DBSJ Japanese Journal
Vol. 16-J, Article No. 8, March 2018

意味の埋め込み表現を用いた文脈の再学習

Re-learning Contexts with Sense Embeddings

小中 史人♡ 白井 匡人♢ 三浦 孝夫♠

Fumito KONAKA Masato SHIRAI
Takao MIURA

語の埋め込み表現には，語が多義語であった場合にそれぞれの意味を区別して表現できないという問題がある．語ではなく意味の埋め込み表現を学習する手法も提案されているが，意味の埋め込み表現のみを用いた文脈の表現・語義の予測はされていない．本研究は，意味の埋め込み表現を用いた文脈表現と語義の予測を目的とする．具体的には，意味の埋め込み表現を用いた多次元正規分布による文脈の表現と語義の予測，並びに文脈の再学習手法を提案する．評価実験では語義曖昧性解消タスクと文の意味的類似度タスクを行い，提案手法の有効性を検証する．

Word embeddings has a problem that it can not be distinguished each sense for a word having multiple meanings. Although methods to learn sense embeddings have been proposed, there is no method for expression of context and prediction of word sense based on sense embeddings. This paper aims context expression and word sense disambiguation with sense embeddings. Specifically, we propose context representation and prediction of word sense by multivariate Gaussian distribution and re-learning of context. In this experiments, we show the effectiveness of our method on the basis of word sense disambiguation task and sentence textual similarity task.

1. 前書き

インターネット上には膨大な数の文書が存在している．今日ではTwitter[1]やFacebook[2], Instagram[3]などのソーシャルネットワークサービス（Social Network Service，SNS）の普及・流行により，あらゆる人々が容易に情報を発信できる．また，過去の書籍の電子化も進んでおり，インターネット上に存在する文書数は日々増加し続けている．これらの文書に含まれる情報の全てを人手で把握することはできないため，計算機技術の発展に伴い，計算機による支援が求められている．

計算機支援の１つとして，文書検索がある．文書検索とは，大量の文書集合からクエリに類似する文書を抽出するタスクである．このタスクでは，文書の情報表現として主にBag-Of-Wordsを用いる．Bag-Of-Wordsはベクトルの各次元を語として考え，各次元の成分を文書内でのその語の出現回数や重みとする．このため，Bag-Of-Wordsで表現された文書はスパースかつ高次元という特徴を持つ．この方法は非常に容易に構築可能という利点を持つ一方で，(1) 短い文書に有効でない，(2) 出現順序情報の損失，(3) 同義語や類義語などの単語同士の意味関係を反映できない，という欠点を持つ．

Bag-Of-WordsではSNS上で生成されるような短い文書を扱えないため，計算機上で文の意味を表現する研究が盛んに行われている [1][2][3][4][5][24]．文の意味を適切に解釈するために，語の並びに注目した手法も提案されている [23]．

語同士の意味関係を扱うためには，外部知識の利用やコーパスを用いた表現の自動獲得が考えられる．前者は，WordNet[16]のような人手によって構築された機械可読な辞書を用いて意味情報を取得する．この方法は非常に細かい粒度で意味を扱うことができるが，意味の関係性は人手によって記述されているため，扱える語数に限りがある．後者は，語の埋め込み表現 (Word Embeddings) と呼ばれる，低次元で密な実数値ベクトルによる単語表現を自動的に獲得する [7][15]．しかし，語が多義語であった場合にそれぞれの語義を区別して表現できない．

ある語が用いている語義を文脈から明らかにするタスクは語義曖昧性解消 (Word Sense Disambiguation, WSD) と呼ばれている [17]．語義曖昧性解消を解くと同時に埋め込み表現を学習する手法も提案されている [18][22] が，予測の際の文脈表現には語の埋め込み表現を利用している．

そこで本研究では，意味の埋め込み表現を用いた文脈表現と語義の予測を目指す．具体的には，意味の埋め込み表現を用いた多次元正規分布による文脈の表現と語義の予測，並びに文脈を再学習する手法を提案する．これにより，意味の埋め込み表現のみを用いた文脈表現が可能となる．語義曖昧性解消に用いられる文脈は多義語を含む可能性があり，各語義を明確に区別して表現することで，より適切な語義の予測が期待できる．

本研究の貢献は，(1) 意味の埋め込み表現による文脈表現の提案，(2) 意味の埋め込み表現による文脈を用いた語義の予測と文脈の再学習手法の提案，の二つからなる．

本稿の構成を述べる．2章では背景と関連研究を示し，本研究の位置づけについて整理する．3章では，意味の埋め込み表現を用いて語義の周辺文脈を表現する提案手法について説明する．4章では評価実験の結果を示し，5章では結論を述べる．

2. 研究目的とその背景

2.1 語の埋め込み表現

語の埋め込み表現は「似た文脈に出現する単語は似た意味を持つ」という分布仮説 [10] に基づく．Bag-Of-Wordsはベクトルの各次元を単語として扱っているため，語数が少ない短い文書は適切に扱うことができない．こうした問題は語の埋め込み表現による解決が期待できる．

Deerwesterらは Bag-Of-Words で表現された単語文書行列を特異値分解することで，低次元で密な行列に圧縮する手法を提案している [7]．単語文書行列は，文書内での語の出現回数や重みをベクトルの成分としている．つまり，文書内全体における共起（以下，大域文脈）を扱える．この手法は単語数の増加と共に次元数が増えるため，大きなコーパスに対して適用する場合は大規模な記憶装置が必要となる．また，行列分解の計算コストが大きいため新しい語の追加も困難である．Mikolovらは，周辺語を予測する Skip-Gram モデルを提案している [15]．この手法は [7] よりも高速に学習できるが，巨大なコーパスを要求する．[7] と比較すると，[15] は単語周辺での共起（以下，局所文脈）のみを考慮している．Penningtonらは局所文脈と大域文脈の両方を用いることで，[7] と [15] 双方の特長を活かす手法を提案している [19]．

♡ 非会員　法政大学理工学研究科
fumito.konaka.2t@stu.hosei.ac.jp
♢ 正会員　島根大学総合理工学研究科
shirai@cis.shimane-u.ac.jp
♠ 正会員　法政大学理工学研究科
miurat@hosei.ac.jp

[1]http://twitter.com
[2]https://www.facebook.com/
[3]https://instagram.com/

また，外部知識を用いて学習済みの語の埋め込み表現を再学習する手法も提案されている [8].

