

Beyond Model Performance: Can Link Prediction Enrich French Lexical Graphs?

Hee-Soo Choi, Priyansh Trivedi, Mathieu Constant, Karën Fort and Bruno Guillaume

LREC-COLING 2024



Motivations

Context:

- ▶ Knowledge graphs and lexical graphs are incomplete
- ▶ Link Prediction task addresses this issue but mostly focuses on model performance and English language

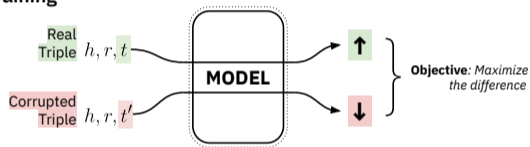
We propose:

- ▶ a resource-oriented approach on two French lexical graphs
- ▶ to extract new relations from a link prediction model to enrich a sparse lexical graph

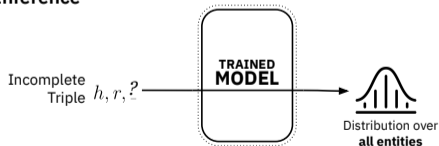
Link prediction task

The link prediction task consists in predicting missing triples in a graph described by a set of triples (h, r, t) for head, relation and tail.

Training



Inference



REZO and JeuxDeMots [Lafourcade and Joubert, 2008]

- ▶ Very dense resource: 6 million nodes and 537 million edges in October 2023
- ▶ Made with GWAPs and semi-automatic mechanisms

The screenshot shows the interface of the JeuxDeMots game. At the top, it says "DONNER DES ASSOCIATIONS D'IDEES AVEC LE TERME QUI SUIT :". Below that, it indicates a record time of "... record à battre de 440 Cr.". The word "consentement" is displayed in large orange letters. On the left, a black ink splat graphic shows the time "Temps 28 s" and a "30s" timer. A search bar contains the text "mettre un terme ici" and has "OK" and "=>" buttons. Below the search bar, it says "Dernier terme proposé : accord • supprimer". A list of "Raffinements possibles" includes: 1. accord (pacte), 2. accord (musique), 3. accord (acceptation), 4. accord (grammaire), 5. accord (droit), and 6. accord (harmonie). At the bottom, there is a note: "Si vous ne savez pas répondre, il faut passer la partie. Si vous pensez qu'il n'y a pas de réponse possible, vous pouvez mettre ***. Vous pouvez supprimer un mot proposé en cliquant dessus dans la liste affichée à droite." On the right side, there are two boxes: "Invité" with a link "Connectez-vous pour plus de détails" and "1/10" with "accord >>".

Datasets for French Link Prediction

- ▶ RezoJDM16k [Mirzapour et al., 2022] and RLF27k
- ▶ Transductive Link Prediction configuration
- ▶ Division into 80%, 10%, 10%

	RezoJDM16k	RLF27k
# nodes	15,746	27,068
# edges	832,093	71,017
# triples Train	665,674	57,643
# triples Valid	83,209	6,674
# triples Test	83,210	6,700

Metrics based on predictions' scores

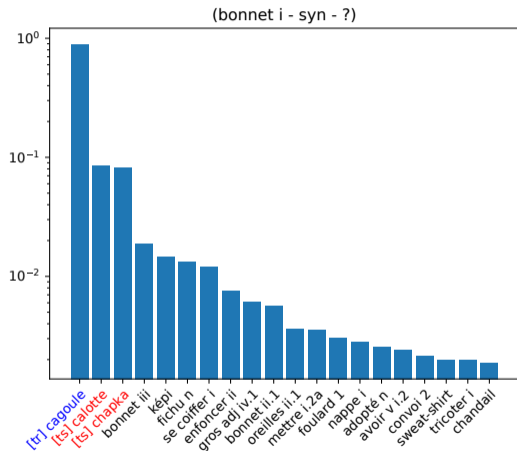
- ▶ MR (Mean Rank): average rank of the positive triples
- ▶ MRR (Mean Reciprocal Rank): average of the reciprocal of ranks of the positive triples
- ▶ Hits@k: proportion of positive triples that appear in the top k of the ranked list of predicted triples.

