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

Hee-Soo Choi[1,2], **Priyansh Trivedi**[4],
Mathieu Constant[1], **Karën Fort**[2,3], **Bruno Guillaume**[2]
[1]ATILF, CNRS, Université de Lorraine, [2]LORIA, INRIA, Université de Lorraine,
[3]Sorbonne Université, [4]Independent researcher
{hee-soo.choi, karen.fort, bruno.guillaume}@loria.fr, mathieu.constant@atilf.fr, mail@priyansh.page

Abstract

This paper presents a resource-centric study of link prediction approaches over French lexical-semantic graphs. Our study incorporates two graphs, `RezoJDM16k` and `RL-fr`, and we evaluated seven link prediction models, with CompGCN-ConvE emerging as the best performer. We also conducted a qualitative analysis of the predictions using manual annotations. Based on this, we found that predictions with higher confidence scores were more valid for inclusion. Our findings highlight different benefits for the dense graph `RezoJDM16k` compared to the sparser graph `RL-fr`. While the addition of new triples to `RezoJDM16k` offers limited advantages, `RL-fr` can benefit substantially from our approach.

Keywords: lexical graphs, French resources, link prediction, graph completion

1. Introduction

Since the early days of Natural Language Processing (NLP), lexical resources have been essential for the development of NLP systems. However, with the recent advancements in language models and deep learning, there has been a gradual shift away from these resources in favor of vast amounts of web-derived text. Nevertheless, lexical resources continue to hold value in related domains like linguistics and language education. In these areas, lexical resources are created manually by domain experts, a process that produces high-quality resources but is nevertheless time-consuming.

Creating machine-readable resources has been a major challenge of the French community of NLP in the last decades, resulting in a large number of freely available resources made by various semi-automatic techniques. For instance, the French lexicon `Le`*`fff`* was created by merging existing resources, automatically acquiring and manually validating corpus data and syntactic information (Clément et al., 2004; Sagot, 2010). Another example is the `RezoJDM` lexical-semantic graph made by crowdsourcing from GWAPs on the JeuxDeMots platform (Lafourcade and Joubert, 2008; Lafourcade and Le Brun, 2020). This approach has enabled the graph to be expanded to several million nodes and relations.

These lexical-semantic resources are often represented as a knowledge graph (KG), due to their interconnected nature. In these graphs, real-world entities are described from a linguistic point of view by specifying the lexical and semantic relations between them. That is, a fact is represented as a triple in the form `(head, relation, tail)` as in `(dog, hyponymy, puppy)`. However, KGs and lexical-semantic graphs are almost always incomplete due to the impossibility of describing the world or a language exhaustively. This incompleteness can be in the form of coverage i.e., not enough entities or relation types exist in the graph, or completeness i.e, there are missing links between existing entities in the graph. The field of Knowledge Graph Completion (see Sec. 3), which has seen a lot of advancement in recent years, aims to reduce this incompleteness in automated or semi-automated manner. It includes among others, the sub-tasks of entity prediction, relation prediction, and link prediction (Chen et al., 2020), the latter of which is the focus of our work.

In this paper, we adopt a computational linguistic point of view to approach the link prediction task. Our primary emphasis lies in the extraction of potential new triples from the model's predictions, as opposed to solely maximizing model performance. Specifically, we focus our efforts improving the completeness of `RL-fr` (Lux-Pogodalla and Polguère, 2011), a relatively sparse lexical-semantic graph created manually by lexicographers. To better contrast the effectiveness of our approach, we also perform our experiments on the sub-graph of `RezoJDM`, `RezoJDM16k` (Mirzapour et al., 2022), which has significantly higher density than `RL-fr`.

The main contributions of our work can be summarized as follows:

1. We studied the effectiveness of seven link prediction models on two French lexical graphs, and provide state-of-the art results.

2. We added a confidence score to the predictions generated by CompGCN-ConvE model

LREC-COLING 2024, pages 2329–2341

to extract potential new triples.

3. Finally, we conducted a qualitative analysis of these predictions with manual annotations.

Our experiments provide encouraging results and chart a new path towards semi-automated lexical-semantic resource enrichment.

2. Link Prediction Task

The link prediction task consists of predicting missing triples in a graph. There are two primary variants of the task: transductive and inductive link prediction. In transductive prediction, both training and inference occur on the same graph. Conversely, in inductive prediction, inference can take place on a different graph, and the test set may include unseen nodes (Galkin et al., 2022). This paper focuses on transductive link prediction.

KGs can be described as a set of interconnected triples, typically denoted as *(h, r, t)* for head, relation, and tail. Given incomplete triples like *(h, r, ?)* or *(?, r, t)*, the model is required to predict the missing entity. To achieve this, neural models are trained to score true triples higher than false ones using negative sampling (Bordes et al., 2013). It consists of corrupting positive triples by replacing the head or the tail entity by another randomly chosen entity. The aforementioned scores are determined according to the model's score function, which depends on the type of model. In this section, we provide broad overview of some of these types. We refer interested readers to the comprehensive review of Chen et al. (2020) for a more in-depth understanding of neural approaches for link prediction.

Translation models These models use translation distance between node and relation embeddings as their score function. TransE (Bordes et al., 2013) uses Euclidean distance, with subsequent models like TransH (Wang et al., 2014), TransR (Lin et al., 2015), and TransD (Ji et al., 2015) offering various extensions. Some, like RotatE (Sun et al., 2019), use complex vector spaces to represent entities, and define relations as rotations between them. This enables them to model more complex relation patterns like symmetry/asymmetry, inversions and compositions.

Semantic-matching models These models use a score function derived from semantic similarity to discover potential semantic connections between entities and relations. Notable examples include RESCAL (Nickel et al., 2011), which can capture pairwise interactions between entities, DistMult (Yang et al., 2015) that reduces the computational overhead but is constrained to symmetrical relations, and ComplEx (Trouillon et al., 2016), which introduces complex vector space based embeddings for broader modeling capacity.

Deep Neural Architectures for Graphs The approaches introduced above are limited to using simple mathematical operations like inner products or matrix multiplications, over entity and relation embeddings. As a result, their modeling capacity can only be increased by changing the embedding dimensions. The application of deep neural networks however on these graphs is non trivial, and was underexplored until the following two developments. First, Dettmers et al. (2018) proposed ConvE which applies convolution layers over the latent embedding space to model entity-relation interactions and uses a dense layer to calculate the score. In parallel, Graph Convolution Networks (GCNs), proposed in Kipf and Welling (2017), enabled information from different nodes to be propagated across paths in the graph, leading to neighbourhood aware representations of entities and relations. Models like R-GCN (Schlichtkrull et al., 2018) and CompGCN (Vashishth et al., 2020) propose further modifications to handle multi-relational graphs i.e., KGs. Note that GCNs themselves do not solve the link prediction task but provide richer ways to encode the graph. In this work, we also include a CompGCN and ConvE based approach for our context.

