概念階層を用いた単語分散表現の多義性評価用データセットの提案

Dataset Based on Concept Hierarchy for Evaluating Word Vectors to Represent Multisense Words

山崎　禎晃[1*]　豊田　哲也[2]　大原　剛三[2]
Tomoaki Yamazaki[1]　Tetsuya Toyota[2]　Kouzou Ohara[2]

[1] 青山学院大学大学院　理工学研究科
[1] Graduate School of Science and Engineering, Aoyama Gakuin University
[2] 青山学院大学　理工学部
[2] College of Science and Engineering, Aoyama Gakuin University

Abstract: Recently, techniques to embed words into a multidimensional space and learn multidimensional vectors that represent them are often used in various tasks in Natural Language Processing. Usually, learning models such as word2vec assign a single vector to each word, but, some of advanced models including sense2vec can give multiple vectors to a single word if it is a multisense word. Although some datasets are publicly available for evaluating word vectors learned, it is difficult to properly evaluate them with the datasets if multisense words appear in the corpus used to learn those vectors. Thus, in this paper, we propose a novel dataset that allows us to evaluate word vectors for multisense words by considering a concept hierarchy in WordNet and BabelNet. We empirically show, with the proposed dataset, we can evaluate word vectors for multisense words more properly than an existing dataset.

1　はじめに

近年，自然言語処理のタスクにおいて，単語を多次元ベクトルで表現する単語分散表現の学習・利用が盛んになっている．この単語分散表現の代表的な学習モデルとしては，Mikolov らの word2vec [1]，Grave らの fastText [2] などが挙げられる．これらのモデルによる得られる単語分散表現は，1 つの単語に対して 1 つのベクトルが対応する 1 対 1 の関係になっており，単語の表現方法として有用ではあるものの，単語がもつ多様な意味を十分に扱えないことが指摘されている．そのような単語分散表現の拡張の 1 つとして，多義性を考慮し，単語が持つ複数の意味に対して異なる分散表現（ベクトル）を学習する手法が提案されている [3, 4, 5]．そのように学習した単語分散表現の評価は，外部的なものと内部的なものに大別できる [6]．外部的な評価では，特定の応用タスクの下，学習した個々の分散表現を分類モデル等の入力として利用し，そのモデルの精度によって単語分散表現の良し悪しを評価する．これに対し内部的な評価では，分散表現間の類似度や分散表現を用いた類推の結果を評価する．単語分散表現自体の評価としては内部的な評価が望ましいが，従来の評価方法 [7, 8] は基本的に単語と分散表現が 1 対 1 の関係となることを想定しており，その関係が 1 対多となり得る多義語を考慮した単語分散表現の評価には適していない．

一方，WordNet [9] に代表される概念階層を持つ知識データベースにおいては，各単語は synset ごとに細かく分類され，各単語の意味は品詞よりもさらに細かい単位で管理されている．ここで，synset とは，単語を synonym（同義語）ごとにグループ化したものであり，1 つの単語は複数の synset に属する場合があり，直感的には 1 つの synset が 1 つの概念に対応する．WordNet においては，synset を一定のルールに基づき人手で分類しており，各 synset は上位・下位関係によって階層構造を形成している．しかし，手動で概念階層を構築することはコストが大きく，保守や更新が難しいことから，Wikipedia などの情報資源を用いて自動的に概念階層を構築する研究が行われている [10]．特に Wikipedia は，語彙網羅性，即時更新性に優れているだけでなく，外部リンクや所属カテゴリなどの情報を持つため，概

*連絡先：青山学院大学大学院理工学研究科
〒 252-5258　神奈川県相模原市中央区淵野辺 5-10-1
E-mail: c5618168@aoyama.jp

念階層とのギャップが小さいという特徴を有する．実際に，そのような Wikipedia を用いて WordNet を拡張した概念階層としては BabelNet [11] があり，その中では Wikipedia 中に出現する単語の意味が網羅されている．また，最近では Wikipedia は学習用コーパスとしても注目されており，単語分散表現の学習にも多く用いられている．

以上のような背景から，本研究では，Wikipedia を学習用コーパスとした単語分散表現を適切に評価するために，概念階層である WordNet と BabelNet に基づいた評価用データセット，およびそのデータセットにおける単語分散表現の評価方法を提案する．提案するデータセットでは，1 つの単語が synset ごとに別の意味をもつことに着目し，synset をその構成単位とすることで単語がもつ多義性を考慮した分散表現の評価を可能にする．また，本稿では，提案するデータセットの妥当性について実験的に検証する．

2 関連研究

2.1 単語分散表現

単語の意味を統計的に学習する場合,「語の意味は周辺の単語すなわち文脈によって決定される」という分布仮説 [12] を定式化した問題を解くことが多い．実際，単語分散表現においても単語の意味を文脈から学習し，その文脈に応じた多次元ベクトルを付与することで，各単語をベクトル空間上に写像する．本節では，単語分散表現の代表的な学習モデルである word2vec，および単語の多義性を考慮したその拡張手法について述べる．

