Boosting Information Extraction through Semantic Technologies: The KIDs use case at CONSOB

Federico Maria Scafoglieri[1], Domenico Lembo[1], Alessandra Limosani[2], Francesca Medda[2], and Maurizio Lenzerini[1]

[1] Sapienza Università di Roma
{lembo,lenzerini,scafoglieri}@diag.uniroma1.it
[2] Commissione Nazionale per le Società e la Borsa
{a.limosani,f.medda}@consob.it

In this paper we report on the initial results of a project concerning the integration of Semantic Technologies with Information Extraction (IE) techniques, jointly carried out by Sapienza University of Rome and CONSOB (Commissione Nazionale per la Società e la Borsa), the Italian public authority responsible for regulating the securities market.

The use case. In the EU, the creators of financial products (a.k.a. financial manufacturers) are obliged by law[3] to make information related to so-called PRIIPs (Packaged Retail Investment and Insurance-based Investments Products) publicly available. The NCAs (National Competent Authorities) have supervisory duties on such products, so that they can be safely placed on the respective national markets. The legislation requires information about PRIIPs to be communicated to NCAs through documents called KIDs (Key Information Documents). In the practice, this means that features to be checked are cast into text reports, typically formatted as pdf files, and extracting structured data from them (to bootstrap control activities), is actually in charge to the authority (In Italy, CONSOB). Due to the massive amount of documents to be analyzed (e.g., ∼700.000 KIDs received by CONSOB in 2019, more than 1 million in 2020), this process cannot be carried out manually, but still it is only partially automated to date.

Objectives. Our main aim is thus to develop a solution to streamline the extraction process and reduce as much as possible (ideally eliminate) the need of manual intervention, still guaranteeing very high accuracy. At the same time, such solution should return a data structure providing a due account of the semantics of the business domain and suited for rich and highly informative post-extraction analysis.

Solution. Given the previously highlighted requirements, the proposed solution aims at constructing a Knowledge Graph (KG), whose intensional component (expressed in OWL) is designed with the help of domain experts, and whose extensional level is automatically created from KIDs through a rule-based IE mechanism. The choice of structuring the extracted data as a KG not only facilitates the integration with other corporate and external data, enabling rich analysis and management at an abstract, conceptual level, but also allows for properly formalizing the conceptual distinction between PRIIPs and KIDs describing them, and the continuous updates which KIDs are subjected to.

Copyright © 2021 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

[3] PRIIPs Regulation n. 1286/2014

CEUR-WS.org/Vol-2980/paper418.pdf

Moreover, the choice to adopt a rule-based approach for IE, instead of a statistical one, lies not only in the great effort required for generating an annotated dataset for training learning algorithms, but also in the lack of transparency, accountability, and human interpretability of Machine Learning solutions, which makes excessively difficult to fully understand the results of the extraction, ill-fitting the financial context of this use case. In our first implementation efforts, we focused on a portion of the information to be extracted, consisting of 12 PRIIPs characteristics (e.g., name of the product, issue date, etc.). We realized two alternative implementations, briefly described below.

First realization. We initially adopted CoreNLP [3], the popular Stanford library for NLP, which is well-documented, fully-supported, and easy-to-use, and provides a rule engine module, called TokenRegex, useful to generate annotations on text via a regex-like rule language. After applying the rules, the generated annotations follow a flow of transformations, realized through components specifically written to translate annotations into facts of the KG. Although the results of our experiments are particularly convincing in terms of precision and recall, both averaging around 99% over a dataset of more than 14.000 KIDs, we encountered two main issues: (*i*) low performance in terms of execution time; (*ii*) complex process to define rules, caused by the low modularity of TokenRegex and the need of complementing rules with ad-hoc (java) code.

Second realization. We have therefore realized a second implementation through MASTRO SYSTEM-T [2], a KG-aided IE tool, which allowed us to solve the above issues, still achieving the same accuracy results. As for issue (*i*), the extraction speed has increased considerably, reducing the time needed to materialize the KG by 46.15%, mainly thanks to the use in MASTRO SYSTEM-T of the highly performing IE tool System-T [1]. Issue (*ii*) has been instead greatly mitigated, by virtue of both the full declarativeness of the language used for the extractors in System-T, and the way in which MASTRO SYSTEM-T casts them into extraction assertions mapping KIDs to KG predicates. This second implementation allowed us also to follow a new approach to access KIDs data. Indeed, MASTRO SYSTEM-T may be used as a Virtual KG engine [4], to perform the extraction at query time, which allows to always get fresh data.

Conclusion. Since all European NCAs need to address the same oversight tasks on PRIIPs as CONSOB, the impact of our research may go fairly beyond the single experience we described, considered also that, to the best of our knowledge, very few authorities have to date developed solutions supporting automatic IE from KIDs. In particular, the adoption of our approach by other authorities is enabled by the fact that both KIDs content and structure must obey to the same common regulation.

References

1. L. Chiticariu, M. Danilevsky, Y. Li, F. Reiss, and H. Zhu. SystemT: Declarative text understanding for enterprise. In *Proc. of NAACL-HLT (Industry Papers)*, pages 76–83, 2018.
2. D. Lembo, Y. Li, L. Popa, K. Qian, and F. Scafoglieri. Ontology mediated information extraction with MASTRO SYSTEM-T. In *Proc. of ISWC (Demos Track)*, pages 256–261, 2020.
3. C. D. Manning, M. Surdeanu, J. Bauer, J. R. Finkel, S. Bethard, and D. McClosky. The Stanford CoreNLP natural language processing toolkit. In *Proc. of ACL*, pages 55–60, 2014.
4. G. Xiao, D. Calvanese, R. Kontchakov, D. Lembo, A. Poggi, R. Rosati, and M. Zakharyaschev. Ontology-based data access: A survey. In *Proc. of IJCAI*, pages 5511–5519, 2018.