局所および大域的特徴量に基づく語義の曖昧性解消

浅川翔[1]，鈴木良弥[2]，李吉屹[2]，福本文代[2]
[1] 山梨大学大学院修士課程工学専攻
[2] 山梨大学大学院総合研究部工学域
{g20tk002, ysuzuki, jyli, fukumoto}@yamanashi.ac.jp

概要

語義の曖昧性解消 (WSD) は，文中に出現する単語の意味を同定するタスクであり, 機械翻訳や情報抽出などの精度向上に貢献できる. 語義の曖昧性解消を高精度で行うためには, 精度に貢献する特徴を学習する必要がある. この問題に対し, 近年, ニューラルネットワークモデルを用いた手法が数多く提案されている. しかし，これらの手法の多くは, 文単位を基本としており，文書単位での特徴抽出を利用した研究は少ない．本研究では，局所的特徴, すなわち文内の単語間の依存関係に基づく単語特徴と大域的な特徴である文書の特徴に注目し，局所及び大域的な特徴を相補的に利用することが WSD において有益であることを示す.

1 はじめに

語義の曖昧性解消 (WSD) は，文中に出現する単語の意味を同定するタスクであり, 機械翻訳や情報抽出などのタスクの精度向上に貢献できる. 従来の多くの WSD システムは，曖昧な単語ごとに独立した特徴を学習している [1]．単語埋め込みをそれらのシステムに統合した手法も提案されている [2]．最近になり，訓練により精度に貢献する特徴抽出を行うためにニューラルモデルが多く利用されている．単語埋め込みに関しては ELMo [3] や BERT [4] などの言語モデルに基づく埋め込み表現 が主流となっている．これは，Word2Vec や GloVe のような静的単語埋め込みとは異なり，文脈に応じた単語表現を獲得することが可能であり，WSD に対して非常に有効に働くためである [5][6]．最近の言語モデルに基づく研究では，文中の単語と文外の単語との間にも依存関係があることに着目し，より長距離の単語間の依存関係を捉える手法が提案されている [7]. しかし，これらの手法の多くは, 文単位でデータを処理しており，文書単位での特徴抽出に関する言及は少ない．本研究では，局所的特徴, すなわち文内の単語間の依存関係に基づく単語特徴と大域的な特徴である文書の特徴に注目し，局所及び大域的な特徴が WSD において有益であることを示す.

2 関連研究

2.1 Gloss 文

Gloss 文 (語釈文) は, テキスト内の単語, またはフレーズの意味を簡潔に表したものであり，語義の曖昧性解消に重要な要素であることが示されている．Luo らは，ニューラル WSD システムの追加入力として Gloss 文を使用し，精度を大幅に向上させた [8]．Kumar らは，WordNet 上に定義されたグラフ構造を用いて Gloss 文を埋め込み，概念ベクトルとして使用した．しかし，GlossEncoder は事前学習されたパラメータで初期化を行い，訓練時はパラメータを固定するため，知識グラフを学習させるためのコーパスが別途必要になる [9]．Blevins らは，ContextEncoder と GlossEncoder の両方を同時に訓練させることにより，GlossEncoder 用の追加コーパスを必要とせずに Gloss の埋め込み表現 (概念ベクトル) を学習する手法を提案した [10]．実験では, BERT と組み合わせることで精度が大幅に向上することを示した．本手法は，対象文と Gloss 文を Transformer ネットワークでエンコードし，各ベクトルの距離に基づきスコアを決定する点で Blevins らの手法と類似している．一方，局所的な特徴量を高精度で抽出するために、エンコーダの最適化を行った点，及び処理単位を文から文書単位に拡張し，大域的特徴量抽出を可能にした点で異なる．

