辞書の階層構造埋め込み学習における日本語辞書定義文の効果的な利用

石井佑樹 佐々木稔
茨城大学工学部情報工学科
{18t4005h, minoru.sasaki.01}@vc.ibaraki.ac.jp

概要

語義曖昧性解消における，既存の知識グラフベースの学習手法は，単語間関係を用いてモデル学習が行われるが，1 単語における語義の階層関係を用いた学習は行われていない．また，日本語辞書の定義文を適用しても語義の階層関係の判定精度が悪く，知識グラフ埋め込み学習による効果を十分に得られないことが課題となっている．本研究では，日本語辞書における語義の階層関係を判定するモデルの判定精度を向上させるための辞書記述の編集方法を分析する．分析の結果，無編集の辞書を学習させた場合の精度は 60.9%であるが，編集を施した場合は精度が 83.3%となりモデルの性能向上が確認できた．このモデルから知識グラフ埋め込みシステムの性能向上が期待できる．

1 はじめに

語義曖昧性解消は，文中の多義語に対して，文中でどの語義が使われているのかを識別するタスクである．現在までに語義ラベル付けのコストを必要としない知識グラフベースによるアプローチが数多く研究されている．これに関連する研究として，GlossBERT[1]，BEM[2]といった語義定義文と教師あり学習を組み合わせた手法，EWISE[3]，EWISER[4]といった知識データの単語間関係を用いた手法が存在する．しかし，これらの手法において，単語における語義の階層関係を用いたモデル学習は行われていない．また，EWISE システムにおいて日本語辞書定義文を原文のまま用いて語義の階層関係を学習させると、生成されたモデルの関係判定精度が低く知識情報を十分に学習できていないことが判明した．岩波国語辞典では「《名詞に付けて》」「((形))」といった用法表現，品詞表現のみ記述された定義文が存在する．このような定義文は辞書内で複数存在し尚且つ内容が重複してしまうため，語義の階層関係の学習を阻害してしまうと考えられる．この課題に対して，既存の単語間関係に加えて語義の階層関係を学習することと，語義の階層関係判定モデルの精度を向上させる日本語辞書内容の編集・変更方法を分析することで，知識グラフ埋め込みシステムの性能向上に繋がることを目的とする．

本研究では，知識グラフの情報をどれ程埋め込むことができるのか調査するため，知識データにおける 3 つ組データの関係判定の精度を求めることのできる EWISE システムの知識グラフ埋め込み学習を利用する．段階的に編集・変更した日本語定義における関係判定精度の比較を行う．

2 語義の階層関係判定システム

日本語辞書における語義の階層関係を判定するのに EWISE システムの知識グラフ埋め込みシステムを用いる．語義の階層関係判定システムを図 1 に示す．

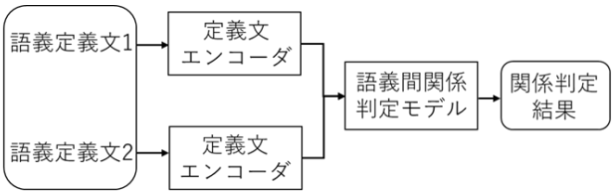

図 1　語義の階層関係判定システム

2.1 定義文エンコーダ

定義文エンコーダでは， BiLSTM Max エンコーダ[5]が採用されている．定義文を BiLSTM に入力し，出力を Max Pooling して得られた固定長の表現が定義文エンコーダの出力となる．

2.2 語義間関係判定モデル

語義間関係判定モデルは，2 つの日本語定義文の分散表現を入力として受け取り定義文間の階層関係を判定するモデルである．この語義間関係判定モデル ConvE[6]が用いられる．ConvE のスコアリング関数$\psi_l(e_h, e_t)$は式(1)で表される．通常知識グラフは，2 つの実体(h, t)と 1 つの関係(l)からなるN個の 3 つ組(h, l, t)の集合Kから構成される．hはヘッドエンティティ，tはテールエンティティである．式(1)は 3 つ組(h, l, t)に対するスコアリング関数である．

$$\psi_l(e_h, e_t) = f\big(\mathrm{vec}\big(f([\overline{e_h}; \overline{e_l}] * w)\big)W\big)e_t \quad (1)$$

