Utilisation de Modèles BERT pour Classer Automatiquement les Concepts de Domaine en Concepts de Haut Niveau DOLCE: Une Étude des Ontologies OAEI

Guilherme Sousa[1], Rinaldo Lima[2], Renata Vieira[3], Cassia Trojahn[1]

[1] IRIT: Institut de Recherche en Informatique de Toulouse, France
[2] Universidade Rural de Pernambuco, Recife, Brazil
[3] CIDEHUS, Universidade de Évora, Portugal

prenom.nom@irit.fr, rinaldo.jose@ufrpe.br, rinaldo.jose@ufrpe.br

1 Introduction

Les ontologies de haut niveau, avec leurs fondements philosophiques bien ancrés, servent d'outils indispensables dans l'ingénierie d'ontologies, facilitant des tâches telles que l'alignement d'ontologies. Cependant, toutes les ontologies ne sont pas ancrées dans des concepts de haut niveau, et certaines sont trop vastes pour une annotation manuelle. Les classifieurs automatiques de haut niveau proposent une solution en associant les ontologies de domaine aux ontologies de haut niveau. Des efforts récents, tels que [1], se sont concentrés sur la construction de jeux de données d'entraînement à partir des entités OntoWordNet, alignées sur les concepts DOLCE, pour évaluer les classifieurs destinés à prédire les concepts de haut niveau des entités. Cet article étend cette recherche en évaluant les performances des modèles de classification et l'impact de l'utilisation des commentaires d'entités seuls en tant que caractéristiques combinées à l'utilisation de modèles de langage plus grands.

Un autre aspect de cette recherche est de traiter les cas de multi-héritage, où différents concepts de haut niveau dans DOLCE peuvent découler de la hiérarchie des entités. Les améliorations dans ce cas impliquent une étape de désambiguïsation et de filtrage des chemins menant à plusieurs concepts. Le classifieur présentant les meilleurs résultats lors de l'entraînement est ensuite utilisé pour analyser la distribution des concepts de haut niveau dans les ontologies des jeux de données OAEI, ainsi que leurs alignements de référence.

2 Matériaux et méthodes

2.1 Jeux de données d'entraînement

Le jeu de données **Lopes22-5c** [1], le jeu de données original, utilisé pour l'entraînement des modèles de prédiction de concepts de haut niveau, comprend 116838 entités dérivées d'OntoWordNet, chacune liée à l'un des cinq concepts de haut niveau de DOLCE (Endurant, Perdurant, Qualité, Situation et Abstrait). Il est organisé en trois colonnes : Concept (concept de haut niveau DOLCE), Label (`rdfs:label`), et Commentaire (`rdfs:comment`). **Sousa23-5c**, une reconstruction de Lopes22-5c, aborde les préoccupations de multi-héritage en filtrant les entités ambiguës, tandis que **Sousa23-6c** aborde le déséquilibre de distribution des concepts en divisant Endurant en deux sous-groupes, créant ainsi un jeu de données plus équilibré avec six concepts. Pour gérer les cas de multi-héritage, deux scénarios sont considérés : l'un dans WordNet et l'autre dans la hiérarchie DOLCE. Des jeux de données de test basés sur les ontologies d'organisation de conférences de l'OAEI ont également été créés pour évaluer les modèles de classification, avec des concepts de haut niveau attribués en utilisant un alignement de référence fourni dans [3], ce qui donne les jeux de données **Conference-5c** et **Conference-6c**.

2.2 Modèles d'apprentissage

Le modèle de prédiction présenté dans [1] utilise à la fois des étiquettes et des commentaires, comprenant un système composé de deux parties. La première partie utilise un réseau de neurones à propagation (FNN), prenant la moyenne des plongements de mots des étiquettes en entrée, tandis que la deuxième partie utilise une architecture BiLSTM pour contextualiser les plongements appris pour chaque mot dans les commentaires du jeu de données. Bien qu'efficace, des architectures plus robustes comme BERT peuvent fournir de meilleurs résultats, comme indiqué dans [2], en raison de la capacité accrue de BERT à gérer le contexte pour de meilleures représentations de texte en langage naturel.

Des enjeux surviennent lorsque des entités partagent la même étiquette mais sont assignées à différents concepts de haut niveau dans le jeu de données Lopes22-5c, ce qui peut potentiellement affecter la capacité du modèle à discerner adéquatement les concepts. De plus, la rareté des commentaires dans les ontologies pose un obstacle supplémentaire, limitant potentiellement la généralisation du modèle lors des tests. Pour répondre à ces problématiques et améliorer la généralisation, une approche d'entrée unifiée incorporant à la fois des étiquettes et des commentaires a été

adoptée, en exploitant BERT avec une tête de classification pour prédire les concepts de haut niveau.

