2017年度　修士論文

語義・概念の分散表現の有効性の検証と適用範囲の拡張

指導教授

林 良彦 教授

小林 哲則 教授

2018年1月30日

早稲田大学 基幹理工学部 情報理工学科
知覚情報システム研究室

5116F026-1

金 田　健 太 郎

目 次

表 目 次

図 目 次

概要

本研究では，単語分散表現と辞書資源から作成された意味（語義・概念）単位の分散表現を，単語間意味関係分類，および Semantic Taxonomy Enrichment というタスクに適用する手法を提案し，単語分散表現を利用した既存手法と比較することで有効性を検証した．

単語の意味を扱うような自然言語処理の問題（タスク）においては Word2Vec に代表される単語の分散表現が広く用いられている．しかし，これら単語の分散表現は一単語につき一つだけ与えられるため，単語の持つ多義性を適切に表現出来ないという問題がある．

そこで本稿では，AutoExtend[35] と呼ばれる手法により，辞書資源（Princeton WordNet: PWN）中の語義・概念（単語の持つ意味）に対して単語と同じ空間に導出された分散表現を利用した．ここで利用する語義・概念の分散表現は，幾つかのタスクにおいて単語分散表現を上回る精度が確認されているが [47, 48, 50, 16]，その適用範囲・性能は辞書資源によって制限されてしまうという問題が指摘されている [49]．

今回扱う単語間意味関係分類，Semantic Taxonomy Enrichment は，どちらも既存の辞書を充実し，適用範囲を広げることに繋がる重要なタスクである．単語間意味関係分類は，与えられた 2 単語中に成立している意味関係を選択肢の中から推定する，他クラス分類タスクである．これが適切に処理できると，辞書で定義された関係性の充実化に貢献することができる．Semantic Taxonomy Enrichment は，ターゲットとなる辞書（PWN）に含まれていない，専門用語や俗語などの単語（未知語）を，手がかりとして与えられた未知語の定義文，品詞を利用して辞

書中の概念へ結びつけるタスクである．これを適切に処理することで，辞書の語彙サイズを拡張することができる．

これらのタスクに対しても単語の分散表現を利用した手法が提案されているが [39, 29]，本研究では，単に単語・語義・概念の分散表現を利用するだけでなく，概念間に辞書上で定義された関係性を利用したアプローチを提案した．

単語間意味関係分類については，類似した意味（概念）を持つ単語ペアは，別の単語に対して類似した関係性を持つ（例：類似した概念に属する語義を持つ 2 単語：car と automobile は，tire に対して似たような関係：全体-部分関係性を持つ）と仮定する．すると，ある単語ペア（w_1, w_2）が与えられた時，そこに特定の意味関係 r が成立する可能性は，w_1 と r の関係にあることが分かっている単語 w_1^r の持つ意味（語義・概念）と，w_2 の持つ語義・概念の類似度によって表されることが期待できる．そこで，各単語に対して辞書上で特定の関係を持つような語義・概念を収集し，それらの類以度，および類以度算出時に用いたベクトルを素性とすることで教師付きの他クラス分類を行った．

Semantic Taxonomy Enrichment に関しては，単語・語義・概念の分散表現が同一空間上に存在していることを利用する．具体的には，未知語の定義文中の単語を用いて未知語の意味を表すようなベクトルを作成し，それと辞書中の語義・概念に対応する分散表現を比較することで候補となる概念を収集する．その上で，概念間に辞書上で成り立つ関係性を利用して候補を絞り込み，最適な概念を選択する手法を提案する．

タスクへの適用評価の結果，単語間意味関係分類については単語分散表現を利用する既存手法を上回る精度が得られた．Semantic Taxonomy Enrichment においては，単語の分散表現や品詞タグを素性として教師あり学習を行う既存手法に迫る性能が得られた．

第1章　序論

本研究では，単語分散表現と辞書資源から作成された意味（語義・概念）単位の分散表現を，単語間意味関係分類，および Semantic Taxonomy Enrichment というタスクに適用する手法を提案し，単語分散表現を利用した既存手法と比較することで有効性を評価する．

