人工知能学会研究会 資料 SIG-SWO-055-07

Wikidata を利用したナレッジグラフ埋め込みの精度向上

Improving the accuracy of knowledge graph embedding using Wikidata

鵜飼孝典 [1*]

[1] 富士通株式会社
[1] Fujitsu Laboratories Ltd.

Abstract: In this paper, we describe a technique for enriching the ontology of background knowledge by connecting to wikidata in order to improve the accuracy of link prediction in knowledge graphs where entities are not typed in detail. We proposed an algorithm to remove entities that are irrelevant to the type of prediction from the training as negative examples. We confirmed that the proposed algorithm improves the accuracy compared to the case without connecting to the existing wikidata.

1 はじめに

Dbpedia[1] など、RDF で表現されたナレッジグラフには、不完全なものが多い. 例えば、Dbpedia のエンティティの 1/3 にはタイプがついていない。そのため、ナレッジグラフから新たな知識を得る場合には、その不完全さをルールなどで補ったり、他のナレッジグラフやデータベースと接続して利用することが多い [2, 3]。

ナレッジグラフ埋め込みは、ナレッジグラフのサブジェクト、プロパティ、オブジェクトの関係性を内積類似度などで近似できるよう、ベクトル表現を与える手法である [4, 5]。ナレッジグラフの不足を予測で補う技術の一つとして、ナレッジグラフ埋め込みが期待され、多くの研究が行われている。しかしながら、例えば、Wikidata[6, 7] は自動でデータを追加する場合は、90% の精度を求めているが、その精度には達していない [4, 2]。

既存のナレッジグラフ埋め込みの技術は、ノードとプロパティの 3 つ組を形式的にニューラル技術を使ってベクトル表現を与えているため、RDF で書かれたナレッジグラフのセマンティクスを表すオントロジーを利用していない。

Wikidata は Wikipedia と同様にインターネットで共同編集されているナレッジベースで、近年多くのデータベースから情報が取り込まれ、エンティティ間の関連付けも充実している。

本論文では、Wikidata のエンティティのタイプ、エンティティ間の関連、階層構造を用いて、不完全なナレッジグラフにおける埋め込み技術を用いたリンク予測の精度向上の方法を提案する。

本論文は、第 2 節で、ナレッジグラフの埋め込み技術やナレッジグラフに外部リソースを接続する技術を紹介する。第 3 節で、外部リソースを接続した場合に効率よく埋め込みを学習するモデルを提案する。さらに第 4 節で、評価データとして、推論チャレンジで用いている推理小説のナレッジグラフと、医薬品の副作用データベースから作成したナレッジグラフを用いて、提案手法の評価を行う。最後に第 5 節で、まとめと今後の課題を述べる。

2 既存研究

2.1 ナレッジグラフの埋め込み技術

ナレッジグラフ埋め込みは、ナレッジグラフのサブジェクト、プロパティ、オブジェクトの関係性を内積類似度などで近似できるよう、ベクトル表現を与える手法である [4, 5]。ナレッジグラフの不足を予測で補う技術の一つとして、ナレッジグラフ埋め込みが期待され、多くの研究が行われている。代表的なナレッジグラフ埋め込み手法である TransE[5] は、ナレッジグラフのサブジェクトとプロパティ、オブジェクトの表現ベクトルをそれぞれ v_s, v_p, v_o としたとき、正例の 3 つ組では、$v_s + v_p = v_o$ の関係が成り立つように学習する。ほかにも関係に応じて空間を制限するモデル (TransH)[8] や行列変換に基づくモデル (TransR)[9]、複素数空間に変換するモデル (ComplEx)[10] など多くの手法が存在する。しかしながら、例えば、Wikidata[6, 7] は自動でデータ