式(1)において，e_h及びe_tは実体のパラメータ，e_lは関係のパラメータ，$\bar{x}$はxの 2 次元変形，wは 2 次元畳み込みフィルタ，$\mathrm{vec}(x)$はxのベクトル化，Wは線形変換，fは正規化線形ユニットである．対象のヘッドエンティティhに対して，グラフ内の各実体をテールエンティティとしてスコア$\psi_l(e_h, e_t)$を計算する．式(2)に示すように，スコアにシグモイド関数を適用することで 3 つ組(h, l, t)の適正性を示す推定値pを得る．

$$p = \sigma\big(\psi_l(e_h, e_t)\big) \quad (2)$$

2.3 モデルの学習

定義文エンコーダ及び語義間関係判定モデルの学習の流れを図 2 に示す．学習データは語義の階層関係を記述した知識データと語義定義文 ID に紐づけられた語義定義文を用いる．語義定義文は形態素に分解され，各形態素の分散表現が定義文エンコーダに入力される．形態素の分散表現は fastText[7]や GloVe[8]を用いて事前学習された表現を利用する．語義間関係は語義間関係判定モデル内の埋め込み層によって埋め込みベクトルに変換される．語義間関係判定モデルのパラメータは語義の階層関係の 3 つ組(h, l, t)のみを学習させた初期モデルのパラメータを初期値とする．また，定義文エンコーダを学習するため式(1)を式(3)に修正する．

$$\psi_l(e_h, e_t) = f\left(\mathrm{vec}\left(f\left(\left[\overline{q(h)}; \overline{e_l}\right] * w\right)\right)W\right)e_t \quad (3)$$

$q(.)$は定義文エンコーダであり，ヘッドエンティティはその実体の定義をエンコードしたものである．式(3)から出力された推定値pとラベルを元にモデルのパラメータを更新する．損失関数には式(4)に示すバイナリ交差エントロピーを用いる．

$$L_C = -\frac{1}{N}\sum_i (t_i \cdot log(p_i) + (1 - t_i) \cdot log(1 - p_i)) \quad (4)$$

t_iは 3 つ組(h, l, t)が適正の場合 1，それ以外の場合は 0 となる．p_iは式(3)で示したスコアの推定値である．

4.1 節に学習データ，形態素の分散表現，4.3 節にモデルの評価方法の詳細を示す．

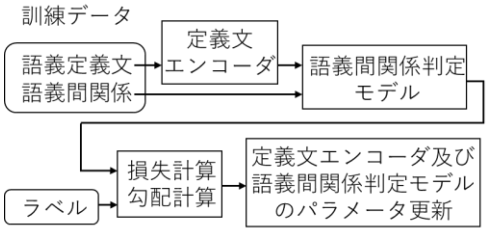

図 2 モデル学習の流れ

3 定義文の編集及び変更

3.1 定義文の編集

EWISE で利用される WordNet[9]の 3 つ組データに対応させるため岩波国語辞典の編集を行う．

3.1.1 定義文の ID 付け及び語義の階層関係

岩波国語辞典の定義文 ID を，複合語番号を除いて見出し語番号(1〜5 桁)・大分類番号(1 桁)・中分類番号(1 桁)・小分類番号(1〜2 桁)をこの順で結合した番号として ID 付けをする．

語義の階層関係は，岩波国語辞典の分類方法から上位概念，下位概念の 2 つの関係を学習させることにする．語義同士の関係は図 3 に示すような木構造の関係にある．木構造の親は子の上位概念となり，子は親の下位概念となる．3 つ組データを記録する際は，辺でつながれているノード同士の関係のみを記録する．

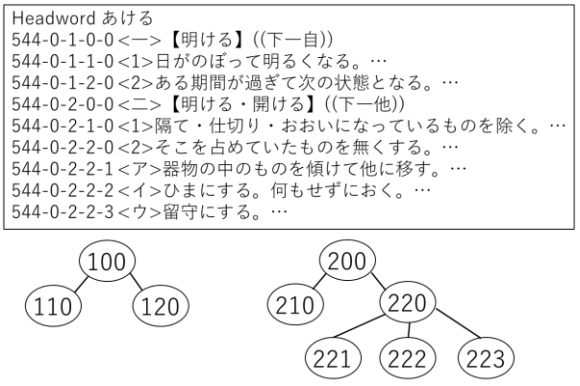

図 3 語義の階層関係の具体例
(ノード内数字は辞書 ID の右三桁と対応する)