語の埋め込み表現は，文脈に依存して変化する語義を区別して表現できない．barrel という名詞を考える．この語は，(1) 大きな樽，(2) 銃身，(3) シリンダー，(4) 時計の香箱，(5) 石油量の単位バレル，(6) 液量の単位バレル，など複数の語義を持っている．barrel は，銃に関する記事では (2) の語義が，石油に関するニュースでは (5) の語義が，ビールに関する記事では (6) の語義が用いられている可能性が高い．また，(2) の語義に類似する語としては銃や弾が考えられるが，(4) の語義については銃よりも CASIO など時計に関係する語が類似すべきである．しかし語の埋め込み表現は，語に対してベクトルを付与しているために，類似すべき語を文脈に応じて区別することができない．この問題は，語ではなく語義に対してベクトルを付与することで解決が期待できる．

2.2 語の埋め込み表現による文の意味的類似度

文の意味的類似度タスク（Semantic Textual Similarity，STS）は 2 つの文に対して類似度を計算し，人手によって付与された類似度と比較するタスクである．表 1 に示すような，2 つの文に対して人手による類似度が付与されたデータセットが公開されている[4].

文の意味的類似度タスクでは，外部知識や固有表現など様々な特徴量を用いて回帰する手法が多く提案されている [1][2][3][4][5][24]．Kenter らは語の埋め込み表現を用いてこのタスクに取り組んでいる [12]．彼らは 2 つの入力文に出現している語の埋め込み表現をそれぞれ平均したベクトルを文の表現と見做す．そして文の表現間のコサイン類似度やユークリッド距離を計算し，それらの値を特徴量として採用している．

文の埋め込み表現を学習するものとしては，段落 ID と語の埋め込み表現を用いて文の埋め込み表現を学習する手法 [13] や，双方向 LSTM (Long Short-Term Memory) に語の埋め込み表現を入力し，文脈の類似度を計算する手法 [23] がある．

文の意味的類似度を精度良く求めるために語義曖昧性解消を用いる手法も存在する．Pilehvar らは，文に含まれている語の語義の最適な組み合わせを探索する手法を提案している [20].

文の意味的類似度を求めるために語の埋め込み表現を用いる手法は数多く提案されている [12][13][23] が，語の埋め込み表現を用いているため，文中の語の多義性を考慮していない．

2.3 関連研究と本研究の位置づけ

語の埋め込み表現は多義語が持つそれぞれの意味を区別して扱えない．この問題を解決するために，語義に対して低次元で密な実数値ベクトルを付与する，意味の埋め込み表現 (Sense Embeddings) が提案されている．Neelakantan らは [15] を拡張し，意味の埋め込み表現を学習する Multi Sense Skip-Gram (MSSG) モデルと Non-Parametric MSSG (NP-MSSG) モデルを提案している [18]．単語あたりの語義数は，MSSG ではパラメータとして与えられるが，NP-MSSG ではノンパラメトリックに決定される．文脈窓幅 2，語義数 3 の MSSG モデルを図 1 に示す．まず語の埋め込み表現による文脈ベクトルと，語義に対応する各クラスタの重心とのコサイン距離を求める．その後，最も距離が近いクラスタを語義として選択し，それを用いて周辺語を予測する．Tian らも [18] と似た発想で，周辺語から語義を確率的に決定し，それを用いて周辺語を予測するモデルを提案している [22].

意味の埋め込み表現は [18] や [22] で，[15] と同程度以上の精度を示すことが報告されており，文の意味的類似度タスクにおいても意味の埋め込み表現が有効に機能することが期待できる．

しかし，これらの手法は周辺語の表現に語の埋め込み表現を用いており，意味の埋め込み表現による文脈の表現や語義の予測は行っていない．つまり，ある単語の周辺に出現している語が多義語であった場合に，その各語義を区別して扱うことができない．このため本研究では，意味の埋め込み表現を用いた文脈表現の獲得と，それを用いた語義の予測を目指す．

[4] http://alt.qcri.org/semeval2016/task1/

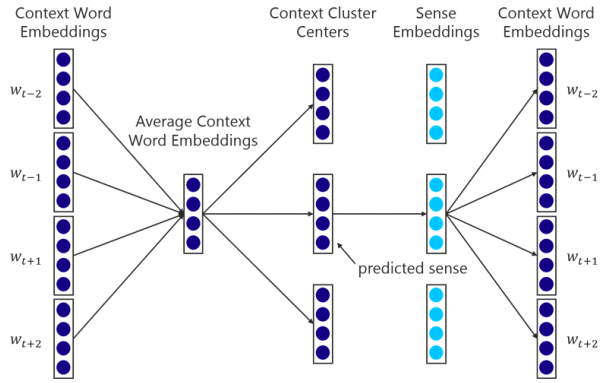

図 1: 文脈窓幅 2，語義数 3 の MSSG モデル．図中の predicted sense は，文脈から予測された語義であることを示す．このモデルは意味の埋め込み表現を学習しているが，周辺語の表現には語の埋め込み表現を用いている．