* 連絡先:富士通株式会社
〒211-8588 神奈川県川崎市中原区上小田中 4-1-1
E-mail: ugai@fujitsu.com

を追加する場合は、90% の精度を求めているが、その精度には達していない [4, 2]。

2.2　他のリソースとの接続

ナレッジグラフから新たな知識を得る場合に、他のナレッジグラフやデータベースと接続して利用することがある。そのため、複数のリソースを接続するため技術の研究もおこなわれている。例えば, 文献 [11] では、DBpedia や Freebase のエンティティをタイプするためにサポートベクターマシンを使う方法を提案しており、ナレッジグラフ間の相互リンクを利用して知識グラフ間の相互リンクを利用し，ある知識グラフのインスタンスを 知識グラフ間の相互リンクを利用し，一方の知識グラフのインスタンスを他方の知識グラフに存在するプロパティに基づいて分類することで のプロパティに基づいて分類することで、カバー率と精度を向上させています。

Nuzzolese ら [12] は、知識の種類を予測するために、Wikipedia のリンクグラフを利用することを提案しています。ウィキペディアのリンクグラフを利用して、k-nearest neighbors 分類器を用いて、ナレッジグラフのタイプを k-nearest neighbors 分類器を用いて知識グラフの種類を予測することを提案している。知識グラフが Wikipedia へのリンクを含んでいる場合、Wikipedia ページ間の相互リンクを利用して、特徴ベクトルを作成する。特徴ベクトルを作成することができる。Wikipedia ページ間のリンクには制約がないため、通常、Wikipedia ページ間の相互リンクは、知識グラフ内の対応するエンティティ間のリンクよりも多いといえる。

3　提案手法

3.1　wikidata との接続

対象となるナレッジグラフのエンティティのラベルと Wikidata の日本語のラベルの一致するものを同一のものとみなして行う。

例えば、図 1 のようなナレッジグラフに対し、次のような手順で wikidata と接続した。

- X 病、副作用 a のような各エンティティについて、そのラベルが一致する wikidata のエンティティを検索する
- 検索の結果見つかったエンティティのタイプと subClassOf のプロパティでつながっているエンティティをナレッジグラフに取りこむ

図 1 に wikidata を接続すると図 2 のようになる。

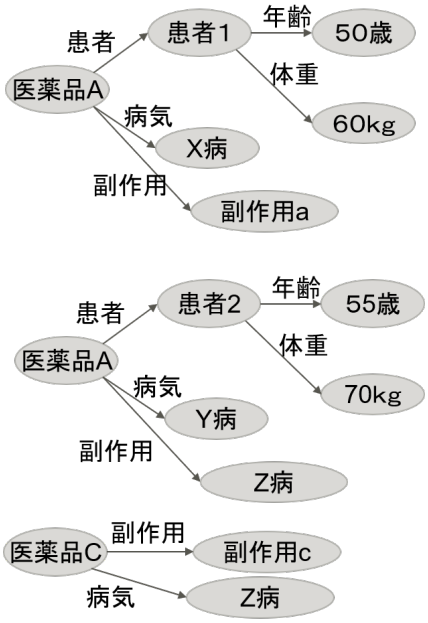

図 1　対象ナレッジグラフの例

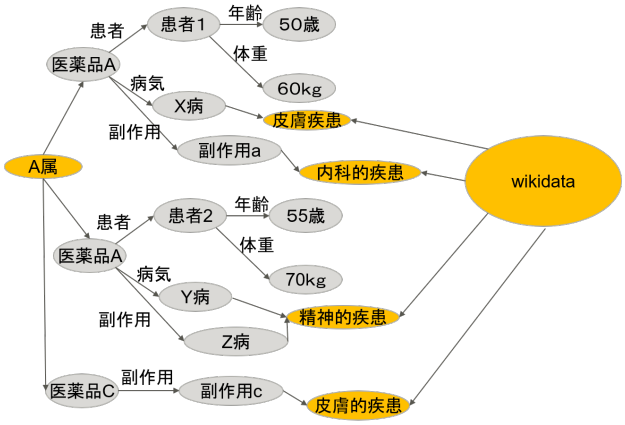

図 2　wikidata との接続

3.2　学習の効率化

タイプが詳細に明らかになったエンティティについて、埋め込みの学習効率を上げるために次のように負例の選択を行った。

- 3 つ組のタイプが同じエンティティのみを学習対象（正例、負例）にする
- 1 つのプロパティのサブジェクトまたはオブジェクトに複数のタイプのエンティティが接続している場合、接続しているすべてのタイプのエンティティを学習対象にする
- 上記外の 3 つ組のタイプの組み合わせは学習対象

