Vol.2015-NL-221 No.4
Vol.2015-SLP-106 No.4
2015/5/25

日本語単語ベクトルの構築とその評価

吉井 和輝[†1]　Eric Nichols[†2]　中野 幹生[†2]　青野 雅樹[†1]

単語ベクトルは，統計的自然言語処理で利用しやすい分散意味表現として近年盛んに用いられるようになってきた．しかしながら，今まで主に英語で評価されてきたため，英語以外の言語での有効性は不明である．本研究では，単語の類推(word analogy)と文完成(sentence completion)の二つの評価タスクを用い，著名なオープンソースツールである word2vec(gensim の再実装)と GloVe を用いて構築した日本語単語ベクトルの評価を行った．単語の類推タスクでは，英語データで公表されている結果に近い結果を得たが，文完成のタスクでは，精度が大幅に減少した．本稿では，両タスクのエラー解析で明らかになった英語の単語ベクトルと日本語の単語ベクトルの性能差や，日本語特有の問題について調査した結果について述べる．

Construction and evaluation of Japanese word vectors

KAZUKI YOSHII[†1]　ERIC NICHOLS[†2]
MIKIO NAKANO[†2]　MASAKI AONO[†1]

Word vectors have been the subject of a great deal of research in recent years, due to their effectiveness at representing meaning in statistical approaches. However, evaluation of word vectors has thus far been limited to a small number of tasks focusing on the English language. This paper aims to fill that gap by providing comprehensive evaluation of Japanese word vectors. We construct datasets for word analogy and sentence completion tasks and compare vectors constructed with two popular tools, word2vec and GloVe. Evaluation on the word analogy task produced comparable results to those reported on English data, while on the sentence completion task, results were significantly lower than those reported on English data. We conduct error analysis for both tasks and discuss potential factors contributing to differences in performance for English and Japanese.

1. はじめに

自然言語処理の分野において，単語の意味を理解することは対話システムや機械翻訳，情報検索，構文解析等の有用で様々なタスクでの活用が期待できるため，重要な課題となっている．単語の意味を計算機で扱うために様々な手法が提案されているが，その中で最も成功しているものの一つに単語の分散意味表現として単語ベクトルを構築する手法がある．

しかし，これまで提案されている手法の多くの評価が英語に限定されているため，他の言語での適用の有効性がまだ不明である．特に，日本語と英語では使用する字種数，語順，文法などに大きな差が存在するため，日本語にこれらの手法を適用できるかは十分な検討が必要となる．

本研究では，厳密に日本語の単語ベクトルを評価することを目標とする．そのため，大規模な学習データを用いて日本語単語ベクトルの構築した．そして，単語の類推と文完成の二つの評価タスクについて新たに評価用データセットを構築し，その評価を行った．この結果から，日本語と英語の単語ベクトル間の差異や日本語特有の問題について考察する．評価には，精度が良いことが知られており，かつオープンソースツールが公開されている word2vec (gensim の再実装)と GloVe の二つの手法を適用した．

†1 豊橋技術科学大学
Toyohashi University of Technology
†2 ホンダ・リサーチ・インスティチュート・ジャパン
Honda Research Institute Japan

2. 関連研究

2.1 単語ベクトル

単語ベクトルは，単語を低次元の意味空間上にて実数値の数値列であるベクトル表現で表したものである．単語ベクトルを構築する手法は古くから研究されており，また様々な種類の手法が提案されているが，主に 2 種類のアプローチに分類できる．

一つはコーパス内での単語の出現頻度や共起などの統計情報をまとめた行列を因子分解し，低次元の単語ベクトルを求めようとする手法で，LSI[1]などがその代表的な手法である．これらの手法はコーパス全体の傾向を考慮して単語ベクトルの構築を行うため，グローバルな手法に分類される．

もう一つは，「似た意味の単語は似た文脈に出現しやすい」という分布仮説をもとに，単語の周辺の文脈情報から単語ベクトルを学習する手法で，確率的ニューラル言語モデル[2]や CBOW モデル、Skip-gram モデル[3]などが代表的な手法である．これらの手法は，文脈というコーパス内の一部分に注目して単語ベクトルの構築を進めていくことから，ローカルな手法に分類される．

ⓒ2015 Information Processing Society of Japan

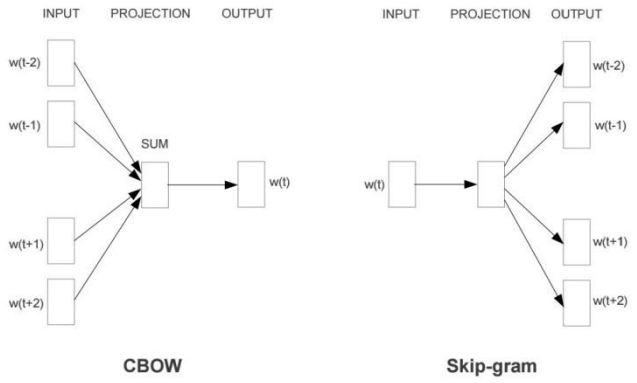

図 1 word2vec のモデルの概念図([3]より抜粋)

2.1.1 word2vec

word2vec は Mikolov らによって提案された，ニューラルネットワークを用いて単語の分散意味表現を獲得する手法，正確には，その手法が実装されたオープンソース実装の名称である[3]．学習する方法として， Continuous Bag-of-Words (CBOW)モデルと Skip-gram モデルの二つのモデルを提案している(図 1).

