Can Knowledge Graph Embeddings Tell Us What Fact-checked Claims Are About?

Valentina Beretta[1], Katarina Boland[3], Luke Lo Seen[2], Sébastien Harispe[1], Konstantin Todorov[2] and Andon Tchechmedjiev[1]
[1]EuroMov Digital Health in Motion, Univ Montpellier, IMT Mines Alès, Alès, France
[2]LIRMM, University of Montpellier, CNRS, Montpellier, France
[3]GESIS, Cologne, Germany
{firstname.lastname}@mines-ales.fr
katarina.boland@gesis.org, todorov@lirmm.fr

Abstract

The web offers a wealth of discourse data that help researchers from various fields analyze debates about current societal issues and gauge the effects on society of important phenomena such as misinformation spread. Such analyses often revolve around claims made by people about a given topic of interest. Fact-checking portals offer partially structured information that can assist such analysis. However, exploiting the network structure of such online discourse data is as of yet under-explored. We study the effectiveness of using neural-graph embedding features for claim topic prediction and their complementarity with text embeddings. We show that graph embeddings are modestly complementary with text embeddings, but the low performance of graph embedding features alone indicate that the model fails to capture topological features pertinent of the topic prediction task.

1 Introduction

Analysing claims shared on social media is of growing interest, from social/political sciences to Artificial Intelligence (AI). Such analyses are often performed with respect to a specific set of topics (e.g. "immigration" or "abortion") that allow carrying out targeted studies of trends, understanding/quantifying hidden biases (Garimella et al., 2018), discovering stances towards those topics (Wang et al., 2018) or their underlying falsehood propagation patterns (Vosoughi et al., 2018). Fact-checking portals offer a wealth of information about claims, their truth values and their sources. To analyse claims about a given topic, scientists need (1) access to heterogeneous repositories of claims and (2) the prior knowledge of which entities are mentioned in claims that belong to a topic (as defined by thematic keywords in each portal).

As a (partial) response to (1), recent work has presented ClaimsKG—a large dynamic knowledge graph (KG) of fact-checked claims harvested from various fact-checking portals (like politifact.com) and their metadata (e.g. truth values, authors, sources, links to DBpedia) (Tchechmedjiev et al., 2019) (cf. Figure 1).[1] ClaimsKG includes thematic *keywords* provided by the fact-checking portals (e.g. "elections" or "taxes"). However, using them to filter claims by topic is problematic as: (1) not all claims are annotated; (2) the keywords are very heterogeneous (granularity or level of abstraction; e.g. "economy" vs. "Kim Kardashian"); (3) there is no standardization within or across portals; (4) there are no links between keywords grouping related concepts and (5) existing annotations are often incomplete. We address this need for normalization and for providing missing topic annotations of claims by investigating representation learning methods for claims.

Representation learning for text (Devlin et al., 2018; Li and Yang, 2018) and graphs (Cai et al., 2018; Goyal and Ferrara, 2018) has been successfully applied to many tasks from entity linking (Radhakrishnan et al., 2018) to link prediction in large KGs (Kazemi and Poole, 2018) allowing for KG completion/fusion. However, the ability of these methods to represent claims and to transfer to other machine learning (ML) tasks (e.g. predicting the topic(s) of a claim) has not been investigated. We evaluate the capability of link prediction graph embeddings to capture pertinent information from the graph structure in order to benefit downstream tasks. We compare the performance resulting from using (1) graph embeddings (CP/N3 model on ClaimsKG enriched with relations between mentions coming from DBPedia) (2) claim textual embeddings, or (3) different combinations thereof, as features in the task of supervised **multi-**

[1]https://data.gesis.org/claimskg/site/

Proceedings of the First Workshop on Insights from Negative Results in NLP, pages 71–75

ISBN 978-1-952148-66-8

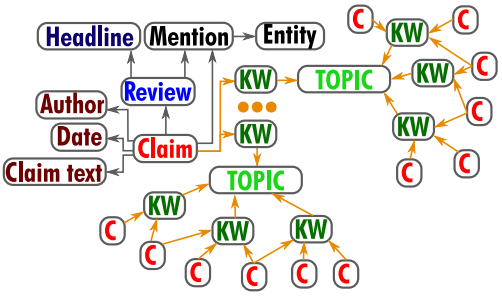

Figure 1: Simplified structure of ClaimsKG and graph baseline structures. KW=Keyword, C=Claim.