Metrics Inspired by information retrieval (IR), the traditional metrics for the link prediction task are based on the ranking of the scores of the correct predictions among all the predictions generated:

- **Mean Rank (MR):** Given a set of predicted triples, MR computes the average rank of the true ones. Lower MR signifies better performance.

- **Mean Reciprocal Rank (MRR):** MRR is the average of the reciprocals of ranks of the true triples. Higher MRR indicates better performance.

- **Hits@k:** This metric computes the proportion of true triples that appear in the top k of the ranked list of predicted triples. A higher Hits@k value denotes superior prediction accuracy for the top k triples.

3. Related Work

While the task of lexical-semantic graph completion remains under-explored, there is a vast literature surrounding the enrichment of lexical graphs.

Broadly, these approaches aim to increase the coverage of the underlying resource i.e., add new nodes to the graph. They rely on new or existing external resources and propose ways to do so. One of the first approaches leveraged statistical co-occurrences to bootstrap existing lexical-semantic graphs (Biemann et al., 2004). These, and other early approaches (Riloff and Shepherd, 1997) required significant manual intervention due to low quality of predictions. Perhaps one of the most important advancement in the field came in the form of `BabelNet` (Navigli and Ponzetto, 2012) which extended `WordNet` (Miller, 1995) by cross referencing word senses with `Wikipedia` articles, and leveraging its multilingual nature. Machine translation models were further used by them, as well as other approaches (Oliver and Climent, 2012; Lam et al., 2014) in varying capacity. Another set of approaches (Taghizadeh and Faili, 2016; Arcan et al., 2016) leverages advancements in Word Sense Disambiguation (WSD), and existence of parallel corpora or dictionaries to enrich non-English Wordnets from English `WordNet`.

On the other hand, approaches which aim to optimize the completeness of these graphs rarely adopt a resource-centric perspective. That is to say that while there are countless approaches which perform transductive link prediction models on these graphs (as Discussed in Sec. 2), very few of these have led to concrete additions to the underlying resource. One exception to this is the work of Fellbaum (1998) on adding links between `WordNet` nodes based on pattern mining from text corpora.

It is worth mentioning that this field has a significant bias towards English resources, as demonstrated by the use of benchmark datasets such as WN18RR (Dettmers et al., 2018) and FB15K-237 (Toutanova and Chen, 2015). While most techniques are in theory, transferable to any graph, very few empirical efforts have been made. For French resources, Mirzapour et al. (2022) study the effectiveness of contemporary link prediction approaches on lexical-semantic graph, although they focus on evaluating models on a French dataset rather than enriching the resource.

This remains an important gap in the field since the effectiveness of graph machine-learning algorithms varies greatly based on the nature of graphs (See Sec. 4.3).

4. French Lexical Graphs

4.1. RezoJDM and RezoJDM16k

`RezoJDM` is a French lexical-semantic network developed using various methods, including GWAPs (Game With A Purpose), contributory approaches, and inference mechanisms (Lafourcade and Joubert, 2008; Lafourcade and Le Brun, 2020). The platform called JeuxDeMots[1] (Games of Words) offers a variety of games with distinct objectives. Some games are designed to expand the network by adding new entries, while others are focused on verifying the information within the network. For instance, the main game in JeuxDeMots prompts players to input terms within a specific time frame based on a given term and relation type (cf. Figure 1). Using GWAPs and crowdsourcing for construction has resulted in a highly dense directed graph. Initiated in 2006, this resource has undergone consistent updates and presently includes more than 537 million relations and six million nodes[2].

There are different types of nodes and relations in `RezoJDM`. Nodes mainly represent terms (type `n_term`) but can also carry other information such as part-of-speech tags (type `n_pos`) or inflected forms (type `n_form`). Relations are divided into three categories: lexical relations (synonymy, antonymy...), ontological relations (hyperonymy, meronymy...) and predicative relations (agent, consequences...). Nodes and relations have the particularity of being assigned weights based on the dynamics of the game. For relations, a positive weight codes a true relation and conversely a negative weight codes a false relation, which is rather uncommon in KGs.

In the resource, the polysemy of a term is expressed by distinguishing a generic node from its refinement nodes. Figure 1 shows the various refinements of the term `accord` (*agreement*) that the player can choose. However, as the game asks the player to enter as many terms as possible in a limited time, players tend not to refine their answers, which leads to a low density of refined nodes. For instance, the generic node `accord` has a degree of 10,549, while the degrees of refined nodes `accord>pacte` (*agreement>pact*) and `accord>acceptation` (*agreement>acceptance*) are 194 and 123 respectively.

Mirzapour et al. (2022) created the sub-graph `RezoJDM16k` to provide a dataset for the link prediction task. It has been created by applying various filters to the nodes and relations of `RezoJDM`. For instance, only nodes of type `n_term` and weight greater than 50 were retained. The same weight filter was applied to the relations and some types were also removed. Furthermore, relation types occurring fewer than 100 times and nodes with a degree less than 45 have been excluded to enhance a better efficiency of the models. At the end of this process, the graph is composed of 15,746 nodes and 832,093 relations and is the one used in the following experiments.

[1] https://www.jeuxdemots.org

[2] As of October 2023.

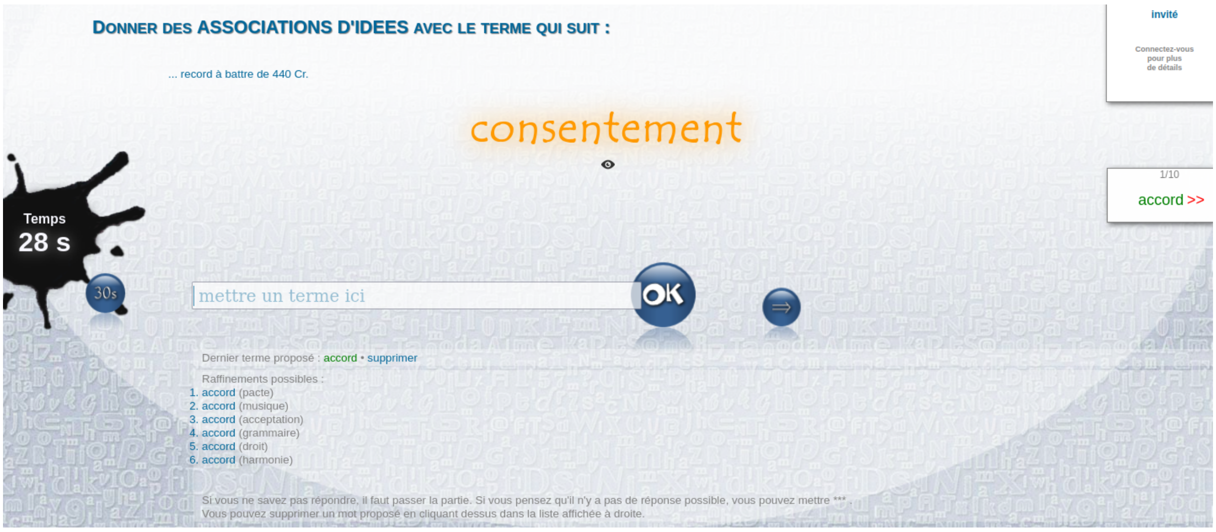

Figure 1: Example of the game JeuxDeMots where the player has to enter terms associated with the term `consentement` (*consent*). The last term entered is `accord` (*agreement*) and the game presents its different meanings (refinements). The player can then refine his answer by choosing one of the meanings.