Skip-gram モデルは, 学習データの文章中に出現するある単語 w_t から，その前後に出現する単語 w_{t-n}, w_{t-n+1}, …, w_{t-1}, w_{t+1},…w_{t+n-1},w_{t+n} を推定するニューラルネットワークを学習する. これは式(1)式のようになり,この式を最大化するように v, u の値の更新を繰り返す．ここで，Tは学習対象としている文中の単語数を，nはウィンドウサイズを，Wは学習データに出現する全単語を，v_{w_t}は単語w_tを表すベクトル，$u_{w_{t+j}}$は単語w_{t+j}の出現を予測するベクトルを表す.

$$\sum_{t=1}^{T}\sum_{-n\le j\le n, j\neq 0}\log P(w_{t+j}|w_t) \qquad (1)$$

$$\log P(w_{t+j}|w_t) = \frac{\exp(v_{w_t}\cdot u_{w_{t+j}})}{\sum_W \exp(v_{w_t}\cdot u_{w_{t+j}})} \qquad (2)$$

学習では，階層的ソフトマックス法やネガティブサンプリング，頻出語のサブサンプリングといった高速化手法が提案されており[4]，他手法に比べ比較的高速に単語ベクトルの構築が行える．また，英語の様々なタスクにおいて他手法より高精度な結果となることが報告されている．

また，Skip-gram モデルで学習した単語ベクトルは，それらの単語ベクトルの演算が単語間の関係を表すような特徴を持つことが報告されている．例えば，king – man + woman という単語ベクトルの演算を行った結果は queen という単語のベクトルと類似度が高くなる．このように，単語の意味や関係を単純な単語ベクトルの演算で表現でき，単語類推のタスクなどで当時の最新の手法よりも良い評価結果が確認された.

2.1.2 GloVe

GloVe(Global Vectors)は Pennington らによって提案された，単語の共起行列を用いて単語の分散意味表現を獲得する手法である[5].

Pennington らは，単語の分散意味表現の獲得方法には 2.1 で述べたようなグローバルな手法とローカルな手法があるとしており，それぞれの手法の長所と短所について考察している．そして，ローカルな手法である Skip-gram モデルにグローバルな手法の長所を組み合わせたモデルを提案している.

コーパス内において，ある単語 i, j が同じ文脈に出現する場合，要素 X_{ij} の値をインクリメントするような単語の共起行列 X を考えたとき，GloVe のモデルは以下の式(3)のように表される．この時，V はコーパス内に出現する単語の種類数を, w_i, w_jは単語 i, j の単語ベクトルを, b_i, b_jは単語 i, j のバイアスを表す. またfは重み付けの関数であり，実験により(4)式が良いことが示されている.

$$J = \sum_{i,j=1}^{V} f(X_{ij})(w_i^T w_j + b_i + b_j - \log Xij)^2 \qquad (3)$$

$$f(x) = \begin{cases} (\frac{x}{X_{max}})^{\frac{3}{4}} & \text{if } x < X_{max} \\ 1 & \text{otherwize} \end{cases} \qquad (4)$$

Pennington らは，Skip-gram モデルの学習について数式的に分析を行い，GloVe は Skip-gram モデルに共起頻度の比による重み付けなどの工夫を加えて改良したものであることを示している．この重み付けにより，文中でよく共起されかつ他の単語とは共起されにくい単語間の関係が学習されやすくなり，より高品質なベクトルが構築される．GloVe の学習は高速であり，構築された単語ベクトルには，Skip-gram モデルと同様に意味の演算が行える性質を持つことが知られている．また，英語での評価では単語類推と固有表現認識のタスクで Skip-gram モデルよりも良い精度が示されている.

2.2 評価尺度

単語ベクトルには決定的な評価方法は存在していないが，1) 単語ベクトル自体の性能に関する評価と，2) 単語ベクトルの何らかのタスクを解く性能の評価の 2 種類に大きく分類することができる.

単語ベクトル自体の性能の評価では，ボキャブラリーのサイズや学習時間，学習進度あたりの学習率などが評価尺度として考えられる．このような評価の結果は，タスクによらない単語ベクトルの実用性などについての参考となる.

