

Business Entity Matching with Siamese Graph Convolutional Networks

Evgeny Krivosheev^{1,2}, Mattia Atzeni^{1,3}, Katsiaryna Mirylenka¹
 Paolo Scotton¹, Christoph Miksovics¹ and Anton Zorin¹

¹ IBM Research - Zurich, Switzerland,

² University of Trento, Italy

³ EPFL, Switzerland

evgeny.krivosheev@unitn.it, {atz, kmi, psc, cmi, zor}@zurich.ibm.com

Abstract

Data integration has been studied extensively for decades and approached from different angles. However, this domain still remains largely rule-driven and lacks universal automation. Recent developments in machine learning and in particular deep learning have opened the way to more general and efficient solutions to data-integration tasks. In this paper, we demonstrate an approach that allows modeling and integrating entities by leveraging their relations and contextual information. This is achieved by combining siamese and graph neural networks to effectively propagate information between connected entities and support high scalability. We evaluated our approach on the task of integrating data about business entities, demonstrating that it outperforms both traditional rule-based systems and other deep learning approaches.

Introduction

Although knowledge graphs (KGs) and ontologies have been exploited successfully for data integration [Trivedi et al. 2018; Azmy et al. 2019], entity matching involving structured and unstructured sources has usually been performed by treating records without explicitly taking into account the natural graph representation of structured sources and the potential graph representation of unstructured data [Mudgal et al. 2018; Gschwind et al. 2019]. To address this limitation, we propose a methodology for leveraging graph-structured information in entity matching. Recently, graph neural networks (GNNs) yielded promising results for building and processing distributed representations of nodes in a graph. This kind of approach has already achieved state-of-the-art performance on several tasks, including link prediction, node classification, and graph generation [Zhang et al. 2018; Ying et al. 2018; Xu et al. 2019; Chami et al. 2019]. The key idea of GNNs is to propagate useful information from a node to its neighbors, allowing us to build discriminative node embeddings from the *data* and *their structure*. This property makes GNNs extremely suited to relational databases as well. In particular, databases that store information about business entities explicitly or implicitly incorporate relations between companies and their attributes (e.g., in terms of ownership, branches, and subsidiaries) [Mirylenka et al. 2019]. We therefore envision that

modeling hidden representations of business entities with GNNs will improve performance of entity matching in this use case and, more generally, in the field of data integration.

To investigate this hypothesis, this paper presents a graph-based entity-matching system that is capable of learning the distance function between business entities using a model based on Siamese Graph Convolutional Networks (S-GCN) [Krivosheev et al. 2020]. The neural network is optimized to produce small distances for nodes belonging to the same business entity and large distances for unrelated nodes. We show the power and flexibility of our S-GCN model in recognizing company names with typos, automatically extracting unseen abbreviations, and performing a semantic search on business entities in a relational database.

Siamese GCN for Entity Matching

Network architecture. We propose a model architecture that combines the advantages of graph convolutional networks (GCNs) [Kipf and Welling 2017] and siamese networks [Bromley et al. 1993] to address the entity-matching task. GCNs are a type of graph neural network that shares filter parameters among all the nodes, regardless of their location in the graph. Our Siamese Graph Convolutional Network (S-GCN) incorporates two identical GCNs, as shown in Figure 1. The primary objective of our S-GCN architecture is to learn discriminative embeddings of nodes in a knowledge graph, in such a way that they can then be used for entity matching with previously unobserved data. In our use case, the input layer of the network expects features extracted from textual attributes of the nodes using a pre-trained BERT model [Devlin et al. 2019]. Next, some GCN layers followed by L_2 normalization are used to produce the output embedding for each node. During the training process, the first GCN focuses on a given node n_1 and its local neighborhood in the KG, to produce an M -dimensional embedding $\gamma_{n_1} \in \mathbb{R}^M$ of node n_1 . Similarly, the same procedure is repeated with the second GCN for a different node n_2 , to create an embedding $\gamma_{n_2} \in \mathbb{R}^M$ for n_2 . Our siamese network is optimized with a *contrastive loss* function [Hadsell, Chopra, and LeCun 2006] between the outputs of the two GCNs, namely γ_{n_1} and γ_{n_2} . Intuitively, this loss will be small for two nodes that are connected in the graph, whereas it will be larger for nodes that are not related to each other. We make use of *Subspace Learning* [Wang, Lan, and Zhang

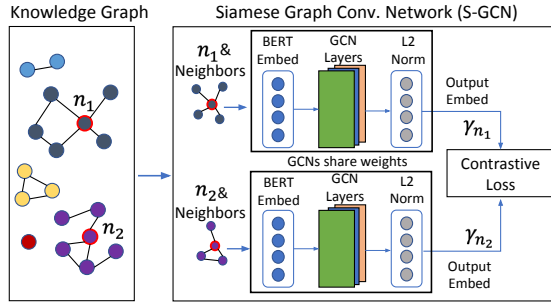


Figure 1: Architecture of the Siamese Graph Conv Network

2017] to carefully and efficiently select informative pair of nodes at training time.

