Vol.2017-NL-232 No.4
2017/7/19

順方向多層 LSTM と分散表現を用いた教師あり学習による語義曖昧性解消

新納 浩幸[1,a)]　古宮 嘉那子[1,b)]　佐々木 稔[1,c)]

概要：
教師あり学習による語義曖昧性解消では，対象単語の周辺文脈をどのようにベクトル化するかが重要な問題である．近年，単語の周辺文脈を双方向の LSTM を用いてベクトル化することが提案され，語義曖昧性解消を含む様々なタスクにおいて有効であることが示された．ただし語義曖昧性解消に限れば，対象単語の語義の選択が，かなり離れた後方位置の単語により影響を受けるとは考えづらい．そこで本論文は逆方向の LSTM は用いずに，後方の文脈は直後数語の分散表現だけを利用する形でベクトル化することを提案する．実験では SemEval-2 の日本語辞書タスクを利用して提案手法の有効性を示す．また本手法において利用する分散表現や LSTM モデルの品質が，どの程度語義曖昧性解消の精度に影響するかを考察する．

Supervised Word Sense Disambiguation using Forward Multi-layered LSTM and Word Embeddings

Shinnou Hiroyuki[1,a)]　Komiya Kanako[1,b)]　Sasaki Minoru[1,c)]

1. はじめに

本論文では順方向の多層 LSTM と分散表現を用いて対象単語の周辺文脈をベクトル化することで語義曖昧性解消を行う手法を提案する．また利用する LSTM のモデルや分散表現の品質が語義曖昧性解消の精度に及ぼす影響を考察する．

語義曖昧性解消は意味解析の最もプリミティブな処理であり，その重要性は明らかである．そのため従来より様々な手法が試みられ，近年は深層学習の技術を利用した研究が活発である．深層学習を利用した語義曖昧性解消は二つのタイプに大別できる．一つは分散表現を利用したものであり，もう一つは LSTM を利用したものである．

分散表現は単語の意味を低次元の密ベクトルで表現したものであり，従来の bag of words のモデルによるベクトル化と比較して，より適切に意味を表現していると考えられる．実際の自然言語処理ステムでは何らかの形で単語をベクトル化する必要があり，近年，単語のベクトル化には，分散表現を用いることが一般的になっている．

分散表現を利用した語義曖昧性解消の研究として，まず，語義の分散表現を求めるものがある．通常，分散表現は単語に対するものであるが，語義の分散表現が構築できれば，対象単語の周辺文脈と語義の分散表現との類似度を求めることで語義曖昧性解消が行える．Neelakantan は単語の語義ごとにベクトルを与えるモデルとして Skip-gram モデルを拡張した MSSG (Multi Sense Skip-gram) モデルを提案し，コーパスから自動で語義の分散表現を構築している [6]．MSSG モデルによる語義の分散表現の構築では，語義数をどのように決定するかが重要になる．Chen は WordNet の辞書データを用いて WordNet の語義ごとの分散表現を学習する手法を提案している [1]．Li は MSSG モデルに CRP (Chinese Restaurant Process) を適応させた NP-MSSG モデルを提案している [4]．これらは基本的に教師なし学習による語義の分散表現を構築しているが，山

[1] 茨城大学 工学部 情報工学科
〒 316-8511 茨城県日立市中成沢町 4-12-1
[a)] hiroyuki.shinnou.0828@vc.ibaraki.ac.jp
[b)] kanako.komiya.nlp@vc.ibaraki.ac.jp
[c)] minoru.sasaki.01@vc.ibaraki.ac.jp

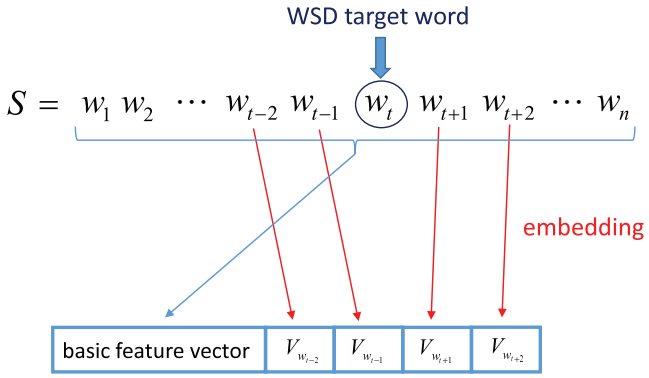

図 1 Sugawara の手法

木は語義のラベル付きデータから語義の分散表現を構築することを試みている [12].