タスクを解くことによる評価については，どのようなタスクを行えばよいかなどの明確な基準はない．しかし，よく対象となるタスクとして,「同義語の判定」タスクがある．これは，意味の似た単語のベクトル同士のコサイン類似度を図り，他の単語に比べ類似度が最も高くなるかを比較するタスクである．その他にも，複数の単語を与えられた時

に，単語間の関係を類推してふさわしい単語を選択するような「単語の類推」タスクや，人名や地名などの固有名詞，日付や時間の表現などを自動的に抽出するような「固有表現認識」タスク[6]など，単語の意味を扱うようなタスクなどで評価されることが多い．また，単語ベクトルの構築過程を言語モデルとして応用できる場合は，「文完成」タスクで評価を行うことができる．このための代表的な英語のデータセットとして，Microsoft Research Sentence Completion Challenge[7]がある．

Skip-gram モデルで構築された英単語ベクトルの評価では同義語判定，単語類推，文完成タスクが，GloVe で構築された英単語ベクトルの評価では単語類推，固有表現認識タスクが用いられている．本研究では，日本語単語ベクトルの単語類推タスクと文完成タスクの性能の評価について取り組む．

3. 日本語単語ベクトルの構築

3.1 データセット

単語ベクトルを構築する際の学習データセットとして，日本語 Wikipedia のテキストデータ，京都テキストコーパス Version4.0 の全テキストデータ，Common Crawl[8][a]の日本語テキストデータを使用した．

収集したデータのサイズはおよそ 20GB であり，総単語数は約 30 億単語であった．

3.2 前処理

単語ベクトルの学習を行うためには，一行毎に学習対象とする一文が記述されているテキストデータが必要となる．収集したデータをそのような形式に変換し，更に効率良く，精度よく学習を行うために，前処理を施す．

日本語 Wikipedia のテキストデータには，データを効率良く活用・管理するために Wikipedia 特有の記法によるタグ等がいくつも付与されている．また，データは xml 形式で配布されるため xml タグなどが多数存在している．これらは単語ベクトルの学習では不要となるので，これらを抽出する正規表現を記述し，不要な文字列の除去を行った．

Common Crawl のデータには，収集元の Web ページの URL が付与されている．また，同じ Web サイトの複数ページがクローリングされる事により，多数の重複文が出現している．これらも不要となるため，重複文，Web リンクが出現する文章の除去を行った．

3.3 データの整形

word2vec や GloVe で単語ベクトルを構築するためには，単語単位に分割された文章が必要となる．英語等の言語はもともと文章が単語毎に区切られているため，自然文をそのまま学習データとして使用できるが，日本語は複数の単語が繋がって一つの文章を構成しているため，そのままでは学習データとして用いることができない．したがって，日本語文を単語単位に分割する処理が必要となる．

文章を単語単位に分割する処理として，分かち書きがある．これは文章に対して形態素解析を行い，形態素毎に文章を分割する処理である．形態素とは言語が意味を持つ最小単位であるため，これを単語として扱うことで学習を行うことが出来る．分かち書きにはオープンソースの日本語形態素解析器である MeCab[b]を使用した．

表 1 学習したモデルのパラメータ

パラメータ	Skip-gram モデル	GloVe
次元数	640	640
ウィンドウサイズ	10	10
最小単語出現数	25	25
ネガティブサンプリング	10	–
学習反復回数	5	25
学習率 α	0.025	0.75
学習率 η	–	0.05

3.4 学習の実施

前処理を行ったデータ全てを分かち書きしたものを最終的な学習データとし，それを用いて日本語単語ベクトルの構築を行った．Skip-gram モデルの単語ベクトル構築には，文の生成確率の導出や単語類推の処理を実装しやすいという理由から Python の gensim モジュール[c][d]を用いた．GloVe での単語ベクトル構築には Pennington らが公開しているオープンソースツール GloVe[e]を用いた．評価では，GloVe での学習結果を gensim で扱うデータ形式で読み込み二つの手法間で同様の評価を行う．

学習には表 1 のパラメータを用いた．ベクトルの次元数やウィンドウサイズは Mikolov らが行った実験の値を参考に決定した．また，最小単語出現数とネガティブサンプリングは実験的に決定し，その他のパラメータは使用したツールのデフォルトの設定で行った．最終的に，辞書サイズが約 42 万単語の日本語単語ベクトルが構築された．

4. 単語類推タスク

4.1 実験の概要

本実験では，単語間の関係を推定し，別の単語にとって同じ関係となる単語を類推する問題を解く性能について評価を行う．この問題では，特定の関係を持つ２つの単語と，

a.) http://commoncrawl.org/

b.) http://mecab.googlecode.com/svn/trunk/mecab/doc/index.html
c.) https://radimrehurek.com/gensim/
d.)予備実験を行い，単語ベクトルの学習時間や精度が word2vec と同等に成ることを確認した。
e.) http://nlp.stanford.edu/projects/glove/

