \textsc{OIETriple}(E1, R, E2) \rightarrow \textsc{RelevantTriple}(E1, R, E2) \quad (3)$$

Rule 1 captures an interaction between the probability of a subject term in a topic model and the relevance of the triple. Rule 2 uses a salience measure of a verb, such as TF-IDF, to support triple relevance. Finally, Rule 3 combines a named-entity recognition system with a learned type affinity for a particular relationship.

7 CONCLUSION AND FUTURE WORK

In this abstract, we highlight several preliminary results from applying open IE and topic models to financial filings. We propose an overarching vision for capturing relevant facts in a knowledge graph of financial entities and relationships using these types of textual signals. Our model can be decomposed into a series of lower level textual statistics and ontological information, which is then integrated using a knowledge fusion model specified using probabilistic soft logic. In future work, we plan to implement the remaining components of this model.

REFERENCES

[1] 2014. *DSMM'14: Proceedings of the International Workshop on Data Science for Macro-Modeling.*

[2] 2016. *DSMM'16: Proceedings of the Second International Workshop on Data Science for Macro-Modeling.*

[3] S. H. Bach, M. Broecheler, B. Huang, and L. Getoor. 2015. Hinge-Loss Markov Random Fields and Probabilistic Soft Logic. arXiv:1505.04406 [cs.LG] (2015).

[4] Xin Dong, Evgeniy Gabrilovich, Geremy Heitz, Wilko Horn, Ni Lao, Kevin Murphy, Thomas Strohmann, Shaohua Sun, and Wei Zhang. 2014. Knowledge Vault: A Web-Scale Approach to Probabilistic Knowledge Fusion. In *KDD*.

[5] Gerard Hoberg and Gordon Phillips. 2016. Text-based network industries and endogenous product differentiation. *Journal of Political Economy* 124, 5 (2016), 1423–1465.

[6] Mausam, Michael D. Schmitz, Robert E. Bart, Stephen Soderland, and Oren Etzioni. 2012. Open Language Learning for Information Extraction. In *EMNLP*.

[7] Jay Pujara, Hui Miao, Lise Getoor, and William W. Cohen. 2015. Using Semantics & Statistics to Turn Data into Knowledge. *AI Magazine* 36, 1 (2015), 65–74.

[8] Daya C. Wimalasuriya and Dejing Dou. 2010. Ontology-based Information Extraction: An Introduction and a Survey of Current Approaches. *J. Information Science* 36, 3 (2010).