人工知能学会研究会 資料 SIG-SWO-051-08

オントロジーを備えた RDF に向いたグラフ埋め込み

Embedding Method for Knowledge Graph with Rich Ontology

鵜飼孝典 [1*] 岡嶋成司 [1]

[1] (株) 富士通研究所
[1] Fujitsu Laboratories Ltd.

Abstract: Knowledge Graph embedding(KGEmbedding) is an expected technology to retrieve a new knowledge with completion of knowledge graph, combining of knowledge graphs and so on. However most of KG Embedding models don't consider ontology part of the knowledge graph. Especially the relationship between properties. We propose a model for knowledge graph with well-described ontology in this article. And we show some results of experiments with basic dataset and our own practical dataset.

1 はじめに

ナレッジグラフ (RDF) の世界では書かれたこと (だけ) が事実であるものとして検索を行う。ところが, 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 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。

そこで本研究では、オントロジーのなかでも、プロパティ間の関係性を利用したナレッジグラフ埋め込みを提案する。既存のナレッジグラフ埋め込み手法は、ナレッジグラフをノード、プロパティ、ノードの 3 つ組の集合と定義しているが、RDF では、プロパティに対する定義を記述することができるため、プロパティがサブジェクト、あるいはノードになることがある。そのため本提案では、プロパティもノードの一部であるとグラフを定義することで、プロパティに関する記述を考慮したモデルを作成した。これにより翻訳などの同じ意味で異なる表現を持つプロパティでつながっているノード同士の関係が密になるように表現される。

本稿は、以下第 2 節で既存のナレッジグラフ埋め込みのオントロジーをうまく利用できていないという課題を代表的な手法である TransE[6] を用いて説明する。第 3 節で我々が提案するナレッジ埋め込み手法 TransU を説明する。第 4 節で既存のナレッジグラフ埋め込み手法と比較した実験結果を示し、第 5 節で考察を述べ、最後にまとめを述べる。

2 既存の手法の問題

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)[7] や行列変換に基づくモデル (TransR)[8]、複素数空間に変換するモデル (ComplEx)[9] など多くの手法が存在する。

これらのモデルでは、グラフは $G = (E, V)$ という形でノードと枝の集合の要素の対の集合と定義されている。また、E と V は独立の集合であるとしている。しかしながら RDF は、枝であるプロパティがノードになることがある。

* 連絡先:(株) 富士通研究所
〒211-8588 神奈川県川崎市中原区上小田中 4-1-1
E-mail: ugai@fujitsu.com

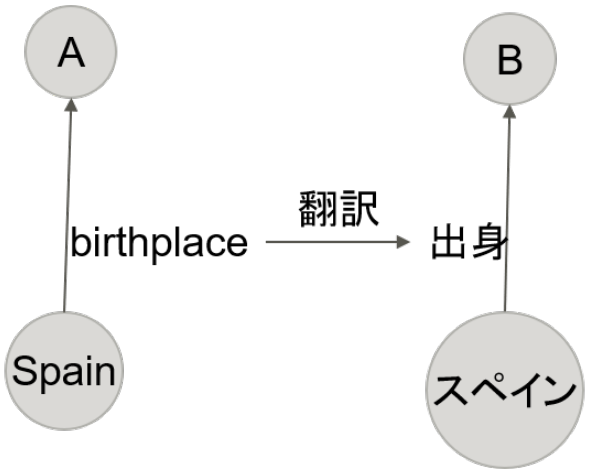

図 1　うまく学習できない例

```
ex:A exp:p1 ex:C .
exp:p1 a exp:T1 .
```

RDF のプロパティは、オントロジーという形で構造が定義されている。例えば、上の記述はプロパティ $p1$ にタイプを定義している。既存のモデルでは、1 行目の $exp:p1$ と 2 行目の $exp:p1$ を異なるものとしてそれぞれ表現ベクトルを定義する。

そのため、既存の埋め込みでは、オントロジーの構造を利用したリンクの予測がうまくいかない例えば、$exp:p1$ にタイプが定義されていなかったときに、タイプを推論することが困難になる。

1 は、既存のナレッジグラフ埋め込み手法でうまく学習できない例の一つである。RDF で表記すると以下のようになる

```
A birthplace Spain .
B 出身スペイン .
Birthplace 翻訳出身
```

この RDF において, birthplace と出身は、翻訳という関係であるため、A の出身地がスペイン、B の birthplace を Spain と予測したい。

2 は、説明を簡便に行うために TransE を用いて、1 の RDF を 2 次元で表現した例である。property の birthplace, 出身とエンティティの birthplace と出身はそれぞれ異なるベクトルで表現されるため、A と スペインに関係があることが学習されない。そのため、A の出身地をスペインと予測することが難しい。

