Automatic Construction of Knowledge Graphs from Text and Structured Data: A Preliminary Literature Review

Maraim Masoud, Bianca Pereira, John McCrae, Paul Buitelaar
(Insight SFI Centre for Data Analytics, NUIG, Ireland)

Introduction

Knowledge graphs have been shown to be an important data structure for many applications. In the enterprise domain, such graphs need to be constructed based on both structured (e.g. databases) and unstructured (e.g.textual) internal data sources; preferentially using automatic approaches due to the costs associated with manual construction of knowledge graphs. However, despite the growing body of research that leverages both structured and textual data sources in the context of automatic knowledge graph construction, the research community has centred on either one type of source or the other. Example of these initiatives:

- Unstructured data initiatives: The SemEval Taxonomy Extraction Task (SIGLEX/SIGSEM)
- Structured data initiatives:: OAEI (ISWC)
- Structured and unstructured initiatives: TAC-KBP and NEEL

Despite all the available initiatives in the automatic construction of knowledge graphs based on different sources, none of these initiatives explicitly focuses on both: **(i) the aggregated use of text and structured data sources,** and **(ii) the generation of a domain-specific knowledge graph**. Therefore, in this work, we conduct a preliminary literature review to investigate approaches that can be used for the integration of textual and structured data sources in the process of automatic knowledge graph construction. We highlight the solutions currently available for use within enterprises and point areas that would benefit from further research

Methodology

The literature review is performed by following these steps:
(i) selection of seed papers
(ii) search
(iii) filtering
(iv) analysis

Selection of seed papers based on a convenience sample. The six seed papers are [1,2,3,4,5,6]. The search of relevant papers is expanded to include also the literature in their list of references. Following the criteria:

Inclusion Criteria	Exclusion Criteria
✓ written in English.	X only (semi-) manual approaches.
✓ published in conference proceedings or in a journal.	X no explanation of the method used.
✓ the abstract and conclusion inculcate the paper is in the topic of enriching the results of automatic extraction of knowledge graph with structured data.	X using only textual data sources.

Dimensions of analysis:

1. **Point of integration**: pre-construction, during construction, and post-construction.
2. **Integration goals**: (i) knowledge graph completion, and (ii) knowledge graph validation.
3. **The format of the structured data**: Date sources can be represented by an entity-relation diagram or converted to a graph.
4. **The format of the output knowledge graph.** (i) term taxonomy, (ii) topic taxonomy, (iii) labelled graph, and (iv) ontologies.

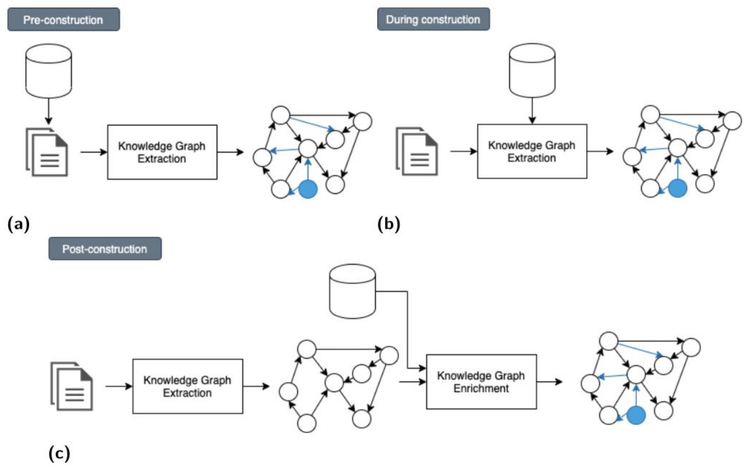

Figure 2: Integration of textual and structured data sources to enrich the automatic construction of knowledge graphs: (a) pre-construction, (b) during construction, and (c) post-construction.

