4W-06

知識グラフ補完性能向上のための同一エンティティ判定を用いた知識グラフ拡充

近辻 脩壱 † 宗像 北斗 † 武田 龍 † 駒谷 和範 †
† 大阪大学 産業科学研究所

1. はじめに

現在，対話システムでの知識グラフの利用が検討されている．知識グラフとはグラフ構造を持つデータベースである．知識グラフ補完によって予測できるエンティティ間の未知の関係性を，対話文生成に利用できる [1]．しかし，データが少なく知識グラフが疎であり，関係の予測に十分な情報量が存在しない場合，知識グラフ補完性能が低下する．補完性能を向上させる手段として，別のデータベースを用いた知識グラフの拡充が挙げられる．

知識グラフの拡充時，データベース間におけるエンティティ名の表記ゆれにより，エンティティ間に新しく拡充できる関係の数が減少する．表記ゆれのある同一エンティティを特定し，表記を揃えることで，エンティティ間の関係をより多く得られる．

表記ゆれ解消に関して，言語モデルを用いてテキストの正規化を行う研究があるが，これらの手法は学習データが必要であり，ドメインごとに人手で用意するには大きなコストを要する [2]．学習データを必要としない手法として，編集距離に基づく同一エンティティ判定手法がある．これに加え，同一性の判定には，各エンティティが持つ意味や属性の情報も有益であると考えられる．

本研究では，事前学習済み言語モデルに基づき，グラフ情報を活用して同一エンティティ判定を行う．提案法ではエンティティ名に加え，グラフ情報を入力として，埋め込み表現を出力する．埋め込み表現の類似度の計算を行うことで同一エンティティ判定を行う．判定を行ったうえで知識グラフの拡充を行い，拡充後に得られる知識グラフ補完性能について評価する．

2. 知識グラフ拡充による補完性能の向上

2.1 問題設定

知識グラフは，トリプル (e_s, r, e_o) の集合データとして表現される．e_s と e_o はエンティティ，r はリレーションである．知識グラフ補完では，$(e_s, r, ?)$ または $(?, r, e_o)$ が与えられた際に?に当てはまるエンティティを予測する．

知識グラフ拡充時は既存のトリプル (e_s, r, e_o) の集合に対し，別のデータベースから新しく得られるトリプル (e'_s, r', e'_o) の集合を追加する．この際，別のデータベースのエンティティ e'_s, e'_o に対する知識グラフの同一エンティティ e_s, e_o を特定し，その表記を知識グラフ側の表記に揃える．本研究では，エンティティの表記ゆれについてのみ考え，リレーション名の表記ゆれは考慮しない．

2.2 編集距離に基づく判定とその問題点

単純な同一エンティティ判定手法として，編集距離の近さに基づく手法がある．この手法では，判定を行う二つのエンティティ e_1, e_2 に対し，次式で表すスコア $f(e_1, e_2)$ を計算する．

$$f(e_1, e_2) = 1 - \frac{d_{1,2}}{\max(\mathrm{len}(e_1), \mathrm{len}(e_2))} \quad (1)$$

ここで，$d_{1,2}$ は e_1, e_2 のローマ字表記での編集距離，$\mathrm{len}(e)$ はエンティティ e のローマ字表記での文字数を表す．このスコアは編集距離が小さいほど 1，大きいほど 0 に近づく．拡充に用いるデータベースの各エンティティに対して，知識グラフの各エンティティとのスコアを計算し，値が最も大きくなったエンティティを同一と判定する．スコアには閾値を設け，スコアの最大値が閾値未満の場合は同一エンティティが見つからなかったとする．

スコアの閾値を 1 付近まで大きく設定した場合は，同一である可能性の高いエンティティの組の集合を得ることができる．しかし，編集距離の大きい同一エンティティや，「たこ焼き」と「たい焼き」のように編集距離は小さいが同一でないエンティティは正しく判定できない．