2.1.1 word2vec

word2vec [1] は Mikolov らによって提案された単語分散表現の学習モデルであり，得られたベクトルの加減算による類推や自然言語処理の応用タスクにおける有用性から注目を浴びている．word2vec には学習手法がいくつか存在し [13]，ここでは，そのうち高速かつ高精度な Skip-Gram Negative Sampling（SGNS）[14] を説明する．学習コーパスを単語列 $w_1, w_2, \cdots, w_T$，文中のある位置 t における単語 w_t に対して，その前後 ws 個の単語列を w_t の文脈 $\boldsymbol{C}_{w_t} = (w_{t-ws}, \cdots, w_{t-1}, w_{t+1}, \cdots, w_{t+ws})$ とする．このとき，通常の Skip-Gram では，$p(w_t|\boldsymbol{C}_{w_t})$ を各文脈語 $c \in \boldsymbol{C}_{w_t}$ から単語 w_t を予測する条件付き確率 $p(w_t|c)$ の積に分解し，式 (1) で定義されるその対数尤度 L を最大化するような単語分散表現を学習する．

$$L \equiv \sum_{t=1}^{T} \sum_{c \in \boldsymbol{C}_{w_t}} \log p(w_t|c) \tag{1}$$

表 1: WordSim353 の例

word1	word2	単語間類似度 [0,10]
cup	coffee	6.58
cup	article	2.40
tiger	tiger	10.00

これに対し SGNS では，計算量軽減のために式 (1) の近似となる式 (2) で定義される L' を目的関数として利用する．

$$L' \equiv \sum_{t=1}^{T} \sum_{c \in \boldsymbol{C}_{w_t}} (\log\sigma(\boldsymbol{v}_c \cdot \tilde{\boldsymbol{v}}_{w_t}) + \sum_{k=1}^{K} \log\sigma(\boldsymbol{v}_c \cdot \tilde{\boldsymbol{v}}_{\check{w}'})) \tag{2}$$

ここで，σ はシグモイド関数，$\tilde{\boldsymbol{v}}_w$ はサンプリング時に使用される各単語 w がもつベクトル，$\check{w}'$ はサンプリングした負例単語，K は負例単語数である．

2.1.2 多義性を考慮した単語分散表現の学習

word2vec では 1 つの単語に 1 つのベクトルしか付与できないため，多義語に関しては，その単語がもつ複数の意味が 1 つのベクトルに集約されてしまい，その多義性を十分に表現できないという問題がある．そのような多義性を持つ単語への対処法としては，1 つの単語が持つ意味ごとに異なるベクトルを割り当てることが考えられる．その 1 つの実現方法として，事前に品詞タグ付けをしたコーパスに word2vec を適用し，品詞ごとに異なる意味ベクトルを学習する sense2vec [3] が提案されている．しかし，この手法では同一の品詞において複数の意味をもつ単語には対応できない．そのため，学習時に単語の意味を自動推定しながら意味ベクトルを学習する手法が研究されている．Neelakantan らの MSSG [4] では，Skip-gram の学習時に単語の意味を推定し，対応した意味ベクトルを適応的に選択することで，各単語の意味ベクトルを学習している．また，WordNet などの概念階層を活用した手法も存在する．Pilehvar らの DeConf [5] は，WordNet の階層構造から各単語に対する synset における代表単語を選出し，選出した単語と事前に学習した単語分散表現から各 synset に対応する分散表現を求めることで，多義性を考慮した単語分散表現を獲得している．

2.2 単語間類似度データセット

単語分散表現の性能は，単語間類似度やベクトル演算による類推に関するデータセットによって評価される．代表的なデータセットとしては WordSimilarity-353（以下，WordSim353） [7] があり，表 1 に例示するよ

表 2: SimLex-999 の品詞分布

名詞	動詞	形容詞–副詞
666	222	111

うな単語のペアとその類似度の組み合わせ 353 個（重複含む）によって構成されている．単語分散表現の評価では，ペアとなっている単語に対する分散表現（ベクトル）間の類似度とデータセットにおける単語間類似度それぞれによって並び替えた単語列に対しての順位相関が用いられる．一方，WordSim353 では類似性と関連性を区別していないという問題が指摘されている [15]．たとえば，WordSim353 では cup と coffee のペアは 6.58 という比較的高い類似度をもつが，実際には cup と coffee は関連性はあるものの類似性はない．類似性という観点では，cup と glass に対する類似度が高くなり，cup と coffee に対する類似度は低くなることが望ましい．