3. 提案手法

本章では新たな文脈の表現と語義の予測，並びに文脈の再学習手法を提案する．これにより，(1) 意味の埋め込み表現による文脈の表現，(2) 意味の埋め込み表現による文脈表現を用いた語義の予測，(3) 意味の埋め込み表現による文脈表現を用いた文脈の再学習，が可能になる．提案手法は意味の埋め込み表現を用いることで，従来の文脈表現では扱えない周辺語の多義性を考慮することができる．また，(1), (3) により適切な文脈表現の獲得と，(2) により多義性を考慮した文脈表現を用いた適切な語義の予測が期待できる．

具体的には，文脈表現と語義予測のために，学習済みの意味の埋め込み表現と多次元正規分布を用いる．多次元正規分布は連続値を扱う確率分布である．埋め込み表現は連続値を持つベクトルであるため，多次元正規分布の導入は妥当である．また多次元正規分布のパラメータである平均と分散により，より柔軟な文脈表現が可能となる．

提案手法のうち，まず文脈の表現方法を述べる．単語 w の語義 s が持つ文脈を，多次元正規分布 $\mathcal{N}_w^s(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ で表現する．$\boldsymbol{\mu}, \boldsymbol{\Sigma}$ はそれぞれ平均と共分散である．意味の埋め込み表現はベクトルであり，その次元同士は独立である．従って，本研究では共分散 $\boldsymbol{\Sigma}$ を対角成分のみを持つ分散行列として扱う．平均 $\boldsymbol{\mu}$ は，語義 s が出現している全ての位置の局所文脈を抽出し，その平均とする．分散 $\boldsymbol{\Sigma}$ は各局所文脈と $\boldsymbol{\mu}$ より求まる．

次に，意味の埋め込み表現による文脈を用いた語義の予測手法を述べる．ある局所文脈 c_L における単語 w が持つ語義 s は，多次元正規分布の確率密度関数 $f(\boldsymbol{x})$ を用いて事後確率を計算し，それが最大となる語義とする．

$$f(\boldsymbol{x}) = \frac{1}{(\sqrt{2\pi})^d \sqrt{|\boldsymbol{\Sigma}|}} \exp\left\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right\} \tag{1}$$

$$s = \max_i P(s_i \mid w, c_L) = \max_i f_w^{s_i}(c_L) \tag{2}$$

最後に，多次元正規分布のパラメータの更新と語義の予測を行う再学習手法の擬似コードを Algorithm 1 に示す．意味の埋め込み表現は既存手法を用いて獲得済みとする．まず，再学習に用いるコーパスを初期化する．その後，各語が持つ語義について，その語義の全ての局所文脈を取得し，$\mathcal{N}_w^s(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ のパラメータを更新す

表 1: STS 用データセットに含まれているデータの例．人手によって [0, 5] の類似度が付与されており，0 に近いほど意味的に類似していないことを，5 に近いほど類似していることを示す．

類似度	文 1	文 2
0.5	A man is smoking.	A man is skating.
1.6	A woman is slicing tomato.	A man is slicing onion.
2.8	A man is playing a guitar.	A girl is playing a guitar.
3.8	An animal run in circles.	A squirrel runs in circles.
5.0	A plane is taking off.	An air plane is taking off.

る．そして，更新されたパラメータを用いて語義を予測する．これを十分な回数繰り返す．

Algorithm 1 Re-learning Contexts with Sense Embeddings

Require: text, sense embeddings, word list of sense embeddings

1: $mean_{word}^{sense}$, $var_{word}^{sense} \leftarrow random$, $random$ # Initialize step 1
2: Predict senses of words in every position # Initialize step 2
3: **for** $iteration = 1, 2, ..., n$ **do**
4: **for** each word w in word list **do**
5: **for** each sense s in w **do**
6: $L \leftarrow getLocalContexts(w, s)$
7: $mean_w^s$, $var_w^s \leftarrow \frac{\sum_i L[i]}{L.length}$, $\frac{\sum_i (L[i]-mean)^2}{L.length}$
8: **end for**
9: **for** each pos p in text if p has w **do**
10: Predict the sense of w in p by equation (2)
11: **end for**
12: **end for**
13: Update senses of words in every position
14: **end for**

例題を用いて，Algorithm 1 の動作を述べる．以下は bank という単語の例文の集合である[5]．

> The boat grounded on a sand **bank**. ... The **bank** gave way in the flood. ... The 130th National **Bank** (Hyakusanju **Bank**) → 11 **banks** including the 130th National **Bank** were merged into the Hozen **Bank** in 1923. ... It is not allowed to climb the **bank**.

初期化ステップ 2 で，何らかの方法により bank の語義曖昧性を解消したものとする．以下はその結果である．

> The boat grounded on a sand **bank_2**. ... The **bank_1** gave way in the flood. ... The 130th National **Bank_1** (Hyakusanju **Bank_1**) → 11 **banks_1** including the 130th National **Bank_1** were merged into the Hozen **Bank_1** in 1923. ... It is not allowed to climb the **bank_1**.

ここで bank_1 は銀行を，bank_2 は土手もしくは堤防という語義を表しているものとする．簡単のため bank のみ意味を区別しているが，他の語についても同様に語義曖昧性を解消する．

次に bank_1 と bank_2 それぞれについて，局所文脈を得る．文脈窓幅は 2 とすると，最初に出現する bank_1 の周辺語として見なされる語は the, gave, way となる．文脈語について意味の埋め込み表現を平均し，それを局所文脈として扱う．bank_1, bank_2 それぞれについて全ての局所文脈を計算し，平均する．これにより多次元正規分布のパラメータが更新される．更新されたパラメータを用いて，式 2 に従い，語義を予測する．以下に示す結果において，斜体は語義が更新された箇所である．

> The boat grounded on a sand **bank_2**. ... The ***bank_2*** gave way in the flood. ... The 130th National **Bank_1** (Hyakusanju **Bank_1**) → 11 **banks_1** including the 130th National **Bank_1** were merged into the Hozen **Bank_1** in 1923. ... It is not allowed to climb the ***bank_2***.