これらのタスクに限らず，自然言語の意味を扱うタスクを計算機によって処理するためには，まず計算機内に適切に言語の意味を表現することが必要である．言語表現を構成する基本要素は単語であるから，単語の意味表現を適切に定めた上で，これを効果的に利用して所望の処理機能を実現することが適切である．

計算機による処理を前提とした単語の意味に関する研究領域：計算論的語彙意味論 (computational lexical semantics) は自然言語処理でも重要な一分野をなしているが，この領域に大きな影響を与えた言語学の考え方に，“a word is characterized by the company it keeps” [11] というフレーズによって表される分布仮説 (distributional hypothesis) がある．これは単語の様々な特性はそれが使用される文脈 (context) に現れることを述べており，この仮説に基づいた単語の意味表現：分布表現を求める手法は，様々に提案されている．

ここで，分布表現は実数値を要素として持つベクトルとして与えられ，単語間の意味的な類似度や関連度は，内積やコサインなどのベクトル演算により表される．これは，単語を単に語彙 (vocabulary) におけるインデックスとして規定する方法 (one-of-k/one-hot encoding と呼ばれる) にはない大きな利点である．単語を扱う際，one-of-k 表現の代わりに分布表現を利用することで，感情分析や係り受け解析等の様々なタスクにおいて性能の向上が確認されている [7, 5].

特に単語の予測モデルに基づき構成された単語のベクトルは，単語の分散表現 (distributed representation)，もしくは，単語埋め込みベクトル (word embedding vector) などと呼ばれる．これらの単語の分散表現の有用性は広く認識され，多くのアプリケーションに適用されている [36, 7, 1, 43] が，その問題として単語が有する多義を考慮していないことが挙げられる．一般に単語は複数の語義 (sense または word sense) をもつ．例えば，“bank” は少なくとも「金融機関としての銀行」，「土手」の意味を持つが，コーパスから構成される “bank” の分散表現には，コーパスにおける出現頻度を反映した形で両語義に対応する成分がミックスされてしまう．

この問題に対するアプローチとして，特定の単語の特定の語義，あるいは語義が指示する概念 (concept) に対して分散表現を与える方法論が近年注目を浴びつつある[1]．

語義・概念に対する分散表現（”sense representation”) を獲得するための方法論は，(1) コーパスにおける出現文脈をクラスタリングすることにより,「語義」を導出する (induction) と同時に，語義に対する分散表現を与える，(2) 語義・概念の関係性を陽に表わす情報構造を利用し，単語の分散表現から数学的手法により語義・概念の分散表現を得る，という 2 つの方向性に大別できる．(1) の手法は分布仮説をさらに追求するものであるが，クラスタリングにより導出される「語義」がネイティブ話者が感覚として有している語義，すなわち，辞書において区分される語義と適合する保証がないという問題がある．この点，(2) の手法は辞書に一般化されて記述されている言語知識を利用するものであるから，問題とはならない．さらに (2) によって得られた分散表現をタスクに適用する際は，分散表現だけではなく辞書から得られる様々な意味関係 (上位・下位，全体・部分，属性，含意など) が利用できる．

本研究では,（２）に区分される手法の一つである AutoExtend [35] により導出

[1]昨年には最初の国際ワークショップ　SENSE2017（https://sites.google.com/site/senseworkshop2017/home）が開催されるなど，議論が行われている．

した語義・概念の分散表現を用いる．この手法は，単語の分散表現を入力とし，WordNet [25] のような単語・語義・概念の関係をグラフ表現した辞書資源を参照する数学的手法 (より具体的には自己符号化ニューラルネットワーク) により語義・概念の分散表現を導出する．