これに対し，類似性をより厳密に評価するためのデータセットとして，Hill らは SimLex-999 [8] を提案している．SimLex-999 は WordSim353 と同様に単語のペアとその類似度の組み合わせを構成単位とし，その数は 999 個となっている．しかし，WordSim353 とは異なり，単語間類似度を決める被験者アンケート時に類似性について掲示することで関連性との差別化を図っている．SimLex-999 の品詞分布は表 2 に示す通りであり，単語間類似度と単語間の上位・下位関係には相関があることが示唆されている．

一方，これらの単語間類似度データセットを用いて単語分散表現を評価する際は，ペアとなる単語それぞれがもつ分散表現ベクトルの類似度を求める必要がある．word2vec のように 1 つの単語 w が 1 つのベクトル $\boldsymbol{v}_w$ のみをもつ場合には，ベクトルの内積あるいはコサイン類似度を計算すればよい．しかし，多義性を考慮した単語分散表現では，単語 w が持つ各意味 s に対応する意味ベクトル $\boldsymbol{v}_{w_s}$ が存在する．このような意味ベクトルを考慮した評価方法としては，単語がもつ意味ベクトルの平均をグローバルベクトル $\boldsymbol{v}_{w_g}$ とし，グローバルベクトルによって類似度を求める globalSim，式 (3) により与えられる各意味ベクトル間の類似度の平均を用いた avgSim，式 (4) により与えられる各意味ベクトル間の類似度の最大値を用いる maxSim などがある．

$$avgSim(w, w') \equiv \frac{1}{S_w \cdot S_{w'}} \sum_{i=1}^{S_w} \sum_{j=1}^{S_{w'}} cos(\boldsymbol{v}_{w_i}, \boldsymbol{v}_{w'_j}) \quad (3)$$

$$maxSim(w, w') \equiv \max_{1 \le i \le S_w, 1 \le j \le S_{w'}} cos(\boldsymbol{v}_{w_i}, \boldsymbol{v}_{w'_j}) \quad (4)$$

ここで，S_w，$S_{w'}$ はそれぞれ単語 w，w' が持つ意味数を表す．

2.3 概念階層

2.3.1 WordNet

本研究で用いる WordNet [9] は，Princeton 大学で作成された，自然言語処理分野において最も有名な知識データベースである．WordNet 2.1 には 155,327 個の単語と 117,597 個の synset（概念）が登録されており，有用な言語資源として知られている（現在は WordNet 3.1）．WordNet における synset は同義語の集合であり，たとえば，舞台（dramatic work）を上位語に持つ play を含む synset は次のようになる．

$$\{play_n^1, drama_n^1, dramatic\ play_n^1\}$$

単語がもつ synset が混同されるのを防ぐため，synset の各要素となる単語には synset ID が振られる．たとえば，上記の $play_n^1$ の “1” は，この synset が単語 play に対する 1 番目の synset であることを意味する．また，下付き文字は品詞を表し，この場合の “n” は名詞（Noun）を意味する．品詞は，主に名詞，動詞，形容詞，副詞に分類される．また，synset 間には，instance-of，is-a，part-of という 3 つの意味的な関係を定義することができる．instance-of 関係は，synset とその例の関係を表し，例えば劇作家（dramatist）に対する Shakespeare の関係がそれにあたる．is-a 関係，part-of 関係はそれぞれ上位語，下位語の関係を表し，$play_n^1$ はその上位語 $dramatic\ work_n^1$ から見て is-a 関係にあり，その下位語 $playlet_n^1$ から見て part-of 関係にある．上位 synset との関係は 1 対 1 であるが，下位 synset との関係は 1 対多であることが多い．このように，WordNet 中の synset は上位・下位関係に基づく階層構造をもっている．また，前述の単語 play は実際には多義語であり，17 個の名詞の synset，35 個の動詞の synset，計 52 個の synset（意味）を持っている．synset 数は単語によって異なり，単一の synset のみをもつ単語も多数存在する．

2.3.2 BabelNet

本研究で用いるもう 1 つの概念階層である BabelNet [11] は，英語の概念階層をもつ WordNet を語彙網羅性や即時更新性といった優れた性質をもつ多言語百科事典である Wikipedia によって拡張した多言語対応可能な概念階層である．WordNet と Wikipedia を結合して作られた BabelNet は，非常に多くの言語への対応と語彙網羅性の面で優れた言語資源となっており，国際ワークショップ SemEval の語義曖昧性解消タスク [16] でも用いられている．WordNet と Wikipedia の結合では，まず，両者に含まれる同様の概念が重複しないようにそれらを統合する．次に，Wikipedia の翻訳ページや機械翻訳によって多言語の語彙情報を収集する．最

後に，収集した語彙に対する synset 間に関係を与えることで WordNet と Wikipedia の結合は完了する．BabelNet は 271 言語へ対応していることも利点の 1 つであるが，Wikipedia の情報を加えたことで，固有名詞や直近で増えた意味にも対応できるという利点がある．たとえば，英語における salad という単語は WordNet 上では食べるサラダという意味しかないが，BabelNet では Salad というバンド名としての意味も含まれている．なお，BabelNet には Wikipedia に登場しない概念は含まれない．