また分散表現を教師あり学習の素性として利用する研究もある [9]. Sugawara は対象単語の前後 2 単語，計 4 単語の分散表現と従来の素性ベクトルに結合して，教師あり学習による語義曖昧性解消を行っている [8]. Sugawara の手法は分散表現を従来のシソーラスの代用としていると考えられる．その場合，分散表現の位置が固定される問題点がある．Yamaki は独自の手法でこの問題に対処している [10].

LSTM (Long Short-Term Memory) は RNN (Recurrent Neural Network) を拡張したものであり，RNN 同様，時系列データを扱うモデルである．簡単に述べれば，LSTM は注目している単語 w_t の直前までの単語列 $w_1 w_2 \cdots w_{t-1}$ の情報を表現したベクトル v_{t-1} を計算するものである．当初 LSTM は言語モデルの構築に利用されていたが，v_{t-1} が単語列 $w_1 w_2 \cdots w_{t-1}$ の意味的情報も含んでいるために応用範囲が広がっている．語義曖昧性解消としては w_t を対象単語とすれば，v_{t-1} が w_t の前文脈を表す．また逆方向の LSTM が構築する単語列 $w_{|w|} w_{|w|-1} \cdots w_{t+1}$ に対するベクトル v'_{t+1} が w_t の後文脈を表す．つまり双方向の LSTM を学習することで対象単語 w_t の周辺文脈を $[v_{t-1}; v'_{t+1}]$ によりベクトル化することで語義曖昧性解消が行える．

Kågebäck は双方向の LSTM の学習時に word dropout [2] と類似の手法を導入して語義曖昧性解消の精度を上げている [3]. Melamud は双方向の LSTM から得られた w_n の周辺文脈のベクトルを語義曖昧性解消を含む様々なタスクに利用している [5].

語義曖昧性解消に双方向の LSTM を用いることは効果的である．ただし語義の選択はほぼ決定的に行えるものであり，後文脈の情報が語義の選択に与えている影響は小さい．特に対象単語からかなり離れた後方位置にある単語が対象単語の語義の選択に影響を与えているとは考えづらい．そこで本論文では逆方向の LSTM は用いずに，後文脈は Sugawara の手法と同じく直後 2 単語の分散表現の連結によりベクトルする．また LSTM を多層化することでより品質の高い LSTM を用いる．

実験では SemEval-2 に日本語辞書タスクのデータ [7] を用い，提案手法の有効性を示す．また本手法は利用する分散表現の質と LSTM の質が精度に大きな影響を与えている．この点を考察する．

2. 分散表現を用いた語義曖昧性解消

本章では教師あり学習による語義曖昧性解消に分散表現を利用した研究として Sugawara の手法 [8] を述べる．

語義曖昧性解消の対象単語を w とするとき，訓練データは w を含む文 s であり，そのラベルは w の語義である．従来の教師あり学習による語義曖昧性解消では，s 内の w の周辺単語から素性リストを作成し，そこから SVM 等の教師あり学習手法により分類器を作成する．ここで素性リストは前後の単語やその品詞やその概念 id などである．係り受けなどの情報も利用することもある．

ここでは学習アルゴリズムとして SVM を用いるとする．この場合，全素性リストの各要素を一つの単語と捉えれば，各素性リストは bag of words のモデルでベクトル化できる．ここではこのベクトルを基本素性ベクトルと呼ぶことにする．

Sugawara の手法は s 中の語義曖昧性解消の対象単語 w の前後 2 単語に対する分散表現を s の基本素性ベクトルに連結し，連結されたベクトルを新たな素性ベクトルと考えて，分類器を作成するものである（図 1 参照）.