ⓒ2015 Information Processing Society of Japan

Vol.2015-NL-221 No.4
Vol.2015-SLP-106 No.4
2015/5/25

表 2 類推を行う単語間の関係

単語の関係	関係の説明	具体例
形容詞 – 副詞 (Q1)	形容詞と同じ意味を持つ副詞	(赤い,赤く), (眠い,眠く)
名詞 – 副詞 (Q2)	漢字 1 文字の名詞と同じ意味を持つ副詞	(小,小さく), (美,美しく)
名詞 – 形容詞(Q3)	漢字 1 文字の名詞と同じ意味を持つ形容詞	(短,短い), (欲,欲しく)
国名 – 首都 (Q4)	国の名前とその国の首都の名前	(日本, 東京), (カナダ,オタワ)
県名–県庁所在地(Q5)	県の名前とその県の県庁所在地の名前	(北海道, 札幌) (愛知, 名古屋)
国名–通貨単位(Q6)	国の名前とその国で使われる通貨の単位	(日本, 円), (中国, 元)
食べ物 – 味 (Q7)	食べ物の名前とその食べ物の味を形容する単語	(カレー, 辛い), (ケーキ, 甘い)
男 – 女 (Q8)	男の名称と同じ意味を持つ女の名称	(兄, 姉), (王, 女王)
名詞の反対語 1 (Q9)	名詞と逆の意味をもつ名詞,かつ,それらがポジティブまたはネガティブな意味を持たないもの	(上, 下), (表, 裏) (北, 南), (大人, 子供)
名詞の反対語 2 (Q10)	名詞と逆の意味をもつ名詞,かつ,それらがポジティブまたはネガティブな意味を持つもの	(光, 闇), (利点, 欠点), (幸福, 不幸), (黒字, 赤字)
形容詞の反対語(Q11)	形容詞と逆の意味を持つ形容詞	(良い, 悪い), (強い, 弱い)
動詞の反対語(Q12)	動詞と逆の意味を持つ動詞	(得る, 失う), (押す, 引く)

それとは関係のない 1 つの単語の合計 3 単語が与えられる．後者の単語に対して，前者の 2 つの単語間の関係を満たすような単語を類推することがこの問題の趣旨である．以下に問題例を示す．

男－女，　王－?　　正答　女王
日本－東京，　アメリカー?　　正答　ワシントン

4.2 データセットの構築

はじめに，類推を行う単語間の関係の定義を行った．Mikolov らが Skip-gram モデルの評価のために作成した評価問題[3][f]を確認し，どのような単語の関係の類推を行っているのかを調査した．この評価問題に使われている単語の関係としては，現在進行形や過去形、複数形といった英単語の文法に関するもの，国名とその首都名や通貨の単位といった事実関係に関するもの，そして男女の名称や単語の反対語といった概念的な関係に関するものに分類できた．そこで，日本語単語ベクトルの評価の場合もこれらの関係について定義し，評価に用いる単語間の関係を定義した．定義した単語間の関係とその具体例を表 1 に示す．表 1 の Q1~Q3 が文法に関する関係，Q4~Q6 が事実に関する関係，Q7~Q12 が概念に関する関係となっている．

表 3 単語類推タスクの実験結果

単語の関係	問題数	Skip-gram モデル		GloVe	
		正解数	正答率	正解数	正答率
Q1	496	349	0.704	375	0.756
Q2	378	52	0.138	104	0.275
Q3	378	29	0.077	123	0.325
Q4	66	30	0.455	28	0.424
Q5	171	109	0.637	119	0.696
Q6	66	8	0.121	13	0.197
Q7	10	0	0.000	1	0.100
Q8	105	87	0.829	84	0.800
Q9	300	117	0.390	101	0.337
Q10	45	22	0.489	24	0.533
Q11	171	65	0.380	53	0.310
Q12	91	10	0.110	4	0.044
Total	**2277**	**878**	**0.386**	**1029**	**0.452**

4.3 実験方法

単語ベクトルを用いて本問題を解くために，単語ベクトルの加減算と類似度の計算を行う．例として，単語 A, B, C が与えられ単語 A, B と単語 C, X が同じ関係となるような単語 X を類推する問題を考える．単語 A, B, C の単語ベクトルをそれぞれ v_A,v_B,v_C として,以下の計算を行う．

$$v_B - v_A + v_C = v_X$$

そして, 辞書内の各単語の単語ベクトルと v_X とのコサイン類似度を計算する．この類似度が最も大きい単語を，単語 X の予測として扱う. 単語 X の予測と実際の単語が一致していた場合，正解とする．

評価尺度には正答率を用いる．正答率は，以下の式で計算する．

$$\text{正答率} = \frac{\text{正答数}}{\text{総問題数}}$$

4.4 実験結果

実験結果を，表 3 に示す．結果の考察については 6.1 節で述べる．

f.) https://code.google.com/p/word2vec/source/browse/trunk/questions-words.txt

ⓒ2015 Information Processing Society of Japan

5. 文完成タスク

5.1 実験の概要

本実験では，ある文章を完成させるためにふさわしい単語や表現を推測する問題を解く性能について評価を行う．この問題では，空白部分を持つ文章と，その空白部分に挿入するための単語（表現）の候補が４つ選択肢として与えられる．４つの候補のうち，いずれか１つのみが実際に空白部分に挿入することで文章として成立するものであり，残りの３つが文章として意味が通らなくなるものである．この時に，正しい文章となる単語，または表現を推測することが問題の趣旨である．

5.2 データセットの構築

本実験の評価データとして,「日本語能力試験(JLPT)」[g]の問題を用いる.「日本語能力試験」は日本語を母語としない人の日本語能力を測定し認定する目的で行われている試験である．この試験で扱われる問題の中に，本実験と同様の形式のものがあるため，それを評価問題として使用することにした[h].