AutoExtend により得られる語義・概念の表現は入力とした単語の分散表現と同じ空間にあるため，異なる言語単位 (例: 単語と語義) の間で相互に比較可能であるという大きな利点がある [47]．この利点を利用し，[35] では語義曖昧性解消に対して適用を行っている．語義曖昧性解消は，ある単語 w とその出現文脈 $c = \{w_1, w_2, ..., w_n\}$ が与えられた時，そこで用いられている w の意味を，与えられた語義 $L_w = \{l_1, l_2, ..., l_m\}$ の中から選択するタスクである．ここでは，L_w 中の各語義が結びついた概念の分散表現の平均と，文脈中の単語から作成した文脈ベクトルの内積を素性として利用した教師あり学習を行うことで，既存手法 [46] を上回る精度が確認されている．また，[47] では単語間類以度・関連度の定量化に対して適用を行っている．単語間類以度・関連度の定量化は，与えられた単語ペアの類似性・関連性の強さを定量化するタスクである．ここでは，各単語の持つ語義および，語義の結びついた概念の分散表現間の類以度を利用することで，単語分散表現 [24] を利用した手法を上回る精度が確認されている．

このように，AutoExtend によって得られた語義・概念の分散表現は幾つかのタスクにおいて既存手法を上回る精度を達成しているが，既存の辞書資源，情報資源を利用するアプローチは，適用範囲・性能が辞書に制限されてしまう問題がある．辞書資源はおおよそ手動によって作られるものであるため，語彙サイズは大規模コーパスから抽出するものに比して小さい傾向にあり，適用範囲が限定されてしまう．また，辞書中で概念間に定義された関係性は不完全であり，欠けが有り得る．しかし，手動によって語彙の拡張，あるいは関係性の充実化を行うには，大きなコストがかかる．

本研究で取り上げる 2 つのタスクは，これらの辞書が抱える問題を自動的に解

決することに貢献する．

単語間意味関係分類は，与えられた 2 単語中に成立している意味関係を選択肢の中から推定する，他クラス分類タスクである．これが適切に処理できると，概念に与えられた分散表現間類以度等を利用し，関連しているとされた単語ペアの持つ関係性が，辞書上のどれにあたるかを推定することができる．これによって，辞書で定義された関係性の充実化が行える．

Semantic Taxonomy Enrichment は，ターゲットとなる辞書（PWN）に含まれていない，専門用語や俗語などの単語（未知語）を，手がかりとして与えられた未知語の定義文，品詞を利用して辞書中の概念へ結びつけるタスクである．これを適切に処理することで，辞書の語彙サイズを拡張することができる．

単語間意味関係分類に関しては，評価データとして BLESS データセット [2] を用いる．このデータセットを扱う従来手法 [29] では，単語ペアが与えられたとき，それぞれの単語に対して対して与えられた単語分散表現の差分，または連結のいずれかを入力に用い，教師付き学習によって多クラス分類を行っている．

単語ペアが与えられた時，意味関係は各単語が持つ特定の意味（語義・概念）間に成立するものであるため，その推定には語義・概念の分散表現を利用するのが適切である．

本研究では，単に分散表現を利用するだけでなく，その収集に辞書構造を利用する．類似した意味（概念）を持つ単語ペアは，別の単語に対して類似した関係性を持つ（例：類似した概念に属する語義を持つ 2 単語：car と automobile は，tire に対して似たような関係：全体-部分関係性を持つ）．このことから，ある単語ペア（w_1, w_2）が与えられた時，そこに特定の意味関係 r が成立する可能性は，w_1 と r の関係にあることが分かっている単語 w_1^r の持つ意味（語義・概念）と，w_2 の持つ語義・概念の類似度によって表されることが期待できる．そこで，各単語に対して辞書上で特定の関係を持つような語義・概念を収集し，それらの類以度，および類以度算出時に用いたベクトルを素性とすることで教師付きの他クラス分類

を試みる．

Semantic Taxonomy Enrichment に関しては，評価データとして SemEval 2016 Task14[15] で公開されたテストデータを用いる．このデータで最高精度を達成している従来手法 [39] は，未知語の定義文中に含まれる単語に結びついた概念を候補集合とした上で，集合中に含まれる概念に単語の分散表現や品詞タグを用いて素性を与えた後に，対応付ける概念の推定をランキング学習により行っている．