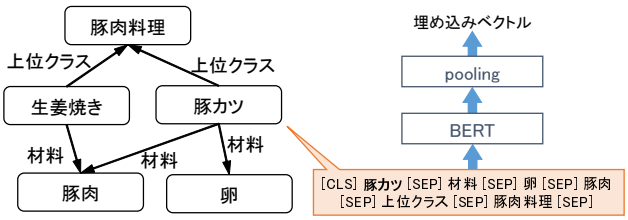

図 1: BERT に対する知識グラフの入力形式

3. 知識グラフ情報を活用した判定

提案法では，知識グラフ内の情報を用いて各エンティティをベクトル表現として埋め込み，それらを基に同一エンティティ判定を行う．大規模言語モデル BERT [3] を用いることで，編集距離では扱えない単語の意味的情報に加え，既存の知識グラフの情報を活用できる．

3.1 エンティティの埋め込みベクトルの活用

BERT にエンティティ名と，そのエンティティを含むトリプルの情報を入力する．入力文は，各エンティティ名とリレーション名を [SEP] トークンで区切った文とする．図 1 に例を示す．例えば，図中のエンティティ「豚カツ」を埋め込む場合，入力文は「[CLS] 豚カツ [SEP] 材料 [SEP] 卵 [SEP] 豚肉 [SEP] 上位クラス [SEP] 豚肉料理 [SEP]」となる．出力される系列ベクトルの平均値をそのエンティティの埋め込みベクトルとして得る．

埋め込みベクトルのコサイン類似度を同一エンティティ判定のスコアとして用いる．出力される埋め込みベクトルは埋め込み空間上の特定の領域に偏ることがある．そのため，事前に全エンティティの埋め込みベクトルの平均値を計算し，スコア計算時に平均値を減算することで偏りを取り除く．2.2 節の手法と同様に，スコアが最大の組を同一と判定し，最大値が閾値未満の場合は同一エンティティは見つからなかったとする．なお，提案法では計算コストを減らすため，推論時に事前に 2.2 節の手法を用い，対象とするエンティティの組のフィルタリン

Knowledge Graph Augmentation with Entity Identification for Improving Knowledge Graph Completion Performance: *Shuichi Chikatsuji, Hokuto Munakata, Ryu Takeda, and Kazunori Komatani (Osaka Univ.)*

グを行った．これにより，同一でない可能性が高い組を効率的に除外した．

3.2 同一エンティティ判定モデルの学習

より良い埋め込み表現を得るため，BERT は事前学習済みモデルを fine-tuning して用いる．学習データには，2.2 節の手法によって得られる，同一である可能性の高いエンティティの組の集合を正例に用いる．目的関数には Triplet Objective Function [4] を用いる．上述の各正例に対し，ランダムに負例をサンプリングして学習を行う．

4. 実験・評価

4.1 設定

4.1.1 使用データ

料理に関する知識グラフに対し，外部のレシピデータから作成した材料に関するトリプル $(e'_s, \text{材料}, e'_o)$ の集合を拡充した．ここでは e'_s は料理，e'_o は材料を表すエンティティである．

知識グラフには，Wikidata [5] から抽出した料理部分グラフを用いた．抽出前のグラフにおけるエンティティ「食品」に対し，リレーション「上位クラス」で 10 ホップ以内にたどり着けるエンティティが存在する．エンティティは 8423 種，リレーションは 110 種含まれる．

テストデータには Wikidata 料理部分グラフの一部を抜き出して用いた．提案法の評価のためには，拡充するトリプルに関連するテストデータを用いる必要がある．そのため，Wikidata 料理部分グラフのトリプルの中で，レシピデータから得たトリプルと完全一致するものをテストデータとした．知識グラフ拡充時には，テストデータに含まれるトリプルは追加しないこととした．含まれるトリプル数は，拡充先データは 14454 個，テストデータは 485 個である．

レシピデータとしては楽天公開データ[*1]中の楽天レシピを用い，含まれる料理名，材料名のデータから拡充用のトリプルを作成した．レシピは約 80 万件あり，約 230 万個の「材料」リレーションを持つトリプルが得られた．例えば，料理名が「豚カツ」のレシピの材料欄に「豚肉」があれば，(豚カツ, 材料, 豚肉) というトリプルを得た．

4.1.2 知識グラフ補完モデルと評価指標

知識グラフ補完モデルには，広く知られている TransE [6] と RotatE [7] を用いた．補完モデルによる埋め込み次元は 200 とした．テストデータの各トリプル (e_s, r, e_o) について，e_s か e_o のどちらかを無作為に予測した際の性能を確かめた．評価指標には Hits@10 と MRR を用いた．各トリプル (e_s, r, e_o) に対する正解エンティティ（e_s か e_o）の予測に関して，Hits@10 は予測順位の上位 10 位以内に正解エンティティが存在する割合を表す．MRR は正解エンティティの予測順位の逆数の平均を表す．