4.2. RL-fr

`RL-fr` (Réseau Lexical du Français) is a French lexical-semantic network created by lexicographers, where nodes correspond to lexical units and edges to semantic or combinatorial lexical relations (Lux-Pogodalla and Polguère, 2011). The resource is based on the principles of Explanatory Combinatorial Lexicology, the lexical component of the Meaning-Text Theory (Mel'čuk, 1996). In this section, we present information regarding the version 2.1 of `RL-fr`, the version used for our experiments.

The lexical units in `RL-fr` are the fundamental entities for lexicographic description and can be either a lexeme or an idiom. Lexemes are monolexemic lexical units and correspond to word senses. Therefore a polysemic word, called a vocable, is represented as collections of lexical units connected through a copolysemy relation. Version 2.1 of the `RL-fr` contains 29,220 lexical units and 18,625 vocables. Here are the lexemes for the vocable `jambe`, *leg* as examples:

- `Jambe I.1`:
  Marc attend patiemment, les **jambes** croisées.
  *Marc waits patiently, his **legs** crossed.*

- `Jambe I.2a`:
  Le cheval s'est blessé à la **jambe**.
  *The horse hurt its **leg**.*

- `Jambe I.2b`:
  Il y a de la **jambe** de porc au menu.
  *There is pork **leg** on the menu.*

- `Jambe II`:
  La **jambe** droite du pantalon est déchirée.
  *The right **leg** of the trouser is torn.*

- `Jambe III`:
  Une des **jambes** de suspension doit être changée.
  *One of the suspension **legs** must be replaced.*

Unlike conventional dictionaries, where the senses of a word are generally only listed, `RL-fr` represents the types of relation between the senses with the notion of copolysemy, described as the relation among various senses of a word, opposed to polysemy which is the property of words to express several meanings (Polguère, 2018). Therefore, several copolysemy relations exist, as shown in Figure 2 for the vocable `jambe`.

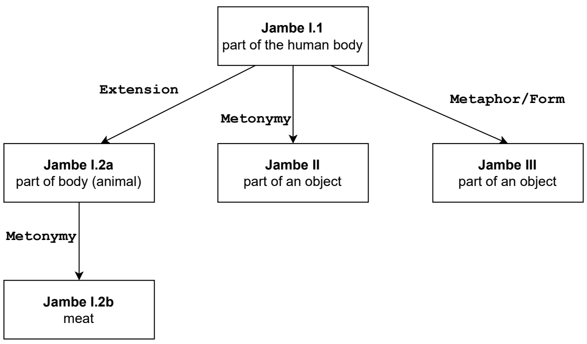

Figure 2: Polysemy structure of the vocable `jambe` *(leg)* in `RL-fr`.

Both paradigmatic and syntagmatic relations are encoded with lexical functions (Mel'čuk, 1996).

Some examples of paradigmatic lexical functions are given below:

- Synonymy (Syn): `vélo` → `bicyclette` *bike* → *bicycle*
- Antonymy (Anti): `accord` → `désaccord` *agreement* → *disagreement*
- Hyperonymy (Gener): `amour` → `sentiment` *love* → *feeling*

Syntagmatic relations include, among others, collocations and support verbs:

- Intensifier (Magn): `boire` → `comme un trou`, *drink* → *like a fish*
- Support verb (Oper): `danger` → `courir` *danger* → *run*, for the expression "courir un danger" which means "take a risk".

Lexical functions can be used to represent simple and complex semantic relations, resulting in 689 different lexical functions, in `RL-fr 2.1`. For our experiments, we used the families of lexical functions to reduce the number of relation types, corresponding to 95 types of lexical functions. For copolysemy relations, we conserved the 11 existing relations. In total, there are 62,641 relations encoded by lexical functions and 9,413 copolysemy relations.

4.3. Different yet complementary graphs

Despite having a common network format, `RezoJDM16k` and `RL-fr` exhibit significant differences in terms of their creation, scope, and the way in which they represent polysemy. Table 1 illustrates the contrast in topology between these two graphs: `RezoJDM16k` is very dense with an average node degree of 105.7 and boasting over 10 times the number of edges compared to `RL-fr`. On the other hand, `RL-fr` presents more nodes, due to the representation of word senses in distinct nodes. Figure 3 shows an overview of the `RL-fr` network around the vocable `accord`, represented with two nodes `accord 1` and `accord 2`[3]. Notice the difference with `RezoJDM16k`, which has a central generic node and refined nodes (cf. Figure 1).

The method of resource creation inherently impacts the graph structure, with manual enrichment being a time-consuming process. However, it ensures controlled, high-quality content verified by domain experts. Regarding `RezoJDM16k`, the quality of the resource is not necessarily lower, as it is contributed by volunteers who are driven by their interest for the language (Lafourcade and Le Brun, 2020), but does present some noise despite the semi-automatic checks.

[3] `https://spiderlex.atilf.fr/fr/q/*accord***`

	`RezoJDM16k`	`RL-fr`
# nodes	15,746	29,220
# edges	832,093	72,054
# edge types	150	106
Mean node degree	105.7	5.2
Min node degree	2	1
Max node degree	10,403	187

Table 1: Statistics on `RezoJDM16k` and `RL-fr`.

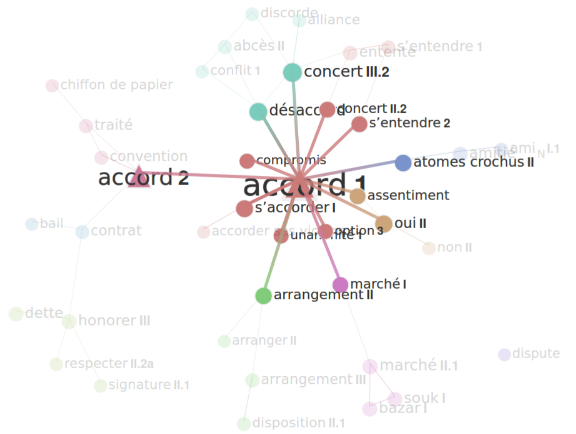

Figure 3: Vocable `accord` *(agreement)* represented in `RL-fr`.