Training Link Prediction models on RezoJDM16k and RLF27k

Model (For RezoJDM16k)	MRR \uparrow	MR \downarrow	Hits@10 \uparrow	Hits@3 \uparrow	Hits@1 \uparrow
TransE [Bordes et al., 2013]	0.180	200.78	0.437	0.242	0.040
TransH [Wang et al., 2014]	0.217	173.28	0.503	0.293	0.064
TransD [Ji et al., 2015]	0.216	168.18	0.500	0.290	0.065
DistMult [Yang et al., 2015]	0.219	194.16	0.446	0.252	0.109
ComplEx [Trouillon et al., 2016]	0.256	190.79	0.539	0.309	0.119
RotatE [Sun et al., 2019]	0.312	177.04	0.587	0.409	0.155
CompGCN-ConvE [Vashishth et al., 2020]	0.461	171.26	0.659	0.514	0.357

Model (For RLF27k)	MRR \uparrow	MR \downarrow	Hits@10 \uparrow	Hits@3 \uparrow	Hits@1 \uparrow
TransE [Bordes et al., 2013]	0.278	2594.24	0.624	0.497	0.033
TransH [Wang et al., 2014]	0.250	2957.59	0.581	0.465	0.011
TransD [Ji et al., 2015]	0.255	2752.03	0.587	0.472	0.016
DistMult [Yang et al., 2015]	0.373	2748.25	0.613	0.502	0.216
ComplEx [Trouillon et al., 2016]	0.413	3447.98	0.593	0.524	0.284
RotatE [Sun et al., 2019]	0.399	3650.92	0.490	0.454	0.336
CompGCN-ConvE [Vashishth et al., 2020]	0.515	2808.68	0.627	0.559	0.450

Analyzing CompGCN-ConvE model's predictions



▶ [tr] cagoule - 0.893

▶ [ts] calotte - 0.085

▶ [ts] chapka - 0.082

▶ Triples not in the graph: < 0.02

→ Function score only can't discriminate relevant new triples

Computing a confidence score with Monte Carlo Dropout

During inference, we apply Monte Carlo Dropout [Gal and Ghahramani, 2016] :

- ▶ Dropout: **Randomly switching off neurons** in a neural network
- ▶ **100 output distributions** for the same input by sampling **different dropout mask**
- ▶ We compute **how many times a prediction appears in the top 10**.
Example: If it appears 60 times in the top 10, the confidence score is 60%.

Extracting candidates triples

- ▶ We generate all possible combinations of triples for RezoJDM16k and RLF27k.
- ▶ Triples already existing in the graphs are removed.
- ▶ For RLF27K, we extract **triples whose entities are not linked by a directed path** in the graph: 95,766 triples.
- ▶ For RezoJDM16k, due to the high density of the graph, we extract **triples whose paths are of length 3 and 4**: 154,168 triples.

Evaluating the confidence score with manual annotations

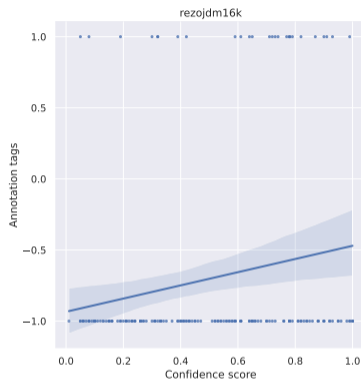
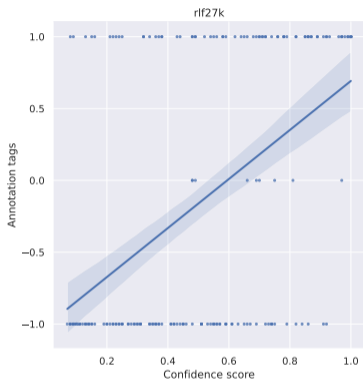
Annotation of 240 triples by 4 annotators for each dataset.

The task is to determine if two entities are linked with semantic or syntactic relation.

Three annotation tags are used:

- ▶ 1: there is a link between the entities
- ▶ -1: there is no link
- ▶ 0: the link is ambiguous or questionable

Correlation between annotations and confidence scores



- ▶ RLF27k: high correlation - triples with high confidence score are relevant
- ▶ RezoJDM16k: poor correlation due to high density of the graph. Maximum path between two nodes is length 4, so two nodes semantically different are related with a relatively short path

Determining a confidence score threshold for RLF27k



→ A confidence threshold of 0.95 results in 100% of triples annotated as correct in RLF27k, which gives us **398 potential good triples** out of the 95,766 candidates.

