Enrichissement de fonctions de perte avec contraintes de domaine et co-domaine pour la prédiction de liens dans les graphes de connaissance

Nicolas Hubert[1,2], Pierre Monnin[3], Armelle Brun[2], Davy Monticolo[1]

[1] Université de Lorraine, ERPI, Nancy, France
[2] Université de Lorraine, CNRS, LORIA, Nancy, France
[3] Université Côte d'Azur, Inria, CNRS, I3S, Sophia-Antipolis, France

{nicolas.hubert, davy.monticolo}@univ-lorraine.fr, armelle.brun@loria.fr, pierre.monnin@inria.fr

Résumé

Les modèles à base d'embeddings pour la prédiction de liens dans les graphes de connaissance sont entrainés avec des fonctions de perte. Les approches traditionnelles considèrent que l'étiquette d'un triplet est soit vraie, soit fausse. Nous affirmons que les triplets négatifs qui sont sémantiquement valides au regard du profil de la relation devraient être traités différemment de ceux sémantiquement invalides. Nous proposons des fonctions de perte guidées par la sémantique. La généralité et la supériorité de notre approche sont clairement établies sur trois jeux de données publics.

Mots-clés

Graphe de connaissance, prédiction de liens, modèle à base d'embeddings, schémas.

Abstract

Knowledge graph embedding models used for link prediction are trained with loss functions w.r.t. a batch of labeled triples. Traditional approaches consider the label of a triple to be either true or false. We posit that negative triples that are semantically valid regarding relation's domain and range should be treated differently from semantically invalid ones. We then propose semantic-driven versions for the three main loss functions for link prediction. The generality and superiority of our approach is clearly demonstrated on three public benchmarks.

Keywords

Knowledge graph, link prediction, embedding models, schemas.

1 Introduction

Cet article a été accepté à ESWC 2024 [1].
Un graphe de connaissance (GC) est une collection de triplets (s, p, o) où le sujet s et l'objet o sont deux entités du graphe, et le prédicat p qualifie la nature de la relation entre ces deux entités. Les GCs sont intrinsèquement incomplets et une des principales tâches est de prédire les liens manquants [1].
La tâche de prédiction de liens (PL) est souvent abordée à l'aide de modèles à base d'embeddings qui représentent les entités et les relations sous forme de vecteurs.
Ces modèles sont entrainés avec des fonctions de perte visant à maximiser les scores assignés aux triplets positifs (présents dans le GC) et à minimiser ceux des triplets négatifs (non présents). Un pan de la littérature étudie l'influence des triplets négatifs [3]. En effet, leur génération (*negative sampling*) se fait en général en remplaçant le sujet ou l'objet d'un triplet positif par une autre entité du GC. Des travaux récents démontrent l'intérêt d'utiliser les informations extraites d'un schéma (ou ontologie) pour générer des triplets négatifs de meilleure qualité, notamment des triplets sémantiquement valides au regard du domaine et co-domaine d'un prédicat [4]. Cependant, ces travaux consistent à contraindre les triplets négatifs lors de l'échantillonnage. La possibilité d'échantillonner des triplets négatifs de tout type (sémantiquement valides ou non) et de les considérer différemment dans la fonction de perte n'a, à notre connaissance, jamais été étudiée.
Dans ce travail, nous proposons des fonctions de perte guidées par la sémantique. Ces fonctions tirent parti de la connaissance des domaines et co-domaines des prédicats extraits d'un schéma. La supériorité de notre approche par rapport aux approches traditionnelles est validée expérimentalement sur plusieurs modèles et de jeux de données.

2 Approche

Les fonctions de perte que nous proposons visent à distinguer les négatifs sémantiquement valides de ceux qui ne le sont pas. Les premiers sont définis comme respectant à la fois le domaine et co-domaine du prédicat, tandis que les seconds violent a minima le domaine ou co-domaine. Le domaine (resp. co-domaine) d'un prédicat est le type d'entités attendu comme sujet (resp. objet). Par exemple, le prédicat `président` attend une `Personne` comme sujet et un `Pays` comme objet.
Nous proposons des versions guidées par la sémantique pour les trois principales fonctions de perte utilisées dans la littérature [5] : la *pairwise hinge loss*, la *1-N binary cross-*

entropy loss et la *pointwise logistic loss*.
Dans ce qui suit, nous étayons le passage de la *pairwise hinge loss* (PHL) telle que définie traditionnellement, à sa version sémantique. La PHL est définie ci-dessous :

$$\mathcal{L}_{PHL} = \sum_{t \in \mathcal{T}^+} \sum_{t' \in \mathcal{T}^-} [\gamma + f(t') - f(t)]_+ \quad (1)$$

où $\mathcal{T}$, f, et $[x]_+$ représentent respectivement un ensemble de triplets, la fonction de score, et la partie positive de x. $\mathcal{T}$ est ensuite séparé en un ensemble de triplets positifs $\mathcal{T}^+$ et un ensemble de triplets négatifs $\mathcal{T}^-$. γ est un hyperparamètre de marge ajustable et qui spécifie à quel point les scores assignés aux triplets positifs doivent être distants des scores assignés aux triplets négatifs correspondants.
La version sémantique de la PHL est alors définie comme suit :

$$\mathcal{L}^S_{PHL} = \sum_{t \in \mathcal{T}^+} \sum_{t' \in \mathcal{T}^-} [\gamma \cdot \ell(t') + f(t') - f(t)]_+$$

$$\text{où } \ell(t') = \begin{cases} 1 \text{ si } t' \text{ est sémantiquement invalide} \\ \epsilon \text{ sinon} \end{cases} \quad (2)$$

