The Challenges of Creating, Maintaining and Exploring Graphs of Financial Entities

Michael Loster, Tim Repke, Ralf Krestel, Felix Naumann
Hasso Plattner Institute
Potsdam, Germany
firstname.lastname@hpi.de

Jan Ehmueller, Benjamin Feldmann
Hasso Plattner Institute
Potsdam, Germany
firstname.lastname@student.hpi.de

Oliver Maspfuhl
Commerzbank
Frankfurt am Main, Germany
oliver.maspfuhl@commerzbank.com

1 OVERVIEW & MOTIVATION

The integration of a wide range of structured and unstructured information sources into a uniformly integrated knowledge base is an important task in the financial sector. As an example, modern risk analysis methods can benefit greatly from an integrated knowledge base, building in particular a dedicated, domain-specific knowledge graph. Knowledge graphs can be used to gain a holistic view of the current economic situation so that systemic risks can be identified early enough to react appropriately. The use of this graphical structure thus allows the investigation of many financial scenarios, such as the impact of corporate bankruptcy on other market participants within the network. In this particular scenario, the links between the individual market participants can be used to determine which companies are affected by a bankruptcy and to what extent.

We took these considerations as a motivation to start the development of a system capable of constructing and maintaining a knowledge graph of financial entities and their relationships. The envisioned system generates this particular graph by extracting and combining information from both structured data sources such as Wikidata and DBpedia, as well as from unstructured data sources such as newspaper articles and financial filings. In addition, the system should incorporate proprietary data sources, such as financial transactions (structured) and credit reports (unstructured). The ultimate goal is to create a system that recognizes financial entities in structured and unstructured sources, links them with the information of a knowledge base, and then extracts the relations expressed in the text between the identified entities. The constructed knowledge base can be used to construct the desired knowledge graph. Our system design consists of several components, each of which addresses a specific subproblem. To this end, Figure 1 gives a general overview of our system and its subcomponents.

2 INGESTING STRUCTURED DATA TO THE KNOWLEDGE BASE

In order to build up a knowledge base we start by integrating heterogeneous structured data sources, such as Wikidata or DBpedia. The main challenge is to resolve entities, i.e., to recognize different representations of the same company, person, product, etc. As a starting point for our knowledge base construction, we use a manually curated dataset of our industrial partner and merge it with all other structured data sources. During this merge process, found matches are used to enrich the existing entities in the knowledge base with new information, while non-existing entities are added to the knowledge base. Currently, we use a combination of traditional similarity metrics, such as MongeElkan [3] and JaroWinkler [4], for the deduplication process and are developing a novel method based on neural networks that can be easily swapped in due to the modular system architecture.

3 KNOWLEDGE BASE ENRICHMENT USING UNSTRUCTURED DATA

The knowledge base generated from structured sources is further enriched by information extracted from unstructured data sources, such as newspaper articles and bank-internal documents. To extract the desired information, we apply natural language processing techniques to each unstructured data source. The text mining component used for this purpose consists of three subcomponents: named entity recognition, entity linking, and relation extraction.

Named Entity Recognition. The named entity recognition (NER) module is responsible for discovering and extracting mentions of financial entities from texts. The extraction technique used for this purpose was developed to recognize company names in texts and uses conditional random fields (CRF) classifier to do so [2]. A key advantage of this method is the possibility to integrate external knowledge in the form of manually created dictionaries into the training process of the classifier. In this way, the classifier is able to recognize company mentions with a precision of 91.11% and a recall of 78.82%.

Entity Linking. Subsequently, the extracted company names are linked to the knowledge base previously constructed from the structured data sources. This task is handled by the entity linking (EL) component. The currently employed EL module uses a simple fuzzy matching approach to link the extracted company names with their corresponding knowledge base entries. However, due to the modular system design, this preliminary linking strategy can be easily replaced by more sophisticated entity-linking approaches such as CohEEL [1].