本実験を行うにあたり, web 上に日本語能力試験の 1, 2, 3 級の対策問題として公開されている問題を手動で収集し，評価用の問題のデータセットを作成した．データセットの総問題数は 605 問となった．

作成した問題の具体例を表 4 に示す.

5.3 実験方法

単語ベクトルを用いて本問題を解くために，問題文に選択肢の単語を当てはめた場合の，文章の生起確率を計算する．この時，GloVe はモデルの設計上言語モデルを生成できないため，Skip-gram モデルでのみ評価を行う．

Skip-gram モデルで単語ベクトルを構築する場合，式(2)を用いて，ある単語が出現した時にその前後に出現する単語の出現確率を最大化するように学習を行う．よって，学習後の言語ベクトルを用いて(2)式を計算することで，ある単語の周囲の単語の出現確率を計算することができる．したがって，文中に出現するすべての単語に対して，その周囲の単語の出現確率を計算してその総和をとり，これをその文章の生起確率として扱う．実験では，上記の方法で空白部分に選択肢の単語を当てはめた文章の生起確率をそれぞれ計算する．その中で，最も生起確率の高くなる選択肢を，単語ベクトルでの推測結果として評価を行う．

評価尺度には正答率を用いる．正答率は，単語類推タスクと同様の計算方法で算出する．

比較手法として，n-gram 言語モデルを用いる．n-gram 言語モデルを用いる場合も同様に，それぞれの選択肢の単語で空白を埋めた問題文の生起確率を計算し，最も確率の高いものを正解として評価を行う．単語ベクトルの構築には SRILM[7]を使用し，N=3,4,5 の場合に Kneser-Ney スムージングを適用した．学習データは日本語 Wikipedia のテキストデータのみを用いた.

表 5 文完成タスクの実験結果

	Skip-gram	3-gramLM	4-gramLM	5-gramLM
正答数	160	400	410	425
正答率	0.264	0.661	0.678	0.702

5.4 実験結果

実験結果を，表 5 に示す． 結果の考察については 6.2 節で述べる.

6. 考察

6.1 単語類推タスク

6.1.1 手法間の正答率の差異について

表 3 より, Glove の方が Skip-gram モデルに比べより良い結果となった．特に大きな差となっているのは「名詞–形容詞」,「名詞–副詞」の問題であるが，この傾向は両手法間の単語ベクトルの学習方法の差異から生じたと考える．

「名詞 – 形容詞」,「名詞 – 副詞」の問題では，使用する単語に必ず「漢字一文字の単語」が含まれる．このような例は語義曖昧性の問題から Skip-gram のようなモデルでは上手く学習することが出来ないが， GloVe では単語の共起頻度の比を重み付けに使うことでそのような単語との学習の優先度が下がり，精度向上に繋がったのではないかと考える．このような例の問題点については，6.1.4 節で詳しく述べる.

6.1.2 英語と日本語の正答率の差異について

英語での実験では，評価用データとして Mikolov らが作成したデータセット[3][e]が使用されている．総問題数は 19554 問で， 14 種類の単語の関係が存在していた．ベクトル構築は約 60 億単語が含まれる学習データが使用され, 実験の結果，Skip-gram モデルで最大 69.1%， GloVe で最大 75.0%の正答率が確認できたことが報告されている[5].

学習データのサイズの差を考慮しても正答率に大きな差が存在しているため，Skip-gram モデルや GloVe では，英語に比べて日本語の分散意味表現を学習することが難しいということがいえる．その原因として，単語ベクトルに存在する既存の問題や日本語特有の問題が考えられるが，それらについては 6.1.4 節で詳しく述べる.手法間の正答率の差については英語と日本語の間で同様の傾向を示すことが確認できた.

g.) http://www.jlpt.jp/
h.) センター試験や大学等の入試問題も使用することを検討したが,単語ベクトルの評価問題としては難易度が高すぎたため,使用していない.