label claim topic prediction on a gold dataset. This task was chosen given that (1) it is significantly more challenging than typical topic classification tasks and (2) we can control the parameters of the evaluation by design and check for desirable properties captured by link prediction graph embeddings. We evaluate the use of claim vectors as features with or without the addition of neighbourhood vectors (outgoing relations and targets). We then perform ablation studies over different features to better characterise what is captured by the graph embeddings. Our results show that state-of-the-art link prediction models fail to capture equivalence structures and transfer poorly.

2 Claim Topic Classification Dataset

We present our semi-automatic approach to build a gold standard dataset of annotated claims for topic classification. Since ClaimsKG covers a wide range of different topics, annotating a random sample of claims would not yield a sufficient number of claims *per* topic. Thus, we identified a set of 7 topics that have a high number of claims in ClaimsKG and are relevant for claim-related studies: "healthcare " (1777), "taxes" (1519), "elections" (1074), "crime" (947), "education" (1263), "immigration" (1147) and "environment" (567). We then automatically identify claims potentially referring to these topics using the keywords assigned by the fact-checking sites. First, we mapped all keywords to common high-level concepts in two upper level taxonomies: the TheSoz thesaurus of social sciences (Zapilko et al., 2013) and the UNESCO Thesaurus[2] employing a dictionary-based entity linking approach (the concepts are noted as `TOPIC` in Figure 1). We then extracted a random subset of claims that are linked to at least one of the chosen topics through their keywords. Note that one claim

[2]http://vocabularies.unesco.org/thesaurus

Graph	Train	Test	Dev	MRR (%)	HITS@1 (%)	HITS@3 (%)	HITS@10 (%)
CKG	9144792	190491	190403	19.45	16.79	20.86	26.36
CKG-KW	9078900	189014	189110	16.77	13.65	17.93	22.98

Table 1: Link prediction performance for ClaimsKG graph embeddings (Standard metrics: Mean Reciprocal Rank (MRR), HITS@1, HITS@3, HITS@10).

can correspond to several concepts thus creating a multi-label dataset. To validate and complete the semi-automatically assigned labels, we finally asked 5 annotators to re-annotate the dataset and assign the claims to all applicable topics. This gold standard, composed of 629 annotated claims, has a Krippendorff's α annotator agreement (Masi distance) (Passonneau, 2006) of 0.75 which is a reasonably high agreement but also shows that the task is not trivial. For example, consider the claim "Nobody is leaving Memphis. That's a myth." uttered by a city councilman, with the keywords "Population" and "Census" assigned by the fact-checking site.[3] At first glance, none of the selected topics seems to apply. However, the claim review explains that this claim had been uttered in context of a debate concerning the fear that a proposed one-time tax for schools might make people leave the city with this claim defending the tax. Thus, this claim may be interpreted as being about "taxes" and even "education", depending on how much of the pragmatic context is taken into account. In the final dataset[4] the topic distribution is the following: "healthcare" (25%), "taxes" (21%), "elections" (17%), "crime" (16%), "education" (13%), "immigration" (12%) and "environment" (10%).

3 Representation Learning and Evaluation Pipeline

Graph embedding models. We train[5] a CANDECOM/PARAFAC model with N3 regularization (CP-N3) (Lacroix et al., 2018).[6] We computed a model for ClaimsKG (CKG) and a variant without keywords (CKG-KW) needed in the ablation studies. The link prediction performance, reported in Table 1, is lower than for YAGO3-10, the standard dataset most similar to ClaimsKG at an equivalent

[3]https://tinyurl.com/y6ysg4ju

[4]https://github.com/claimskg/claim_topics_dataset

[5]Code: https://github.com/twktheainur/kbc

[6]Current SOTA. Optimal parameters within hardware constraints (GeForce 2080Ti with 11GB VRAM) – CP Model, Rank 50, Adagrad optimizer, 0.1 learning rate, N3 regularizer with coefficient 0.005, 30 epochs max, batch size 150 – Approx. 3h/epoch $\times$ 3 models $\times$ 30 epochs $\times 275W \simeq 270h \times 275W \simeq 74.25KWh@\$0.31/KWh \simeq \$23$

rank (MRR = 0.54, HITS@[1, 3, 10]=[47%, 58%, 68%]): ClaimsKG is larger and sparser (fewer triples *per* relation, more disconnected structure), which could explain this.