教師付き学習は一般に良い精度をもたらすが，学習データを必要とする問題がある．また，未知語定義文に含まれる単語に結びついた概念中に，候補として選択するのに適切な概念が含まれている保証はない．例えば名詞の未知語に対して something (形容詞）という形で定義文が与えられた場合は，適切な概念を収集することが不可能である．

そこで本研究では，単語・語義・概念の分散表現が同一空間上に存在していることを利用し，未知語の定義文中の単語を用いて未知語の意味を表すようなベクトルを作成し，それと辞書中の語義・概念に対応する分散表現を比較することで候補となる概念を収集する．その上で，概念間に辞書上で成り立つ関係性を利用して候補を絞り込み，最適な概念を選択する手法を提案する．提案手法では学習を必要としないため，学習を必要とする従来手法と比して，日々増える新語などに対して迅速に意味的基盤を与えることが期待できる．

本論文の以下では，上記で概説した関連概念，関連研究の整理 (第 2 章) を行った後，**単語間の意味関係の推定** (第 3 章)，**Unsupervised 手法による Taxonomy Enrichment** (第 4 章) のそれぞれについて，アプローチと実験による評価結果・考察を述べる．また，残された課題とそれに対するアプローチについても議論する (第 5 章)．

第2章　関連研究

ここでは，提案手法で用いる辞書資源を概説し，単語・語義・概念の分散表現に関する研究を整理する．

2.1　辞書資源

辞書資源は，単語が持つ意味を端的な文章（定義文）によって定義し，意味間に成り立つ関係性 (lexical semantic relation) を記述した知識資源の一つである．主に人手で作成されているため，人間の観点が明示的に反映されているのが特徴である．

辞書資源として広く利用されているものに，Princeton WordNet: PWN [25] がある．PWN は英語の意味的な構造を定義した辞書である．PWN において，単語はいくつかの意味（語義）を持つとされ，同じ意味を持つ語義が概念 (synset) を構成し，概念間には対義関係 (antonymy)，同義関係 (synonymy)，上位–下位関係 (hypernymy,hyponymy)，全体–部分関係 (holonymy,meronymy) などの関係性 (lexical semantic relation) が定義される．ここで，単語表記と語義の組を語彙素 (lexeme) という．なお以下では，誤解のない範囲で語彙素のことを語義と書く．

上位–下位関係とは上位概念と下位概念の関係性であり，例えば lawsuit, suit, case という単語の「訴訟」を表す語義が属する概念と，countersuit という単語の「反訴」を表す語義が属する概念は上位–下位関係にある．前者から見て後者は下位の概念（hyponym）であり，後者から見た前者は上位の概念（hypernym）である．全体–部分関係とはある概念がペアとなる概念の部分を構成するような関係性

であり，例えば「自動車」を表す語義が属する概念と,「ハンドル」を表す語義が属する概念は全体–部分関係にある．ここで，共通する上位語を持つ単語同士に成り立つ関係は，兄弟関係（coordinate）と呼ばれる．

辞書資源は，語義・概念の分散表現作成以外にも，語義曖昧性解消 [27] や感情分析 [8]，情報検索 [42] 等，様々な自然言語処理分野のタスクに適用されている．しかし辞書資源をタスクに適用する際，その適用範囲は辞書の語彙サイズに制限されるという問題がある．

また，人手で作られたものである以上，辞書上で定義された関係性は不完全である [4]．すなわち，本来特定の関係性が成立しうる概念間に，その関係性が記述されていない場合がある．

2.2　分布仮説に基づく単語のベクトル表現

ここでは単語のベクトル表現作成手法のうち，分布仮説に基づくものを整理する．

これらの手法によって，各単語には実数値を要素として持つベクトルが与えられる．これらベクトル間の類以度（コサイン類以度）は様々な関係性を区別せずに扱うことになり，**単語間の意味的な関連性の指標**となる．これは，単語を単に語彙 (vocabulary) におけるインデックスとして規定する方法 (one-of-k/one-hot encoding と呼ばれる) にはない大きな利点である．

分布仮説に基づいてコーパスから単語のベクトル表現を導出する方法は，次の２つに大別できる．

2.2.1　単語間の共起に基づく方法

一つはコーパスから得られる単語の共起情報を利用して単語共起行列や単語文書行列を構成し，それを行列分解等によって圧縮することで密なベクトルを得る手法である [41, 33]．特に PMI で重み付けした単語共起行列を行列分解によって低ランク近似する場合，得られる分散表現は後述する予測モデルによって得られ

る分散表現と等価な表現能力を持つことが証明されている [20].

