Relation Clustering in Narrative Knowledge Graphs

Simone Mellace , **Vani K** and **Alessandro Antonucci***
IDSIA - Lugano (Switzerland)
{simone,vanik,alessandro}@idsia.ch

Abstract

When coping with literary texts such as novels or short stories, the extraction of structured information in the form of a knowledge graph might be hindered by the huge number of possible relations between the entities corresponding to the characters in the novel and the consequent hurdles in gathering supervised information about them. Such issue is addressed here as an unsupervised task empowered by transformers: relational sentences in the original text are embedded (with SBERT) and clustered in order to merge together semantically similar relations. All the sentences in the same cluster are finally summarized (with BART) and a descriptive label extracted from the summary. Preliminary tests show that such clustering might successfully detect similar relations, and provide a valuable preprocessing for semi-supervised approaches.

1 Introduction

Recent applications in the field of *Natural Language Processing* (NLP) are exploiting data-driven techniques from the general area of *Machine Learning* (ML). These are typically *Deep Learning* (DL) systems based on multi-layer neural networks fitted with the input text data, to be converted in numerical objects by some embedding scheme. Such DL-NLP systems are successful in extracting knowledge from natural language and capturing the underlying *narratives*.

As a matter of fact, most of these NLP efforts are focused on a few mainstream applicative areas, such as biomedical literature [Zhang *et al.*, 2018; Lv *et al.*, 2016] or news and social media [Trieu *et al.*, 2017; Ghosh and Shah, 2018]. Other inputs such as *literary text* in the form of novels or short stories received less attention [Wohlgenannt *et al.*, 2016; Volpetti *et al.*, 2020]. This is unfortunate as literary texts might exhibit high complexity in the narrative plots, while also lacking explicit annotations, thus making the knowledge extraction process very challenging. Handling such complexities, helps in evaluating the models Natural Language Understanding and creating benchmarks for these low-resource domains. This could in turn be helpful for common sense reasoning, reading comprehensions and enhance NLP applications such as summary generation, machine translations and question answering.

Despite their astonishing applications in NLP, e.g., [Zhu *et al.*, 2019; Paulus *et al.*, 2017], DL models are typically based on discriminative functions with a huge number of parameters, whose interpretation is often problematic. This prevents both the *explainability* of the results and the possibility of doing *reasoning* over the model entities. For this reason, alternative approaches to NLP, based on so-called *Knowledge Graphs* (KGs), i.e., relational ontologies providing interlinked descriptions of the entities involved in a text, are also popular. Despite the existence of techniques for automatic KG extraction acting at the syntactic level [Tang *et al.*, 2016; Ruan *et al.*, 2016], most of the approaches require supervision in the form of manual annotations or access to knowledge bases, such as UMLS[1], for higher level descriptions.

Of course these two orthogonal perspectives, say DL and KGs, can be combined. DL models can be trained from KGs [Socher *et al.*, 2013; Li and Mao, 2019] and used for ML, and, *vice versa*, DL models such as embeddings can be used to predict missing links of the KG, classify relations, or align entities from different KGs [Liu *et al.*, 2019; Lin *et al.*, 2015].

Here, we follow such an integrated point of view, being motivated by specific features of literary text understanding. In fact, for this kind of text, the KG entities are typically the characters in the plot, and no serious alignment issues appear, while the classification of the relations becomes much more challenging because of the lack of supervision. In other domains the number of possible relations is typically limited (e.g., in [Chen *et al.*, 2010], few relations such as *binding*, *expression*, *protein interaction* and few others), while in the literary case the possible relations between characters (e.g., Table 1) can be much more. Accordingly, we explore some directions for an unsupervised approach to the identification of relations in KGs obtained from literary texts. The goal is to cluster semantically equivalent relational sentences including

*Contact Author

In: A. Jorge, R. Campos, A. Jatowt, A. Aizawa (eds.): Proceedings of the first AI4Narratives Workshop, Yokohama, Japan, January 2021, published at http://ceur-ws.org

[1] https://uts.nlm.nih.gov

descriptions of relations between the characters of a novel. Our preliminary tests seem to be promising with respect to the proper identification of similar relations, while also giving directions about the most suitable clustering strategies as well as further development of semi-supervised tools.

The paper is organized as follows. In Section 2, we summarize the existing literature in the field. Section 3 describes our workflow, which is demonstrated by applicative examples in Section 4. Conclusions and outlooks are in Section 5.