Feature Fusion and Evaluation Pipeline. The graph embeddings are used as features along with text embeddings in a multi-class, multi-label topic classification task. Given the small size of the dataset it was difficult to use supervised neural encoding architectures to learn intermediary representations, e.g. Bi-LSTM or Transformer, (no meaningful convergence), we rather used a classical machine learning pipeline with standard classifiers from Scikit-learn (+grid-search on held-out training data and 10-fold cross-validation).[7] Text embeddings for claims were computed through a SOTA unsupervised pooling method (Akbik et al., 2019) implemented in the `flair`[8] library on the basis of language models from the `transformers` repository. We tested most base and large models: DistilRoberta (base models) and GPT-2 (large models) consistently performed best and were retained in the evaluation.

4 Evaluation and Discussion

Comparison and combination of graph and text embeddings. We explore the performance of graph embedding vs. text embedding features and whether there is any complementarity of the two. We train and evaluate a ridge classifier (bayesian ridge regressor used as a classifier)[9] as per section 3 by using [(1) CKG] graph embedding features (claim left-hand side vector), [(2) TEDR, (3) TEGPT2] text embedding features (pooled token vectors) from DistilRoberta (DR) and GPT2, [(1) & (2), (1) & (3)] the combination of both (concatenation), as reported in the first segment of Table 2. We use the topic associations extracted from the graph for the construction of the dataset (pre human-annotation) as a baseline. If graph embeddings can capture the equivalence structures that were used to create the baseline effectively, we expect that using them as features for the topic classification task will allow us to reach similar performance to that of the baseline.

Graph embeddings alone lead to poor performance, but there is a small complementarity with text embeddings. Adding graph embeddings to GPT-2 Large lowers performance: it is possible that most of the claims and associated reviews are part of GPT2's training data, thus making any information captured from the metadata superfluous. The baseline being the basis for the gold annotations prior to human annotation, it is expected to achieve a very high performance: given the poor performance of graph embedding features alone, it is likely that the model fails to capture these equivalence structures effectively.

Setting	Accuracy	F_1 mi.	F_1 Ma.
Complementarity of graph and text embedding features			
(1) CKG	36.40	51.77	42.84
(2) TEDR	69.60	82.80	79.38
(3) TEGPT2	74.20	86.16	84.43
↗(1) CKG & (2) TEDR	72.00	84.79	81.76
↘(1) CKG& (3) TEGPT2	68.60	81.61	79.57
Graph Baseline	81.00	89.00	88.88
Impact of neighbourhood features			
↗(4) CKG Flat concat	44.79	61.96	56.92
↗(5) CKG Triple concat	44.00	62.22	58.32
↗(4) CKG Flat concat & (2) TEDR	74.80	86.19	83.81
↗(4) CKG Flat concat & (3) TEGPT2	74.80	86.43	84.62
↗(5) CKG Triple concat & (2) TEDR	70.40	84.04	81.62
↗(5) CKG Triple concat & (3) TEGPT2	74.80	86.43	84.62
Ablation studies – Using only keywords for the text embeddings			
↘(6) TEDR KW Only	32.20	45.86	45.39
↘(7) TEGPT2 KW Only	34.00	46.40	45.35
↗(1) CKG & (6) TEDR KW Only	43.60	60.25	57.83
↗(1) CKG & (7) TEGPT2 KW Only	44.60	60.07	57.58
↗(4) CKG Flat concat & (6) TEDR KW Only	47.00	63.81	61.81
↗(4) CKG Flat concat & (7) TEGPT2 KW Only	47.60	63.23	61.24
Ablation studies – Graph embedding model without keywords			
↘(8) CKG No KW	0.60	1.01	0.73
↘(9) CKG Flat concat No KW	11.00	19.90	16.37
↗(8) CKG No KW & (6) TEDR KW Only	34.40	49.19	47.10
↗(8) CKG No KW & (7) TEGPT2 KW Only	34.40	47.08	46.33
↘(9) CKG Flat concat No KW & (6) TEDR KW Only	30.79	46.17	44.84
↘(9) CKG Flat concat No KW & (7) TEGPT2 KW Only	28.40	44.63	44.02
Ablation studies – Text embeddings of all text properties			
↗(10) TEDR All text	71.20	84.40	81.50
↗(11) TEGPT2 All text	72.80	84.61	80.19
↘(1) CKG& (10) TEDR All text	70.80	84.00	81.23
↗(1) CKG& (11) TEGPT2 All text	76.20	86.58	84.42
↘(4) CKG Flat concat & (10) TEDR All text	67.40	81.20	78.92
↗(4) CKG Flat concat & (11) TEGPT2 All text	73.20	84.65	82.37