また Sugawara は対象単語の前後 2 単語の分散表現に組み合わせる基本素性ベクトルとして，いくつかのタイプを試した結果，単純な bag of words で基本素性ベクトルを作成することを提案しているが，ここでは基本素性ベクトルとしてシソーラスの情報も含めた標準的な素性を利用する．

3. LSTM を用いた語義曖昧性解消

言語モデルを学習する LSTM の時刻 t 時の入出力を表したネットワークを図 2 に示す．時刻 t で単語 w_t が入力

され，それを w_t の分散表現のベクトルに変換し，その分散表現のベクトルを LSTM ブロックに入力する．LSTM ブロックでは次の時刻 $t+1$ への LSTM ブロックへ w_0 から w_t の単語列の情報を圧縮したベクトル h_t と記憶セル c_t を渡す．同時に y_t を出力し，それを線形作用素 W で one-hot 形式のベクトルに直すことで次に現れる単語を予測する．学習時には Wy_t と w_{t+1} との誤差からネットワークの重みを学習する．

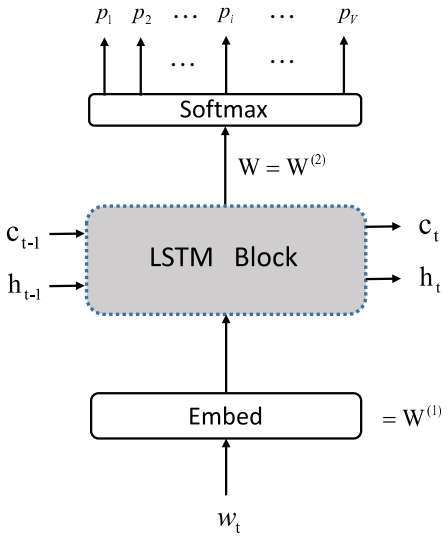

図 2 LSTM の時刻 t の入出力

上記は通常の順方向の LSTM であるが，逆方向の LSTM は単に入力単語列を逆順に並べた LSTM である．通常，学習は順方向ものと逆方向のものが同時に行われるが，順方向の LSTM と逆方向の LSTM は独立なので，別個に学習させても本質的な違いはない．

対象単語が w_t である語義曖昧性解消に LSTM を用いる場合，順方向の LSTM が出力する h_t(以下 h_t^f) と逆方向の LSTM が出力する h_t(以下 h_t^b) とを連結したベクトル $[h_t^f : h_t^b]$ を素性ベクトルとし，SVM などの学習アルゴリズムから分類器を作成する．

4. 提案手法

基本的に Sugawara の手法と双方向 LSTM による手法を組み合わせる．Sugawara の手法における基本素性ベクトルを LSTM から得られる素性ベクトルに設定する．ただしここでは逆方向の LSTM が出力する h_t^b を利用しない．これは語義曖昧性解消では対象単語 w_t からかなり離れた後方位置にある単語が w_t の語義の決定に影響を与えているとは考えづらいからである．ここでは w_t の語義の決定に影響を及ぼす w_t の後方文脈は直後数語と考える．また LSTM は通常，多層にした方が品質が向上です．このためここでは 2 層の LSTM を用いることにする．語義曖昧性解消では 2 層目の LSTM ブロックが出力する h_t を利用する．

以上より本論文では語義曖昧性解消の対象単語が w_t であるとき，その素性ベクトルを以下で設定する．ここで v_w は単語 w の分散表現を表す．

$$[h_t : v_{w_{t-2}} : v_{w_{t-2}} : v_{w_{t+1}} : v_{w_{t+2}}]$$

5. 実験

5.1 分散表現の準備と LSTM モデルの学習

利用する分散表現は nwjc2vec である．これは国立国語研究所が構築した超大規模日本語コーパスから word2vec[*1] を用いて構築された分散表現である [13]．また分散表現の次元は 200 次元となっている．

LSTM の学習用コーパスとしては毎日新聞'07 年度版からランダムに取り出した 50 万文（約 190MB のテキスト）とした．LSTM は 2 層のものを用い，分散表現は nwjc2vec に固定し LSTM の学習対象から外した．また dropout の比率は 0.5 とした．同じ設定で逆方向の LSTM も学習する．語義曖昧性解消の実験には 10 epoch 後のモデルを利用する．

5.2 SemEval-2 データによる評価

語義曖昧性解消のデータセットとしては SemEval-2 の日本語辞書タスク [7] のデータセットを用いる．このタスクでは 50 単語の対象単語が設定され，各対象単語に対して，50 用例の訓練データと 50 用例のテストデータが与えられている．手法の評価は対象単語に対するテストデータの正解率の平均（つまりマクロ平均）により行う．