3 提案データセット

3.1 提案データセットの概要

本研究では，Wikipedia に含まれる概念のみからなる評価用データセットを構築し，多義性を考慮した単語分散表現を適切に評価することを目的としている．Wikipedia 上の意味を網羅し，概念階層に基づいた信頼性の高い分散表現間類似度の評価を可能にするため，データセットの構築には概念階層をもつ WordNet と BabelNet を利用する．

提案データセットは，各単語がもつ synset（意味）ごとに，単語，品詞，同義語群（synset），上位語群 1，上位語群 2，上位語群 3 から構成される．ここで，上位語群 n は，対象単語を含む synset に対して概念階層を n ステップだけ上位に遡った先の synset に含まれる単語群を指す．表 3 にその一部を例示する．表 3 に示すように，データセットの各行が単語の 1 つの synset に対応しており，たとえばこの表の 1 行目からは，単語 heactare の品詞は Noun（名詞）であり，同義語に ha や hm2 が含まれ，概念階層を 1 つ上位に移動した synset には単語 metric が含まれることがわかる．また，単語 heactare に対する行が 1 行しかないことから，この単語のもつ意味は 1 つだけとなる．これに対し，accomplish，month，announce は複数の行から構成されることから，その行の数だけの意味をもつ多義語であることがわかる．また，BabelNet の情報を追加したことで，WordNet に多く見られる time_unit のような複合語だけでなく，12ヶ月のすべての頭文字を取った JFMAMJJASOND といった固有名詞も month の同義語に含まれている．なお，JFMAMJJASOND は本来大文字表記であるが，提案データベースではすべての単語を小文字化して扱っている．

表 3: 提案データセットの例

単語	品詞	同義語群	上位語群 1	···
hectare	Noun	ha, hm2, ...	metric, ...	···
halve	Verb	bisect	divide, ...	···
accomplish	Verb	action, ...	effect, ...	···
accomplish	Verb	achieve, ...	win, ...	···
month	Noun	jfmamjjasond, ...	period, ...	···
month	Noun	-	time_unit, ...	···
announce	Verb	proclaim, ...	inform, ...	···
announce	Verb	declare	state, ...	···
announce	Verb	-	name, ...	···
announce	Verb	herald, ...	tell	···

表 4: 提案データセットにおける品詞分布

意味数	Noun	Verb	Noun と Verb
1	416	84	-
2 以上	317	63	120

3.2 データセット構築手順

提案データセットには，2018 年 11 月 20 日付の英語版 Wikipedia, WordNet 3.0, BableNet 4.0.1 を言語資源として用いた．形容詞，および副詞は上位 synset を安定して取得できないため除外し，名詞（Noun）のみの synset をもつ単語，動詞（Verb）のみの synset をもつ単語，および名詞と動詞の両方の synset をもつ単語のみを対象とした．最終的な提案データセットにおける品詞分布を表 4 に示す．

データセットの構築は，対象単語の synset に対する上位語の探索および不要な上位・下位関係にある synset の組合せの排除の 2 段階で構成される．これらの作業は，Wikipedia 上に存在する意味のみを取得し，構築するデータセットにおいて同一の単語においても synset ごとに意味を区別することを目的としている．

1. **上位語の探索**
   WordNet と BabelNet の概念階層を用いて，各単語の synset ごとに上位語を探索する．この時，上位 synset は 3 階層まで遡り，上位 synset がない場合はその階層で探索を終了する．また，上位関係を探索して得られたすべての上位語は小文字化する．登録する単語は SimLex-999 に含まれる単語を優先して選出し，残りは Wikipedia に含まれる単語からランダムに選出した．また，複数の synset をもつ単語（多義語）の Wikipedia での出現頻度は，1 つの synset のみをもつ単語の出現頻度と比べて必然的に高くなるため，そのような学習コーパス中の出現頻度の偏りを軽減するため，Wikipedia における単語の出現頻度に基づいて単語の選択範囲を 3 つに分けた．具体的には，WordNet において synset を 1 つもつ単語と複数もつ単語それぞれの Wikipedia における出現頻度の中央値が 132, 2011 であったため，それらを閾値とし，最低出現頻度を 10 として，Wikipedia での出現頻度の範囲が 10〜131，132〜2010，2011〜の 3 つのグループに単語を分け，

各グループ内で単一の synset をもつ単語と複数の synset をもつ単語の数が同程度になるように単語を選択した．ただし，出現頻度が 2011 回以上ある動詞で 1 つの synset しかもたない単語は十分な数を取得できなかったため，必要数の半分以上を 132〜2011 の範囲で代替した．

