7C-04

# ドメイン文書による知識グラフ埋め込みを利用したオントロジーの関係抽出手法の提案

丸古凌介 † 大木環美 † 駒井雅之 † 野村雄司 † 稲葉陽子 †

† 株式会社エヌ・ティ・ティ・データ

## 1 はじめに

近年，知識グラフは質問応答や情報検索の要素技術として利用されている．本稿では対象ドメインの知識を体系化したオントロジーと実データの情報を構造化したインスタンスを結びつけたものを知識グラフと定義する．実ビジネスで知識グラフを利用するためには，既存のオントロジーが存在しないことが多く，手動で構築する必要がある．手動の構築には膨大な時間とドメインに関する専門的な知識を要するといった課題が存在する．そこで，我々はオントロジーの半自動構築技術の開発を進めている．オントロジー半自動構築にはいくつかの要素技術が存在するが，本稿では特にオントロジーの概念間の関係を抽出する技術について述べる．関係抽出技術において，オントロジーとインスタンスをベクトル空間に埋め込み，相互に情報補完をしながら関係を抽出する先行研究 JOIE[1] がある．ドメイン文書からオントロジーを構築する場合の JOIE の課題として，実データではドメイン特化ではない一般的な関係が数多く出現し，重要な関係が抽出できない点が挙げられる．本稿では多様なインスタンスに係るドメイン特化ではない関係をグラフ理論に基づいて除去し，またドメイン文書における用語の重みを考慮した抽出手法を提案する．

## 2 先行研究と課題

知識グラフは有向グラフの形で表され，ノードが概念・インスタンスを，エッジが関係を示す．2 つの概念もしくはインスタンス $h$,$t$ とそれらの関係 $r$ が存在するとき，トリプル ($h$,$r$,$t$) が知識グラフには含まれている．関係抽出は関係を予測する知識グラフ構築技術の要素技術であり，本稿では特にオントロジーの概念間での関係を予測する技術について述べる．オントロジーでの関係抽出には様々な手法があり，ドメインコーパスを利用して頻度と確率により抽出する手法，アプリオリアルゴリズムを応用する手法，そして知識グラフ埋め込み技術を応用した手法が先行研究として挙げられる．埋め込み技術は，頻度やアプリオリアルゴリズムを応用した手法と比較して，暗黙的に意味の近しい関係を予測できることで，オントロジーに出現していない関係を他のドメイン文書から抽出することができる．先行研究の一つに知識グラフ埋め込み技術を利用した手法 JOIE がある．JOIE はオントロジーとインスタンスの両者を異なるベクトル空間にそれぞれ埋め込み，別空間への写像関数も同時に学習することで相互に情報補完する手法である．JOIE での埋め込みにはインスタンスがどの概念に紐づくかが情報として必要であり，この情報を用いて写像関数を学習する．概念をインスタンス空間に写像することで，インスタンス空間に出現している関係のうち，確からしいものを概念間の関係として予測することができる．

JOIE でのオントロジーとインスタンスの埋め込みには知識グラフ埋め込み技術の手法を利用でき，本稿では RotatE[2] と呼ばれる手法を採用している．RotatE は知識グラフの補完技術として利用されており，知識グラフ内の欠落を予測することができる手法の一つでもある．RotatE はオイラーの定理を利用して複素ベクトル空間に知識グラフを埋め込み，ベクトル間の距離を関係のベクトルとして表現する．知識グラフに含まれているトリプル ($h$,$r$,$t$) では，$\mathbf{h}$ と $\mathbf{r}$ のベクトルの計算結果が $\mathbf{t}$ のベクトルに近しくなるように学習する．RotatE において関係ベクトル $\mathbf{r}$ は回転の操作と同義であり，各要素が式（1）を満たすベクトルとなる．

$$r_i = (\cos\theta_i, \sin\theta_i), where|r_i| = 1 \quad (1)$$

下記の式 (2) において $\circ$ はベクトル要素ごとの積を実施するアダマール積を表す．

$$d(\mathbf{h}, \mathbf{r}, \mathbf{t}) = ||\mathbf{h} \circ \mathbf{r} - \mathbf{t}|| \quad (2)$$