比較手法として以下を試す．

baseline SemEval-2 のベースラインの手法であり，前後の単語や品詞，対象単語 w_t の周辺の自立語の分類語彙表 ID といった標準的な素性ベクトルを用いたもの．ここでは，この素性ベクトルを B_{w_t} と名付ける．

Sugwara 対象単語が w_t のとき基本素性ベクトルを B_{w_t} とした Sugawara の手法．素性ベクトルとしては $[B_{w_t} : v_{w_{t-2}} : v_{w_{t-1}} : v_{w_{t+1}} : v_{w_{t+2}}]$ となる．

Bi-LSTM 対象単語が w_t のとき順方向 LSTM から得られる h_t^f と逆方向 LSTM から得られる h_t^b を連結したベクトル $[h_t^f : h_t^b]$ を素性ベクトルとして用いるもの．

Bi-LSTM+ 単純に Sugawara の手法と Bi-LSTM の手法を組み合わせたもの．素性ベクトルとしては $[h_t^f : h_t^b : v_{w_{t-2}} : v_{w_{t-1}} : v_{w_{t+1}} : v_{w_{t+2}}]$ となる．

OurMethod ここでの提案手法．Bi-LSTM+ から h_t^b を除いたもの．対象単語が w_t のとき素性ベクトルとして $[h_t^f : v_{w_{t-2}} : v_{w_{t-1}} : v_{w_{t+1}} : v_{w_{t+2}}]$ を用いる．

実験結果を表 1 と図 3 に示す．提案手法の平均正解率が最

[*1] https://github.com/svn2github/word2vec

表 1　平均正解率 (%)

baseline	Sugawara	Bi-LSTM	Bi-LSTM+	OurMethod
76.92	77.28	72.84	77.96	**79.20**

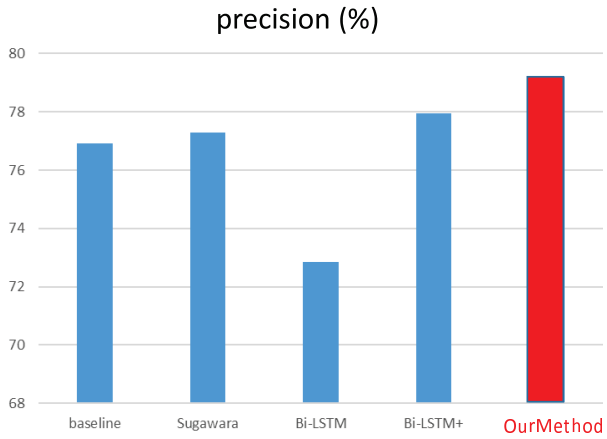

図 3　実験結果

も高く，提案手法の有効性が確認できる．特に Bi-LSTM+ と提案手法の差は逆方向の LSTM から得られる後文脈に相当する h_t^b の有無だけであり，h_t^b を利用する効果がないことも確認できる．

6. 考察

6.1 分散表現の品質の影響

語義曖昧性解消に分散表現を利用する場合，その分散表現の品質が精度に大きく影響する．この点を確認するために，分散表現のみで語義曖昧性解消を行ってみる．具体的には Sugawara の手法で基本素性ベクトルを用いずに，$[v_{w_{t-2}} : v_{w_{t-1}} : v_{w_{t+1}} : v_{w_{t+2}}]$ を素性ベクトルとして語義曖昧性解消を行う．

また本論文で利用した分散表現 nwjc2vec との比較のため，ここでは新たに二つの分散表現を構築した．一つは毎日新聞'93 年度版から '99 年度版の 7 年分の記事から取り出した 6,791,403 文を用いて word2vec により構築した分散表現 mai2vec であり，もう一つは日本語 wikipedia から取り出した 23,500,060 文を用いて word2vec により構築した分散表現 wkpd2vec である．どちらも word2vec に与えるオプションは nwjc2vec 構築のもの [13] と合わせた．