2.2.2　言語モデルに由来する手法

もう一方は，言語モデルに由来する手法群である．言語モデルとは，文脈が与えられた時，次に来る単語を予測するようなモデルである．

その中でも広く利用されている Word2Vec[24] は，ニューラルネットワークを利用した言語モデルの中間層を無くし，Negative Sampling や Hierarchical Softmax を導入することで計算量を削減し，大量のデータを用いた学習を現実的な時間で可能にしている．

Word2Vec においては,（前後の）文脈が与えられたときに，その中心に来る単語を予測するモデル（Continuous Bag of Words）と，ある単語が与えられた時に，その周囲に現れる単語（文脈）を予測するモデル（Skip-gram）が提案されている．

Skip-gram に修正を加えた手法もいくつか提案されており，[19] では扱う文脈を周囲に現れる単語ではなく，係り受け解析によって得られた依存木から得ている．また，[22] では，語順を考慮して周囲の文脈を推定している．

これらの分散表現は感情分析 [7]，情報検索 [43]，固有表現抽出 [1] クエリ拡張 [36] 等のタスクにおいて広く応用され，有効性が確認されている．しかしながら，いくつかの問題点も指摘されている [18, 35].

そのうちの一つとして，単語の分散表現では多義性を考慮していないため，多義語の持つ複数の意味が単一のベクトル中で混合されてしまうという問題がある．

この問題に対し，単語の分散表現と同様にコーパスから得られる情報を利用しながら，単語の持つ意味一つ一つに対して分散表現を作成する手法が提案されている．

2.3　語義・概念のベクトル表現

単語の持つ意味一つ一つに分散表現を与える際，コーパス上には明示的に現れることのない単語の持つ意味（語義・概念）をどのように定義するかが問題となる．その方法は，次の 2 つに大別される．

2.3.1　教師無しで語義を誘導（induction）する方法

ひとつは，与えられた単語が出現する文脈や類似する単語をクラスタリングし，そのクラスタ一つ一つを語義とするという手法である．

[13] では，分布意味論をもとにした「異なる意味（語義）で使われる単語は，異なる文脈で現れる」という仮定のもとで，一度全単語に対して単語分散表現を与えた後，それぞれが出現する文脈を，文脈中に含まれる単語の分散表現によって表し，クラスタリングする．ここで得られたクラスタを，各単語の持つ語義としている．[30] では，Word2Vec の Skip-Gram を拡張し，各出現単語に対するクラスタリングと各クラスタに対応する分散表現の学習を同時に行う手法を提案している．[3, 40] では，ノンパラメトリックベイズ [3]，および混合ベイズモデル [40] によって出現した単語がどの語義で使われているかを推定した上で，各語義に対応する分散表現の学習を行っている．

[32] では，単語に意味表現を与え，その単語と類似した表現を持つ単語をクラスタリングすることで語義を誘導している．

これらの手法を用いた場合，理論上はコーパスに出現する全単語に対して，意味レベルの分散表現を与えることができる．ただし，誘導された語義（相当のもの）と人間の知覚する語義が対応する保証がないという問題がある．

2.3.2　辞書上で定義された語義・概念を利用する手法

辞書に定義された語義・概念は，人の感覚に対応している．また，これらに対して導出された分散表現をタスクに適用する際は，分散表現だけでなく，辞書で概念間に定義された意味的ネットワークから得られる情報を利用することができる．