2.2 文法と位置情報

シソーラス辞書である WordNet では，例えば動詞「evoke」などのように概念の説明文である Gloss の中に目的語の属性を示すものがあるなど，概念と文

表 1 長距離依存関係の例: 表中, 黄色は生物関連の単語を示す．mouse の概念には，「動物のマウス」，「コンピュータマウス」，「目をぶつけたときにできるあざ」，「静かな人」があり，対象単語「mouse」が属する文書内に生物関連のキーワードがある場合，対象単語の語義は「動物のマウス」である可能性が高くなる．この例の場合，対象単語「mouse」と生物関連の単語「animal」は文外に出現している．

Sentence: We classify mice as “ straight haired ” or “ wavy haired ” , but a hairless mouse appears .
Sense: mouse%1:06:00::
Gloss: any of numerous small rodents typically re-sembling diminutive rats ... hairless tails
Text: Some experiments ... We classify mice as “ straight haired ” or “ wavy haired ” , but a hairless mouse appears . We can escape from such a diffi-culty by ruling out the animal as not constituting a trial , but ... experiment has the value * * f .

法が深く関わる例が多く見受けられる (例. evoke#1’s Gloss: call forth **(emotions, feelings, and responses)**). すなわち, 文法情報は WSD の精度貢献に大きく関わる．最近の研究では，BERT[4] を用いることにより文法やフレーズを考慮した埋め込み表現を獲得できることから, これを用いて入力文や Gloss のエンコードを行う手法が主流となっている [11] [10]. Yang らは，BERT の絶対位置情報や MASK トークンを用いた事前学習タスクの改善を図り，相対位置情報の使用や Permutation Language Modeling という新たな事前学習タスクを用い，言語モデルの精度を向上させた [12]．Song らは，絶対位置情報と相対位置情報を相補的に扱うことでトークン間の依存関係をより正確に表現することにより，これが言語モデル (MPNet) の精度に貢献することを示した [13]. MPNet は文法情報が重要な要素となるタスク CoLA において，以前の言語モデルの精度を大幅に上回っていることから，文法やフレーズを考慮した埋め込み表現の獲得に優れていることが分かり，WSD にとっても有益であると考えられる．本手法では，MPNet をニューラル WSD システムに統合し，局所的特徴量をより正確に抽出することで，さらなる精度向上を図った．

2.3 文書中の長距離依存関係

従来の WSD は語義の解消を文単位で行う手法が多い．本手法は，ニューラル WSD システムを文単位の処理から文書単位の処理へ拡張することにより精度向上を目指す．表 1 に，文外の依存関係の例を示す．文書単位の処理へ拡張することにより，語義を同定する上で重要な特徴を，周辺文から抽出することができるようになる．

2.4 訓練コーパスの追加

最近の研究では，訓練コーパスとして SemCor に加えて WordNet Gloss Corpus(WNGC) と WordNet Example Sentence Corpus(WNEX) を用いることにより，精度が向上することが報告されている [6][11]. 本手法においても WNGC と WNEX を訓練コーパスとして追加し，さらなる精度向上を図った．

3 提案手法

本手法を図 1 に示す．本手法は, 対象単語を文中の周辺単語に応じて表現する役割を担う SentenceEncoder と，文書全体をベクトルとして表現する役割を担う TextEncoder，及び語義の定義文である Gloss 文をベクトルとして表現する役割を担う GlossEncoder で構成されている．各エンコーダの深層ネットワーク部分は，事前学習により獲得した一般的な単語の埋め込む表現を用いるために，MPNet により初期化する．