語義が更新されたため，更なるパラメータの更新が可能になる．

4. 評価実験

本章では，語義曖昧性解消と文の意味的類似度の 2 つの実験を行う．語義曖昧性解消タスクでは，提案している意味の埋め込み表現による文脈表現が，適切な語義予測を実現できることを確認する．文の意味的類似度タスクでは，適切な語義の予測が，適切な文の表現を実現できることを確認する．

4.1 語義曖昧性解消タスク

4.1.1 データセット

語義曖昧性解消タスクでは，評価のための様々なデータセットが提案されている [9][11][21]．しかし，それらの多くは単語ペアと人手によるスコアのみが与えられている．本研究では意味の埋め込み表現を用いた文脈表現とそれによる語義の予測を提案しているため，そのようなデータセットは評価に適さない．本研究における評価実験では，文脈を用いた語義曖昧性解消の精度を評価するため，Huang らが提案している SCWS データセット [11] を用いる．表 2 に SCWS データセットに含まれているデータの一例を示す．ベースライン，提案手法ともに，語義曖昧性を解消する単語の前後 5 語を文脈として用いる．

4.1.2 評価指標

評価はデータセットに付与された人手によるスコアと，予測されたスコアとのスピアマン順位相関係数 ρ で行う．単語間の類似度指標は，[18] と同様に，$avgSimC$，$localSim$ を用いる．

$$avgSimC(w, w') = \sum_{j=1}^{K} \sum_{i=1}^{K} P(w, c, i) P(w', c', j) \times sim(SE(w, i), SE(w', j)) \quad (3)$$

$$localSim(w, w') = sim(SE(w, i), SE(w', j)) \quad (4)$$

ここで $P(w, c, i)$ は，局所文脈 c の下で単語 w が i 番目の語義を選択する確率である．また，$SE(w, i)$ は単語 w の i 番目の語義に対応する意味の埋め込み表現であり，sim は類似度関数である．本

[5] Weblio 英語例文 (http://ejje.weblio.jp/sentence/) より引用．

実験では類似度関数としてコサイン類似度を採用する．*avgS imC* は，ある文脈における各語義の生起確率で重み付けされた類似度指標である．一方 *localS im* は，文脈から予測された意味の埋め込み表現のみを用いて類似度を算出する．

4.1.3 モデル

ベースラインとして，NP-MSSG[18] を採用する．このモデルは文脈表現に語の埋め込み表現を用いているため，意味の埋め込み表現で文脈を表現する提案手法との比較に適している．NP-MSSG による意味の埋め込み表現の学習には ukWaC[6] の 100 万語コーパスを用い，出現回数 10 回以上の語を対象に，次元数を 150，文脈窓幅を 5，最大語義数を 5，エポック数は 15 とする．結果として 42,802 個の意味の埋め込み表現が獲得される．*avgS imC* の計算のための $P(w, c, i)$ には，[18] に従い，コサイン距離の逆数を用いる．

提案手法のパラメータは，ベースラインとして獲得された意味の埋め込み表現を用いて ukWaC の 100 万語コーパスの文脈を再学習して獲得する．Algorithm 1 における初期化ステップ 2 は [18] に従い，Context Cluster Centers と局所文脈とのコサイン距離を計算し，最も距離が近いクラスタに対応する語義を選択する．*avgS imC* の計算のための $P(w, c, i)$ には，多次元正規分布の確率密度を用いる．また SCWS データセットに対しても同様に，まず Context Cluster Centers を用いて周辺語の語義を予測する．その後，予測された語義に対応する意味の埋め込み表現を用いて確率密度を計算する．

4.1.4 実験結果

表 3 に，実験結果を示す．*avgS imC* の観点では，反復回数が偶数の時に提案手法がベースラインよりも強い相関を示していることがわかる．また，反復回数が偶数回の時には 0.31 程度の相関を示している．一方，奇数回の時には 0.22 程度の相関を示している．これより，提案手法では偶数回の反復で得られる分布と奇数回で得られる分布に規則性があると思われる．*localS im* の観点では，提案手法は全ての分布においてベースラインよりも強い相関を示しており，ベースラインよりも適切に文脈を表現できていると言える．

4.1.5 考察

まず，提案手法において反復回数 1 回の分布が *localS im* の観点で最も強い相関を示す理由を述べる．評価実験において，局所文脈の意味の埋め込み表現は NP-MSSG によって予測されたものである．反復回数 1 回の分布も同様に，NP-MSSG によって予測された語義から得たパラメータである．ゆえに，反復回数 1 回の分布が最も強い相関を示していると考えられる．

次に，*avgS imC* の観点で提案手法が示す相関に規則性が見られる理由を検討する．図 2 に，各反復における KL ダイバージェンスを示す．n 回目の反復で得られた分布の KL ダイバージェンスは，下式で算出している．

$$D_{KL}(P_n||P_{n-1}) = \frac{1}{2}\left\{\log\frac{|\Sigma_{n-1}|}{|\Sigma_n|} + \mathrm{tr}(\Sigma_{n-1}^{-1}\Sigma_n)\right.$$
$$\left.+(\mu_n - \mu_{n-1})^{\mathrm{T}}\Sigma_{n-1}^{-1}(\mu_n - \mu_{n-1}) - d\right\} \quad (5)$$