2 Existing Work

As discussed in the previous section, DL tools such as sequence and self-attention models as well as transformers (e.g., BERT, Xlnet, BART) have been widely and successfully used in NLP for word and sentence encoding [Peters *et al.*, 2018; Devlin *et al.*, 2018; Yang *et al.*, 2019; Lewis *et al.*, 2019]. These models can be fine-tuned and used for various tasks such as classification, summarization and sentiment analysis. This also concerns KGs, where DL models are used for embedding the triplet information and used for tasks such as link predictions and KG completion [Lin *et al.*, 2019; Yao *et al.*, 2019], while other researchers worked on the training of embedding from KGs [Ji *et al.*, 2015; Wang *et al.*, 2014; Lin *et al.*, 2015; Bordes *et al.*, 2013].

None of these application was concerned with literary text. Despite some attempts to apply ML and DL models in the field [Worsham and Kalita, 2018; Short, 2019; Labatut and Bost, 2019; K and Antonucci, 2019; Volpetti *et al.*, 2020] to analyze character relations, sentiments and visualizations, a connection with KGs still remains under-explored. The goal of this paper is to fill this gap by providing an unsupervised alternative to the *relational* classifiers recently developed for supervised tasks in [K *et al.*, 2020].

3 Workflow

Figure 1 depicts the workflow of the approach we propose for the unsupervised identification of similar relations in the KGs obtained from literary text. This involves a NLP part for preprocessing (entity recognition, sentence tokenization, detection of relational sentences and triplet generation) corresponding to the red blocks, a DL abstraction level (sentence embedding and summarization) corresponding to the blue blocks, as well as classical ML techniques (characters de-aliasing, sentence clustering, semi-supervised extension) associated with green blocks. These steps in their sequential order are described here below together with the main challenges they present. The tool is available as a free software.[2]

Named Entity Recognition (NER). The very first step is the identification of the entities to be associated with the KG nodes. These are detected by a custom version of the Stanford NER Tagger[3] such that consecutive entities in a sentence (i.e. words tagged as *PERSON*), are detected as a unique element (e.g., *Harry James Potter*).

[2]https://github.com/IDSIA/novel2graph

[3]https://nlp.stanford.edu/software/CRF-NER.html

Dealiasing. As a same character can be termed with different *aliases* in the same novel, a de-aliasing might be required. This issue has been already addressed in [K *et al.*, 2020], where a satisfactory solution based on ML and NLP has been found. Here we adopt a similar strategy based on the classical DBSCAN clustering ($\epsilon = .3$ and Levenshtein string distances), together with a number of manual adjustments. In our approach in fact, we first perform separate pre-clustering over entities starting with the same letter (e.g., *Hermione* and *Hermione Granger* are identified as a cluster while *Harry*, *Harry Potter* and *H. Potter* as another one), and then adding similar but unassigned names to a cluster (e.g., *Granger* assigned to *Hermione*'s cluster and *Potter* to *Harry*). All the occurrences of the aliases in the same cluster are finally replaced by identifiers (e.g., CHAR0 replaces *Harry*, *Potter*, *Harry Potter* and so on).

Tokenization. Embeddings based on tranformers are based on contextual information. Since, sentences are considered as the simplest logical and meaningful unit that provides a semantic intuition of the context, we rely on a segmentation at this level.

Relational Sentence Identification. Let us call *relational* a sentence including two or more characters. We extract relational sentences from the de-aliased and tokenized text, by also evaluating whether or not the text between the two character occurrences is a simple proposition or not (e.g., *Harry and Ron were having good time* and *Harry looked at Ron*). If this is the case we call the relation *symmetric* and we generate two distinct input for the pipeline. Note also we only use sentences containing exactly two characters and excluding self-relations (e.g., *Harry, I am Harry Potter*).

Sentence Embedding. To identify the relations between entities, we embed the relational sentences using Sentence BERT (SBERT) [Reimers and Gurevych, 2019]. SBERT uses a Siamese network structure [Schroff *et al.*, 2015] to reproduce meaningful encodings. The method was specifically modelled for clustering and semantic search. SBERT adds a pooling operation on top of BERT to derive these embeddings. SBERT is fine tuned on SNLI [Bowman *et al.*, 2015] and MNLI [Williams *et al.*, 2018] datasets with a three-way soft-max classifier objective function for one epoch with the default pooling strategy MEAN (computing the mean of all output vectors).