3.1 WSD タスクおよび本手法の枠組み

本手法は，対象単語, 対象単語が属する文と文書，及び対象単語の候補語義に対応付けられた Gloss 文の埋め込み表現を用いる．入力文 $c = c_0, c_1, ..., w, ..., c_m$($w$ は曖昧性を解消する対象単語) と，入力文 c が属する文書 d，及び対象単語 w の候補語義 S_w の Gloss 文 $g = g_0, g_1, ..., g_l$ が与えられ，語義 $s = s^0_{c_0}, s^1_{c_0}, ..., s^i_w, ..., s^j_{c_m}$ を出力する (i^{th} は対象単語の i 番目の候補語義を示す). **SentenceEncoder** T_c は，式 (1) のように入力文 c 中の対象単語 w をサブトークン $w_l = w_j, ..., w_k$ 毎にエンコードし，その平均を w の表現 r_w として出力する．

$$r_w = \frac{1}{k-j}\sum_{l=j}^{k}(T_c(c)[l]) \tag{1}$$

TextEncoder T_d は，式 (2) で示されるように対象単語 w が属する文書 d をエンコードし，先頭のトークン (特殊トークン [CLS] に対応) を d の表現 r_d として出力する．

$$r_d = T_d(d)[0] \tag{2}$$

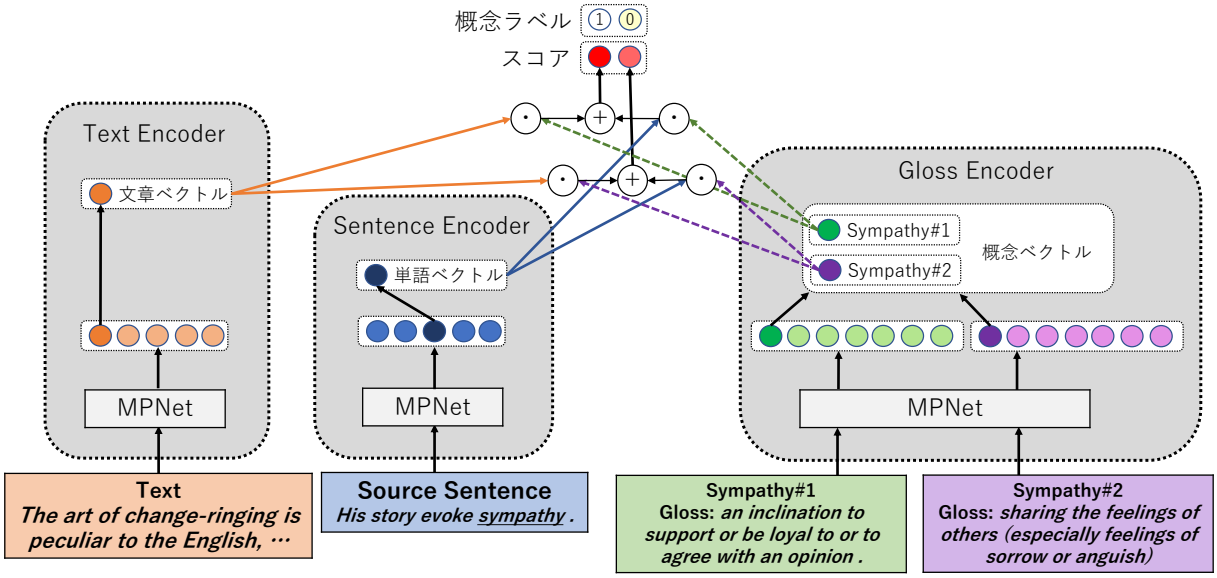

図 1 textBEM の枠組み: ContextEncoder に対象単語が属する文を，TextEncoder に対象単語が属する文書を，GlossEncoder に候補語義の Gloss を入力する．各エンコーダの深層ネットワーク部分は MPNet を用い初期化する．

GlossEncoder T_g は，対象単語の候補語義 $s \in S_w$ に対応付けられた Gloss $g = g_0, g_1, ..., g_l$ をエンコードし，先頭のトークン (特殊トークン [CLS] に対応) を s の表現 r_s として出力する．$i = 0, ..., |S_w|$ とした時，式 (3) で示すように対象単語の語義 s は，対象単語 w と候補語義 s_i のエンコードされた表現 r_w と r_{s_i} の内積及び，対象単語が属する文書 d と候補語義 s_i のエンコードされた表現 r_d と r_{s_i} の内積の平均で計算される．

$$\Phi\ (w, d, s_i) = \frac{1}{2}(r_w \cdot r_{s_i} + r_d \cdot r_{s_i}) \tag{3}$$