3.2 定義文の変更

3 つ組の知識データの関係判定精度を高めるために，以下の設定 A・B・①～⑤を提案する．

3.2.1 語義説明の無い定義文の削除 A・B

定義文が 2 重括弧「(( ))」, 2 重山括弧「《 》」, 亀甲括弧「〔 〕」の表現のみで記述される単語定義があり，これらは単語語義を説明しないため定義文そのものを削除する．削除した場合，以降の節の変更と異なって 3 つ組の知識データの総数に影響を与える． 岩波国語辞典を編集しない A, 単語語義を説明しない定義文を削除した設定 B について実験を行い, 精度の高い設定を以降の節の変更に引き継ぐ. 具体例として，設定 B においては図 5 に示した赤字の部分を削除することになる．

3.2.2 辞書内表現・記号の削除・置換，全角英字を半角英字に置換，用例文内のハイフン「—」の置換(ひらがな置換)①・①*

2 重括弧「(( ))」表現，2 重山括弧「《 》」表現，山括弧「< >」，これら 3 種類の表現は見出し語の属性，用法，分類を示すものであり定義文から削除した．亀甲括弧「〔 〕」表現については，WordNet の定義文に同様の表現が存在するが，カテゴリ名が必要かどうかを確認するため削除した設定①*を設ける．定義文の分散表現を得る際に記号は余分な単語表現となってしまうため削除，置換した．削除した記号は「↓」「△」「×」の 3 種類で，置換した記号は「▽」を「。」，「【】」を「『』」の 2 種類である．また，辞書内の全角英字を半角にした．辞書内では用例文を記述しているが，用例文に含まれている見出し語の部分がハイフン「—」で置き換えられている．このハイフン「—」を見出し語(ひらがな)で置換した．

3.2.3 用例文内のハイフン「—」の置換(一対一置換)②

ひらがなの場合は同音意義の単語が存在するため，正確な分散表現を得るには漢字に変換して記述するのが好ましい．見出し語は同訓異字が多く，対応する漢字が複数存在するため，一対一で置換できる単語のみを漢字に置換した．

3.2.4 追記情報の削除③

定義文内では，定義文の追記情報は単語の直接の語義になりえない表現が多く存在する．以下に削除した内容を示す．

- 「▽」から続く文は，対象単語の語義を超えた範囲での説明がなされる．
- 派生｜から続く文は送り仮名の派生を示す．
- 読み仮名を削除するため，全角括弧「（ ）」で囲まれた内容がひらがなのみの場合削除する．
- 半角括弧「( )」で囲まれた内容は，辞書内の図番号及び注釈番号を示す．

3.2.5 全角括弧「（ ）」表現を削除④

④では，内容に構わず全角括弧「（ ）」で囲まれた内容を削除する．

3.2.6 部分的な用例文の削除⑤

岩波国語辞典における鍵括弧「「」」で囲まれた用例文は，見出し語が取り除かれた文章となっている場合が存在する．取り除かれている場合は，文を補完することが難しく不完全な用例文が残ってしまう．不完全な用例文を削除するため，定義文を句点で区切った一文が鍵括弧「「」」で囲まれている場合，その用例文を全て削除する．

4 実験

4.1 実験データ

・語義の階層関係を記述した知識データ

岩波国語辞典より，定義文の ID(h,t)及び語義の階層関係(l)をh,l,tの順で記録した 3 つ組の学習データを利用する．岩波国語辞典から階層関係を記録した 3 つ組は無編集の設定 A の場合 24040 組，単語語義説明の無い定義文を削除した設定 B の場合 10842 組得られた．それぞれ得られた 3 つ組学習データを 8 : 1 : 1 の割合で，訓練，開発，テストデータに分けている．

・語義定義文

知識データの定義文 ID に紐づけた語義の定義文を実験データとして利用している．定義文の分かち書きは MeCab と mecab-ipadic-NEologd を用いた．また，定義文における形態素の分散表現は 2021 年 10 月 10 日に更新された日本語 Wikipedia の全ページ記事のダンプデータをコーパスに fastText, GloVe を用いて事前学習された表現を利用する．