Sentence Clustering. Since, these embeddings encode semantic and contextual information, sentences with similar vector representations are supposed to share similar relations. Hence, we adopt a simple clustering approach to group the sentences with similar relations. The distances between the vectors returned by SBERT are assumed to reflect the semantic similarity between the corresponding sentences and hence the relations included in these sentences. Classical clustering methods such as k-means or DBSCAN can be therefore used to create groups of sentences and hence triplets with the same relation. We considered the Euclidean distance, as well as the classical cosine distance. Even though clustering the entire sentence may not explicitly cluster the relationships, the sentences that fall into similar semantic spaces can provide us a coarse-grained grouping of relations.

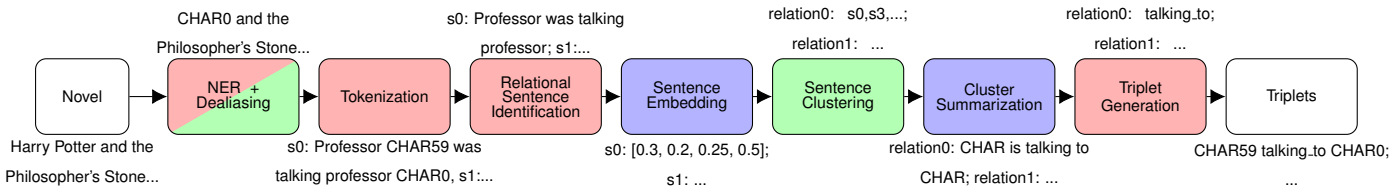

Figure 1: The workflow of the triplet generation process, with relation clustering.

Cluster Summarization. After the clustering of the relational sentences, we might want to represent these relations as a summary of the sentences involved in the cluster. To achieve that we adopt the BERT summarization pipeline based on the BART [Lewis *et al.*, 2019] model. This includes an encoder like BERT and a decoder like GPT [Radford *et al.*, 2019] and it is trained on CNN/Daily Mail dataset with learning rate $3 \cdot 10^{-5}$ (Adam optimizer). This performs extractive summarization, giving most suitable representative sentences of each cluster. Although training is not in-domain, as news articles are also narratives, we use this for a preliminary set-up.

Triplet Generation. Once the extractive summary is produced, for asymmetric relations we extract the phrase which comes between the two reference characters. We then extract only the verbs from these phrases, which are the considered as part-of-speech tags that could convey some information about the type of relations.

From Unsupervised to Semi-Supervised Learning. The overall procedure described in this section is purely unsupervised. Yet, the clusters of relational sentences are described by the summaries, first, and then labels generated by the system. This might be the basis of a system where, part of those clusters are manually inspected and their summaries/labels validated or fixed by a human annotator. This would turn the system into a semi-supervised one, where the annotated clusters can be used as classifier of the relations.

Book	Sentence
HP	*Dumbledore* smiled at the look of amazement on *Henry*'s face
HP	*Ron* grinned at *Henry*
LW	*Brooke* smiling at *Meg* as if everything had become possible him now
HP	*Henry* stared as *Dumbledore* sidled back into the picture . . . gave him a small smile

Table 1: Relational sentences from the same cluster.

4 Experiments

For a first empirical validation of our pipeline we process, in a single run, two novels, namely *Harry Potter and the Philosopher's Stone* (HP) by J. K. Rowling and *Little Women* (LW) by Louisa May Alcott. 1307 suitable sentences out of 32365 are identified and grouped in 200 clusters (i.e. different relations types). As the characters of the two books are distinct, the system generates a KG with two disconnected components (see Figure 2). Yet, the relation clustering is able to detect similarities between sentences in the two books. E.g., sentences in Table 1 are related to smiling actions. For that cluster the extractive summarization returns the first sentence as a summary and, finally, the triplet generation mechanism return *smile* as representative label.

Concerning sentence clustering we considered both DBSCAN and k-means algorithms both paired with Euclidean and cosine distance. In the considered setup we did not found significant differences with the two metrics. Regarding the algorithms, an observed issue with DBSCAN was a sudden transition from a huge number of single-sentence clusters to very large clusters. Both these extreme scenarios prevent a meaningful identification of relations. Yet, it was not possible to automatically decide the number of clusters with k-means, as the silhouette analysis returned monotone results.