本手法は，モデルの訓練のために cross-entropy loss を使用した．(対象単語, 対象単語の属する文書, 語義) の組 (w, d, s_i) が与えられた時の損失関数を式 (4) に示す．

$$L(w, d, s_i) = -\Phi\ (w, d, s_i) + log \sum_{j=0}^{|S_w|} exp(\ \Phi\ (w, d, s_j)) \tag{4}$$

4 実験

4.1 データセットおよび評価方法

実験では，Raganato ら [14] の評価手法を用い本手法の検証を行った．訓練コーパスとして，SemCor，WNGC，及び WNEX を用いた．SemCor および WNGC は WordNet 上の語義に対し，人手により注釈付けされたコーパスである．WNEX は，WordNet 上の各語義の例文に自動で注釈付されたコーパスである．検証セットは，SemEval-2015 (S15) [15] を用いた．テストセットは，Senseval2 (S2) [16], Senseval-3 (S3) [17], SemEval-2007(S7) [18], 及び SemEval-2013 (S13) [19] を用いた．

4.2 実験結果

4.2.1 ベースラインおよび関連研究との比較

表 2 に提案手法である **textBEM** および最近の state-of-the-art システムの結果を示す．**EWISE** は，WordNet で定義された語義間の関係性に基づいた構造化ネットワークを，WSD モデルに統合したモデルである [9]．EWISER は，概念ベクトルを事前学習により得たモデルであり，データセットの拡張やグラフ構造を統合している [11]．**BEM** は，ContextEncoder と GlossEncoder を同時に訓練するバイエンコーダモデルである [10]．本手法は，BEM がより広範囲の情報を捉えることを可能にしたモデルであるため，BEM を本研究のベースラインに設定した．**BEM-MPNet** は，BEM における各エンコーダを初期化するための事前学習済みパラメータを，BERT-base から MPNet-base に置き換えたモデルである．表 3 に，提案手法に入力する文書別の結果を示す．**前一文**は，対象単語が属する文の直前の 1 文を用いた結果を示す．**前全文**は，対象単語が属する文以前の全ての文を用いた結果を示す．**全文**は，対象単語が属する文書全体を用いた結果を示す．

表 2 English all-words WSD タスクにおける F1-score(%): **ALL** は，全テスト用データセットを結合して検証した場合の結果．**S**，**G**，**E** はそれぞれ SemCor，WNGC，及び WNEX を示す．**T** は, 入力の単位が文書であることを示す. "*"は，スコアが最良のものと比較して統計的に有意であることを表す（有意水準 5%での t 検定に基づく）.

System	S	G	E	T	ALL	S15	S2	S3	S7	S13	N	V	A	R
EWISE [9]	✓				71.8	74.5	73.8	71.1	69.4	74.5	74.0	60.2	78.0	82.1
EWISER [11]	✓	✓	✓		80.1	81.8	80.8	79.0	**75.2**	80.7	82.9	69.4	83.6	87.3
BEM [10]	✓				79.0	81.7	79.4	77.4	74.5	79.7	81.4	68.5	83.0	87.9
BEM-MPNet	✓	✓	✓		81.2	84.4	81.9	79.1	74.7	**82.1**	83.8	70.7	**84.7**	**89.9**
textBEM	✓	✓	✓	✓	**81.5***	**85.5***	**82.2***	**79.6***	74.1	**82.1**	**83.9***	**71.6***	84.5	89.3

表 3 入力文書別の平均 F1-score: "*"は，スコアが最良のものと比較して統計的に有意であることを表す（有意水準 5%での t 検定に基づく）.