A の出身地がスペイン、B の birthplace を Spain と予測できるためには、3 のように birthplace と出身地が近いベクトルに表現される必要がある。

3 では A と B は同じ出身地なので近いベクトルで表現され、Spain とスペインは、翻訳関係で近いベクトルで表現される。結果として、A とスペインも近いベクトルで表現され、ベクトルの距離などを用いて A の出身地がスペインであると推論できる。

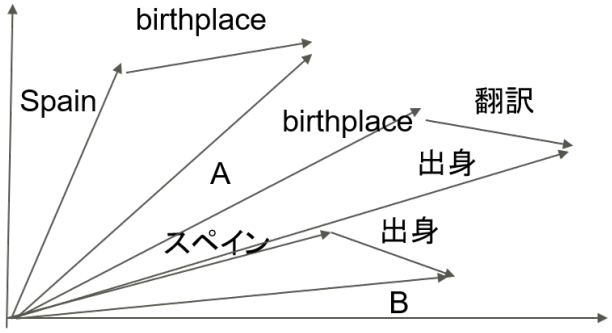

図 2　2 次元ベクトルで表現した例

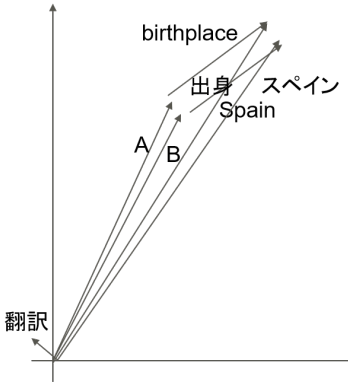

図 3　目指すベクトル表現

3　提案: TransU

本稿で提案する埋め込み手法を以下 TransU と呼ぶ

TransU では 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、すべて合わせたものをエンティティ集合 E とし、サブジェクトとオブジェクトを合わせたエンティティを $E1$, プロパティを $E2$ とし、ナレッジグラフを $G = (E1, E2)$、$E1 \subset E, E2 \subset E1$ と定義し、プロパティをエンティティの部分集合と定義する。

表現ベクトルの学習アルゴリズムは、既存のナレッジグラフ埋め込み手法のものを用いる。ただし、プロパティは、エンティティとして使われる場合も同じベクトルとして計算する。

以上の定義から、TransU と組み合わせる場合、学習アルゴリズムとして用いる既存のナレッジグラフ埋め込み手法においてエンティティとプロパティの次元を同じものにする必要がある。

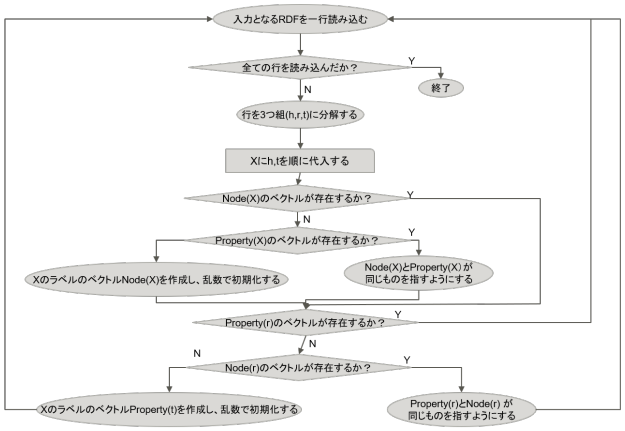

図 4　ベクトル表現の初期化アルゴリズム

既存の手法と組み合わせたときに、表現ベクトルの初期化部分だけ、4 に示したアルゴリズムを用いる。本手法を用いて、プロパティがエンティティとして使われるとき、プロパティを表現するベクトルがエンティティの計算に用いられる。そのため、プロパティ間の関係、プロパティとエンティティの関係がプロパティのベクトルに反映されるプロパティの構造が予測に利用され、精度が上がる

4　評価実験:結果

評価実験は、ナレッジグラフの埋め込み手法の評価によく使われる FB15[10] とプロパティの定義が充実していると考えた推論チャレンジ [11] 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。

FB15K は、Feebase から抽出されたナレッジグラフで 592213 の 3 つ組をもち、14951 のエンティティと 1345 のプロパティでできている。プロパティ間の関係は含まれていない。FB15K については、TransU を TransE と組み合わせたものと、TransH と組みあわせたもので評価した。