ベースラインは拡充を行わない場合および 2.2 節の編集距離に基づく手法を用いた場合とした．また，グラフ情報を入力文に用いることの有効性を検証するため，グラフ情報を入力せずにエンティティ名のみ入力した場合についても性能比較した．

4.1.3 拡充時のパラメータ設定

拡充時に用いるパラメータは手動で設定した．提案法では，編集距離のスコアの閾値を 0.9 として集めたエンティティの組を，BERT の学習データに用いた．事前フィルタリングは，編集距離のスコアの閾値を 0.5 として行った．コサイン類似度のスコアの閾値は 0.4 とした．BERT モデルには日本語 BERT[*2]を用いた．ベースラインの編集距離に基づく判定は，スコアの閾値を 0.9 とした．

表 1: 判定手法ごとの知識グラフ補完性能

手法	TransE		RotatE	
	Hits@10	MRR	Hits@10	MRR
拡充なし	0.062	0.025	0.085	0.035
編集距離	0.347	0.154	0.394	0.209
提案法 (グラフ情報なし)	0.447	0.237	0.482	0.279
提案法 (グラフ情報あり)	**0.557**	**0.330**	**0.612**	**0.390**

表 2: 判定手法ごとの拡充後トリプル数

手法	トリプル数
拡充なし	14454
編集距離	54927
提案法 (グラフ情報なし)	290238
提案法 (グラフ情報あり)	225780

4.2 結果・考察

同一エンティティ判定手法ごとに得られた知識グラフ補完性能を表 1 に示す．拡充を行った場合，行わなかった場合に比べて大きく補完性能が向上した．提案法ではグラフ情報を用いた場合，編集距離に基づく判定を用いた場合と比較し，いずれの指標においても高い性能を示し，Hits@10 ではそれぞれ 0.2 ポイント以上向上した．さらにグラフ情報を用いなかった場合に比べ，用いた方が高い補完性能を示した．

各手法による拡充後の知識グラフのトリプル数を表 2 に示す．編集距離に基づく判定に比べ，提案法を用いた場合はトリプル数が大幅に増加した．提案法では，グラフ情報を用いた場合，用いなかった場合に比べてトリプル数は減少した．これはグラフ情報の活用によって同一エンティティ判定の誤りを減らせたことを示唆している．

5. おわりに

本研究では，編集距離を基に得たデータで BERT を学習し，グラフ情報を活用して同一エンティティ判定を行った．拡充後の知識グラフ補完性能について，提案法では編集距離のみで判定した場合より高い性能を示した．また，提案法はグラフ情報を有効に活用できたことを示した．今後は，編集距離が大きい表記ゆれについても正しく判定することを目指す．

参考文献

[1] K. Komatani, et al. Knowledge graph completion-based question selection for acquiring domain knowledge through dialogues. In *Proc. IUI*, pp. 531–541, 2021.

[2] I. Saito, et al. Improving neural text normalization with data augmentation at character-and morphological levels. In *Proc. IJCNLP*, pp. 257–262, 2017.

[3] J. Devlin, et al. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proc. NAACL-HLT*, pp. 4171–4186, 2019.

[4] N. Reimers, et al. Sentence-BERT: Sentence embeddings using siamese BERT-networks. In *Proc. EMNLP-IJCNLP*, pp. 3982–3992, 2019.

[5] D. Vrandečić, et al. Wikidata: a free collaborative knowledgebase. *Commun. ACM*, Vol. 57, No. 10, pp. 78–85, 2014.

[6] A. Bordes, et al. Translating embeddings for modeling multi-relational data. In *Proc. NIPS*, pp. 2787—-2795.

[7] Z. Sun, et al. RotatE: Knowledge graph embedding by relational rotation in complex space. In *Proc. ICLR*, pp. 1–18.

[*1] https://rit.rakuten.com/data_release_ja/

[*2] https://huggingface.co/cl-tohoku/bert-base-japanese-whole-word-masking