La fonction de perte dans l'Équation (2) comporte désormais un exposant *S* pour clarifier qu'il s'agit de la version sémantique. Un choix de $\epsilon < 1$ conduit le modèle à appliquer une marge plus élevée entre les scores des triplets positifs et des triplets sémantiquement invalides qu'entre les triplets positifs et les triplets sémantiquement valides. Pour un triplet positif donné, cela permet de maintenir les scores de ses contreparties négatives sémantiquement valides relativement plus proches par rapport aux scores de ses contreparties sémantiquement invalides. Intuitivement, lorsque le modèle produit des prédictions erronées, un plus grand nombre d'entre elles sont néanmoins censées respecter les contraintes de domaine et de co-domaine imposées par les relations. Ainsi, les prédictions erronées sont supposés être sémantiquement plus proches du triplet positif.
Ce guidage sémantique est permis par l'introduction d'un terme ϵ, que nous appellons colloquialement le *facteur sémantique* et qui a pour but de rapprocher le score des négatifs sémantiquement valides de ceux des triplets positifs. Il est important de souligner le fait que ce facteur sémantique s'insère dans les trois fonctions de perte mentionnées ci-haut (et non seulement la PHL comme détaillé ci-dessus). Ceci démontre la généralité de notre approche, qui peut être étendue à d'autres fonctions de perte. Les définitions des fonctions de perte traditionnelles et de celles guidées par la sémantique sont détaillées dans l'article original [1].
Enfin, il convient de rappeler que notre apprche ne contraint aucunement le processus de génération de triplets négatifs. Au lieu de cela, notre approche distribue dynamiquement les triplets négatifs à différentes parties d'une même fonction de perte. Cela conduit à un traitement différencié selon la nature des triplets négatifs, pour un coût computationnel moindre que les approches les plus sophistiquées de *negative sampling* [1].

3 Résultats

Les jeux de données, expériences et résultats sont détaillés dans l'article original [1]. Le code source est également disponible [1]. Nous étudions spécifiquement notre approche sur 3 jeux de données et 8 modèles différents. A chaque fois, nous comparons les résultats obtenus en utilisant la fonction de perte originale du modèle et en utilisant notre version guidée par la sémantique, au regard de Sem@K [2] ainsi que des métriques traditionnelles basées sur le rang (MRR, Hits@K). Nous observons une nette amélioration des capacités sémantiques des modèles dans la quasi totalité des cas, ainsi qu'une amélioration satisfaisante des résultats en termes de MRR et Hits@K dans la majorité des cas. Ces résultats démontrent que notre approche n'améliore pas seulement la validité sémantique des prédictions, mais est également pertinente au regard des métriques traditionnelles basées sur le rang.

4 Conclusion

Ce travail se concentre sur les principales fonctions de perte utilisées pour la prédiction de liens dans les GCs. En nous appuyant sur l'hypothèse que tous les triplets négatifs ne sont pas égaux pour apprendre de meilleures représentations, nous proposons de les différencier en fonction de leur validité sémantique par rapport au domaine et co-domaine des prédicats lors de l'entrainement des modèles. Notre approche conduit les modèles à faire des prédictions sémantiquement plus plausibles et améliore également leur capacité à assigner un score plus élevé au triplet authentique.

Références

[1] Nicolas Hubert, Pierre Monnin, Armelle Brun, and Davy Monticolo. Treat different negatives differently : Enriching loss functions with domain and range constraints for link prediction. In *The Semantic Web - 21st International Conference, ESWC 2024, Proceedings*.

[2] Nicolas Hubert, Pierre Monnin, Armelle Brun, and Davy Monticolo. Sem@k : Is my knowledge graph embedding model semantic-aware ? volume 14, pages 1–37, 12 2023.

[3] Bhushan Kotnis and Vivi Nastase. Analysis of the impact of negative sampling on link prediction in knowledge graphs. *arXiv preprint 1708.06816*, 2017.

[4] Denis Krompaß, Stephan Baier, and Volker Tresp. Type-constrained representation learning in knowledge graphs. In *The Semantic Web - 14th International Semantic Web Conference (ISWC)*, volume 9366, pages 640–655, 2015.

[5] Andrea Rossi, Denilson Barbosa, Donatella Firmani, Antonio Matinata, and Paolo Merialdo. Knowledge graph embedding for link prediction : A comparative analysis. *ACM Transactions on Knowledge Discovery from Data*, 15(2) :14 :1–14 :49, 2021.

1. https://github.com/nicolas-hbt/semantic-lossfunc/