2. **不要な上位関係の排除**
概念階層では，$evaluate_v^1$ の上位語に意味の異なる $evaluate_v^2$ が含まれるような場合がある．そのように同一の単語が異なる意味で異なる階層に現れる場合，その単語に対する意味ベクトルの評価が正しくできない可能性があるため，WordNet 内でこのような上位・下位関係が見られる場合はデータセットから除外し，BabelNet 内で見られた場合は WordNet における関係を優先して残した．また，$play_n^8$ と $play_n^{14}$ の上位 synset がともに $diversion_n^1$ を含むような場合もある．このように単語 w の異なる synset が別の単語 u の特定の意味を上位 synset 中にて共に含む場合も，w に対するそれらの意味（synset）に対応する意味ベクトルを適切に評価できない可能性があるため，そのような単語 w はデータセットに含めないようにした．また，Wikipedia 上に出現しない概念がデータセットに含まれることを避けるため，BabelNet 上に含まれない synset をもつ単語もデータセットには含めないものとした．

3.3 構築したデータセットの統計情報

前述のように単語の選択範囲を 3 分割したことで，synset 数の違いによる出現頻度の偏りを抑えることができ，構築したデータセットに含まれる synset 数が 1 つの単語と複数の単語それぞれの Wikipedia での出現頻度の中央値は 327，572 となった．最終的に構築したデータセットにおける単語の選択範囲に対する synset（意味）数の平均と単語数，Wikipedia での出現頻度を表 5 に示す．この表からは，単語の出現頻度の増加に伴い，意味数が増加する傾向にあることがわかる．また，品詞ごとの分布を表 6 に示す．この表からわかるように，名詞・動詞の synset の両方を持つ単語はその他の単語よりも synset 数が多くなった．

3.4 提案データセットの評価方法

多義性を考慮した分散表現では，異なる複数の意味を正しく学習できていることを示すために，多義語が持つ各意味ベクトルに対する近傍の単語を呈示し多義語を学習した証拠とすることがしばしばある．しかし，いくつかの単語に対してのみ例を示すことでは単語分散表現全体を評価することはできないため，網羅的かつ定量的な評価が必要とされる．そこで，ここでは，提案データセットにおける同義語群と上位語群の和集合である関連単語群 T_w と単語 w が持つ意味 s に対する意味ベクトルから求まる N 近傍の単語集合 N_{w_s} の適合率（Precision）を用いた評価方法を提案する．これは，1 つの単語に対し複数の意味ベクトルを付与した単語分散表現を学習する場合，正しく学習ができていれば，各意味ベクトルに対応する synset は異なる上位語群をもち，学習結果の意味ベクトルの近傍に含まれる単語もそれに応じて異なると考えられるためである．具体的には，データセット中の各単語に対する評価値は式 (5) で求め，分散表現全体の評価は式 (6) に示すようにデータセット中のすべての単語 W についての評価値の平均とする．

$$score_w = \frac{\max_{1\leq s\leq S_w, 1\leq d\leq S_d} Precision@N(N_{w_s}, T_w)}{\max(S_w, S_{d_w})} \tag{5}$$

$$score = \frac{1}{|W|}\sum_{w\in W} score_w \tag{6}$$

ここで，$Precision@N(N_{w_s}, T_w)$ は $|N_{w_s} \cap T_w|/N$ で与えられる適合率であり，S_{d_w} は対象単語が提案データセット中にもつ synset 数，S_w は単語 w がもつ意味ベクトル数である．また，今回の実験では $N = 1$，5，10，100 の近傍単語を対象とした．

式 (5) の分子で適合率の最大値を取るのは，単語が複数の意味ベクトルをもつ場合，単語がもつ同義語群・上位語群に対し適合率の高い意味ベクトルがその synset の意味として妥当であると考えられるためである．また，多義性を考慮した単語分散表現では，学習コーパスに存在する意味数より多くの意味ベクトルを学習してしまうことがある．これは，同じ意味に対して異なる

表 5: 提案データセットにおける単語選択範囲に対する単語数分布

範囲	平均意味数	単語数	出現頻度
all	1.92	1,000	431
10-131	1.65	332	46.5
132-2010	1.81	351	465.0
2011-	2.32	317	4647.0

表 6: 提案データセットにおける品詞ごとの単語分布数

品詞	平均意味数	単語数	出現頻度
Noun	1.65	733	434.0
Verb	1.78	147	380.0
Noun, Verb	3.71	120	489.0

ベクトルを学習する場合や，学習がうまくできずに不要なベクトルを生成する場合があるためである．そのような冗長な意味ベクトルを学習している場合は，単語分散表現の評価を下げる必要があるため，式 (5) の分母では S_{d_w} と S_w のうち大きいほうの値で適合率を割っている．