表 1 実験結果

設定/結果詳細	fastText				GloVe			
	単語ベクトル数	全語彙数	割合	精度	単語ベクトル数	全語彙数	割合	精度
A	60067	72619	0.8272	0.60853	60066	72618	0.8272	0.60957
B	60147	72752	0.8267	0.82286	60146	72751	0.8267	0.82203
B①*	66813	81039	0.8245	**0.83420**	66812	81038	0.8245	0.82640
B①②	68311	79958	0.8543	0.82890	68310	79957	0.8543	0.82590
B①②③	62275	71860	0.8666	0.83137	62274	71859	0.8666	0.82874
B①②③④	61449	70805	0.8679	0.83177	61449	70805	0.8679	0.82928
B①②③④⑤	43267	48612	0.8900	0.82251	43266	48611	0.8900	0.82098
B①*③④	60417	72181	0.8370	0.83341	60416	72180	0.8370	**0.83135**

全語彙(定義文に含まれる語彙の総数)，単語ベクトル数(分散表現が得られた単語数),割合(語彙ベクトル数/全語彙)

4.2 各学習のハイパーパラメータ

以下各学習のハイパーパラメータよりデフォルト値から変更したものを示す.

定義文エンコーダ及び語義間関係判定モデル学習時は，バッチサイズを 64 に変更，学習回数は 160 回に設定．初期モデルの学習回数は 300 回に設定，GloVe の事前学習は学習を考慮する周辺単語数を 10，分散表現の次元数を 300 に設定した.

4.3 評価方法

語義間関係判定モデルの評価方法には MRR（Mean Reciprocal Rank）を用いている．MRR はランキング評価指標の一つであり，語義間関係判定モデルから出力された関係推定値を対象としてモデルを評価する．今回の場合 MRR は式(5)で求められる.

$$MRR = \frac{1}{|U|}\sum_{u\in U}\frac{1}{k_u} \quad (5)$$

uは対象となる 3 つ組，Uは全 3 つ組，k_uは対象となる 3 つ組に対して関係が正しく判定された実体が出現した順位を表す.

5 実験結果

実験結果を表 1 に示す．単語語義を説明しない定義文の削除が効果的であり，辞書内表現の削除，追記情報の削除，等の変更を加えた場合にも僅かに判定精度が向上した.

単語語義を説明しない定義文を削除した場合に精度向上が大きく向上した．このような定義文は活用形を示す内容が大半で，一字一句内容が同じ定義文が複数存在することになる．定義文の重複を避けることで精度が向上したと考えられる.

定義文を変更することで，定義文全体に含まれる語彙数及び単語分散表現の初期値を得られる単語数に影響があるが，単語ベクトル数や語彙数の多さにかかわらず対象の語義の説明に不必要だと判断できる表現は取り除いても構わないと考えられる.

特に精度が下がった設定⑤では，補完不可な用例文の削除が難しく精度が下がってしまったと考えられ，完全な辞書の定義文を利用するには人為的な修正が必要になる.

また， fastText の単語分散表現で学習させることで，GloVe と比べ僅かに精度が向上したことから，サブワード情報を用いて学習された単語表現が大域的な共起情報を用いて学習した単語表現より判定精度を上げることが分かった.

6 おわりに

実験の結果，定義文の語義に対して説明の役割を持たない表現を削除・変更し学習することで，階層関係判定モデルの性能向上に有益であることが明らかになった．今後は，本研究で得られた階層関係判定モデルから知識グラフ埋め込みシステムの性能向上が期待できるため，知識グラフ埋め込みシステムを利用した語義曖昧性解消の実験を行うことで，この手法がどれほど有効か研究を行う予定である.

参考文献

1. Huang, L., Sun, C., Qiu, X. and Huang, X. “GlossBERT: BERT for Word Sense Disambiguation with Gloss Knowledge.” In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pp. 3509-3514, 2019.
2. Blevins, T. and Zettlemoyer, L. “Moving Down the Long Tail of Word Sense Disambiguation with Gloss Informed Bi-encoders”, In Proceedings of the 58th Association for Computational Linguistics (ACL2020), pp. 1006-1017, 2020.
3. Sawan, K. Sharmistha. J. Karan, S. and Partha T."Zero-shot word sense disambiguation using sense definition embeddings." In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pp. 5670-5681, 2019.
4. Bevilacqua M. and Navigli R. “Breaking Through the (80%) Glass Ceiling: Raising the State of the Art in Word Sense Disambiguation by Incorporating Knowledge Graph Information.” In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL2020), pp. 2854-2864, 2020.
5. Alexis, C. Douwe, K. Holger S. Loïc, B. and Antoine, B. “Supervised learning of universal sentence representations from natural language inference data.” In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pages 670–680, Copenhagen, Denmark. Association for Computational Linguistics, 2017.
6. Tim, D. Pasquale, M. Pontus, S. and Sebastian, R. “Convolutional 2d knowledge graph embeddings.” In Thirty-Second AAAI Conference on Artificial Intelligence, 2018.
7. Piotr, B. Edouard, G. Armand, J. and Tomas, M. “Enriching Word Vectors with Subword Information.” In Transactions of the Association for Computational Linguistics, Volume 5, pp. 135–146, 2016.
8. Jeffrey, P. Richard, S. and Christopher, M. “GloVe: Global Vectors for Word Representation.” In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 1532–1543, 2014.
9. George, M. “Wordnet: a lexical database for english.” Communications of the ACM, 38(11):39– 41, 1995.