1 は、FB15K のデータを用いた評価実験の結果である。組み合わせ対象の埋め込み手法については、元論文に記載のスコアと我々が組み合わせるために独自に実装したもので測定した結果を示している。ベクトルの学習時は、エンティティとプロパティは、同じ空間のベクトルとして学習しているが評価時は、エンティティとプロパティを区別して、エンティティにならないプロパティは、結果から除いて正解の順位を計算している。TransU と組み合わせた実験の結果は、10 回行った平均のスコアとその 10 回のベストスコアになっている。

10 回の平均では、TransU と組み合わせないものよりややスコアが低くなった。

「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。

2 は「まだらの紐」を用いた評価実験の結果である。いずれの場合も表現ベクトルは 200 次元とした。 ComplEx も実数、虚数それぞれ 100 次元とした。学習係数は TransE, TransH が 0.001, ComplEx は最適化関数に Adam を用いて学習係数を 0.01 としている。学習回数はいずれも 1000 回である。 3 つの組み合わせ対象のアルゴリズムに対し TransU を組み合わせたほうが良い結果が得られている。

5　評価実験:考察

FB15K を用いた実験では、10 回の平均では、TransU と組み合わせないものよりややスコアが低くなったが、FB15K には、プロパティに関するオントロジー、すなわち、プロパティがエンティティになる 3 つ組が含まれないため、TransU を組み合わせない場合とそれほど変わらない結果が得られたと考えている。また、先に示したように評価時は、エンティティとプロパティを区別して、エンティティにならないプロパティは、結果から除いて正解の順位を計算している。学習時にエンティティとプロパティを区別していないためにエンティティとプロパティのすべての組み合わせが正解候補となり、学習の結果、エンティティ、エンティティ、プロパティの組が偶然、ある 3 つ組とベクトル空間上の計算において近くなることがある。今回の評価では、このような組み合わせは、ノイズとして除いた。

まだらの紐のデータを用いて行った実験では、プロパティはエンティティになるが、エンティティはプロパティにならないという取り扱いを評価時に行った。そのため、エンティティのタイプがリテラルというような RDF としては不自然な 3 つ組も、正解候補として正解を探す候補に含まれている。サブジェクトとしてリテラルが入ったり、オントロジーによって、レンジやドメインのタイプが決められているときに、それのみを候補にすれば、さらに精度が高まると考えている。

表 1　FB15K を用いた評価実験の結果

Model	MeanRank(Raw)	MeanRank(Filter)	Hit@10(Raw)	Hit@10(Filter)
TransE(paper)	243	125	34.9	47.1
TransH(paper)	212	87	45.7	64.4
TransE(Base)	244	132	32.6	46.1
TransH(Base)	202	84	42.3	54.4
TransU(TransE):Avg	260	30.3	42.3	52.0
TransU(TransE):Best	202	28.2	40.0	42.4
TransU(TransH):Avg	220	48.3	66.6	70.8
TransU(TransH):Best	198	40.4	38.0	40.2

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
TransU(TransE)	2.00	74
TransU(TransH)	1.98	82
TransU(CompEx)	1.42	92

6　おわりに

本稿では，オントロジーのなかでも、プロパティ間の関係性を利用したナレッジグラフ埋め込みを提案した。本提案では、プロパティもノードの一部であるとグラフを定義することで、プロパティに関する記述を考慮したモデルを作成した。これにより翻訳などの同じ意味で異なる表現を持つプロパティでつながっているノード同士の関係が密になるように表現される。本稿で提案するモデルは、既存の埋め込み手法と組み合わせて用いる。評価実験により、プロパティが充実しているナレッジグラフでは既存の埋め込み手法より良い精度を得ることができたが、プロパティ間の関係をあまり持たないナレッジグラフでは、既存の手法とほとんど変わらないことがわかった。本提案の手法は、プロパティ間の関係は、翻訳や異表記などの近しい関係と、反意などの遠い関係を同じように扱っている。これらのプロパティ間の関係の意味を反映したモデルであれば、より意味を反映した推論が可能になると考えている。また RDF では、リテラルはサブジェクトにならないため、エンティティにサブジェクトになるものとならないものがある。これも区別することでより精度を高めることができると考えている。

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] Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko. Translating embeddings for modeling multi-relational data. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pp. 2787–2795. Curran Associates, Inc., 2013.

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

[8] Yankai Lin, Zhiyuan Liu, Maosong Sun, Yang Liu, and Xuan Zhu. Learning entity and relation embed-

dings for knowledge graph completion. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, AAAI ' 15, p. 2181–2187. AAAI Press, 2015.

[9] Théo Trouillon, Johannes Welbl, Sebastian Riedel, Éric Gaussier, and Guillaume Bouchard. Complex embeddings for simple link prediction, 2016.

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

[11] 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.