4 提案データセットによる単語分散表現の評価

4.1 単語分散表現の学習と取得

本研究では，word2vec と sense2vec を用いて Wikipedia から新たに学習した単語分散表現，その他の手法により学習済みの単語分散表現を提案データベースと SimLex-999 を用いて評価した．学習コーパスとする Wikipedia は表記ゆれを修正しなかったもの，nltk [17] の WordNetLemmatizer によって表記ゆれを修正したうえで複合語の表記を提案データセット中の表記とあわせたものの 2 つを用意した．以下，それぞれ wiki_{nl}, $\text{wiki}_{\text{multi}}$ とする．また，sense2vec で利用する品詞情報は，Universal Part-of-Speech Tagset に基づいた nltk により付与した．word2vec と sense2vec の学習には gensim [18] の 3.6.0 を使用し，次元数を 300，window size を 5，min count を 10，negative sampling 数を 5 として単語分散表現を学習した．学習済みの単語分散表現としては，Wikipedia を学習コーパスとして MSSG [1]により学習したもの，Google News を学習コーパスとして word2vec [2]で学習したもの，およびその word2vec の学習結果を DeConf [3]で拡張したものを用いた．

4.2 提案データセットによる評価結果

各分散表現を提案データセットにより評価した結果を表 7 に示す．提案データセットの適合率の計算において，単語 w の関連語群 T_{w_s} と一致する意味ベクトルの近傍単語は最近傍である場合が多く，近傍単語数 N を 100 まで増やしても適合率の上昇は見られなかった．そのため，N が小さい値ほど高い評価結果になる傾向を確認した．また，Wikipedia を学習コーパスに用いた単語分散表現では，複合語を考慮した $\text{wiki}_{\text{multi}}$ を word2vec で学習した場合が最も高評価となった．この結果は，学習用コーパス中の複合語に関しては適切に学習できるような前処理が重要であり，そうでない場合は適切な意味ベクトルの学習が困難であることを示唆するものである．

[1]http://iesl.cs.umass.edu/downloads/vectors/release.tar.gz
[2]https://code.google.com/archive/p/word2vec/
[3]https://pilehvar.github.io/deconf/

表 7: 適合率を用いたデータセットの評価結果

単語分散表現	適合率@N	単語数
word2vec(wiki_{nl})	0.182@1	1000
word2vec(wiki_{nl})	0.085@5	1000
word2vec(wiki_{nl})	0.058@10	1000
word2vec(wiki_{nl})	0.011@100	1000
sense2vecwiki_{nl})	0.109@1	1000
sense2vec(wiki_{nl})	0.053@5	1000
sense2vec(wiki_{nl})*	0.146@1	1000
sense2vec(wiki_{nl})*	0.074@5	1000
word2vec($\text{wiki}_{\text{multi}}$)	**0.200@1**	988
word2vec($\text{wiki}_{\text{multi}}$)	**0.106@5**	988
sense2vec($\text{wiki}_{\text{multi}}$)	0.114@1	1000
sense2vec($\text{wiki}_{\text{multi}}$)	0.061@5	1000
MSSG(wiki)	0.160@1	501
MSSG(wiki)	0.083@5	501
word2vec(google)	0.191@1	748
word2vec(google)	0.102@5	748
DeConf(google)	**0.381@1**	882
DeConf(google)	**0.234@5**	882

一方，sense2vec については多くの単語において品詞タグ付けに誤りが見られた．そのため，提案データセットにおける品詞と対応する意味ベクトルのみを評価対象とした結果を sense2vec*として表 7 に示している．提案データセットにおける品詞情報を考慮しなかった場合の評価結果よりも評価値が上がっていることから，sense2vec は正しい品詞に対応したベクトルの学習にはおおよそ成功しているものの，誤った品詞タグ付けにより余分な意味ベクトルを学習し，その結果，評価が下がっていると考えられる．また，Google News を学習コーパスとした単語分散表現では，DeConf が word2vec の評価を上回った．DeConf では，WordNet を参考に各意味ベクトルを学習しているため，提案データセットにおける意味（synset）の分類に近い意味ベクトルの学習に成功し，その結果，word2vec の評価を上回ったと考えられる．

4.3 SimLex-999 との比較

次に，提案データセットにおける多義語の評価の妥当性を検証するために，SimLex-999 と提案データセットに共通する 50 単語について，それぞれにおける評価を比較した．SimLex-999 の評価では，それら 50 個の単語の一方が含まれる 77 個の単語ペアを使用した．共通する 50 単語のうち 47 単語が複数の synset をもつ多義語であった．

表 8: SimLex-999 と提案データセットの評価結果

単語分散表現	SimLex-999		提案データセット	
	avgSim	maxSim	適合率@1	適合率@5
word2vec(wiki_{nl})	**0.379**	**0.375**	**0.131**	0.098
word2vec(multi)	0.366	0.366	0.111	0.079
sense2vec(wiki_{nl})	0.066	0.020	0.093	0.076
sense2vec(multi)	0.102	0.047	0.070	0.062
MSSG(wiki)	-0.023	0.002	**0.131**	**0.103**
word2vec(google)	**0.490**	**0.490**	**0.173**	0.084
DeConf(google)	0.101	0.100	0.162	**0.149**