DSMM'18, Houston, TX, USA

DOI: 10.1145/3220547.3220553

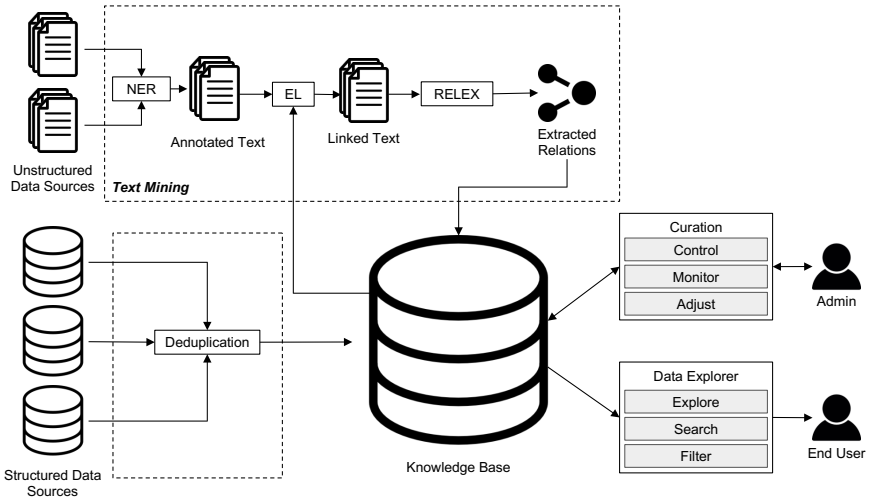

Figure 1: Overview of the system architecture

Relation Extraction. The task of the relation extraction (RELEX) component is to detect relationships between the found financial entities. The RELEX component currently in use extracts co-occurrence relationships between individual financial entities found within the same sentence. Initial analysis has shown that co-occurring financial entities often imply a business relationship.As with the other modules, it is possible to replace the currently used RELEX module with more advanced extraction techniques, such as the technique presented by Zuo et al. [6] or even more advanced neural network based techniques [5]. In particular, the approach proposed by Zuo et al. is promising, as it is able to extract directional relationships where the arguments are of the same type, in our case, company to company relationships. The resulting knowledge base can then be used to create a knowledge graph which can be used to analyze complex issues such as the spread of risk factors.

4 USER INTERFACES, INTERACTIONS AND SYSTEM SPECIFICATIONS

The entire system runs on a distributed platform to manage the processing of ever growing data volumes. A key aspect of the system is its modular structure so that each of the subcomponents can be easily interchanged to benefit from future advancements in the respective research fields. A novel aspect of the system is to separate the end-user interface from a curation interface that is specially designed to monitor the overall process and to take corrective actions at different levels. On a fine-grained level, an administrator can make minor changes, e.g. correcting a company name or an address, while it is also possible to do larger operations, such as reverting an entire deduplication run so that the knowledge base is returned to its original state. The curation interface is essentially intended as a tool for monitoring and controlling the knowledge base construction process and thus does not address the needs of the end user. It is primarily designed to inspect the results of each individual processing step (duplication, entity linking, etc.) and make corrections to mismatched entities using a specially designed view.

The Data Explorer focuses on the needs of the end user (e.g., a credit risk officer) and is thus designed for exploring, inspecting and searching the generated knowledge graph. Addressing these needs, it allows the user to search for individual nodes of interest, to further explore the graph, and to display the associated node information from the knowledge base. Furthermore, the user is able to filter the displayed knowledge graph by node and edge types, such as displaying only nodes of type "company" or only edges of type "is partner of". Another important feature is the ability to make suggestions for data changes in order to enable a continuous improvement process of the knowledge base.

5 CONCLUSIONS

We presented our architecture for a exploratory financial information system based on a custom-built knowledge graph. We further described the components of the processing pipeline in use to populate the knowledge base. As a next step, we plan to evaluate the system as a whole and do a fine-grained analysis of the results of the individual components in the context of a use case from our financial partner.

REFERENCES

[1] Toni Grütze, Gjergji Kasneci, Zhe Zuo, and Felix Naumann. 2016. CohEEL: Coherent and efficient named entity linking through random walks. *Journal of Web Semantics* 37–38 (2016), 75–89.

[2] Michael Loster, Zhe Zuo, Felix Naumann, Oliver Maspfuhl, and Dirk Thomas. 2017. Improving Company Recognition from Unstructured Text by using Dictionaries. In *Proceedings of the International Conference on Extending Database Technology (EDBT)*. 610–619.

[3] Alvaro E. Monge and Charles P. Elkan. 1996. The Field Matching Problem: Algorithms and Applications. 267–270.

[4] William E. Winkler and Yves Thibaudeau. 1991. An application of the Fellegi-Sunter model of Record Linkage to the 1990 US decennial census. *US Bureau of the Census* (1991), 1–22.

[5] Peng Zhou, Wei Shi, Jun Tian, Zhenyu Qi, Bingchen Li, Hongwei Hao, and Bo Xu. 2016. Attention-Based Bidirectional Long Short-Term Memory Networks for Relation Classification. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*. 207–2013.

[6] Zhe Zuo, Michael Loster, Ralf Krestel Krestel, and Felix Naumann. 2017. Uncovering Business Relationships: Context-sensitive Relationship Extraction for Difficult Relationship Types. In *Proceedings of the Conference on "Lernen, Wissen, Daten, Analysen" (LWDA)*. 271–283.