5. Experiments

5.1. Datasets pre-processing

Transductive link prediction implies that all the nodes in the validation and test sets must be present in the training set. To do so, Mirzapour et al. (2022) checked whether the head and the tail of each triple were involved in at least one other triple. If so, the triple was placed in the validation set. Once the desired validation set size was reached, the same process is repeated on the remaining triples to obtain the test and the training sets.

Despite ensuring the inclusion of all nodes of the graph in the training set, this method is not suitable for sparse graphs such as `RL-fr`. Hence, we decided to split randomly the triples into training, validation and test sets (80%, 10%, 10%) and check if all the nodes in the validation and test sets were present in the training set. If not, we removed the triple from the set. For a highly dense graph such as `RezoJDM16k`, no triple was lost even after ten different random splits. On the other hand, a sparse graph such as `RL-fr` implies edge and node losses (- 2,152 nodes, - 1,037 edges). Therefore, we obtained a final graph of 27,068 nodes and 71,017 edges, which we call `RLF27k`. Table 2 presents `RezoJDM16k` and `RLF27k` and their respective training, validation and test sets.

	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

Table 2: Statistics on training, validation and test sets of RezoJDM16k and RLF27k after a 80%, 10%, 10% division.

5.2. Comparing models performance

We evaluated the performance of link prediction models on both datasets using the same six models as described by Mirzapour et al. (2022). We replicated their experiments on RezoJDM16k and conducted them on RLF27k. Additionally, we trained a ConvE model with CompGCN (Vashishth et al., 2020) encoder (referred to as CompGCN-ConvE hereafter) to explore the efficacy of a graph neural network approach on these French datasets.

The results for RezoJDM16k are presented in Table 3. We find that CompGCN-ConvE outperforms all the other models in almost all metrics. Notably, it achieves an Hits@1 of 0.357, which is more than twice that of RotatE. Its Hits@3,10 and MRR scores are consistently best as well and the MR closely matches the best score achieved by TransD.

Table 4 displays the results for RLF27k. Once again, CompGCN-ConvE stands out with an MRR of 0.515, and highest Hits@k scores. However, it's worth noting the high MR, which suggests a notable disparity in the ranks of correct triples: 60% of them rank within the top 10, while the remaining 40% are positioned significantly lower.

5.3. Confidence-aware predictions

With CompGCN-ConvE being the best model on both datasets, we performed a more in-depth analysis of its predictions. Beyond assessing the model's performance in predicting the test set triples, we examined all predictions made by the model for a given head entity and relation by ranking the predictions according to their score.

In the example in Figure 4, we generated the top-20 predictions for the head entity bonnet i and the relation synonym of RLF27k dataset. The model logically predicts a triple present in the training set with a high score (0.893), then predicts two triples from the test set with scores around 0.08. We assume that the remaining triples, which do not exist in the original graph, might be potential candidates. Nevertheless, the score function does not provide a meaningful way to evaluate the relevance of these triples, given that all scores are very low (around 0.01).

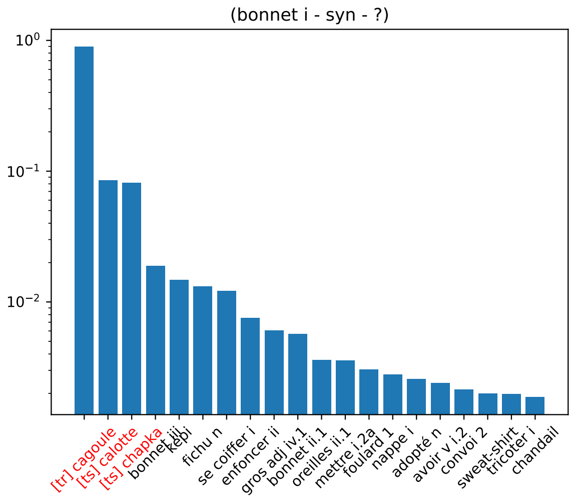

Figure 4: Predictions for the head node bonnet i (*hat*) and the relation synonymy. Triples that already exist in the dataset RLF27k are in red ([tr] for training set and [ts] for test set). Note that the y-axis is scaled logarithmically for better visibility.

Based on this, we aim to approximate the predictive distribution for a given point from our parameterized model. The predictive distribution $p(\mathbf{y}^*|\mathbf{x}^*, \mathbf{X}, \mathbf{Y})$ for a new data point x^* given a dataset $(\mathbf{X}, \mathbf{Y})$ and a model parameterized by θ is:

$$p(\mathbf{y}^*|\mathbf{x}^*, \mathbf{X}, \mathbf{Y}) = \int p(\mathbf{y}^*|\mathbf{x}^*, \theta) p(\theta|\mathbf{X}, \mathbf{Y})\, d\theta \quad (1)$$

Here, the first term is the likelihood of observing the output y^* given the input x^* and model parameters θ, and second term is the posterior distribution of the parameters given the data. Computing the latter is intractable for deep neural networks due to its high dimensional parameter space. Monte-Carlo (MC) Dropout (Gal and Ghahramani, 2016) allows us to approximate this intractable integral. By performing T stochastic forward passes through the network with dropout enabled, we obtain T predictions for each test input $\mathbf{x}^*$. The empirical distribution of these predictions approximates the predictive distribution $p(\mathbf{y}^*|\mathbf{x}^*, \mathbf{X}, \mathbf{Y})$. Mathematically, this can be expressed as:

$$p(\mathbf{y}^*|\mathbf{x}^*, \mathbf{X}, \mathbf{Y}) \approx \frac{1}{T} \sum_{t=1}^{T} p(\mathbf{y}^*|\mathbf{x}^*, \theta_t) \quad (2)$$

More practically, while we typically disable dropout during inference, to ensure deterministic and less noisy predictions. However, under MC Dropout based inference, we draw multiple predictions for the same input, by sampling a different dropout mask each time. This yields a predictive distribution for a given model, and inputs, which

Model	MRR	MR	Hits@10	Hits@3	Hits@1
TransE	0.180	200.78	0.437	0.242	0.040
TransH	0.217	173.28	0.503	0.293	0.064
TransD	0.216	**168.18**	0.500	0.290	0.065
DistMult	0.219	194.16	0.446	0.252	0.109
ComplEx	0.256	190.79	0.539	0.309	0.119
RotatE	0.312	177.04	0.587	0.409	0.155
CompGCN-ConvE	**0.461**	171.26	**0.659**	**0.514**	**0.357**

Table 3: Results of KGE link prediction models on `RezoJDM16k`.