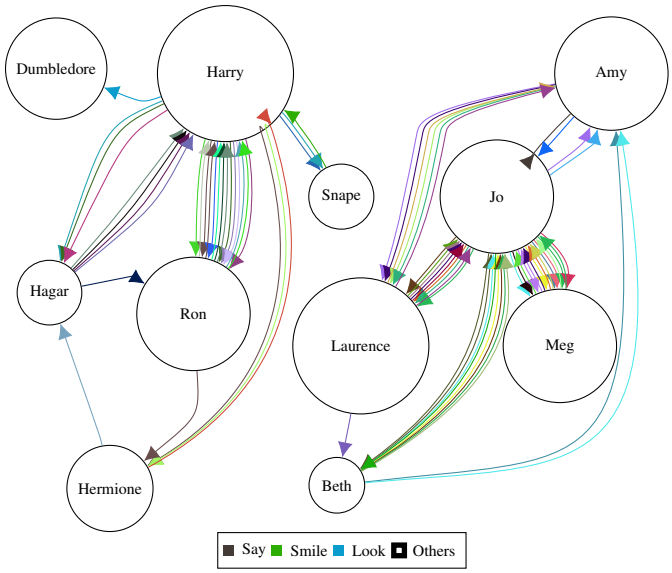

Figure 2: Narrative KGs of HP (left) and LW (right). Nodes corresponds to de-aliased characters and arcs to clustered relations.

5 Conclusions

An unsupervised approach to KG extraction from narrative texts has been proposed. The procedure exploits transformer models to detect similar relations in the triplets, then generates summaries and representative labels for these clusters of similar relations. This represent a pre-processing step for a semi-supervised approach where the representative labels are validated by human annotators and used as a relational classifier. Validated clusters can define relational classifiers, while the automatically generated labels are used for the others. As a future work we want to apply our pipeline to a corpus of literary texts and validate the clusters. This being a starting point for the creation of a knowledge base for literary texts.

References

[Bordes *et al.*, 2013] Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko. Translating embeddings for modeling multi-relational data. In *Advances in Neural Information Processing Systems*, pages 2787–2795, 2013.

[Bowman *et al.*, 2015] Samuel Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, 2015.

[Chen *et al.*, 2010] Bin Chen, Xiao Dong, Dazhi Jiao, Huijun Wang, Qian Zhu, Ying Ding, and David J Wild. Chem2Bio2RDF: a semantic framework for linking and data mining chemogenomic and systems chemical biology data. *BMC bioinformatics*, 11(1):255, 2010.

[Devlin *et al.*, 2018] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[Ghosh and Shah, 2018] Souvick Ghosh and Chirag Shah. Towards automatic fake news classification. *Proceedings of the Association for Information Science and Technology*, 55(1):805–807, 2018.

[Ji *et al.*, 2015] Guoliang Ji, Shizhu He, Liheng Xu, Kang Liu, and Jun Zhao. Knowledge graph embedding via dynamic mapping matrix. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 687–696, 2015.

[K and Antonucci, 2019] Vani K and Alessandro Antonucci. NOVEL2GRAPH: Visual summaries of narrative text enhanced by machine learning. In *Text2Story@ ECIR*, pages 29–37, 2019.

[K *et al.*, 2020] Vani K, Simone Mellace, and Alessandro Antonucci. Temporal embeddings and transformer models for narrative text understanding. In *Proceedings of Text2Story - Third Workshop on Narrative Extraction From Texts*, 2020.

[Labatut and Bost, 2019] Vincent Labatut and Xavier Bost. Extraction and analysis of fictional character networks: A survey. *ACM Computing Surveys (CSUR)*, 52(5):1–40, 2019.

[Lewis *et al.*, 2019] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*, 2019.

[Li and Mao, 2019] Pengfei Li and Kezhi Mao. Knowledge-oriented convolutional neural network for causal relation extraction from natural language texts. *Expert Systems with Applications*, 115:512–523, 2019.

[Lin *et al.*, 2015] Yankai Lin, Zhiyuan Liu, Maosong Sun, Yang Liu, and Xuan Zhu. Learning entity and relation embeddings for knowledge graph completion. In *Twenty-ninth AAAI conference on artificial intelligence*, 2015.

[Lin *et al.*, 2019] Bill Yuchen Lin, Xinyue Chen, Jamin Chen, and Xiang Ren. Kagnet: Knowledge-aware graph networks for commonsense reasoning. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2822–2832, 2019.