ⓒ2015 Information Processing Society of Japan

表 6 単語類推タスクの解答例

	問題例	正答	Skip-gram モデル			GloVe		
			@1	@2	@3	@1	@2	@3
1	日本-円，中国-?	元	億	ドル	万	000	万	800
2	安全-危険，光-?	闇	闇	影	暗闇	影	闇	放つ
3	男-女，王-?	女王	王妃	女王	太子	女王	熟	遊戯
4	カナダ-カナダドル,スイス-?	フラン	スイスフラン	英ポンド	オーストラリアドル	スイスフラン	英ポンド	オーストラリアドル
5	カレー-辛い，塩-?	しょっぱい	硫酸	アンモニウム	ナトリウム	つらい	苦しい	辛く
6	寒-寒く，優-?	優しく	エミリ	愛梨	優希	駿	あびる	悠
7	大きい-小さい，強い-?	弱い	強く	弱く	弱い	強く	弱い	弱く
8	得る-失う，買う-?	売る	買い	買っ	売る	買っ	買お	買わ

6.1.3 単語の関係毎の正答率について

単語の関係毎に結果を比較すると，正答率の差が大きくなり，日本語単語ベクトルで推測しやすい問題やそうでない問題があることが明らかになった．

正答率が高かった関係は「形容詞–副詞」,「国名–首都」,「県名–県庁所在地」,「男–女」などであった．「形容詞–副詞」を除くと，いずれも固有名詞の対の関係となっており，その他の関係を見ても名詞の対の関係は他の関係に比べ高い正答率となった．これは，日本語では名詞の前後に出現する単語に傾向が現れやすく，名詞の単語の分散意味表現の学習が精度よく行えたためと考える．

正答率が低かった関係は「国名–通貨単位」,「食べ物–味」,「動詞の反対語」などであった．「国名–通貨単位」では，通貨単位の前後に出現する単語として数値や数の単位が多いことから，表 6 の 1 のように，数値や単位を表す単語を選ぶ誤答が多かった．「食べ物–味」では，正答とは異なる味覚の表現や別の食べ物の名前を誤答する例が多かったが，他と比べ問題数が少ないため，信頼できる結果でないと考える．「動詞の反対語」では，問題の単語の異なった活用形が推測されることが多かった．これについては 6.1.4 節で詳しく述べる．

6.1.4 実際の解答例とそれに対する考察

今回の実験の結果，日本語単語ベクトルでの単語類推タスクで良く失敗する例をいくつか確認できた．

はじめに，今回用いた評価方法では日本語単語ベクトルは正解の判定が難しい，もしくは正解を不正解としてしまうという例である．表 6 の 2 の問題は，「光」の反対語を求める趣旨であるが「光」の反対語には「闇」と「影」の 2 種類が存在し，そのどちらも解答として問題の趣旨を満たす．しかし今回の実験では「闇」を正答と設定していたため，「影」と解答した GloVe では不正解と判定された．表 6 の 3 の問題も同様で，正答の「女王」に対して Skip-gram モデルの「王妃」という解答も問題の趣旨として正しく，不正解としてよいか難しい．表 6 の 4 の問題では，正答を「フラン」としていたが，どちらの手法も「スイスフラン」と解答した．これも，正答として良い解答例である．

以上のように，同一なモノを表現する方法が多い日本語では，同義語や表記ゆれが評価結果に良くない影響を与えることがわかった．よって，日本語単語ベクトルは今回の実験結果よりも良い性能を持っている可能性が高いと考える．この例は日本語特有の問題であるとかんがえる．また，より厳密に日本語単語ベクトルの評価を行う場合，上記の点を考慮した問題設定，評価方法が必要となると考える．

次に，複数の意味を持つ単語が問題に含まれる場合に，正答率が悪くなるという例である．この問題は，一文字の単語でも多く発生した．表 6 の 5 の問題は「塩」の味の表現を答える趣旨の問題であるが，Skip-gram モデルでは，この単語は調味料としての「塩(しお)」でなく化学の分野で良く現れる「塩(えん)」としての意味も含まれており，化学物質に関係する単語が解答された．また GloVe では，問題に出現した「辛い(からい)」という単語に別の読みである「辛い(つらい)」という単語の意味も含まれており，その単語に似ているネガティブな単語が解答された．

表 6 の 6 の問題は「優」という単語が人名に良く使われることから，どちらの手法でも人名に使われる単語が解答された．

複数の意味を持つ単語や一文字の単語は異なった文脈に多数出現するが，その場合前後に出現する単語が文脈毎に変わってしまうため，その単語が多く出現する文脈の意味に偏って学習されてしまう．このような語義曖昧性の問題は，Skip-gram モデルや GloVe に存在する既存の問題である．日本語には漢字の読み方や漢字一文字による同綴異義語が多く，これらの分散意味表現は Skip-gram モデルや GloVe では上手く学習できないと考える．近年，Skip-gram モデルの拡張として，ある単語の周辺の文脈をコンテキストベクトルで表現して意味の異なる文脈毎に学習するベクトルを使い分ける事により、一つの単語に複数の意味ベクトルを付与する手法[10]が提案されている．この問題の対策として，このような語義曖昧性を改善するような手法を

ⓒ2015 Information Processing Society of Japan

表 7 文完成タスクの解答例

	問題文	正答	選択肢 1	選択肢 2	選択肢 3	Skip-gram	5-gramLM
1	なぜか分からないが,旅行の日に( )天気が悪くなる.	かぎって	とって	めぐって	ばかり	ばかり	かぎって
2	夢を( )させるには努力が必要だ.	実現	現実	現象	表現	現象	実現
3	ダイエット中なので,甘いものは( )と思うのですが,なかなか実行できません.	食べまい	食べたい	食べない	食べよう	食べない	食べない

1. http://jlpt.u-biq.org/2g1.html (12)より引用 (c)2009 U-biq All rights reserved.