上式において，変数の下に添えられた n は n 回目の反復で得たパラメータであり，d は多次元正規分布の次元数である．図 2 より，収束性は確かめられない．また，奇数回目の反復で得た分布は，偶数回目の反復で得た分布と比べ，分布が大きく変化した語義が少ないことがわかる．表 5 に，偶数回目／奇数回目の反復で得た分布における，各分布の KL ダイバージェンス降順トップ 100 語義と，前回の分布におけるトップ 100 語義との共通語義数と共通する語義の一例を示す．表 5 より，奇数回目の反復で得た分布では偶数回目の反復で得た分布よりも多くの共通語義を有していることが確認できる．これは，文脈表現を更新する提案手法が意味の埋め込み表現自体の再学習をしていないために，文脈表現が振動していると考えられる．

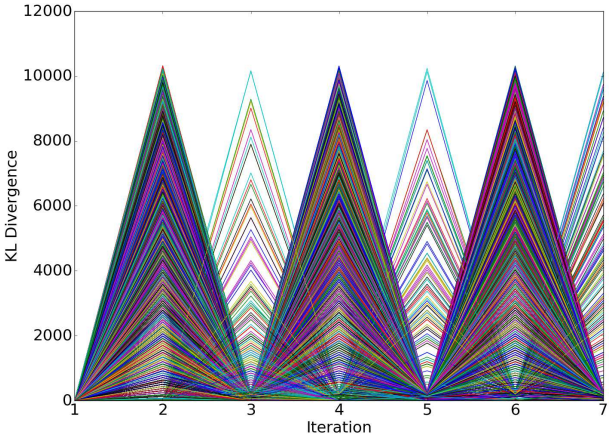

図 2: $n-1$ 回の反復で得た分布で求めた KL ダイバージェンス．奇数回目の反復で得た分布は，偶数回目の反復で得た分布と比べ，分布が大きく変化した語義が少ないことがわかる．

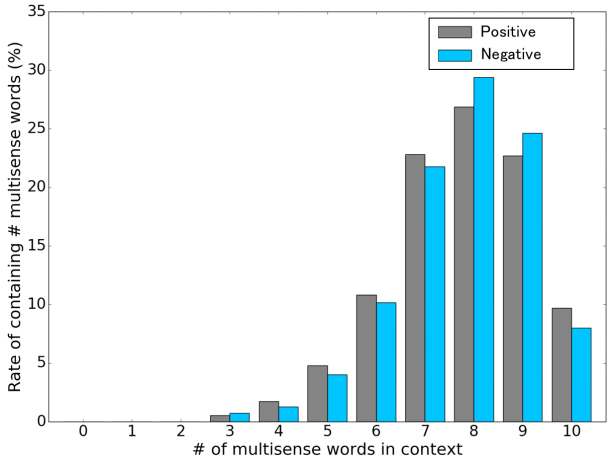

図 3: Low の場合の Positive/Negative のそれぞれのペアにおける，文脈中に含む多義語数の分布．後者は前者よりも多くの多義語を文脈中に含んでいることがわかる．

更に，*localS im* の観点で提案手法がベースラインよりも強い相関を示す理由を考察する．SCWS データセットには 2,003 件の評価用データが含まれている．このうち 783 件は，未知語が語義曖昧性解消の対象であるために類似度を計算できていない．類似度が求まった 1,220 件について，表 4 に，人手によるスコアが [0,5) の場合は Low，[5,10] の場合は High，反復 1 回で得た分布における *localS im* が [-1,0) の場合は Negative，[0,1] の場合は Positive とした際の混同行列を示す．High の場合は Positive, Negative ともに NP-MSSG と提案手法に差は見られない．しかし Low の場合では，提案手法が NP-MSSG よりも多くのペアの類似性を適切に判別できていることがわかる．図 3 に，人手によるスコアが Low の場合のそれぞれのペアにおける，文脈中に含む多義語数の分布を示す．図 3 より，Negative は Positive よりも多くの多義語を文脈中に含んでいることがわかる．これより，提案している文脈表現がより精度良い語義の予測を実現していると考えられる．

最後に，類似度を求めることができた実例を検討する．表 6 に，判定できた単語ペアの一例を示す．多義語の判定は WordNet 3.0 と NP-MSSG を用いて行っている．表 6 より，文脈中に多くの多

表 2: SCWS データセットに含まれているデータの一例．ボールド体は語義曖昧性を解消する単語であることを示す．単語ペアには [0,10] の人手による類似度スコアが付与されている．この例では ability と know-how がペアになっており，この単語間の類似度スコアを予測する．

Word 1	... However it is believed that sentient beings' karmas limit the **ability** of the Buddhas to help them. ...
Word 2	... Although Carey stated that the necessary **know-how** was at hand due to the many relations ...

表 3: 各手法のスピアマン順位相関係数 ρ による比較．提案手法がベースラインよりも正方向に強い相関を示している．

	$avgSimC$	$localSim$
NP-MSSG	.27	.166
反復回数 1	.257	**.222**
2	.313	.216
3	.224	.217
4	**.314**	.214
5	.226	.219
6	.311	.21
7	.221	.215

表 4: 人手によるスコアが [0,5) の場合は Low，[5,10] の場合は High，$localSim$ が [-1,0) の場合は Negative，[0,1] の場合は Positive とした際の混同行列．High の場合はベースラインと提案手法に差は見られないが，Low の場合は提案手法がベースラインよりも適切に区別できていることがわかる．

		High	Low
NP-MSSG	Positive	331	535
	Negative	82	272
提案手法	Positive	326	458
	Negative	87	349