入力文書	ALL	S15	N	V	A	R
入力なし	81.0	84.3	83.6	70.7	84.1	**89.1**
前一文	80.8	84.9	83.4	70.7	83.6	88.9
前全文	80.9	85.0	83.5	70.8	84.3	88.9
全文	**81.3***	**85.1**	**83.8***	**71.4***	**84.5**	89.0

表 4 エンコーダ初期化用の言語モデル別の平均 F1-score: 訓練コーパス SemCor のみを用いて訓練した．本手法は，バイエンコーダ型を採用し，各エンコーダは BERT-base 及び MPNet-base で初期化した．"*"は，スコアが最良のものと比較して統計的に有意であることを表す（有意水準 5%での t 検定に基づく）.

System	ALL	S15	N	V	A	R
BERT	78.4	81.5	81.2	67.7	81.4	**87.5**
MPNet	**79.7***	**82.2***	**82.7***	**68.8***	**82.2***	87.3

4.2.2 語義の解消結果

表 2 より，EWISER は，EWISE と比較し ALL スコアが 8.3%高く, 大幅な精度の向上が認められる. BEM は，概念ベクトルを同時に学習するバイエンコーダモデルにしたことで，概念ベクトルを事前学習により取得する EWISE の精度を上回る．また，EWISER と同様のデータセットを使用して学習することにより，EWISER と同等の精度が得られている．さらに，BEM のエンコーダ初期化用の事前学習済みパラメータを BERT から MPNet に変更することにより，さらに優れた精度が得られている．このことから，MPNet により局所的特徴量をより正確に抽出することは，ニューラル WSD モデルにおいて有益であると考えられる．本手法である textBEM は，ALL 評価セットで最高精度が得られている．特に名詞と動詞に精度の貢献が確認できる．このことから，より広範囲の文書情報を用いることがニューラル WSD モデルにおいて有益であると言える．本研究では，MPNet-base を用いている. 一方で MPNet-large を用いることでさらなる精度向上も期待できる．しかし，本研究では文法情報をより正確に捉えられる MPNet と WSD との親和性, さらにニューラル WSD モデルの局所的特徴量を高精度で抽出することが目的であるため，メモリ効率の高い MPNet-base を用いた.

表 3 は, 入力文書の違いによる精度を示す. 表 3 より, 対象文が属する文書全て (全文) を用いた場合, ALL で 81.3%であり最高精度であった． 特に名詞と動詞の精度が高いことがわかる. 一方で「入力なし」と「前一文」, 及び「入力なし」と「前全文」の ALL スコアには有意差がなかった．このことから, 入力文書の範囲は textBEM の精度向上に関係すること, さらにより広範囲にした方がモデルの精度に貢献すると言える. 表 4 にエンコーダ初期化用の言語モデル別の結果を示す. MPNet-base を用いた場合が ALL 評価セットで BERT よりも優れており, 特に．名詞，動詞及び形容詞で顕著であった．このことから，絶対位置情報と相対位置情報を相補的に扱う MPNet は，絶対位置情報のみを扱う BERT と比較し，文法情報が重要な WSD タスクとの親和性が高いと言える.

5 結論

本研究では，局所的な特徴量を高精度で抽出するためのエンコーダの最適化に加え，バイエンコーダモデルに TextEncoder を統合し，大域的な特徴量を学習する手法を提案した．実験の結果，ベースモデルの ALL スコアと比較し，2.5%の精度向上が見られ，エンコーダの最適化および大域的特徴量の抽出手法の有効性が示せた.

謝辞

本研究の一部は, 科研費 17K00299, 及び JKA 補助事業の助成を受けたものである.

参考文献

[1] Razvan Bunescu Hui Shen and Rada Mihalcea. Coarse to fine grained sense disambiguation in wikipedia. **Lexical and Computational Semantics**, Vol. 1, pp. 22–31, 2013.

[2] Sascha Rothe and Hinrich Schutze. Autoextend: Extending word embeddings to embeddings for synsets and lexemes. **Association for Computational Linguistics**, Vol. 53, pp. 1793–1803, 2015.