2. http://news.lnzsks.com/htm/IMS_20100507_14252.htm より引用 Copyright (c) 1998 - 2014 LNZKS.com, All Rights Reserved3. http://www.nihongo-pro.com/quiz/0d95a7958b/essential-jlpt-n2-grammar (3)より引用 (c)2010-2014 Horizon Web Services LLC. All rights reserved.

用いることが考えられる.

最後に，活用形が存在する単語が問題に存在する場合，正答率が悪くなるという例である．表 6 の 7 の問題では，どちらの手法でも問題として使用されている単語「強い」の活用形である「強く」が解答された．表 5 の 8 の問題でも同様に「買う」という単語の活用形で「買い」「買っ」「買お」などが解答された.

このように，形容詞や副詞，動詞といった活用形によって形が変わる単語は精度が悪く，特に上記の様に，形容詞の反対語を同じ意味の副詞に誤答する例や動詞の反対語を同じ意味の動詞の別の活用形に誤答する例が多く見られた．今回の結果から，日本語に Skip-gram モデルや GloVe を適用する場合，活用形間関係を考慮するための対策を考える必要がある.

6.2 文完成タスク

6.2.1 比較手法との正答率差について

表 4 より，比較手法である 5-gram 言語モデルでは 70%を超える正答率となっているが，Skip-gram モデルでは 26.4%と低い結果となった．特に，比較手法では学習データとして Skip-gram モデルの 1/4 ほどのデータしか使用していないため，実際の正答率以上に性能に差がある可能性が高いと考える.

6.2.2 英語と日本語の正答率の差異について

英語での実験では，評価用データとして Microsoft Research Sentence Completion Challenge[7]が使用されている．このデータはシャーロック・ホームズの小説に登場する文章から作成されており，選択肢の単語は，同じ出現頻度で出現する複数の単語を選出し，そこから人手で問題に割り当てることで作成されている．総問題数は 1040 問で， 選択肢は 5 つとなっている．実験の結果，4-gram 言語モデルで 39%の正答率, Skip-gram モデルで 48.0%の正答率が報告されている[3].

日本語での実験結果と比較すると，正答率は日本語での 5-gram 言語モデルが最も高くなっている．しかし実際は，使用された題材や選択肢の数から評価用データの難易度が異なるため，単純に結果を比較することはできない．英語データは問題文が古い小説から抽出された文章で選択肢が 5 つなのに対して，日本語データは外国人向けの試験問題の文章で選択肢は 4 つであるため，難易度は英語の方が高い．したがって，本稿の実験に比べて英語データの結果が悪くなることは不自然ではないと考える．評価をより英語のものへ近づけるには，青空文庫で公開されている小説から同様の手順で問題を作成する方法が考えられる.

また，英語の実験では，今回比較手法として設定した n-gram 言語モデルの正答率を上回っている．そのため，英語の場合 n-gram 言語モデルよりも英単語ベクトルは文完成タスクを解く性能が高いといえ、日本語単語ベクトルに比べ文完成のタスクを解く性能に優れているということもいえる．すなわち，日本語単語ベクトルで正答率が悪くなってしまった原因は，日本語特有の何らかの問題にあると考える．この原因の予想については 6.2.3 節で述べる.

6.2.3 実際の解答例とそれに対する考察

5-gram 言語モデルの場合，表 7 の 3 のような特定の文脈に使われる言い回しなどは正答率が低い傾向にあった．しかし, Skip-gram モデルの解答について正解例と不正解例を比較したが，そこに明確な差は見られなかった．また，Skip-gram モデルにはすべての選択肢で出現確率の総和が似ているという特徴があった．したがって，特定の形式の問題の正答率が悪いということではなく，日本語単語ベクトルそのものにこのタスクを行う上での問題点が存在することになる.

原因として，単語の分割単位が細かすぎるということが考えられる．3.3 節でも述べたとおり，日本語単語ベクトルを構築するためには文章の分かち書きを行い，生成された形態素を単語として扱う必要がある．例えば，表 7 の 1

の問題を分かち書きした結果は以下の様になる．

夢，を,(　),さ，せる，に，は，努力，が，必要，だ，.