Entity matching with S-GCN. The network is trained on a reference KG (which comprises business entities in our use case) so that the structure of the graph guides the training process. After the S-GCN model has been trained, it projects the KG into an embedding space, where nodes that represent the same entity are mapped to similar vectors. At test time, the trained model can be used to project an unseen entity mention into the same embedding space. Then, K -NN search can be applied to retrieve the closest business entity.

System Evaluation

Data. We set up a data-integration task on a proprietary relational database that contains fine-grained information about business entities (i.e. companies). Companies represented in the database may have different branches, subsidiaries and headquarters, which can be grouped together hierarchically into a tree that represents a single entity. Therefore, we can model our reference database as a graph composed of several disjoint trees. Each node in the graph has a textual attribute that provides the precise name of the branch associated to that node. The resulting training graph contains approximately 40k nodes organized into 1.7k business entities. As a data-augmentation step, we generate an additional canonical or normalized version of a company name and link it to the real name in the graph, using conditional random fields, as described in [Gschwind et al. 2019]. This step yields an enriched training graph with 70k nodes. We are interested in linking unseen company names to an entity in the reference database, and we evaluate our proposed approach using a test set with 1.8k company names.

Scalability. The linkage task involves searching for the nearest-neighbors of a vector in $\mathbb{R}^M$. This is a computationally difficult task that has resulted in the development of *approximate nearest-neighbors* (ANN) algorithms. [Aumüller, Bernhardsson, and Faithfull 2017] have evaluated a set of state-of-the-art ANN algorithms under different conditions on a variety of datasets. They show that it is possible to retrieve ten nearest-neighbors (Euclidean distance) in less than 100 μ s in a 960-dimensional embedding space with recall of 1.00 on servers with Intel Xeon Platinum 8124M CPUs.

Experiments. We compared our S-GCN model against three baselines, namely (i) a record-linkage system (RLS) designed for company entities [Gschwind et al. 2019], (ii) a

Algorithm	KG w/o Augmentation	Augmented KG
RLS	0.71	0.78
NN	0.71	0.75
GCN	0.56	0.70
S-GCN	0.75	0.85

Table 1: Accuracy of entity matching on the test set

feed-forward neural network (NN), and (iii) a model based on graph convolutional networks (GCN). Both the GCN and NN models use BERT features as input and a softmax output layer. We evaluated these algorithms in terms of their accuracy on the test set and we report experimental results in Table 1. As we can see, our S-GCN model consistently outperforms the baselines both on the original KG and on the data augmented with artificially generated canonical names [Gschwind et al. 2019]. Moreover, performing data augmentation on the KG greatly improves the performance of both models based on graph convolutions. Adding such augmented names allows S-GCN to effectively propagate information from the most descriptive parts of the original names. We also performed some experiments using our S-GCN model with fine-tuned BERT embeddings. The fine-tuning allowed us to largely improve the performance of the final system, reaching an accuracy of 0.90 on the test set.

Demonstration

We deployed the network as a service on Kubernetes and we built a user interface to easily search company names and retrieve the match provided by our model. The demonstration will assess the robustness of the system and showcase some interesting properties learned implicitly by our S-GCN architecture. As a first example, the neural network learned to recognize some abbreviations of company names. Both the names “IBM” and “International Business Machines” are matched correctly to the same company, as shown in. Moreover, we applied different kinds of perturbations to company names (e.g., swapping, inserting, replacing, or removing characters) and we noticed that the system was still able to retrieve the correct entity. This shows that the model has implicitly learned an appropriate string similarity measure for our use case. As an example of such ability, we mention that some wrong input names like “Glecor”, “Ilgencore”, “AGlencore”, “Glenocre”, and “Glencorh” were all matched correctly to the most similar company in the KG, namely “Glencore PLC”. In some cases, when the number of typos introduced in the input name was large enough to deviate the prediction of the system, the model still provided a company similar to the input text. Moreover, we also observed that the system was able to perform a semantic search on the reference database. For instance, if provided with the input word “food”, the model retrieves the company “Valentino Gastronomie AG”, which operates in the food domain. Similarly, the word “communications” is matched to the company “Schweizerische Radio- und Fernsehgesellschaft”. This ability is probably given by high-level semantic knowledge incorporated in pre-trained BERT embeddings. A video that demonstrates the examples discussed in this section and many other capabilities of the model is available at <https://youtu.be/LDzCZNRcm3o>.