Model	MRR	MR	Hits@10	Hits@3	Hits@1
TransE	0.278	**2594.24**	0.624	0.497	0.033
TransH	0.250	2957.59	0.581	0.465	0.011
TransD	0.255	2752.03	0.587	0.472	0.016
DistMult	0.373	2748.25	0.613	0.502	0.216
ComplEx	0.413	3447.98	0.593	0.524	0.284
RotatE	0.399	3650.92	0.490	0.454	0.336
CompGCN-ConvE	**0.515**	2808.68	**0.627**	**0.559**	**0.450**

Table 4: Results of KGE link prediction models on `RLF27k`.

provides a richer set of information about model predictions such as the ability to compute confidence scores for any prediction.

Consider for example the incomplete triple `(bonnet i, syn, ?)`. Given the original predictive distribution of the model, as shown in Figure 4, we apply MC Dropout to generate $n = 100$ new distributions for the triple. Then, we set an inclusive criterion to turn real valued predictions to binary decisions, i.e., we check whether the predicted entity is in the top-10 scored predictions. Finally, we calculate the confidence score as the ratio of the number of times the predicted entity appears in the top-10 to the number of predictive distributions n.

This enables us to draw confidence scores about model predictions, and make claims like "according to this model, the tail entity `kepi` is in the top-10 predictions for the input head entity `bonnet i` and relation `synonymy` with 75% confidence". We outline the use of these confidence scores in Sec. 6.2.

6. Qualitative analysis

6.1. Extracting candidates

Our goal is to determine whether this confidence score can be used to identify relevant triples that could be incorporated into graphs.

First, we generated all the possible combinations of triples for both datasets and we removed triples already in the graphs. In total, 533,551 predictions were generated for `RLF27k` and 1,720,454 for `RezoJDM16k`. We specifically focused on triples whose entities are not linked by any directed path. For `RLF27k`, we obtained 95,766 final triples.

In the case of `RezoJDM16k`, due to the high density of the graph, all entities are linked via shortest paths ranging from 1 to 4. To apply a similar methodology as for `RLF27k`, we kept triples with paths of lengths 3 and 4, resulting in a total of 154,168 triples.

6.2. Manual annotations

To evaluate the confidence score, 240 triples per dataset were annotated by four annotators, each of whom annotated 120 triples, thus obtaining two annotations for each triple and calculating an inter-annotator agreement (IAA). Confidence scores are homogeneously represented in the samples of each annotator, grouped by 0.1.

The annotation task consists of determining whether a semantic or syntactic link exists between two entities, without taking into account the relation. Annotators have no information outside the triple. Three annotation tags are possible:

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

Figures 5 and 6 show the pairwise IAA with a Cohen's kappa, respectively in `RezoJDM16k` and `RLF27k` samples, with A1, A2, A3 and A4 referring to the four annotators. The IAA appears notably stronger in the `RLF27k` sample with the highest agreement at 0.84 against 0.61 for `RezoJDM16k`. The minimum agreement in `RezoJDM16k` is particularly low at 0.1, in contrast to `RLF27k`'s 0.49.

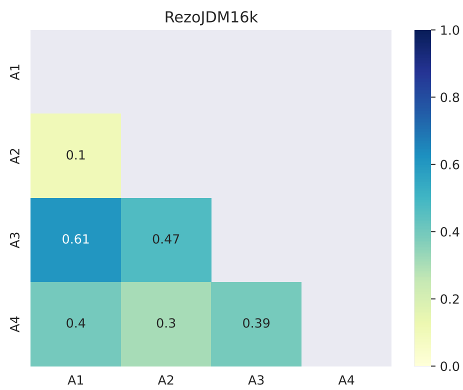

Figure 5: Cohen's kappa values on `RezoJDM16k`.

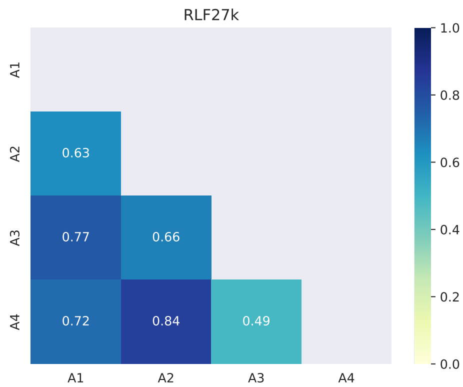

Figure 6: Cohen's kappa values on `RLF27k`.

6.3. Results discussion

The difference in the IAA between `RezoJDM16k` and `RLF27k` can be attributed to the unequal distribution of annotation tags, more specifically to the overrepresentation of the tag -1 in `RezoJDM16k`. Of the 240 triples, the annotators agreed on 183. Of these 183 triples, 85% (156 triples) were annotated as -1 and 15% (27) as 1. The high density of the graph implies that nodes semantically unrelated are connected by a relatively short maximum path of 4, which explains the high proportion of -1 annotations. Figure 7 presents the comparison between the manual annotations and the confidence scores of the triples in the sample. Notably, rare triples that were manually annotated as accurate tend to exhibit high confidence scores. However, due to the prevalence of -1 annotations, a robust correlation is challenging to establish.

On the other hand, in the `RLF27k` sample, out of the 200 triples where both annotators agreed, 56.5% (113) were annotated as -1, 39% (78) as 1 and 4.5% (9) as 0. We can note that the annotation tags in `RLF27k` exhibit a more balanced distribution compared to `RezoJDM16k`. Figure 8 shows a correlation between the annotation tags and the confidence scores on the `RLF27k` triples where the two annotators agree. We can observe that triples annotated as -1 have a low confidence score, while those annotated as 1 have higher confidence score.

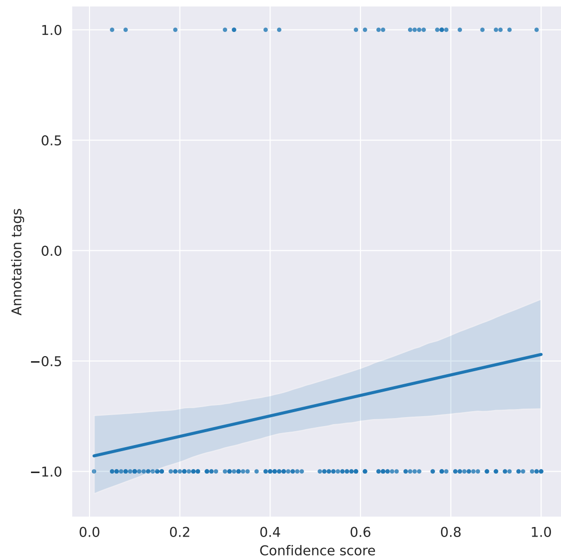

Figure 7: Correlation between annotation values and confidence scores in `RezoJDM16k` sample. The triples considered are those where both annotators agreed.