知識グラフに含まれているトリプル ($h$,$r$,$t$) では式 (2) が 0 に近くなり，含まれないトリプルでは式の値が大きくなるような目的関数を作成し，埋め込みを学習する．

ドメイン文書から構造化したインスタンスとオントロジーの概念を紐づけた知識グラフに，先行研究である JOIE を適用した際に課題が存在する．様々なインスタンス間で多用される関係や一般的な関係が埋め込み時に重要視されてしまい，ドメイン特有の抽出すべき関係が重要視されない点である．ドメイン文書から構造化したインスタンスは論文で利用されている WordNet や DBpedia といった知識グラフと比較して整備されていないためである．そこで次章でグラフ理論を用いて関係をフィルタリングする手法とドメイン用語の重みを用いてドメイン特有の関係を抽出する手法を提案する．

## 3 提案手法

グラフ理論を用いたフィルタリング手法では媒介中心性を利用して多くのインスタンスと関連している関係を算出する．本稿では中心的なノードではなくエッジを算出したいため，ノードとエッジの役割を逆転させることで関係の媒介中心性を計算する．次に関係の媒介中心性を計算するアルゴリズムを述べる．(i) 各関係で結合

---

Relation Extraction for Enterprise Ontology with Knowledge Graph Embedding

† Ryosuke Maruko, Megumi Ohki, Masayuki Komai, Yuji Nomura, Yoko Inaba

† NTT DATA Corporation

表 1　精度評価用インスタンスデータ

| インスタンス数 | 関係数 | トリプル数 |
|---|---|---|
| 7397 | 1851 | 58835 |

しているインスタンスを記録する，(ii) 同じインスタンスに繋がっている関係をノードとしてインスタンス名のエッジで結合する，(iii) 各インスタンスでの関係の次数が 2 以上である場合にはダミー関係のノードを作成し，オリジナルの関係と結合する．出力されるグラフは無向グラフとする．

上記アルゴリズムを適用して構築したグラフで媒介中心性を計算し，閾値よりも高くなった関係を含むトリプルを元のグラフより削除する．

ドメイン用語の重みを考慮する抽出手法としては Relevance 法 [3] を用いた重みづけ埋め込みを行う．Relevance 法は一般的なコーパスとドメインコーパスでの TF-IDF を考慮した計算をすることでドメインコーパス特有の用語を抽出する手法である．本稿では一般的なコーパスを Wikipedia の文書とする．フィルタリングを適用した知識グラフで埋め込みを行った JOIE での予測結果のスコアと動詞・形容詞の Relevance 法のスコアを利用し，ドメイン関係を抽出する．本稿ではベクトルとして学習時の関係に重みづけを行う手法を式 (3) と式 (4) を用いて提案する．式 (3) はトリプルごとの重みを示しており，式 (2) すなわちトリプルの距離の値が大きい際には重みを考慮せず，小さい際には関係 $r$ の Relvance 法によるスコア $s_r$ を重みとする．また $X$ はインスタンスのトリプル集合とする．式 (4) においては複素ベクトルの要素ごとの積において回転角度に重みを考慮させたものである．これらを学習時に適用することでドメイン用語でない関係の重要度を下げる埋め込みとする．

$$d_r = \begin{cases} s_r & (\frac{d(\mathbf{h},\mathbf{r},\mathbf{t})}{\max\{d(\mathbf{h}',\mathbf{r}',\mathbf{t}'),(\mathbf{h}',\mathbf{r}',\mathbf{t}')\in X\}} < 0.5) \\ 1.0 & (otherwise) \end{cases} \quad (3)$$

$$R(h_i, r_i, d_r) = h_i \cdot \{\cos(\theta \cdot d_r), \sin(\theta \cdot d_r)\} \quad (4)$$