義語を含んでいることがわかる．このため，多義性を考慮した文脈表現が有効に機能していると考えられる．

4.1 文の意味的類似度タスク

4.2.1 データセット

本実験では SemEval 2012[1], SemEval 2013[2], SemEval 2014[3], Sentences Involving Compositional Knowledge (SICK)[14], Paraphrase and Semantic Similarity in Twitter (PIT)[24] の 5 種類のデータセットのテストデータを用いる[6]．表 7 に各データセットのデータ数を示す．

4.2.2 評価指標

評価は，データセットに付与された人手によるスコアと，予測されたスコアとのピアソン相関係数 r，スピアマン順位相関係数 ρ で行う．文ペアの類似度関数はコサイン類似度とする．

$$sim(S_1, S_2) = cos(\sum_{w_1 \in S_1} w_1, \sum_{w_2 \in S_2} w_2) \quad (6)$$

4.2.3 モデル

ベースラインには，NP-MSSG を用いる．まず文中の各語について語義曖昧性解消を行う．そして予測された語義に対応する意味の埋め込み表現の平均を文の表現とする．

提案手法は，$localSim$ の観点で最も強い相関を示した反復 1 回目で得た分布を用いる．まず NP-MSSG で語義曖昧性解消を行う．その後，予測された語義に対応する意味の埋め込み表現を用いて，再度，語義曖昧性解消を行う．そしてベースラインと同様に，予測された語義の平均を文の表現とする．

4.2.4 実験結果

表 8 に実験結果を示す．殆どのデータセットにおいて，ピアソン相関係数，スピアマン順位相関係数ともに提案手法がベースラインよりも強い相関を示していることが確認できる．

4.2.5 考察

提案手法がベースラインよりも強い相関を示す理由を考察する．図 4 に，提案手法の適用によりベースラインとは異なる語義が予測された語が文ペア中に占める割合の分布を示す．pairs: baseline はベースラインが提案手法よりも人手によるスコアに近いスコアを予測したペアを，pairs: proposed method は提案手法がベースラインよりも人手スコアに近いスコアを予測したペアを表す．横軸の割合は，下式で算出している．

$$rate = \frac{S_1.diff + S_2.diff}{S_1.length + S_2.length} \quad (7)$$

ここで，$S.diff$ はある文において異なる語義が予測された語の総数であり，$S.length$ は未知語を除いた文中の総単語数である．また縦軸は，2 種類のペアそれぞれにおける割合を示す．図 4 より，pairs: proposed method は pairs: baseline よりも，多くの異なる語義が予測されていることがわかる．これより，意味の埋め込み表現による文脈表現が適切な予測を実現し，結果としてより良い文の表現を獲得できていると考えられる．

表 9 にデータセット中の文ペアの一例と，各手法により予測された語義番号を示す．表 9 より，Sentence 1 の Two, floating, water はベースラインとは異なる語義が予測されていることがわかる．これより，ある語義の変更が他の語の語義予測に影響を与えていることが確認できる．

5. 結論

本研究では意味の埋め込み表現を用いた適切な文脈の表現と語義の予測を目的に，多次元正規分布による文脈表現と語義予測，並びに文脈の再学習手法を提案した．多次元正規分布のパラメータである平均と共分散は，より柔軟な文脈の表現を可能にする．提案手法は学習済みの意味の埋め込み表現を用いてパラメータを獲得し，そのパラメータに基いて語義の予測を行う．そして，新たに予測された語義に対応する意味の埋め込み表現を用いてパラメータを更新する．

評価実験では，語義曖昧性解消タスクと文の意味的類似度タスクの 2 種類の実験を行った．前者では，提案手法はスピアマン順位相関係数の観点でベースラインよりも正方向に強い相関を示し，ベースラインよりも適切に語義を予測できることが確かめられた．

[6] SemEval 2015 データセット [4]，SemEval 2016 データセット [5] には欠落している箇所があるため，本実験では採用していない．

表 5: 偶数回目の反復で得た分布／奇数回目の反復で得た分布における，各分布の KL ダイバージェンス降順トップ 100 語義と，前回の分布におけるトップ 100 語義との共通語義数と共通する語義の一例．奇数回目の反復で得た分布は，偶数回目の反復で得た分布よりも多くの共通語義を有していることがわかる．これより，文脈表現の振動が考えられる．

	偶数回目の反復で得た分布 $D_{KL}(4\|\|2), D_{KL}(6\|\|4)$	奇数回目の反復で得た分布 $D_{KL}(3\|\|1), D_{KL}(5\|\|3)$	奇数回目の反復で得た分布 $D_{KL}(5\|\|3), D_{KL}(7\|\|5)$
KLD トップ 100 との共通語義数	12	27	54
語義の例	band_5 click_5 policy_2	action_1 house_4 styles_3	action_1 house_4 styles_3

表 6: SCWS データセットに含まれているペアの一例．ボールド体は曖昧性解消の対象の単語である．人手によるスコアは 1.8 となっている．ベースラインはこのペアを類似度 0.06 と判定したが，提案手法では –0.07 と判定している．多義語の判定は WordNet 3.0 と NP-MSSG を用いて行っている．文脈中に多義語が出現しており，多義性を考慮した文脈表現が機能していると考えられる．

	Word 1										
	to	that	history	the	country	**role**	in	the	region	as	indeed
WordNet	-	-	○	-	○	○	○	-	○	○	○
NP-MSSG	○	○	○	-	○	○	○	-	○	○	○
	Word 2										
	as	a	conjuror	takes	his	**hat**	he	produced	an	endless	swarm
WordNet	○	○	○	-	-	○	○	-	-	○	○
NP-MSSG	○	○	-	○	○	○	○	○	○	○	-