3 Évaluation expérimentale

En utilisant trois ensembles de données (Lopes22-5c, Sousa23-5c et Sousa23-6c), une validation croisée est utilisée après le sous-échantillonnage des instances de concepts majoritaires pour normaliser le jeu de données avant l'entraînement. Nous avons choisi Glove 6B pour les plongements de mots en raison de son équilibre entre performance et taille du modèle. Plusieurs modèles de base sont testés, notamment Bernoulli Naive Baye (BNB), Réseau de Neurones à Propagation Avant (FNN), Naive Bayes Gaussien (GNB), Arbre de Décision (DT), Forêt Aléatoire (RF), Régression Logistique (LR), et Machine à Vecteurs de Support (SVM). Les modèles proposés Model-Lopes et BERT sont entraînés avec des optimiseurs et des hyperparamètres spécifiques, évalués en utilisant la métrique micro-F1, et testés avec différentes combinaisons d'entrées. Notamment, l'utilisation des seuls commentaires tend à donner de meilleurs résultats dans tous les classifieurs, sauf pour le Naive Bayes Gaussien dans les ensembles de données Lopes22-5c et Sousa23-5c. Le modèle BERT surpasse constamment les autres, atteignant même des résultats significatifs lors de l'utilisation de l'entrée étiquette+commentaire. Les matrices de confusion pour BERT révèlent des erreurs de classification notables, notamment entre les Perdurant et Situation.

4 Application du Meilleur classifieur : Une Évaluation sur les Jeux de Données OAEI

Cette section se penche sur une analyse des concepts de haut niveau à travers divers axes OAEI [1], ainsi qu'un examen des caractéristiques des commentaires au sein des entités ontologiques. Tout d'abord, la distribution des concepts de haut niveau dans les ontologies est analysée en utilisant les estimations du modèle BERT. Les entités dans chaque axe d'ontologie, à l'exclusion des nœuds vides et des propriétés, sont analysées, avec des étiquettes collectées à partir de prédicats d'étiquetage, ou des identifiants de ressources. Les ontologies Complex, Food et BioDiv présentent des concentrations significatives d'Endurants, contrastant avec des concepts plus uniformément distribués dans d'autres axes. De même, la distribution par le modèle entraîné Sousa23-6c met en évidence des concentrations prononcées dans les ontologies d'Anatomie, d'Alimentation, de BioML et de KG. Des divergences entre les deux modèles sont observées, notamment dans les distributions d'entités de Qualité à travers les axes.

La cohérence des alignements entre les entités correspondantes est également évaluée, révélant des similitudes entre les deux modèles dans plusieurs axes et un nombre plus élevé de correspondances du même type dans l'axe de la Conférence, cependant, avec des divergences dans l'Anatomie, reflétant des différences de distribution sous-jacentes. BioDiv, en particulier, met en évidence des enjeux dans l'alignement en raison de classifications conflictuelles. Notamment, les modèles à 5 et 6 classes rencontrent des difficultés avec les correspondances dans MSE, attribuables à des informations d'entité éparses. Enfin, la discussion s'étend à la distribution terminologique, où la rareté des commentaires dans les axes d'ontologie pose des défis pour la généralisation du modèle.

1. Les descriptions détaillées de ces axes peuvent être trouvées sur `https://oaei.ontologymatching.org/2022/` (consulté le 01/07/23)

5 Conclusion et Travaux Futurs

Dans cette étude, la prédiction de concepts de haut niveau est explorée, générant des ensembles de données Sousa23-5c et Sousa23-6c avec 5 et 6 concepts, respectivement, dérivés d'OntoWordNet tout en abordant l'enjeu du multi-héritage lors de la génération d'ensemble de données. Les résultats de ce travail soulignent l'importance de `rdfs:comment` pour la compréhension automatisée des concepts par le système. De plus, le modèle offrant les meilleures performances est appliqué pour estimer les distributions de concepts dans les ontologies à partir des jeux de données OAEI. L'analyse des alignements de référence a relevé une forte proportion de correspondances partageant le même type.

Pour les travaux futurs, l'expérimentation par la mise en place de nouvelles architectures d'apprentissage profond est envisagée pour améliorer les résultats. De plus, l'exploitation de la structure ontologique en tant qu'information contextuelle dans les modèles de classification peut améliorer la prédiction de concepts de haut niveau, en particulier dans les cas d'ambiguïté d'étiquetage. De plus, la prévalence de correspondances partageant le même type dans certaines axes OAEI suggère la possibilité d'améliorer les performances du système de correspondance en alignant les entités avec des types de haut niveau similaires.

Références

[1] Alcides Gonçalves Lopes Junior, Joel Luis Carbonera, Daniela Schimdt, and Mara Abel. Predicting the top-level ontological concepts of domain entities using word embeddings, informal definitions, and deep learning. *Expert Syst. Appl.*, 203 :117291, 2022.

[2] Alcides Lopes, Joel Luis Carbonera, Daniela Schmidt, Luan Fonseca Garcia, Fabrício Henrique Rodrigues, and Mara Abel. Using terms and informal definitions to classify domain entities into top-level ontology concepts : An approach based on language models. *Knowl. Based Syst.*, 265 :110385, 2023.

[3] Daniela Schmidt, Cássia Trojahn, Renata Vieira, and Mouna Kamel. Validating top-level and domain ontology alignments using wordnet. In *Proceedings of the IX ONTOBRAS Brazilian Ontology Research Seminar, Curitiba, Brazil, October 3rd, 2016*, volume 1862 of *CEUR Workshop Proceedings*, pages 119–130. CEUR-WS.org, 2016.