Results

From the seed papers and their references, we ended up with **131** papers. From those**, 97 papers** were removed due to our exclusion criteria**.,** resulting in **one seed paper** and further **34 papers** analysed**.** Results of the analysis according to each of the dimensions chosen for the literature review:

1. **Point of integration**: The approaches used for *pre-construction* leverage the knowledge graph extraction algorithm in one of three ways: (i) by combining both textual and structured information into a single semantic space, (ii) by generating a profile for each term using the metadata of the documents associated to the term (e.g.popularity of a document), and (iii) by using the structured data source to generate an initial graph that connects the different terms from text. Regarding the integration *during construction*, the structured data source is provided as training data for the detection of relations between terms extracted from text. This detection is based on either: (i) relation classification (is achieved by using neural networks, or probabilistic methods), or (ii) relation prediction (via embeddings or tensors). In post-construction integration, the knowledge graph built from text can be integrated with other knowledge graphs by: (i) graph alignment, (ii) graph fusion, or (iii) logical inference.
2. **Integration goals**: 33 papers out of 40 focused exclusively on *knowledge graph completion*.. Few papers focused only on the *validation of knowledge graphs*, and they are limited to verifying the correctness of entities and relations but not correcting the detected errors.
3. **The format of the structured data**: *Tables* and *value-key pairs* are provided as metadata to textual documents in pre-construction approaches either by the using explicit links between entities, or by the inference of these links via analysis of user interactions with both data sources. *Graphs*, on the other hand, are a dominant format in the analysed literature
4. **The format of the output knowledge graph.** The analysed literate have a strong focus on *labelled graph*, while the extraction of *taxonomies* and *ontologies* is underrepresented.

Conclusion

The goal of this work is to provide knowledge on what is available in the literature for use by enterprises wishing to generate knowledge graphs based on their own internal data sources. For that, we present a preliminary literature review that investigates approaches used for the integration of textual and structured data sources in the process of automatic knowledge graph construction. Based on this analysis, we conclude the following:

- Enterprises have a range of approaches available if aiming at the generation of a labelled knowledge graph that aggregates data from both textual and structured sources, where the structured data source used has a graph-like structure and the integration between textual and structured source is done only after a knowledge graph has been extracted from text (post-construction).
- The integration of data sources before they are used for automatic knowledge graph construction, as well as the use of tables or key-value pairs as structured data sources are still areas with possibility for further research

Learned Lessons and Future Work

Many lessons can be drawn from this specific analysis in terms of our conceptual framework, survey analysis and findings.

1. *Our categorization*, while specific, has shown to be useful in classifying available approaches for constructing a domain-specific knowledge graph by;
   - **(i)** categorizing similar approaches based on the selected dimensions,
   - and **(ii)** displaying the patterns that influence the decision to adapt a specific variation of each dimension as discussed in the results section.
2. The value of this classification is that it provides enterprises with a clear set of approaches for constructing domain-specific knowledge graphs from structured and unstructured data sources.
3. From the survey perspective, there is an opportunity to future investigate the research and the application of knowledge graphs in the enterprise domain
4. From the perspective of survey results, there is a range of options for generating knowledge graphs by aggregating structured and unstructured sources. According to our findings, integration using a graph-like structure is a popular approach in comparison to tables and key-value pairs. The later formats are currently in the tentative stage as outlined in the results section.

Future work will include expanding the survey to a systematic review with keyword-based seed papers. We envision this as a large-scale study that will examine the enterprise knowledge graph integration from different perspectives and demonstrate use cases from various application domains.

Reference

- [1] Paulheim, H., 2017. Knowledge graph refinement: A survey of approaches and evaluation methods. Semantic web, 8(3), pp.489-508.
- [2] *Shang, J., Zhang, X., Liu, L., Li, S. and Han, J., 2020, April. Nettaxo: Automated topic taxonomy construction from text-rich network. In Proceedings of The Web Conference 2020 (pp. 1908-1919).*
- *[3] Cai, H., Zheng, V.W. and Chang, K.C.C., 2018. A comprehensive survey of graph embedding: Problems, techniques, and applications. IEEE Transactions on Knowledge and Data Engineering, 30(9), pp.1616-1637.*
- *[4] Nguyen, H.L., Vu, D.T. and Jung, J.J., 2020. Knowledge graph fusion for smart systems: A Survey. Information Fusion.*
- *[5] Zhao, X., Jia, Y., Li, A., Jiang, R., Xie, H., Song, Y. and Han, W., 2019, June. Multi-source knowledge fusion: a survey. In 2019 IEEE Fourth International Conference on Data Science in Cyberspace (DSC) (pp. 119-127). IEEE.*
- *[6] W. Shen, J. Wang and J. Han, 2015, February. Entity Linking with a Knowledge Base: Issues, Techniques, and Solutions. In IEEE Transactions on Knowledge and Data Engineering, 27(2), pp. 443-460.*