具体的には，ある文脈下での語義の生起確率で重み付けされた類似度指標では最大で 0.314 の相関を達成し，ベースラインと比べ最大 0.044 の相関の向上を確認できた．予測された語義で計算される類似度指標では最大で 0.222 の相関を達成し，ベースラインと比べ最大 0.056 の相関の向上を確認できた．また，後者においても提案手法は殆どのデータセットにおいて，ピアソン相関係数，スピアマン順位相関係数ともにベースラインよりも正方向に強い相関を示し，ベースラインよりも適切に文を表現できることが確かめられた．具体的には，ピアソン相関係数の観点では最大で 0.455 の相関を達成し，ベースラインと比べ最大 0.136 の相関の向上を確認できた．スピアマン順位相関係数の観点では最大で 0.456 の相関を達成し，ベースラインと比べ最大 0.144 の相関の向上を確認できた．

今後は，巨大なコーパスを用いて学習することで語義曖昧性解消タスクにおいてより高い相関を実現したい．さらに，文の意味的類似度タスクで語の埋め込み表現を用いた文の表現と比較し，意味の埋め込みがより適切に文を表現できることを確認したい．また，提案手法は文脈表現を更新しているが意味の埋め込み表現の学習は行っていないため，提案手法の意味の埋め込み表現の学習への組み込みも検討したい．

[文献]

[1] Eneko Agirre, Daniel Cer, Mona Diab, and Aitor Gonzalez-Agirre (2012): ”SemEval-2012 Task 6: A Pilot on Semantic Textual Similarity”, in Proceedings of the 6th International Workshop on Semantic Evaluation (SemEval), pp. 385-393.

[2] Eneko Agirre, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, and WeiWei Guo (2013): ”*SEM 2013 shared task: Semantic Textual Similarity”, in Proceedings of the 7th International Workshop on Semantic Evaluation (SemEval), pp. 32-43.

[3] Eneko Agirre, Carmen Banea, Claire Cardie, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Weiwei Guo, Rada Mihalcea, German Rigau, and Janyce Wiebe (2014): ”SemEval-2014 Task 10: Multilingual Semantic Textual Similarity”, in Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval), pp. 81-91.

[4] Eneko Agirre, Carmen Banea, Claire Cardie, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Weiwei Guo, Inigo Lopez-Gazpio, Montse Maritxalar, Rada Mihalcea, German Rigau, Larraitz Uria, and Janyce Wiebe (2015): ”SemEval-2015 Task 2: Semantic Textual Similarity, English, Spanish and Pilot on Interpretability”, in Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval), pp. 252-263.

[5] Eneko Agirre, Carmen Banea, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Rada Mihalcea, German Rigau, and Janyce Wiebe (2016): ”Semeval-2016 Task 1: Semantic Textual Similarity, Monolingual and Cross-Lingual Evaluation”, in Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval), pp. 497-511.

[6] Marco Baroni, Silvia Bernardini, Adriano Ferraresi, and Eros Zanchetta (2009): ”The WaCky Wide Web: A Collection of Very Large Linguistically Processed Web-Crawled Corpora”, Language Resources and Evaluation 43 (3), pp. 209-226.

[7] Scott Deerwester, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, and Richard Harshman (1990): ”Indexing by latent semantic analysis”, Journal of the American society for information science, 41(6), pp. 391-407.

表 7: 各テストデータのデータ数.

データセット		データ数
SemEval 2012	MSRpar	750
	MSRvid	750
	SMTeuroparl	459
	OnWN	750
	SMTnews	399
SemEval 2013	FNWN	189
	headlines	750
	OnWN	561
SemEval 2014	deft-forum	450
	deft-news	300
	headlines	750
	images	750
	OnWN	750
	tweet-news	750
SICK		4,927
PIT		972
合計		14,257

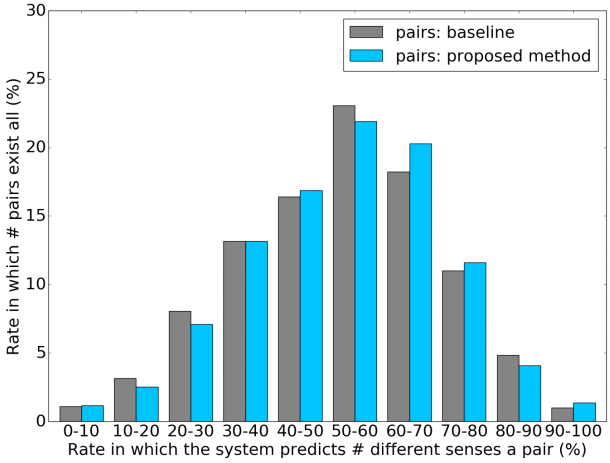

図 4: 提案手法の適用によりベースラインとは異なる語義が予測された語が文ペア中に占める割合の分布．提案手法がベースラインよりも人手によるスコアに近いスコアを予測したペアでは，より多くの語義が変更されていることがわかる．

[8] Manaal Faruqui, Jesse Dodge, Sujay Kumar Jauhar, Chris Dyer, Eduard Hovy, and Noah A. Smith (2015): ”Retrofitting Word Vectors to Semantic Lexicons”, in Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT), pp. 1606-1615.

[9] Lev Finkelstein, Evgeny Gabrilovich, Yossi Matias, Ehud Rivlin, Zach Solan, Gadi Wolfman, and Eytan Ruppin (2001): ”Placing search in context: The concept revisited”, in Proceedings of the 10th international conference on World Wide Web (WWW), ACM, pp. 406-414.

[10] Zellig S. Harris (1954): ”Distributional structure”, Word, Vol. 10, pp. 146-162.

[11] Eric H. Huang, Richard Socher, Christopher D. Manning, and Andrew Y. Ng (2012): ”Improving Word Representations via Global Context and Multiple Word Prototypes”, in Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (ACL), pp. 873-882.