付録

本文 3.2 節で示した設定の具体例を以下に示す．これらの語義定義文は岩波国語辞典第五版から引用した．

Headword あかず【飽かず】
288-0-0-0-0 ((連語))《副詞的に》
288-0-0-1-0 <1>あきもせずに。「一見入る」
288-0-0-2-0 <2>そうするのはあきたりないのに、やむなく。「一別れる」

図 A 設定 B の具体的な変更例(赤字を削除する)

Before
Headword こぼれる
18050-0-0-1-0 <1>【△零れる・×溢れる】((下一自))余って漏れ出る。
18050-0-0-1-1 <ア>液体や粒状のものなどが、あふれて落ちる。「涙が一」。…
18050-0-0-1-2 <イ>あり余って外に出る。あふれる。…
18050-0-0-2-0 <2>【×毀れる】((下一自))欠けたりくずれたりして、完全な姿を失う。「刃が一」▽(1)に対する他動詞は「こぼす」、…

↓

After
Headword こぼれる
18050-0-0-1-0 『零れる・溢れる』余って漏れ出る。
18050-0-0-1-1 液体や粒状のものなどが、あふれて落ちる。「涙がこぼれる」。…
18050-0-0-1-2 あり余って外に出る。あふれる。…
18050-0-0-2-0 『毀れる』欠けたりくずれたりして、完全な姿を失う。「刃がこぼれる」。(1)に対する他動詞は「こぼす」、…

図 B 設定①の具体的な変更例

Before
Headword しだい【次第】
…
21323-0-1-3-0 <3>経過。成行き。
21323-0-1-3-1 <ア>物事の事情。「こういう一だ」「事と一によっては」
21323-0-1-3-2 <イ>由来。「一書（がき）」
…

↓

After
Headword しだい【次第】
…
21323-0-1-3-0 <3>経過。成行き。
21323-0-1-3-1 <ア>物事の事情。「こういう次第だ」「事と次第によっては」
21323-0-1-3-2 <イ>由来。「次第書（がき）」
…

図 C 設定②の具体的な変更例

Headword けっこう【結構】
…
14890-0-0-2-2 <イ>気だてがよい。「一な御仁（ごじん）」▽古風な言い方。
14890-0-0-2-3 <ウ>これ以上は、いらない。「もう一（です）」▽ウは、主として言い切りの形で使う。派生|さ
…
Headword こて【鏝】
…
17794-0-0-2-0 <2>こて(1)に形の似た、熱して使う道具。
…

図 D 設定③の具体的な変更例(赤字を削除する)

Headword いたむ【痛む・傷む・悼む】
…
2216-0-1-2-1 <ア>（食品が）くさる。「りんごが一」
2216-0-1-2-2 <イ>（器物・建物などが）破損する。「ペン先が一」
…

図 E ④の具体的な変更例(赤字を削除する)

Headword あ【亜】
…
1-0-0-2-1 <ア>「亜細亜（アジア）」の略。「亜欧・東亜」
…
Headword いたい
…
2161-0-1-1-0 <1>…外力・病気で肉体や精神が苦しい。「くもない腹をさぐられる」（思ってもみない事で邪推される）「そんな事はくもかゆくもない（＝少しもこたえない。↓いたしかゆし）」
…

図 F ⑤の具体的な変更例(赤字を削除する)

単語分散表現の事前学習に利用した Wikipedia コーパスは以下から入手した．
https://ja.wikipedia.org/wiki/Wikipedia:%E3%83%87%E3%83%BC%E3%82%BF%E3%83%99%E3%83%BC%E3%82%B9%E3%83%80%E3%82%A6%E3%83%B3%E3%83%AD%E3%83%BC%E3%83%89