Relevant new triples for RLF27k

- ▶ (kidnappeur, Syn, ravisseur I) (*kidnapper, Syn, abductor I*)
- ▶ (marchande, Syn, débitante) (*merchant, Syn, retailer*)
- ▶ (motocycliste n-fem, Syn, motarde) (*motorcyclist n-fem, Syn, biker*)

Refined triples in RezoJDM16k

- ▶ 31% of the edges in RezoJDM16k are the general relation associated

In RezoJDM16k	In CompGCN-ConVE's predictions
(infirmière, associated, personne)	(infirmière, is_a, personne)
(herpès, associated, médecine)	(herpès, domain, médecine)
(ouvrir, associated, fermer)	(ouvrir, antonym, fermer)

Conclusion

Contributions:

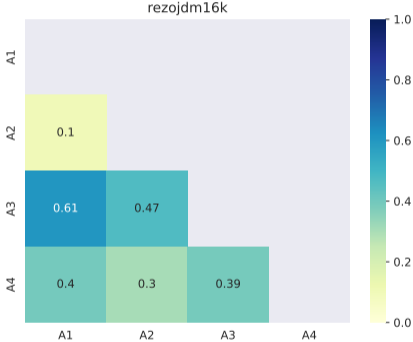
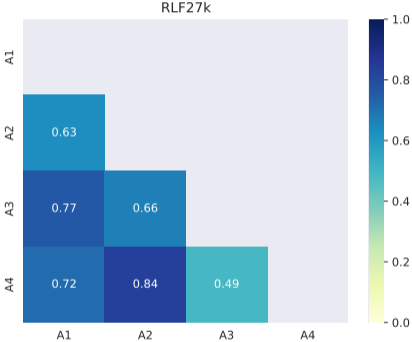
- ▶ Link prediction on 2 French lexical semantic graphs with 7 models
- ▶ Addition of a confidence score to CompGCN-ConvE model's predictions
- ▶ Qualitative analysis of predictions based on manual annotations
- ▶ Extraction of new triples in RL-fr




Limitations:

- ▶ Need for manual verification of candidate triples
- ▶ Influence of the representation of polysemy in different nodes in the RL-fr

Thank you for your attention
Questions?


Inter-annotators agreements





-  Bordes, A., Usunier, N., Garcia-Durán, A., Weston, J., and Yakhnenko, O. (2013). Translating embeddings for modeling multi-relational data. In Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2, NIPS'13, page 2787–2795, Red Hook, NY, USA. Curran Associates Inc.
-  Gal, Y. and Ghahramani, Z. (2016). Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In Balcan, M. F. and Weinberger, K. Q., editors, Proceedings of The 33rd International Conference on Machine Learning, volume 48 of Proceedings of Machine Learning Research, pages 1050–1059, New York, New York, USA. PMLR.
-  Ji, G., He, S., Xu, L., Liu, K., and Zhao, J. (2015). 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, Beijing, China. Association for Computational Linguistics.

-  Lafourcade, M. and Joubert, A. (2008).
JeuxDeMots : un prototype ludique pour l'émergence de relations entre termes.
In
JADT'08 : Journées internationales d'Analyse statistiques des Données Textuelles,
pages 657–666, France.
-  Lux-Pogodalla, V. and Polguère, A. (2011).
Construction of a French Lexical Network: Methodological Issues.
In First International Workshop on Lexical Resources, WoLeR 2011, pages 54–61,
Ljubljana, Slovenia.
-  Mel'čuk, I. (1996).
Lexical functions in lexicography and natural language processing.
Lexical Functions: A Tool for the Description of Lexical Relations in the Lexicon,
pages 37–102.
-  Mirzapour, M., Ragheb, W., Saedizade, M. J., Cousot, K., Jacquenet, H., Carbon, L., and Lafourcade, M. (2022).
Introducing RezoJDM16k: a French KnowledgeGraph DataSet for link prediction.

In Proceedings of the Thirteenth Language Resources and Evaluation Conference, pages 5163–5169, Marseille, France. European Language Resources Association.

 Sun, Z., Deng, Z., Nie, J., and Tang, J. (2019).
Rotate: Knowledge graph embedding by relational rotation in complex space.
In 7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019. OpenReview.net.


 Trouillon, T., Welbl, J., Riedel, S., Gaussier, E., and Bouchard, G. (2016).
Complex embeddings for simple link prediction.
In Balcan, M. F. and Weinberger, K. Q., editors, Proceedings of The 33rd International Conference on Machine Learning, volume 48 of Proceedings of Machine Learning Research, pages 2071–2080, New York, New York, USA. PMLR.

 Vashishth, S., Sanyal, S., Nitin, V., and Talukdar, P. P. (2020).
Composition-based multi-relational graph convolutional networks.
In 8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020. OpenReview.net.

 Wang, Z., Zhang, J., Feng, J., and Chen, Z. (2014).

Knowledge graph embedding by translating on hyperplanes.

Proceedings of the AAAI Conference on Artificial Intelligence, 28.

-  Yang, B., Yih, W., He, X., Gao, J., and Deng, L. (2015).
Embedding entities and relations for learning and inference in knowledge bases.
In Bengio, Y. and LeCun, Y., editors, 3rd International Conference on Learning
Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference
Track Proceedings.