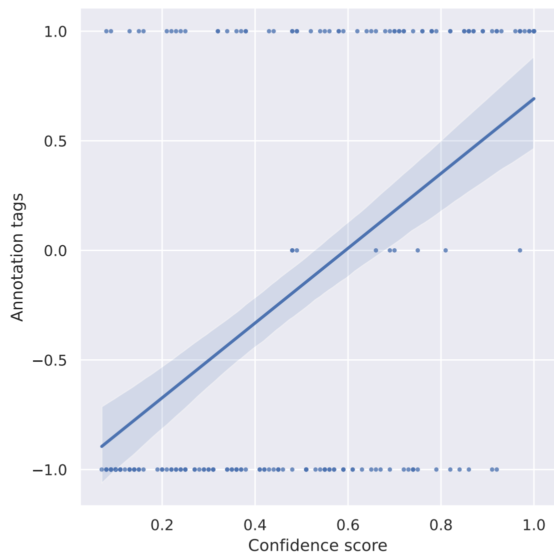

Figure 8: Correlation between annotation values and confidence scores in `RLF27k` sample. The triples considered are those where both annotators agreed.

Figure 9 presents the ratio of predictions annotated as correct by both annotators according to the confidence threshold. We can notice that in-

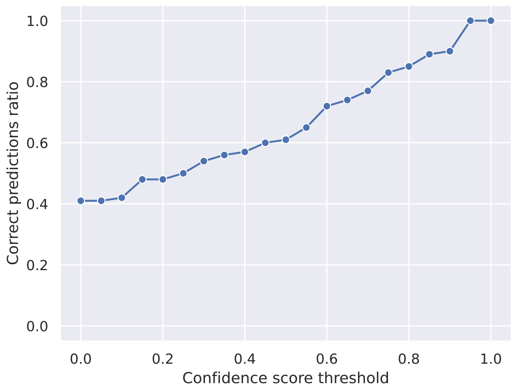

Figure 9: Correct (annotated as 1) predictions ratio of `RLF27k` according to the confidence threshold.

creasing the confidence threshold leads to a higher proportion of correct triples. For instance, choosing a confidence threshold of 0.95 would result in all the triples being correct. On the complete `RLF27k` graph, there are 95,766 triples whose entities are not connected by a directed path and 398 of them have a confidence score greater than 0.95, which can therefore be considered as potential new triples. Here are some examples of relevant candidates:

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

The limitation in `RLF27k` lies in the fine-grained representation of polysemy. A large part of the candidates are triples where the model predicted the wrong sense of the target entity, which is probably due to the copolysemy relations between the senses of a vocable. Thus, another experiment would be to train a model without these relations to reduce this phenomenom.

7. Discussion

In this study, we aimed to identify potential new triples for two lexical-semantic graphs, `RezoJDM16k` and `RLF27k`, using link prediction models. Initially, we assessed seven KGE models on both the dense graph `RezoJDM16k` and the sparser graph `RLF27k`. The model CompGCN-ConvE yielded the best performance on these datasets, showing state-of-the art results for French datasets. We further studied the use of MC Dropout to generate confidence-aware predictions. Based on this, we chose predictions from the best performing model (CompGCN-ConvE), as potential triples to be added to the graph and finally, we evaluated the confidence score using manual annotations.

There is a marked difference between the two graphs in terms of density. For a dense graph like `RezoJDM16k`, adding new triples might not yield significant benefits, given that even entities with slight semantic proximity are connected via short paths. Yet, our approach may prove useful for refining general relations and identifying errors in the graph, especially since it derives its data from a less controlled process of GWAPs. Notably, `RezoJDM16k` heavily features the general `associated` relation, constituting 31% of the edges, which could be refined through the model's predictions. For instance: `(infirmière, associated, personne)`, *(nurse, associated, person)* is refined as `(infirmière, is_a, personne)`, `(herpès, associated, médecine)`, *(herpes, associated, medecine)* as `(herpès, domain, médecine)` and `(ouvrir, associated, fermer)`, *(open, associated, close)* becomes `(ouvrir, antonym, fermer)`.

On the other hand, for a sparse graph created manually by linguists such as `RL-fr`, our approach holds greater significance as it offers a valuable way to enrich the resource. In a final step, we asked annotators to verify the predictions yielded from our aforementioned technique. Based on their results, we find that predictions with high confidence scores are more likely to be selected as valid additions to the graph. As a result, several triples could potentially be integrated into the resource, subject to expert validation.

While the results of our approach are promising, we maintain that manual verification is an important step, as the representation of polysemy into separate nodes directly affects predictions. Given that the model only relies on the graph's structure and neighboring nodes to grasp semantics, there are inherent limitations in predicting the precise entity among the various senses of a vocable.

In subsequent research, beyond intrinsic evaluation methodologies, we intend to conduct extrinsic evaluation using augmented or corrected lexical-semantic graphs in NLP tasks, notably Word Sense Disambiguation (WSD). Furthermore, from a resource-centric perspective, we aim to explore the potential mutual benefit and enhancement between the two French lexical-semantic networks.

All associated code for our experiments and manual annotations are available in this repository: https://github.com/hschoi4/fr-link-prediction.

8. Ethical Considerations

In order to run all the experiments in this study, we estimated that a total of 485 kWh of electricity were used. This gives about 14.5 kg CO_2, using the estimation of carbon intensity (30 g CO_2 eq/kWh). These values were collected with a carbon tracker (Anthony et al., 2020).

We also want to emphasize that while we tried to make this study as reproducible as possible, we are aware that this may not be practical due to the necessary hardware resources. To train the models, a computer with at least 96 GB of system memory (or GPU memory) was needed and we used a cluster of 113 computers with 18-core CPUs during 10 hours to compute all the predictions. This also required the storage of 2 TB of data.

9. Acknowledgments

We thank the reviewers for their useful comments which allowed to improve this paper. We also thank Vincent Tourneur (LORIA, INRIA) for all his help with technical manipulations. Experiments presented in this paper were carried out using the Grid'5000 testbed, supported by a scientific interest group hosted by Inria and including CNRS, RENATER and several Universities as well as other organizations (see `https://www.grid5000.fr`). This work has been partially funded by the French Agence Nationale pour la Recherche, through the SELEXINI project (ANR-21-CE23-0033-01).

10. Bibliographical References

Anne Abeillé, Lionel Clément, and François Toussenel. 2003. Building a treebank for French. In Anne Abeillé, editor, *Treebanks*, pages 165 –187. Kluwer, Dordrecht.