にしない

このアルゴリズムにより、負例として選択されるエンティティが同じタイプのものに絞り込まれる。多数のクラスで構成され、グラフとしての密度が低い場合に負例が「接続していない」という情報をより強く保持することになる。

4 評価

4.1 利用したデータ

評価のために、国内の副作用報告とナレッジグラフ推論チャレンジの題材として公開されている短編推理小説のナレッジグラフの 2 つを用いた。Wikidata との接続は、対象となるナレッジグラフのエンティティのラベルと Wikidata の日本語のラベルの一致するものを同一のものとみなして行った。

4.1.1 副作用データ

医薬品の利用において、想定以外の作用（副作用）が起こることは患者の命にかかわることで、副作用をあらかじめ予測することは重要なことである。医薬品の製造販売業者及び医薬関係者等は、副作用によるものと疑われる症例等を知ったときは、医薬品、医療機器等の品質、有効性及び安全性の確保等に関する法律により厚生労働大臣に対して報告することが義務づけられている。医薬品医療機器総合機構は、製造販売業者から報告された国内の副作用報告をオープンデータとして公開している。さらに、この報告データに基づいた分析、予測に関する研究が行われている [**?**]。

一件の報告事例に対し、患者の性別、年齢、身長、体重の属性、服用していた複数の医薬品、原疾患、起こった副作用、転帰などで構成される。

wikidata との接続は、病名、医薬品名、副作用の症状が対象になり、wikidata の分類により類似の病気や医薬品が識別されることが期待される。

本実験では、公開されている CSV 形式のデータを独自に RDF に変換したものを用いている。約 50 万レコードで、6500 万の 3 つ組の RDF に変換された。今回の実験では、ここから 100 分の 1、5000 レコード分を用いて実験を行った。

検証用に報告事例と起こった副作用のリンクの 10% を削除してものを学習用のデータとして用いて、残りの 10% 分の事例と副作用リンクを予測した。

表 1　副作用データを用いた評価実験の結果

Model	MeanRank	Hit@10
TransE	7.20	66
TransH	5.23	78
ComplEx	2.88	82
ComplEx+wikidata	2.22	84

4.1.2 まだらの紐ナレッジグラフ

2018 年から，人工知能学会のセマンティック Web とオントロジー研究会がナレッジグラフ推論チャレンジを開催している [13]。このチャレンジは、シャーロック・ホームズの短編推理小説を題材とし，事件や背景，人物像を知識化したナレッジグラフで与えられる情報に基づいて事件の犯人を正しく突き止め，その理由（証拠やトリックなど）を適切に説明するタスクとしている。今回は、推論チャレンジ [13] のために作成された推理小説のナレッジグラフから「まだらの紐」を用いた。「まだらの紐」のナレッジグラフは、4342 の 3 つ組を持ち、2234 のエンティティと、43 のプロパティでできている。プロパティが、サブジェクトまたは、オブジェクトになっている 3 つ組が 436 ある。このデータセットについては、学習データとテストデータを分離せず、全体を学習データとして、同じく全体をテストデータとして用いた。2234 のエンティティのうち 1241 のエンティティに wikipedia のエンティティが接続され、wikipedia のクラス階層、エンティティ間の階層を含めて、全体で 4450 のエンティティで構成されるナレッジグラフを作成し、実験に用いた。

4.2 利用したモデル

ベースとなるモデルとして、TransE, TransH, ComplEx[10] を用いた。ComplEx は、ナレッジグラフ埋め込み手法の一つであり、ナレッジグラフのサブジェクトとプロパティ、オブジェクトの複素数空間上の表現ベクトルをそれぞれ v_s, v_p, v_o としたとき、正例の 3 つ組では、 $v_s + v_p = v_o$ の関係が成り立つように学習する。実験では、ComplEx と ComplEx に第 3 節で述べた学習アルゴリズムの比較を行った。