SimLex-999 と提案データセットに対する評価結果を表 8 に示す．SimLex-999 による評価においては，avgSim，maxSim のいずれを用いた場合でも word2vec の学習結果が最高の評価となり，提案データセットでは適合率の計算で $N = 1$ の最近傍の単語だけを用いた場合は word2vec と MSSG，$N = 5$ の場合は MSSG と DeConf が最高の評価となった．また，SimLex-999 において評価がほぼ 0 となる単語分散表現であっても，提案データセットでは高い評価値となっていることがわかる．

これらの結果について考察するため，2 つの多義語 accomplish と announce に関して，その意味ベクトルの 5 近傍単語をそれぞれ表 9，10 に示し，提案データセットにおける synset ごとの関連語の一部を表 11 に示す．なお，これらの単語に関しては複合語を考慮した $\text{wiki}_{\text{multi}}$ を学習コーパスに用いても評価値は上がらなかったため，ここではその結果は含めていない．また，sense2vec では announce が正しく学習できていなかったため，その結果も割愛している．表 9 に示した accomplish に関しては，word2vec の学習結果に対する 5 近傍単語には achieve と fulfil が含まれるが，これらは表 11 から異なる synset の上位語となっており，意味が混ざっていることがわかる．そのような場合でも，提案する評価法ではいずれかの synset に関する適合率だけを考慮するため，不当に評価が高くなることはない．一方，DeConf や sense2vec の学習結果では，異なる意味ベクトルの近傍にそれぞれが現れており，適切に多義語に対する意味ベクトルを学習していることが確認できる．

また，word2vec では synset ごとに 1 つの上位語しか現れないのに対し，DeConf や sense2vec では 1 つの synset に対応した上位語が複数存在している．これらのことは，表 10 に示した announce に対する結果についても同じことが言える．このことから，synset の評価においては，最近傍単語のみを対象とするよりも 5 近傍もしくはより大きな範囲の単語を対象とする方が意味をより正確に評価できるものと思われる．また，多義性を考慮した単語分散表現，特に sense2vec の学習結果には不必要な意味ベクトルが存在していることが確認できる．前述のように，このような不必要な意味ベクトルが含まれる場合，提案データセットでは評価が下がるようになっており，その意味で，適切な評価ができていると言える．

表 9: accomplish に対する各単語分散表現の 5 近傍単語

単語分散表現	5 近傍単語
word2vec(wiki_{nl})	accompishing, **achieve**, **fulfill**, accomplishes, ninestrikethreestrikeout
sense2vec(wiki_{nl})	can..., understand, mean, creatively, say-
	achieve, accompishing, accomplished, accomplishes, ninestrikethreestrikeout
	accomplish, **fulfill** accomplishing, accomplishes, **fulfil**
	amplication, costwise, timephased, degreeday, ecoefficient
MSSG(wiki)	**achieve**, accompishingm accomplished, **fulfill**, accomplishes
word2vec(google)	accomplishing, accomplished, **attain** accomplishes, Accomplisihing
DeConf(google)	accomplish, **carry_out**, **fulfil**, **carry_through**, myrmidon
	accomplish, **achieve**, by_luck, haply, by_chance

表 10: announce に対する各単語分散表現の 5 近傍単語

単語分散表現	5 近傍単語
word2vec(wiki_{nl})	announcing, announces, announced, **proclaim**, announcement
MSSG	announces, informed, **declare**, tearful, announcing
	ask, **inform**, listen, **declare**, predict
	announced, announcing, announcement, announce, announced
word2vec(google)	accnounced, announces, annouce, announcing, unveil
DeConf	announce, announce, announce, announced, blazon_out
	announce, announce, announced, **declare**, announcement
	announce, announce, announce, announced, blazon_out
	hearld, announce, announce, announce, **foretell**

表 11: 提案データセットにおける関連語の例

単語	関連語
accomplish	**carry_out**, **carry_through**, **fulfil**, **fulfill**, execute
	achieve, **attain**, reach, win, succeed
announce	**proclaim**, **inform**, denote, exclaim, enuciate
	declare, say, state, tell, express
	dentify, name, denote, refer, intend
	foretell, **herald**, **inform**, annuciate, harbinger

最後に，提案データセットの利点と欠点について考察する．利点としては，SimLex-999 を用いられる avgSim や maxSim のように 1 つの単語に対する複数の意味ベクトルを組み合わせて評価値を算出しないため，意味の分類が曖昧になることを回避できるという点が挙げられる．たとえば，DeConf の学習結果においては，announce に対する意味ベクトルの数は 4 個，declare に対する意味ベクトルの数は 8 個あるため，それらの