Lasse F. Wolff Anthony, Benjamin Kanding, and Raghavendra Selvan. 2020. Carbontracker: Tracking and predicting the carbon footprint of training deep learning models. *CoRR*, abs/2007.03051.

Mihael Arcan, John Philip McCrae, and Paul Buitelaar. 2016. Expanding wordnets to new languages with multilingual sense disambiguation. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 97–108, Osaka, Japan. The COLING 2016 Organizing Committee.

Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary Ives. 2007. Dbpedia: A nucleus for a web of open data. In *Proceedings of the 6th International The Semantic Web and 2nd Asian Conference on Asian Semantic Web Conference*, ISWC'07/ASWC'07, page 722–735, Berlin, Heidelberg. Springer-Verlag.

Sushil Awale and Abhik Jana. 2022. Enriching Hindi WordNet using knowledge graph completion approach. In *Proceedings of the Workshop on Resources and Technologies for Indigenous, Endangered and Lesser-resourced Languages in Eurasia within the 13th Language Resources and Evaluation Conference*, pages 81–85, Marseille, France. European Language Resources Association.

Collin F. Baker, Charles J. Fillmore, and John B. Lowe. 1998. The Berkeley FrameNet project. In *36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, Volume 1*, pages 86–90, Montreal, Quebec, Canada. Association for Computational Linguistics.

Chris Biemann, Sa-Im Shin, and Key-Sun Choi. 2004. Semiautomatic extension of CoreNet using a bootstrapping mechanism on corpus-based co-occurrences. In *COLING 2004: Proceedings of the 20th International Conference on Computational Linguistics*, pages 1227–1232, Geneva, Switzerland. COLING.

Kurt Bollacker, Robert Cook, and Patrick Tufts. 2007. Freebase: A shared database of structured general human knowledge. In *Proceedings of the 22nd National Conference on Artificial Intelligence - Volume 2*, AAAI'07, page 1962–1963. AAAI Press.

Antoine Bordes, Nicolas Usunier, Alberto Garcia-Durán, Jason Weston, and Oksana Yakhnenko. 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.

Marie Candito, Pascal Amsili, Lucie Barque, Farah Benamara, Gaël De Chalendar, Marianne Djemaa, Pauline Haas, Richard Huyghe, Yvette Yannick Mathieu, Philippe Muller, Benoît Sagot, and Laure Vieu. 2014. Developing a French FrameNet: Methodology and First results. In *LREC - The 9th edition of the Language Resources and Evaluation Conference*, Reykjavik, Iceland.

Marie Candito and Djamé Seddah. 2012. Le corpus sequoia : annotation syntaxique et exploitation pour l'adaptation d'analyseur par pont lexical (the sequoia corpus : Syntactic annotation and use for a parser lexical domain adaptation method) [in French]. In *Proceedings of the Joint Conference JEP-TALN-RECITAL 2012, volume 2: TALN*, pages 321–334, Grenoble, France. ATALA/AFCP.

Zhe Chen, Yuehan Wang, Bin Zhao, Jing Cheng, Xin Zhao, and Zongtao Duan. 2020. Knowledge graph completion: A review. *IEEE Access*, 8:192435–192456.

Laurence Danlos, Takuyua Nakamura, and Quentin Pradet. 2014. Toward a French VerbeNet (vers la création d'un verbnet français) [in French]. In *TALN-RECITAL 2014 Workshop FondamenTAL 2014 : Ressources lexicales et TAL - vue d'ensemble sur les dictionnaires électroniques de Jean Dubois et Françoise Dubois-Charlier (FondamenTAL 2014 : Lexical Resources and NLP)*, pages 103–108, Marseille, France. Association pour le Traitement Automatique des Langues.

Tim Dettmers, Pasquale Minervini, Pontus Stenetorp, and Sebastian Riedel. 2018. Convolutional 2d knowledge graph embeddings. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence*, AAAI'18/IAAI'18/EAAI'18. AAAI Press.

Boyang Ding, Quan Wang, Bin Wang, and Li Guo. 2018. Improving knowledge graph embedding using simple constraints. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 110–121, Melbourne, Australia. Association for Computational Linguistics.

Christiane Fellbaum. 1998. *WordNet: An Electronic Lexical Database*. The MIT Press.

Yarin Gal and Zoubin Ghahramani. 2016. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *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.

Mikhail Galkin, Max Berrendorf, and Charles Tapley Hoyt. 2022. An open challenge for inductive link prediction on knowledge graphs.

Guoliang Ji, Shizhu He, Liheng Xu, Kang Liu, and Jun Zhao. 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.

Shaoxiong Ji, Shirui Pan, Erik Cambria, Pekka Marttinen, and Philip S. Yu. 2022. A survey on knowledge graphs: Representation, acquisition, and applications. *IEEE Transactions on Neural Networks and Learning Systems*, 33(2):494–514.

Yantao Jia, Yuanzhuo Wang, Xiaolong Jin, Hailun Lin, and Xueqi Cheng. 2017. Knowledge graph embedding: A locally and temporally adaptive translation-based approach. *ACM Trans. Web*, 12(2).

Pengcheng Jiang, Shivam Agarwal, Bowen Jin, Xuan Wang, Jimeng Sun, and Jiawei Han. 2023. Text augmented open knowledge graph completion via pre-trained language models. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 11161–11180, Toronto, Canada. Association for Computational Linguistics.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, online. Association for Computational Linguistics.

Thomas N. Kipf and Max Welling. 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.

Mathieu Lafourcade and Alain Joubert. 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.

Mathieu Lafourcade and Nathalie Le Brun. 2020. Jeuxdemots : Un réseau lexico-sémantique pour le français, issu de jeux et d'inférences. *Revue Lexique*, 27:47–86.

Khang Nhut Lam, Feras Al Tarouti, and Jugal Kalita. 2014. Automatically constructing Wordnet synsets. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 106–111, Baltimore, Maryland. Association for Computational Linguistics.

Jaejun Lee, Chanyoung Chung, and Joyce Jiyoung Whang. 2023. Ingram: Inductive knowledge graph embedding via relation graphs. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 18796–18809. PMLR.

Yankai Lin, Zhiyuan Liu, Maosong Sun, Yang Liu, and Xuan Zhu. 2015. Learning entity and relation embeddings for knowledge graph completion. *Proceedings of the AAAI Conference on Artificial Intelligence*, 29(1).

Veronika Lux-Pogodalla and Alain Polguère. 2011. Construction of a French Lexical Network: Methodological Issues. In *First International-Workshop on Lexical Resources, WoLeR 2011*, pages 54–61, Ljubljana, Slovenia.