References

- Aumüller, M.; Bernhardtsson, E.; and Faithfull, A. 2017. ANN-Benchmarks: A Benchmarking Tool for Approximate Nearest Neighbor Algorithms. 34–49. ISBN 978-3-319-68473-4. doi:10.1007/978-3-319-68474-1_3.
- Azmy, M.; Shi, P.; Lin, J.; and Ilyas, I. F. 2019. Matching Entities Across Different Knowledge Graphs with Graph Embeddings. *CoRR* abs/1903.06607. URL <http://arxiv.org/abs/1903.06607>.
- Bromley, J.; Guyon, I.; LeCun, Y.; Säckinger, E.; and Shah, R. 1993. Signature Verification Using a Siamese Time Delay Neural Network. In Cowan, J. D.; Tesauro, G.; and Alspector, J., eds., *Advances in Neural Information Processing Systems 6, [7th NIPS Conference, Denver, Colorado, USA, 1993]*, 737–744. Morgan Kaufmann. URL <http://papers.nips.cc/paper/769-signature-verification-using-a-siamese-time-delay-neural-network>.
- Chami, I.; Ying, Z.; Ré, C.; and Leskovec, J. 2019. Hyperbolic Graph Convolutional Neural Networks. In Wallach, H. M.; Larochelle, H.; Beygelzimer, A.; d’Alché-Buc, F.; Fox, E. B.; and Garnett, R., eds., *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, 8-14 December 2019, Vancouver, BC, Canada*, 4869–4880. URL <http://papers.nips.cc/paper/8733-hyperbolic-graph-convolutional-neural-networks>.
- Devlin, J.; Chang, M.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Burstein, J.; Doran, C.; and Solorio, T., eds., *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, 4171–4186. Association for Computational Linguistics. doi:10.18653/v1/n19-1423. URL <https://doi.org/10.18653/v1/n19-1423>.
- Gschwind, T.; Mikšovic, C.; Minder, J.; Mirylenka, K.; and Scotton, P. 2019. Fast Record Linkage for Company Entities. In *2019 IEEE International Conference on Big Data (Big Data)*, 623–630. ISSN null. doi:10.1109/BigData47090.2019.9006095.
- Hadsell, R.; Chopra, S.; and LeCun, Y. 2006. Dimensionality Reduction by Learning an Invariant Mapping. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2006), 17-22 June 2006, New York, NY, USA*, 1735–1742. IEEE Computer Society. doi:10.1109/CVPR.2006.100. URL <https://doi.org/10.1109/CVPR.2006.100>.
- Kipf, T. N.; and Welling, M. 2017. Semi-Supervised Classification with Graph Convolutional Networks. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net. URL <https://openreview.net/forum?id=SJU4ayYgl>.
- Krivosheev, E.; Atzeni, M.; Mirylenka, K.; Scotton, P.; and Casati, F. 2020. Siamese Graph Neural Networks for Data Integration. *CoRR* abs/2001.06543. URL <https://arxiv.org/abs/2001.06543>.
- Mirylenka, K.; Scotton, P.; Mikšovic, C. A.; and Schade, A. 2019. Similarity matching systems and methods for record linkage. US Patent App. 15/980,066.
- Mudgal, S.; Li, H.; Rekatsinas, T.; Doan, A.; Park, Y.; Krishnan, G.; Deep, R.; Arcaute, E.; and Raghavendra, V. 2018. Deep Learning for Entity Matching: A Design Space Exploration. In Das, G.; Jermaine, C. M.; and Bernstein, P. A., eds., *Proceedings of the 2018 International Conference on Management of Data, SIGMOD Conference 2018, Houston, TX, USA, June 10-15, 2018*, 19–34. ACM. doi:10.1145/3183713.3196926. URL <https://doi.org/10.1145/3183713.3196926>.
- Trivedi, R.; Sisman, B.; Dong, X. L.; Faloutsos, C.; Ma, J.; and Zha, H. 2018. LinkNBed: Multi-Graph Representation Learning with Entity Linkage. In Gurevych, I.; and Miyao, Y., eds., *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, 252–262. Association for Computational Linguistics. doi:10.18653/v1/P18-1024. URL <https://www.aclweb.org/anthology/P18-1024/>.
- Wang, C.; Lan, X.; and Zhang, X. 2017. How to Train Triplet Networks with 100K Identities? In *2017 IEEE International Conference on Computer Vision Workshops, ICCV Workshops 2017, Venice, Italy, October 22-29, 2017*, 1907–1915. IEEE Computer Society. doi:10.1109/ICCVW.2017.225. URL <https://doi.org/10.1109/ICCVW.2017.225>.
- Xu, K.; Hu, W.; Leskovec, J.; and Jegelka, S. 2019. How Powerful are Graph Neural Networks? In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net. URL <https://openreview.net/forum?id=ryGs6iA5Km>.
- Ying, Z.; You, J.; Morris, C.; Ren, X.; Hamilton, W. L.; and Leskovec, J. 2018. Hierarchical Graph Representation Learning with Differentiable Pooling. In Bengio, S.; Wallach, H. M.; Larochelle, H.; Grauman, K.; Cesa-Bianchi, N.; and Garnett, R., eds., *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, 3-8 December 2018, Montréal, Canada*, 4805–4815. URL <http://papers.nips.cc/paper/7729-hierarchical-graph-representation-learning-with-differentiable-pooling>.
- Zhang, M.; Cui, Z.; Neumann, M.; and Chen, Y. 2018. An End-to-End Deep Learning Architecture for Graph Classification. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, 4438–4445. AAAI Press. URL <https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/17146>.