類似度を avgSim で計算する場合，合計 32 の組み合わせに対する類似度の平均を取ることになり，妥当な評価とは言い難い．一方，提案データセットにおける評価方法では，1 つの意味ベクトルに対する近傍単語のみを用いて評価するため，そのような問題を回避できる．欠点としては，階層が大きく異なる上位語群が評価に反映されにくく，近傍単語による評価は学習対象となる単語数に依存するという点が考えられる．

5 おわりに

本研究では，多義性を考慮した分散表現の評価のため，WordNet における概念階層を用いて新たなデータセットを構築した．概念階層を用いたことで，各単語が持つ多義性を既存のデータセットを利用した評価方法よりも正確に扱うことができることを実験的に確認した．提案するデータセットの構築方法は人手を介さないため，概念階層を伴う様々なコーパスへの適用や多言語化への応用が考えられる．

今後の課題としては，適合率（Precision）を計算する際の上位語群の階層構造を評価に反映するために，Label Ranking Average Precision （LRAP）や ImageNet [19] の評価などに用いられる mean Average Precision（mAP）の導入が考えられる．また，SimLex-999 に用いられている無関係な単語ペアの挿入や BabelNet が持つ誤った階層構造の排除がより良いデータセットの構築のために必要であると考えられる．

参考文献

[1] T. Mikolov, K. Chen, G. Corrado, and J. Dean, “Efficient estimation of word representations in vector space,” *arXiv preprint arXiv:1301.3781*, 2013.

[2] E. Grave, T. Mikolov, A. Joulin, and P. Bojanowski, “Bag of tricks for efficient text classification,” *ACL*, pp. 3–7, 2017.

[3] A. Trask, P. Michalak, and J. Liu, “sense2vec - A Fast and Accurate Method for Word Sense Disambiguation In Neural Word Embeddings,” *arXiv e-prints arXiv:1511.06388*, 2015.

[4] A. Neelakantan, J. Shankar, A. Passos, and A. McCallum, “Efficient non-parametric estimation of multiple embeddings per word in vector space,” *EMNLP*, pp. 1059–1069, ACL, 2014.

[5] M. T. Pilehvar and N. Collier, “De-conflated semantic representations,” *EMNLP*, pp. 1680–1690, ACL, 2016.

[6] A. Bakarov, “A survey of word embeddings evaluation methods,” *arXiv preprint arXiv:1801.09536*, 2018.

[7] L. Finkelstein, E. Gabrilovich, Y. Matias, E. Rivlin, Z. Solan, G. Wolfman, and E. Ruppin, “Placing search in context: The concept revisited,” *ACM TOIS*, vol. 20, no. 1, pp. 116–131, 2002.

[8] F. Hill, R. Reichart, and A. Korhonen, “Simlex-999: Evaluating semantic models with (genuine) similarity estimation,” *Computational Linguistics*, vol. 41, no. 4, pp. 665–695, 2015.

[9] C. Fellbaum, “Wordnet and wordnets,” *Encyclopedia of Language and Linguistics* (A. Barber, ed.), pp. 2–665, Elsevier, 2005.

[10] R. Navigli, “Babelnet and friends: A manifesto for multilingual semantic processing,” *Intelligenza Artificiale*, vol. 7, no. 2, pp. 165–181, 2013.

[11] R. Navigli and S. P. Ponzetto, “Babelnet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network,” *Artificial Intelligence*, vol. 193, pp. 217–250, 2012.

[12] Z. S. Harris, “Distributional structure,” *Word*, vol. 10, no. 2-3, pp. 146–162, 1954.

[13] X. Rong, “word2vec parameter learning explained,” *arXiv preprint arXiv:1411.2738*, 2014.

[14] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, “Distributed representations of words and phrases and their compositionality,” *NIPS*, pp. 3111–3119, 2013.

[15] E. Agirre, E. Alfonseca, K. Hall, J. Kravalova, M. Paşca, and A. Soroa, “A study on similarity and relatedness using distributional and wordnet-based approaches,” *NAACL*, pp. 19–27, ACL, 2009.

[16] R. Navigli, D. Jurgens, and D. Vannella, “Semeval-2013 task 12: Multilingual word sense disambiguation,” ** SEM*, vol. 2, pp. 222–231, 2013.

[17] S. Bird, E. Klein, and E. Loper, *Natural language processing with Python: analyzing text with the natural language toolkit.* ” O’Reilly Media, Inc.”, 2009.

[18] R. Řehůřek and P. Sojka, “Software Framework for Topic Modelling with Large Corpora,” *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, (Valletta, Malta), pp. 45–50, ELRA, May 2010.

[19] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, *et al.*, “Imagenet large scale visual recognition challenge,” *IJCV*, vol. 115, no. 3, pp. 211–252, 2015.