## 4　精度評価と考察

精度評価は独自に整備した自動車の構造を表すオントロジーを用いて検証を行った．用いる自動車ドメインのオントロジーは 153 個の概念で構成されており，概念間の関係は上位下位を表す階層情報だけが付与されている．インスタンスはインターネットからクローリングしたドメイン文書を構造化したものを用いる．インスタンスデータセットの詳細は表 1 に記載している．オントロジーの概念とインスタンスの結合情報，正解となるオントロジーの概念間の関係は人手で作成したものを使用する．正解がない，すなわち関係が存在しない概念の組み合わせもあり，関係を予測してしまった場合には不正解とする．評価指標として二値分類の指標に FR という指標を追加した．正解とは異なる関係を予測した数を FR，前述の関係のない概念間に関係を予測してした数を FP として，Precision と Recall，f 値を算出する．Precision と Recall について，Precesion を $TP/(TP+FP+FR)$，Recall を $TP/(TP+FN+FR)$ とし，f 値は従来の算出と同様とした．

表 2 に先行研究である JOIE，グラフ理論によるフィルタリングを適用した JOIE$_G$，JOIE$_G$ にドメイン用語の重みを考慮した JOIE$_{GR}$ について精度評価を実施した．Precision，Recall，f 値ともに JOIE$_{GR}$ が最も良い手法となった．

表 2　精度評価結果 (%)

| 手法 | Pre | Rec | f 値 |
|---|---|---|---|
| JOIE | 13.2 | 35.4 | 19.2 |
| JOIE$_G$ | 55.7 | 67.7 | 61.1 |
| JOIE$_{GR}$ | 57.1 | 67.7 | 62.0 |

JOIE ではさまざまなインスタンス間で多用される「has-a」などの関係が，「has-a」関係にない概念間でも抽出されてしまっていたが，提案手法によりドメイン関係を正しく抽出でき，精度が向上した．JOIE$_{GR}$ では JOIE$_G$ と比較して，多用されていない一般的な関係よりもドメイン特有となる関係を抽出でき性能の向上が見られた．例として「クラッチ」と「クラッチカバー」間において，JOIE$_G$ が「なる」と抽出候補に現れていたのに対して，JOIE$_{GR}$ は「劣化する」と抽出候補に出すことができていた．「なる」は多様なインスタンスに結び付かないが Relevance 法のスコアを考慮したことにより重要度を下げることができている．課題として JOIE$_{GR}$ において「劣化する」を抽出候補に出せたものの，最終的な抽出は「関係なし」となってしまった．今回の手法ではドメイン用語としてスコアの高い関係もすべて距離の値が大きくなってしまう数式となっており，ドメイン用語の距離の値も大きくなってしまう．その結果，関係しない場合のスコアの考慮ができておらず，関係がない場合の重要度を反映することができなかった．Relevance 法のスコアが著しく高いドメイン用語では距離の値を小さくする重みの組み込みを行う改善が必要だと考える．

## 5　おわりに

本稿ではオントロジー関係抽出技術をドメイン文書に適用した際に生じる課題を解決する手法を提案した．また，自動車ドメインのデータセットに対して精度検証を実施したことで，先行研究と比較してドメイン特有の関係を抽出できることを実証した．2 つの知識グラフの片方の関係を他方に写像することができるため，知識グラフの統合・拡張のユースケースにも提案手法が使えるのではないかと考えられる．

## 参考文献

[1] Junheng Hao, Muhao Chen, Wenchao Yu, Yizhou Sun and Wei Wang. Universal Representation Learning of Knowledge Bases by Jointly Embedding Instances and Ontological Concepts. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery and data mining*. ACM, 2019.

[2] Zhiqing Sun, Zhi-Hong Deng, Jian-Yun Nie and Jian Tang. RotatE: Knowledge Graph Embedding by Relational Rotation in Complex Space. ICLR 2019

[3] Anselmo Peñas, Felisa Verdejo and Julio Gonzalo. Corpus-based terminology extraction applied to information access. In *Proceedings of the Corpus Linguistics*, vol. 2001.