実験結果を表 2 に示す．

表 2　分散表現の比較 (%)

baseline	nwjc2vec	mai2vec	wkpd2vec
76.92	**77.71**	77.07	70.52

nwjc2vec を用いた場合が最も精度が高く，nwjc2vec の品質が高いことがうかがえる．特に nwjc2vec は baseline を大きく超えている．この点は特筆すべきものである．SemEval-2 の日本語辞書タスクでは baseline がかなり高く，通常のリソースを使う限りでは baseline を超えることは困難である．実際に SemEval-2 の参加システムで baseline を超える正解率を出したシステムはなかった．また新納はこのタスクにおいて様々なシソーラスの情報を試したが，baseline を 0.2% 以上改善できるものはなかった [11]．そこではシソーラスをアンサンブルすることで 77.28% まで改善しているが，nwjc2vec はこの値よりも 0.43% 高い．Yamaki は wikipedia から構築した分散表現と独自の手法を利用して，77.10% の正解率を出したが [10]，この値は mai2vec と同程度である．mai2vec も nwjc2vec も baseline を超えているので，どちらの分散表現もかなり品質が高いといえる．つまり語義曖昧性解消に関しては，高品質の分散表現を用いれば，前後 2 単語の分散表現だけでも従来の標準的な素性を用いた場合以上の精度が得られると考えられる．

また各分散表現を用いた場合の Sugawara の手法も試した．この結果を表 3 に示す．

表 3　Sugawara の手法における分散表現の違いの比較 (%)

baseline	nwjc2vec	mai2vec	wkpd2vec
76.92	77.28	**77.44**	77.28

mai2vec や wkpd2vec は基本素性ベクトルを合わせて利用することで精度が上がったが，nwjc2vec では逆に精度が下がっている．このことから nwjc2vec による前後 2 単語の情報は，従来の標準的な素性が持つ語義識別に関する情報をほぼ含んでいると考えられる．

6.2 LSTM の品質の影響

本論文では LSTM モデルの学習を 10 epoch で止めている．この 10 という数に根拠はない．

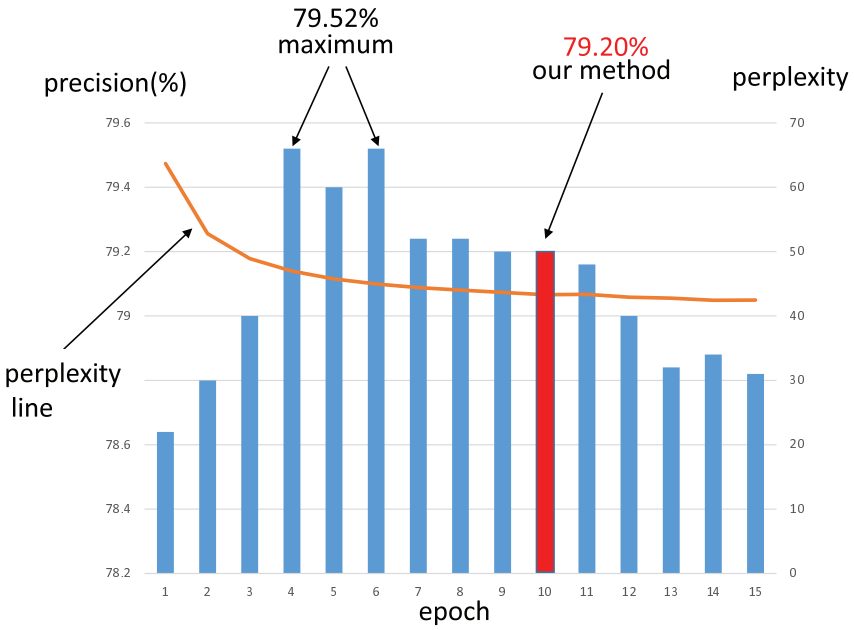

図 4 順方向 LSTM モデルのパープレキシティと語義曖昧性解消の平均正解率

ここでは順方向 LSTM の品質と本手法における精度との関係を見るために，各 epoch の学習終了毎に順方向 LSTM のモデルを保存し，そのモデルのパープレキシティを測った．なおパープレキシティを測るには，評価用コーパスが必要であるが，ここでは毎日新聞'08 年度版から 1 万文をランダムに取り出したものを利用した．また同時に，そのモデルを用いた提案手法による語義曖昧性解消の実験を行い，平均正解率を算出した．