Table 2: Results for topic classification (10-fold): avg. accuracy, F_1 micro/macro. Top – Complementarity of graph and text embedding features, Middle – Impact of different feature extraction strategies from graph embeddings, Bottom – Ablation studies.

Impact of neighbourhood features. The LHS claim embeddings did not capture much useful information for the task. Given the local nature of the link prediction training criterion, do we need to consider the embeddings of the neighbourhood to find useful features that capture the equivalence structures of the baseline? For each neighbour (author, date, sources, mentions in review and claim), we retrieve the RHS and relation vectors. We aggregate by (1) flat concatenation (Flat Concat.); (2) concat. of triple vectors (claim LHS×relation×neighbour RHS – Triple Concat.). Table 2 presents the results: using the neighbourhood brings a small

[7] Code: https://github.com/claimskg/claimskg-embeddings
[8] https://github.com/zalandoresearch/flair
[9] We evaluated several classifiers from scikit-learn, but report only RidgeClassifier as it consistently led to better average accuracy by a significant margin

improvement (+8.39/CKG, +2.80/CKG+TEDR, +0.60/CKG+GPT2), compared to CKG alone or in combination with text embeddings, particularly using concatenation, although we are far from the baseline.

Ablation studies. For the link prediction models, the most informative features arise from the claim/keyword/topic equivalence structures, as they are used to generate the graph baseline (81% accuracy). To understand if those structures are captured beyond relying on classification performance, we investigate three settings: (1) text embeddings of keywords only (KW only) (2) graph embedding without the keyword subgraph (no keywords, no topic concepts, in green in Figure 1 – CKG No KW); (3) Text embedding of all text fields (claim, review headline, author, keywords, date). Table 2 presents the results. When we remove the keyword subgraph, the graph embedding features become irrelevant for the task (0.60% for CKG No KW). Text embeddings of only keywords lead to a classification performance similar to CKG embeddings with keywords (-4.20/DR, -2.40/GPT2), but capture somewhat different information as their combination leads to an improvement over CKG alone (+10.60 with CKG+GPT2). Concatenating neighbourhood vectors for CKG without keywords leads to lower performance, meaning that the information captured that is useful for this task is captured from the keyword structures. In the last setting, we can verify if this additional information captured by claim graph embeddings is similar to what we get from augmented text embeddings that include all the text from the immediate neighbourhood: the results indicate a small complementary with GPT2 (best overall result at 76.2% accuracy), but degraded performance with DR.

Discussion. We have been able to determine, as hypothesized that most of the useful information learned by the link prediction graph embeddings comes from the subgraph pertaining to keywords (green nodes in Figure 1), however the overall resulting classification performance with only embedding features is low (with or without neighbourhood), especially compared to the baseline. One hypothesis could be that the structure of the keyword subgraph is captured to some extent in the embeddings of claims and in the neighbourhood, but since the link prediction performance itself is low compared to standard graphs, there is only some part of the structure that the graph embedding model manages to capture. Of course, the size of the topic classification dataset plays a role in the classification performance, however if the representations learned on CKG (which is in no-way a small dataset by link prediction standards) were able to capture the relevant structures, we should be able to reach results closer to the baseline and to text embedding features (on the same dataset).