[3] Mohit Iyyer Matt Gardner Christopher Clark Kenton Lee Luke Zettlemoyer Matthew E. Peters, Mark Neumann. Deep contextualized word representations. **Association for Computational Linguistics**, Vol. 1, pp. 2227–2237, 2018.

[4] Kenton Lee Kristina Toutanov Jacob Devlin, Ming-Wei Chang. Bert: Pre-training of deep bidirectional transformers for language understanding. **Association for Computational Linguistics**, Vol. 1, pp. 4171–4186, 2019.

[5] Daniel Loureiro and Aliıpio Jorge. Language modelling makes sense: Propagating representations through wordnet for full-coverage word sense disambiguation. **Association for Computational Linguistics**, Vol. 57, pp. 5682–5691, 2019.

[6] Benjamin Lecouteux Loıc Vial and Didier Schwab. Sense vocabulary compression through the semantic knowledge of wordnet for neural word sense disambiguation. **Global Wordnet Association**, Vol. 10, pp. 108–117, 2019.

[7] Iz Beltagy, Matthew E. Peters, and Arman Cohan. Longformer: The long-document transformer. **arXiv:2004.05150**, 2020.

[8] Zexue He Qiaolin Xia Zhifang Sui Baobao Chang Fuli Luo, Tianyu Liu. Leveraging gloss knowledge in neural word sense disambiguation by hierarchical co-attention. **Association for Computational Linguistics**, pp. 1402–1411, 2018.

[9] Karan Saxena Partha Talukdar Sawan Kumar, Sharmistha Jat. Zero-shot word sense disambiguation using sense definition embeddings. **Association for Computational Linguistics**, Vol. 57, p. 5670—5681, 2019.

[10] Terra Blevins and Luke Zettlemoyer. Moving down the long tail of word sense disambiguation with gloss informed bi-encoders. **Association for Computational Linguistics**, Vol. 58, pp. 1006–1017, 2020.

[11] Michele Bevilacqua and Roberto Navigli. Breaking through the 80%glass ceiling: Raising the state of the art in word sense disambiguation by incorporating knowledge graph information. **Association for Computational Linguistics**, Vol. 58, pp. 2854–2864, 2020.

[12] Yiming Yang Jaime Carbonell Ruslan Salakhutdinov Quoc V. Le Zhilin Yang, Zihang Dai. Xlnet: Generalized autoregressive pretraining for language understanding. **Neural Information Processing Systems**, pp. 5754–5764.

[13] Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. Mpnet: Masked and permuted pre-training for language understanding. **Neural Information Processing Systems**, Vol. 34, , 2020.

[14] Roberto Navigli Alessandro Raganato, Jose Camacho-Collados. Word sense disambiguation: A unified evaluation framework and empirical comparison. **Association for Computational Linguistics**, Vol. 15, pp. 99–110, 2017.

[15] Andrea Moro and Roberto Navigli. Semeval-2015 task 13: Multilingual all-words sense disambiguation and entity linking. **Semantic Evaluation**, Vol. 9, pp. 288–297, 2015.

[16] Scott Cotton Lauren Delfs Hoa Trang Dang Martha Palmer, Christiane Fellbaum. English tasks: All-words and verb lexical sample. **Semantic Evaluation**, Vol. 2, pp. 21–24, 2001.

[17] Benjamin Snyder and Martha Palmer. The english all-words task. **Semantic Evaluation**, pp. 41–43, 2004.

[18] Dmitriy Dligach Martha Palmer Sameer Pradhan, Edward Loper. Semeval-2007 task-17: English lexical sample, srl and all words. **Semantic Evaluations**, Vol. 4, pp. 87–92, 2007.

[19] Daniele Vannella Roberto Navigli, David Jurgens. Semeval-2013 task 12: Multilingual word sense disambiguation. **Semantic Evaluation**, Vol. 2, pp. 222–231, 2013.