4.3 実験結果

4.3.1 副作用データを用いた実験

表 1 は、副作用データを用いた実験の結果である。リンク予測の対象になる副作用は 42 種類である。Com-

表 2　まだらの紐を用いた評価実験の結果

Model	MeanRank	Hit@10
TransE	2.10	89
TransH	2.02	84
ComplEx	1.47	92
ComplEx+wikidata	1.40	92

plEx に wikidata を接続して作成したモデルが最も良い精度になった。

4.3.2 「まだらの紐」データを用いた実験

表 2 のデータは、一部、文献 [14] に掲載されたものを用いている。ComplEx に wikidata を接続して作成したモデルが最も良い精度になった。副作用データに比べて、エンティティの種類が多いが、全データを学習対象にして、全データを予測対象にしたため精度としては高い結果になったと考えている。

5　まとめと今後の課題

本論文では、エンティティに詳細なタイプ付けが行われていないナレッジグラフのリンク予測の精度向上を目的として、wikidata と接続して、背景知識となるオントロジーを充実させる技術について述べた。予測対象になるタイプと無関係なエンティティを負例としての学習対象から外すアルゴリズムを提案し、既存の wikidata と接続しない場合と比較して、精度の向上が得られることを確認した。

今後は、表層的な文字列だけによる接続だけでなく、文献 [12] のようなより意味的な接続を行って、より精度の向上を行いたいと考えている。

参考文献

[1] Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary Ives. Dbpedia: A nucleus for a web of open data. In *The Semantic Web*, pp. 722–735, Berlin, Heidelberg, 2007. Springer Berlin Heidelberg.

[2] Heiko Paulheim. Knowledge graph refinement: A survey of approaches and evaluation methods. *Semantic web*, Vol. 8, No. 3, pp. 489–508, 2017.

[3] AnHai Doan, Alon Halevy, and Zachary Ives. *Principles of Data Integration*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1st edition, 2012.

[4] Dat Quoc Nguyen. An overview of embedding models of entities and relationships for knowledge base completion. *CoRR*, Vol. abs/1703.08098, , 2017.

[5] 拓男濱口, 秀和大岩, 仁新保, 裕治松本. 未知エンティティを伴う知識ベース補完: グラフニューラルネットワークを用いたアプローチ. 人工知能学会論文誌, Vol. 33, No. 2, pp. F–H72_1–10, 2018.

[6] Denny Vrandečić and Markus Krötzsch. Wikidata: A free collaborative knowledgebase. *Commun. ACM*, Vol. 57, No. 10, pp. 78–85, September 2014.

[7] Denny Vrandecic. The rise of wikidata. *IEEE Intelligent Systems*, Vol. 28, No. 4, pp. 90–95, 2013.

[8] Zhen Wang, Jianwen Zhang, Jianlin Feng, and Zheng Chen. Knowledge graph embedding by translating on hyperplanes. In *AAAI*, pp. 1112–1119, 2014.

[9] Yankai Lin, Zhiyuan Liu, Maosong Sun, Yang Liu, and Xuan Zhu. Learning entity and relation embeddings for knowledge graph completion. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, AAAI'15, pp. 2181–2187. AAAI Press, 2015.

[10] ThéoTrouillon, Johannes Welbl, Sebastian Riedel, Éric Gaussier, Guillaume Bouchard. Complex embeddings for simple link prediction, 2016.

[11] Jennifer Sleeman and Tim Finin. Type prediction for efficient coreference resolution in heterogeneous semantic graphs. In *2013 IEEE Seventh International Conference on Semantic Computing*, pp. 78–85, 2013.

[12] Andrea Nuzzolese, Aldo Gangemi, Valentina Presutti, and Paolo Ciancarini. Type inference through the analysis of wikipedia links. *CEUR Workshop Proceedings*, Vol. 937, , 01 2012.

[13] Takahiro Kawamura, Shusaku Egami, Koutarou Tamura, Yasunori Hokazono, Takanori Ugai, Yusuke Koyanagi, Fumihito Nishino, Seiji Okajima, Katsuhiko Murakami, Kunihiko Takamatsu, Aoi Sugiura, Shun Shiramatsu, Shawn Zhang, and Kouji Kozaki. Report on the first knowledge graph reasoning challenge 2018 – toward the explainable ai system, 2019.

[14] 鵜飼孝典, 岡嶋成司. オントロジーを備えた rdf に向いたグラフ埋め込み. 人工知能学会第二種研究会資料, Vol. 2020, No. SWO-051, p. 08, 2020.