In the setting of this controlled topic classification task, the structures in question are the equivalence cliques between claims, keywords and topic concepts, which are more complex than the direct links that the local link prediction objective is meant to capture. Although recent advanced in link prediction make models capable of capturing specific formal properties of a relation (transitive, reflexive, anti-symmetric, etc.) in multi-relational graphs, they do not go beyond direct links. Given that such models are increasingly used to infer new relations in complex KGs (e.g., in biomedical informatics), this is a significant limitation of using these approaches for the inference of complex relations or for a downstream classification task.

5 Conclusion and Future Work

We evaluated the effectiveness of claim embeddings as features in a topic classification dataset, produced specifically to allow probing how specific features impact classification performance. We evaluate several strategies for feature retrieval from graph embeddings and combine them with text embedding features (flair + DistilRoberta/GPT2). We found a small complimentary between the features, however, the low accuracy resulting from using graph embeddings alone (compared to the baseline) and the ablation studies show that the graph embedding model's reliance on a local link prediction objective likely limits the ability of the model to capture more complex relationships (e.g. equivalence cliques between claims, keywords and topic concepts). This echoes some of the open-problems identified in the 2019 Graph Representation Learning workshop at NeurIPS (Sumba and Ortiz, 2019). Given that link prediction models are increasingly used with complex KGs to infer new relations (KG completion), this limitation is something to keep in mind and should drive researchers working on knowledge graphs to explore more general graph representation learning approaches such as graph neural networks or random-walk approaches.

References

Alan Akbik, Tanja Bergmann, and Roland Vollgraf. 2019. Pooled contextualized embeddings for named entity recognition. In *NACACL: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 724–728, Minneapolis, Minnesota. Association for Computational Linguistics.

Hongyun Cai, Vincent W Zheng, and Kevin Chen-Chuan Chang. 2018. A comprehensive survey of graph embedding: Problems, techniques, and applications. *IEEE Transactions on Knowledge and Data Engineering*, 30(9):1616–1637.

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

Kiran Garimella, Gianmarco De Francisci Morales, Aristides Gionis, and Michael Mathioudakis. 2018. Quantifying controversy on social media. *ACM Trans. on Soc. Comp.*, 1(1):3.

Palash Goyal and Emilio Ferrara. 2018. Graph embedding techniques, applications, and performance: A survey. *Knowledge-Based Systems*, 151:78–94.

Seyed Mehran Kazemi and David Poole. 2018. Simple embedding for link prediction in knowledge graphs. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31*, pages 4284–4295. Curran Associates, Inc.

Timothee Lacroix, Nicolas Usunier, and Guillaume Obozinski. 2018. Canonical tensor decomposition for knowledge base completion. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 2863–2872, Stockholmsmässan, Stockholm Sweden. PMLR.

Yang Li and Tao Yang. 2018. Word embedding for understanding natural language: a survey. In *Guide to Big Data Applications*, pages 83–104. Springer.

Rebecca Passonneau. 2006. Measuring agreement on set-valued items (MASI) for semantic and pragmatic annotation. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, Genoa, Italy. European Language Resources Association (ELRA).

Priya Radhakrishnan, Partha Talukdar, and Vasudeva Varma. 2018. Elden: Improved entity linking using densified knowledge graphs. In *NACACL: Human Language Technologies, Volume 1 (Long Papers)*, pages 1844–1853.

Xavier Sumba and José Ortiz. 2019. Between the interaction of graph neural networks and semantic web. In *Proceedings of the 2019 NeurIPS Workshop on Graph Representation Learning*.

Andon Tchechmedjiev, Pavlos Fafalios, Katarina Boland, Stefan Dietze, Benjamin Zapilko, and Konstantin Todorov. 2019. Claimskg - a knowledge graph of fact-checked claims. In *International Semantic Web Conference*. Springer.

Soroush Vosoughi, Deb Roy, and Sinan Aral. 2018. The spread of true and false news online. *Science*, 359(6380):1146–1151.

Xuezhi Wang, Cong Yu, Simon Baumgartner, and Flip Korn. 2018. Relevant document discovery for fact-checking articles. In *WWW*, pages 525–533.

Benjamin Zapilko, Johann Schaible, Philipp Mayr, and Brigitte Mathiak. 2013. Thesoz: A skos representation of the thesaurus for the social sciences. *Semantic Web*, 4(3):257–263.