[Liu *et al.*, 2019] Weijie Liu, Peng Zhou, Zhe Zhao, Zhiruo Wang, Qi Ju, Haotang Deng, and Ping Wang. K-bert: Enabling language representation with knowledge graph. In *Proceedings of AAAI*, 2019.

[Lv *et al.*, 2016] Xinbo Lv, Yi Guan, Jinfeng Yang, and Jiawei Wu. Clinical relation extraction with deep learning. *International Journal of Hybrid Information Technology*, 9(7):237–248, 2016.

[Paulus *et al.*, 2017] Romain Paulus, Caiming Xiong, and Richard Socher. A deep reinforced model for abstractive summarization. *arXiv preprint arXiv:1705.04304*, 2017.

[Peters *et al.*, 2018] Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, 2018.

[Radford *et al.*, 2019] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. *OpenAI Blog*, 1(8):9, 2019.

[Reimers and Gurevych, 2019] Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3973–3983, 2019.

[Ruan *et al.*, 2016] Tong Ruan, Mengjie Wang, Jian Sun, Ting Wang, Lu Zeng, Yichao Yin, and Ju Gao. An automatic approach for constructing a knowledge base of symptoms in chinese. In *2016 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 1657–1662. IEEE, 2016.

[Schroff *et al.*, 2015] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823, 2015.

[Short, 2019] Matthew Short. Text mining and subject analysis for fiction; or, using machine learning and information extraction to assign subject headings to dime novels. *Cataloging & Classification Quarterly*, 57(5):315–336, 2019.

[Socher *et al.*, 2013] Richard Socher, Danqi Chen, Christopher D Manning, and Andrew Ng. Reasoning with neural tensor networks for knowledge base completion. In *Advances in Neural Information Processing Systems*, pages 926–934, 2013.

[Tang *et al.*, 2016] Zhiyuan Tang, Dong Wang, and Zhiyong Zhang. Recurrent neural network training with dark knowledge transfer. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5900–5904. IEEE, 2016.

[Trieu *et al.*, 2017] Lap Q Trieu, Huy Q Tran, and Minh-Triet Tran. News classification from social media using twitter-based doc2vec model and automatic query expansion. In *Proceedings of the Eighth International Symposium on Information and Communication Technology*, pages 460–467, 2017.

[Volpetti *et al.*, 2020] Claudia Volpetti, Vani K, and Alessandro Antonucci. Temporal word embeddings for narrative understanding. In *ICMLC 2020: Proceedings of the Twelfth International Conference on Machine Learning and Computing*, ACM Press International Conference Proceedings Series. ACM, 2020. ISBN: 978-1-4503-7642-6.

[Wang *et al.*, 2014] Zhen Wang, Jianwen Zhang, Jianlin Feng, and Zheng Chen. Knowledge graph embedding by translating on hyperplanes. In *AAAI*, 2014.

[Williams *et al.*, 2018] Adina Williams, Nikita Nangia, and Samuel Bowman. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, 2018.

[Wohlgenannt *et al.*, 2016] Gerhard Wohlgenannt, Ekaterina Chernyak, and Dmitry Ilvovsky. Extracting social networks from literary text with word embedding tools. In *Proceedings of the Workshop on Language Technology Resources and Tools for Digital Humanities (LT4DH)*, pages 18–25, 2016.

[Worsham and Kalita, 2018] Joseph Worsham and Jugal Kalita. Genre identification and the compositional effect of genre in literature. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1963–1973, 2018.

[Yang *et al.*, 2019] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. Xlnet: Generalized autoregressive pretraining for language understanding. In *Advances in neural information processing systems*, pages 5754–5764, 2019.

[Yao *et al.*, 2019] Liang Yao, Chengsheng Mao, and Yuan Luo. KG-BERT: BERT for knowledge graph completion. *arXiv preprint arXiv:1909.03193*, 2019.

[Zhang *et al.*, 2018] Yijia Zhang, Hongfei Lin, Zhihao Yang, Jian Wang, Shaowu Zhang, Yuanyuan Sun, and Liang Yang. A hybrid model based on neural networks for biomedical relation extraction. *Journal of Biomedical Informatics*, 81:83–92, 2018.

[Zhu *et al.*, 2019] Xuelin Zhu, Biwei Cao, Shuai Xu, Bo Liu, and Jiuxin Cao. Joint visual-textual sentiment analysis based on cross-modality attention mechanism. In *International Conference on Multimedia Modeling*, pages 264–276. Springer, 2019.