[12] Tom Kenter and Maarten de Rijke (2015): ”Short Text Similarity with Word Embeddings”, in Proceedings of the 24th ACM International Conference on Information and Knowledge Management (CIKM), pp. 1411-1420.

[13] Quoc Le and Tomas Mikolov (2014): ”Distributed Representations of Sentences and Documents”, in Proceedings of the 31st International Conference on Machine Learning (ICML), pp. 1188-1196.

[14] Marco Marelli, Stefano Menini, Marco Baroni, Luisa Bentivogli, Raffaella Bernardi, and Roberto Zamparelli (2014): ”A SICK cure for the evaluation of compositional distributional semantic models”, in Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC), pp. 216-223.

[15] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dea (2013): ”Distributed representations of words and phrases and their compositionality”, in Proceedings of Advances in Neural Information Processing Systems (NIPS), pp. 3111-3119.

[16] George A. Miller (1995): ”WordNet: a lexical database for English”, Communications of the ACM, 38.11: 39-41.

[17] Roberto Navigli (2009): ”Word sense disambiguation: A survey”, ACM Computing Surveys (CSUR) 41.2, 10.

[18] Arvind Neelakantan, Jeevan Shankar, Alexandre Passos, and Andrew McCallum (2014): ”Efficient non-parametric estimation of multiple embeddings per word in vector space”, in Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 1059-1069.

[19] Jeffrey Pennington, Richard Socher, and Christopher D. Manning (2014): ”Glove: Global Vectors for Word Representation”, in Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 1532-1543.

[20] Mohammad Taher Pilehvar, David Jurgens, and Roberto Navigli (2013): ”Align, Disambiguate and Walk: A Unified Approach for Measuring Semantic ”, in Proceedings of the 51th Annual Meeting of the Association for Computational Linguistics (ACL), pp. 1341-1351.

[21] Herbert Rubenstein and John B. Goodenough (1965): ”Contextual correlates of synonymy”, Communications of the ACM, 8.10, pp. 627-633.

[22] Fei Tian, Hanjun Dai, Jiang Bian, Bin Gao, Rui Zhang, Enhong Chen, and Tie-Yan Liu (2014): ”A Probabilistic Model for Learning Multi-Prototype Word Embeddings”, in Proceedings of the 25th International Conference on Computational Linguistics (COLING), pp. 151-160.

[23] Shengxian Wan, Yanyan Lan, Jiafeng Guo, Jun Xu, Liang Pang, and Xueqi Cheng (2016): ”A Deep Architecture for Semantic Matching with Multiple Positional Sentence Representations”, in Proceedings of the 30th AAAI Conference on Artificial Intelligence (AAAI), pp. 2835-2841.

[24] Wei Xu, Chris Callison-Burch, and William B. Dolan (2015): ”SemEval-2015 Task 1: Paraphrase and Semantic Similarity in Twitter (PIT)”, in Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval), pp. 1-11.

表 8: 各手法のピアソン相関係数 r とスピアマン順位相関係数 ρ による比較．殆どのデータセットにおいて，提案手法がベースラインよりも正方向に強い相関を示している．

	SemEval 2012										SemEval 2013					
	MSRpar		MSRvid		SMTeuroparl		OnWN		SMTnews		FNWN		headlines		OnWN	
	r	ρ	r	ρ	r	ρ	r	ρ	r	ρ	r	ρ	r	ρ	r	ρ
NP-MSSG	.234	.258	.084	.14	**.205**	.251	.371	.414	.227	**.21**	**.274**	**.282**	.332	.329	.179	.205
提案手法	**.279**	**.272**	**.201**	**.228**	.196	**.286**	**.452**	**.451**	**.264**	.208	.181	.193	**.432**	**.434**	**.218**	**.256**

	SemEval 2014												SICK		PIT	
	deft-forum		deft-news		headlines		images		OnWN		tweet-news		SICK		PIT	
	r	ρ	r	ρ	r	ρ	r	ρ	r	ρ	r	ρ	r	ρ	r	ρ
NP-MSSG	.204	.251	.393	.385	.316	.312	.231	.257	.373	.408	.418	.397	.35	.37	**.279**	**.301**
提案手法	**.273**	**.289**	**.436**	**.417**	**.455**	**.456**	**.288**	**.327**	**.406**	**.449**	**.517**	**.52**	**.398**	**.405**	.248	.24

表 9: データセット中の文ペアの一例と，各手法により予測された語義番号．表中の - は未知語であることを示す．このペアに付与された人手によるスコアは 1.8 である．ベースラインはこのペアを類似度 3.35 と判定しているが，提案手法では類似度 3.77 と判定している．表より，提案手法とベースラインで異なる語義が予測されていることを確認できる．

	Sentence 1							
	Two	men	on	boat	floating	on	the	water
NP-MSSG	5	5	2	3	3	1	1	1
提案手法	1	5	2	3	4	1	1	5

	Sentence 2					
	A	wooden	yacht	on	the	ocean
NP-MSSG	1	3	-	2	1	5
提案手法	1	3	-	2	1	5

小中 史人　Fumito KONAKA
2017 年法政大学理工学研究科修士課程修了
白井 匡人　Masato SHIRAI
2016 年法政大学理工学研究科博士課程修了．博士 (工学)．現在，島根大学総合理工学研究科特任助教．機械学習，自然言語処理の研究に従事．
三浦 孝夫　Takao MIURA
京大・理卒．工博 (東京大学)．現在，法政大学理工学部創生科学科教授．データモデル，知識表現，演繹データベース，複合オブジェクトなどの分野の研究に従事．ACM，情報処理学会各会員．