この実験結果を図 4 に示す．図 4 内の曲線は各 epoch の毎の順方向 LSTM のモデルのパープレキシティを示している．また図 4 内の棒グラフは各 epoch の毎のモデルを用いた提案手法による語義曖昧性解消の実験での平均正解率を示している．4 あるいは 6 epoch 後のモデルを用いた場合の平均正解率 79.52% は，我々が知る限りでのこのタスクの state of the art の平均正解率 79.5%[14] とほぼ同じである．

図 4 からパープレキシティは学習が進むに連れて徐々に下がってきておりモデルの品質自体は向上していることがわかる．ただし図 4 から，モデルの品質の向上に応じて，語義曖昧性解消の実験での平均正解率が上がっているわけではないこともわかる．

どの段階のモデルを利用すべきかの適切な判断方法については，今後の課題である．

6.3 LSTM の多層化の効果

提案手法では 2 層の LSTM を用いている．提案手法の順方向 LSTM の部分を 1 層にした場合の結果を図 5 に示す．図 5 は LSTM のモデルを学習する際に各 epoch 毎にモデルを保存して，そのモデルを利用して提案手法により語義曖昧性解消の実験での平均正解率を求めたものである．比較として 2 層の LSTM の場合の平均正解率も示した．

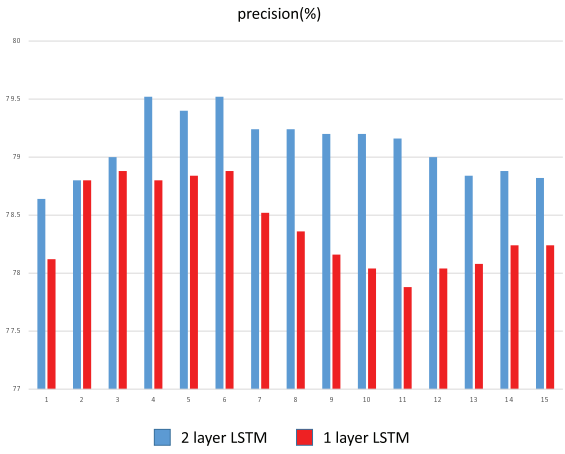

図 5 2 層 LSTM と 1 層 LSTM の違いによる語義曖昧性解消の平均正解率

図 5 から明らかに 2 層にした効果が確認できる．ただし 3 層以上の LSTM で提案手法よりも平均正解率が高くなるとは限らない．おそらくモデルの複雑さから 2 層が最適だと予想している．

6.4 逆方向 LSTM モデルの品質

本実験では逆方向の LSTM を利用する効果がなく，むしろ利用することで悪影響が出ている．この原因の 1 つとして逆方向 LSTM モデルの品質が低いことが考えられる．

ここでは 6.2 の章で利用した LSTM モデルの評価用コーパスを再び利用して，順方向及び逆方向の LSTM モデルのパープレキシティを調べた．その結果を図 6 に示す．図 6 の横軸は epoch 数を示す．

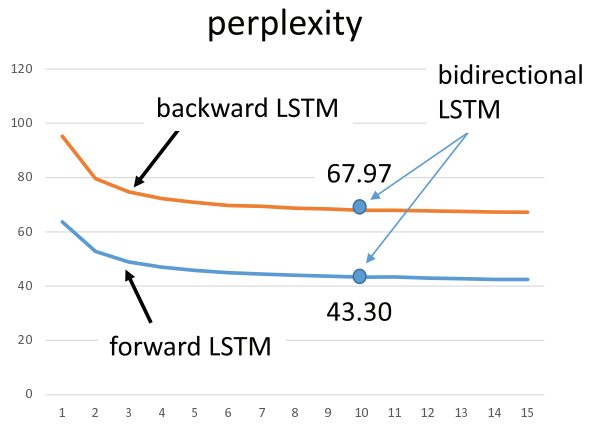

図 6 順方向 LSTM と 逆方向 LSTM のパープレキシティ

逆方向の LSTM モデルも学習が進むに従い，パープレキシティが下がっており，学習は適切に行われていると考えられるが，順方向の LSTM モデルと比較すると，パープレキシティの値は高い．同一の訓練コーパスと同一の評価用コーパスを利用していることを考えると，この差は非常に大きい．この結果は直感的にも明らかである．テキストを前から読んで後ろの単語を推測するよりも，テキストを後ろから読んで前の単語を推測することは難しいからである．