この時,「を」,「さ」,「に」,「は」など接続詞や助詞などの機能語などが細かく分割されていることがわかる．これらの語は日本語のあらゆる文脈に出現するため，Skip-gram モデルで学習を行うと多くの単語で出現確率が同程度になることが予想できる．そのため，選択肢間で出現確率の総和に差が見られず，正答率が低くなったと考える．

単語の分割単位が細かすぎることによるもうひとつの影響として，考慮したい表現が分割されてしまうという問題も挙げられる．

表 7 の 2 の問題の選択肢の単語，例えば「かぎって」という単語は,分かち書きの処理がされると「かぎっ」と「て」という２単語に分割されてしまう．この場合，以下のような文章の生起確率を求めることを行う．

なぜ，か，分から，ない，が,、，旅行，の，日，に，
かぎっ，て，天気，が，悪く，なる,。

生起確率を求めるために「かぎっ」や「て」という単語の前後に「天気」や「日」,「が」,「に」などの単語が出現する確率，また，これらの単語の前後に「かぎっ」や「て」という単語が出現する確率を計算することになる．このような場合,「かぎって」という表現がこのような文脈での出現する確率を考えることができず，問題を正しく評価することができなくなってしまう．

以上の理由により，現状の日本語単語ベクトルでは文完成のタスクを解くことは難しいと考える．対策として，前処理の段階で単語の分割単位を調整し，より大きなまとまりでの表現をベクトルとして学習することが考えられる．

7. まとめ

本研究では，Skip-gram モデル及び GloVe を用いて日本語の単語ベクトルを構築し，単語類推タスク，文完成タスクについて評価を行った．単語類推タスクでは，同義語や表記ゆれなど日本語に多い存在する問題のため英語と同様の評価を行うことが難しいことが明らかになった．また，単語ベクトルに存在する語義曖昧性の問題や日本語特有の単語の活用形間の学習に関する問題の影響により，英単語ベクトルに比べて正答率が低くなることがわかった．しかし，手法間の正答率に英単語ベクトルと同様の傾向を確認することができた．文完成タスクでは，日本語の単語の分割単位が細かくなると周辺単語の出現確率が上手く行えなくなることがわかった．また，考慮したい表現が複数語に分割されてしまうことで，問題文を正しく捉えられなくなるという考察が得られた．以上の理由から，日本語単語ベクトルでの正答率が悪くなるという結果となった．

今後の課題として，本稿で明らかになった問題を改善していくことが挙げられる.単語類推の評価方法の問題では,評価を正解か不正解かだけでなく正解に似ている単語を定義しておき，それを解答した場合にも多少の加点を行うような形式にするなどが考えられる．これにより厳密な評価が行えるのではないかと考える．語義曖昧性の問題は，一つの単語に複数の意味ベクトルを与える手法[10]を適用することで多少改善できると考えられる．単語の分割単位の問題は，word2vec と共に公開されている word2phrase[i]というツールを活用する方法や，分割された単語の品詞に着目して一定のルールで単語を再結合する方法[j]が考えられる．

参考文献

[1] S. Deerwester, S. Dumais, G. W. Furnas, T. K.Landauer, R. Harshman.: Indexing by Latent Semantic Analysis. Journal of the American Society for Information Science 41 (6), pp.391–407. (1990).

[2] Y. Bengio, R. Ducharme, P. Vincent. :A neural probabilistic language model. Journal of Machine Learning Research, 3:1137-1155. (2003).

[3] T. Mikolov, K. Chen, G. Corrado, and J. Dean.: Efficient Estimation of Word Representations in Vector Space. In ICLRWorkshop. (2013).

[4] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean.: Distributed representations of words and phrases and their compositionality. In NIPS, pp.3111–3119. (2013).

[5] J. Pennington, R. Socher, and C. D. Manning.:glove: global vectors for word representation. In EMNLP2014. pp. 1532–1543. (2014).

[6] Turian, Joseph, L. Ratinov, Y. Bengio.:Word representations: a simple and general method for semi-supervised learning. Proceedings of the 48th annual meeting of the association for computational linguistics. (2010.)

[7] G. Zweig, C.J.C. Burges. The Microsoft Research Sentence Completion Challenge, Microsoft Research Technical Report MSR-TR-2011-129. (2011).

[8] C.Buck, K. Heafield, B.V. Ooyen.: N-gram Counts and Language Models from the Common Crawl. Proceedings of the Language Resources and Evaluation Conference. (2014).

[9] A. Stolcke.:SRILM -- An Extensible Language Modeling Toolkit. Proc. Intl. Conf. on Spoken Language Processing, vol. 2, pp. 901-904. (2002).

[10] A. Neelakantan, J. Shankar, A. Passos,A. McCallum.: Efficient nonparametricestimation of multiple embeddings per word in vector space. EMNLP2014. (2014).

i.) https://code.google.com/p/word2vec/
j.) http://www.slideshare.net/piroyoung/word2vec-40436293 pp28-30

ⓒ2015 Information Processing Society of Japan