Shiheng Ma, Jianhui Ding, Weijia Jia, Kun Wang, and Minyi Guo. 2017. Transt: Type-based multiple embedding representations for knowledge graph completion. In *Machine Learning and Knowledge Discovery in Databases*, pages 717–733, Cham. Springer International Publishing.

Farzaneh Mahdisoltani, Joanna Biega, and Fabian M. Suchanek. 2013. YAGO3: A Knowledge Base from Multilingual Wikipedias. In *CIDR*, Asilomar, United States.

Igor Mel'čuk. 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.

Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. 2013. Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 746–751, Atlanta, Georgia. Association for Computational Linguistics.

Mehdi Mirzapour, Waleed Ragheb, Mohammad Javad Saeedizade, Kevin Cousot, Helene Jacquenet, Lawrence Carbon, and Mathieu Lafourcade. 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.

Maximilian Nickel, Volker Tresp, and Hans-Peter Kriegel. 2011. A three-way model for collective learning on multi-relational data. In *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, ICML '11, pages 809–816, New York, NY, USA. ACM.

Antoni Oliver and Salvador Climent. 2012. Parallel corpora for wordnet construction: Machine translation vs. automatic sense tagging. In *Computational Linguistics and Intelligent Text Processing*, pages 110–121, Berlin, Heidelberg. Springer Berlin Heidelberg.

Ciyuan Peng, Feng Xia, Mehdi Naseriparsa, and Francesco Osborne. 2023. Knowledge graphs: Opportunities and challenges. *Artificial Intelligence Review*, 56(11):13071–13102.

Sabine Ploux and Bernard Victorri. 1998. Construction d'espaces sémantiques à l'aide de dictionnaires de synonymes. *Revue TAL*, 39:161–182.

Alain Polguère. 2018. A Lexicographic Approach to the Study of Copolysemy Relations *. *Russian Journal of Linguistics = Vestnik Rossijskogo universiteta družby narodov. Seriâ Lingvistika*, 22(4):788 – 820.

Ellen Riloff and Jessica Shepherd. 1997. A corpus-based approach for building semantic lexicons. In *Second Conference on Empirical Methods in Natural Language Processing*.

Tara Safavi, Danai Koutra, and Edgar Meij. 2020. Evaluating the Calibration of Knowledge Graph Embeddings for Trustworthy Link Prediction. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8308–8321, Online. Association for Computational Linguistics.

Benoît Sagot and Darja Fišer. 2008. Building a free French wordnet from multilingual resources. In *OntoLex*, Marrakech, Morocco.

Michael Schlichtkrull, Thomas N. Kipf, Peter Bloem, Rianne van den Berg, Ivan Titov, and Max Welling. 2018. Modeling relational data with graph convolutional networks. In *The Semantic Web*, pages 593–607, Cham. Springer International Publishing.

Fabian M. Suchanek, Gjergji Kasneci, and Gerhard Weikum. 2007. Yago: A core of semantic knowledge. In *Proceedings of the 16th International Conference on World Wide Web*, WWW '07, page 697–706, New York, NY, USA. Association for Computing Machinery.

Zhiqing Sun, Zhi-Hong Deng, Jian-Yun Nie, and Jian Tang. 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.

Nasrin Taghizadeh and Hesham Faili. 2016. Automatic wordnet development for low-resource languages using cross-lingual wsd. *J. Artif. Int. Res.*, 56(1):61–87.

Kristina Toutanova and Danqi Chen. 2015. Observed versus latent features for knowledge base and text inference. In *Proceedings of the 3rd Workshop on Continuous Vector Space Models and their Compositionality*, pages 57–66, Beijing, China. Association for Computational Linguistics.

Théo Trouillon, Johannes Welbl, Sebastian Riedel, Eric Gaussier, and Guillaume Bouchard. 2016. Complex embeddings for simple link prediction. In *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.

Karel Van den Eynde and Piet Mertens. 2006. Le dictionnaire de valence dicovalence : manuel d'utilisation.

Shikhar Vashishth, Soumya Sanyal, Vikram Nitin, and Partha P. Talukdar. 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.

Kai Wang, Yu Liu, and Quan Z. Sheng. 2021. Neighborhood intervention consistency: Measuring confidence for knowledge graph link prediction. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, pages 2090–2096. International Joint Conferences on Artificial Intelligence Organization. Main Track.

Quan Wang, Zhendong Mao, Bin Wang, and Li Guo. 2017. Knowledge graph embedding: A survey of approaches and applications. *IEEE Transactions on Knowledge and Data Engineering*, 29(12):2724–2743.

Yanjie Wang, Daniel Ruffinelli, Rainer Gemulla, Samuel Broscheit, and Christian Meilicke. 2019. On evaluating embedding models for knowledge base completion. In *Proceedings of the 4th Workshop on Representation Learning for NLP (RepL4NLP-2019)*, pages 104–112, Florence, Italy. Association for Computational Linguistics.

Zhen Wang, Jianwen Zhang, Jianlin Feng, and Zheng Chen. 2014. Knowledge graph embedding by translating on hyperplanes. *Proceedings of the AAAI Conference on Artificial Intelligence*, 28.

Han Xiao, Minlie Huang, and Xiaoyan Zhu. 2016. TransG : A generative model for knowledge graph embedding. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2316–2325, Berlin, Germany. Association for Computational Linguistics.

Bishan Yang, Wen-tau Yih, Xiaodong He, Jianfeng Gao, and Li Deng. 2015. Embedding entities and relations for learning and inference in knowledge bases. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

11. Language Resource References

Clément, Lionel and Sagot, Benoît and Lang, Bernard. 2004. *Morphology Based Automatic Acquisition of Large-coverage Lexica*. European Language Resources Association (ELRA).

Lafourcade, Mathieu and Joubert, Alain. 2008. *JeuxDeMots : un prototype ludique pour l'émergence de relations entre termes*.

Lafourcade, Mathieu and Le Brun, Nathalie. 2020. *JeuxDeMots : Un réseau lexico-sémantique pour le français, issu de jeux et d'inférences.*

Lux-Pogodalla, Veronika and Polguère, Alain. 2011. *Construction of a French Lexical Network: Methodological Issues*.

Miller, George A. 1995. *WordNet: A Lexical Database for English*.

Mirzapour, Mehdi and Ragheb, Waleed and Saeedizade, Mohammad Javad and Cousot, Kevin and Jacquenet, Helene and Carbon, Lawrence and Lafourcade, Mathieu. 2022. *Introducing RezoJDM16k: a French KnowledgeGraph DataSet for Link Prediction*. European Language Resources Association.

Roberto Navigli and Simone Paolo Ponzetto. 2012. *BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network*.

Sagot, Benoît. 2010. *The Lefff, a freely available and large-coverage morphological and syntactic lexicon for French*.