逆方向 LSTM モデルの品質が低いために，語義曖昧性解消のタスクには，それを利用する効果がなかったと考えられる．ただし，それ以外にも原因はあると考えられるので，この点は今後も調べていく必要がある．

7. おわりに

本論文では教師あり学習による語義曖昧性解消を目的として，順方向 2 層の LSTM と対象単語の前後 2 単語の分散表現を利用することを提案した．従来の双方向の LSTM を用いる手法が提案されていたが，逆方向の LSTM を用いることは語義曖昧性解消には有効でないという考えを基本としている．SemEval-2 の日本語辞書タスクを用いた実験により，本手法の有効性を示した．また分散表現や LSTM モデルの品質が語義曖昧性解消の精度にどのように影響するかを考察した．今後は逆方向 の LSTM モデルが語義曖昧性解消に有効でない理由と LSTM モデルの品質と語義曖昧性解消の精度の関係を調べたい．

謝辞

本研究の一部は国立国語研究所の共同研究プロジェクト「all-words WSD システムの構築及び分類語彙表と岩波国語辞典の対応表作成への利用」の研究成果を報告したものである．

参考文献

[1] Chen, X., Liu, Z. and Sun, M.: A Unified Model for Word Sense Representation and Disambiguation, *EMNLP-2014*, pp. 1025–1035 (2014).
[2] Iyyer, M., Manjunatha, V., Boyd-Graber, J. and Daumé III, H.: Deep Unordered Composition Rivals Syntactic Methods for Text Classification, *ACL-2015* (2015).
[3] Kågebäck, M. and Salomonsson, H.: Word Sense Disambiguation using a Bidirectional LSTM, *5th Workshop on Cognitive Aspects of the Lexicon (CogALex)*, Association for Computational Linguistics (2016).
[4] Li, J. and Jurafsky, D.: Do Multi-Sense Embeddings Improve Natural Language Understanding?, *EMNLP-2015*, pp. 1722–1732 (2015).
[5] Melamud, O., Goldberger, J. and Dagan, I.: context2vec: Learning Generic Context Embedding with Bidirectional LSTM, *CoNLL-2016*, pp. 51–61 (2016).
[6] Neelakantan, A., Shankar, J., Passos, A. and McCallum, A.: Efficient Non-parametric Estimation of Multiple Embeddings per Word in Vector Space, *EMNLP-2014*, pp. 1059–1069 (2014).
[7] Okumura, M., Shirai, K., Komiya, K. and Yokono, H.: On SemEval-2010 Japanese WSD Task, 自然言語処理, Vol. 18, No. 3, pp. 293–307 (2011).
[8] Sugawara, H., Takamura, H., Sasano, R. and Okumura, M.: Context Representation with Word Embeddings for WSD, *PACLING-2015*, pp. 108–119 (2015).
[9] Taghipour, Kaveh and Ng, Hwee Tou: Semi-Supervised Word Sense Disambiguation Using Word Embeddings in General and Specific Domains, *HLT-NAACL*, pp. 314–323 (2015).
[10] Yamaki, S., Shinnou, H., Komiya, K. and Sasaki, M.: Supervised Word Sense Disambiguation with Sentences Similarities from Context Word Embeddings, *PACLIC-30*, pp. 115–121 (2016).
[11] 新納浩幸，佐々木稔，古宮嘉那子：語義曖昧性解消におけるシソーラス利用の問題分析，言語処理学会第 21 回年次大会発表論文集，pp. 59–62 (2015).
[12] 山木翔馬，新納浩幸，古宮嘉那子，佐々木稔：教師データを用いた語義の分散表現の構築，言語処理学会第 23 回年次大会発表論文集，pp. 78–81 (2017).
[13] 浅原正幸，岡　照晃：nwjc2vec:『国語研日本語ウェブコーパス』に基づく単語の分散表現データ，言語処理学会第 23 回年次大会発表論文集，pp. 94–97 (2017).
[14] 藤田早苗，Duh, K.，藤野昭典，平　博順，進藤裕之：日本語語義曖昧性解消のための訓練データの自動拡張，自然言語処理，Vol. 18, No. 3, pp. 273–291 (2011).