語義・概念に対し分散表現を与える手法として，[14] はコーパスを既存手法 [26] により語義解消して単語を語義に置き換えた後，単語分散表現の作成手法を適用することで，語義の分散表現を導出している．[6] では，事前に単語分散表現を用意した上で，各概念に対し，定義文中に含まれる単語の分散表現の和によってベクトルを与え，ベクトルのコサイン類以度を用いてコーパスの語義解消を行った後に単語分散表現の作成手法を適用している．これらの手法により得られる分散表現の正当性は，語義解消の精度に依存する．また，[14] に関しては語義の分散表現のみを作成することになるため，異なる言語単位間で分散表現を比較することが不可能である．

語義解消に依らない手法として，[34] は，各語義の意味を表すような単語群を辞書のネットワークを用いて収集し，その分散表現の重み付き和によって語義の分散表現を導出している．この手法においては，複数の意味が混同（conflate）して表現されうる単語の分散表現の和によって語義の分散表現が与えられているため，その中で語義に合致した意味だけが適切に表現されている保証はない．

2.3.3　AutoExtend

本稿で語義･概念の分散表現作成手法として用いる AutoExtend[35] は，Word2Vec などの手法により得られた単語の分散表現を辞書構造を反映したオートエンコーダーへ入力することで，単語の分散表現と同じ空間上に語義・概念の分散表現を導出する．

2.3.2 節で挙げた語義・概念の分散表現作成手法と比較した際の特徴は，導出過程に辞書の情報構造が反映されている，すなわち，単語の分散表現を語義の分散

表現に分解し，語義の分散表現によって概念の分散表現が構成されていることにある．

具体的には，単語 w^i の概念 s^j に属する語義を $l^{(i,j)}$ としたとき，それぞれのベクトルに関して次の関係性が成り立つようにオートエンコーダーを構成する（図 2.1）．

$$\vec{w^i} = \sum_j l^{(\vec{i,j})} \quad , \quad \vec{s^j} = \sum_i l^{(\vec{i,j})}$$

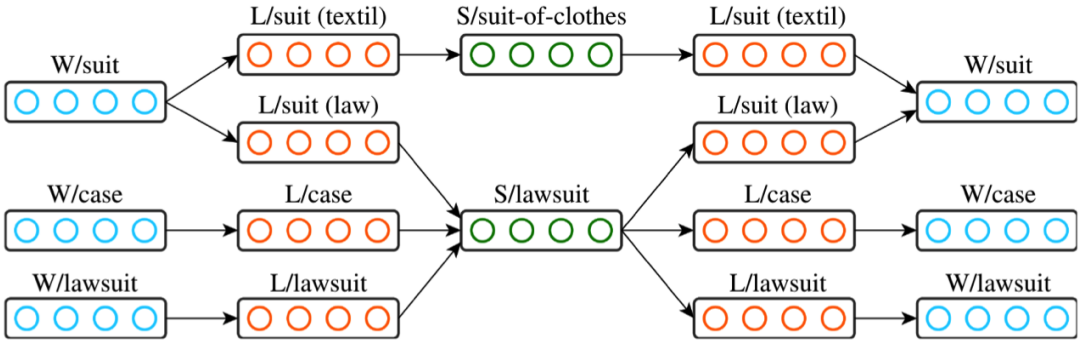


図 2.1:　構成されるオートエンコーダーの一部（[35] より抜粋）

2.4　分散表現の評価

様々な言語単位に対して与えられる分散表現は，意味の理解，および，意味の理解に基づく何らかの処理機能の実現を目的としている．そのため，導出された分散表現は次の 2 通りの観点から評価を行う必要がある．

- 適切に意味が埋め込まれているか；人間の感覚が反映された表現になっているか (intrinsic evaluation)
- 実際に言葉の意味を扱うようなタスクへ適用した際，有効に機能するか (extrinsic evaluation)

前者の観点から行われる評価方法としては，単語間類以度 [10]・関連度の定量化 [12] やアナロジー [24] のタスクへの適用評価が広く利用されている．

後者の観点から行う評価については，意味を扱うようなあらゆる自然言語処理タスクへの適用が想定されるが，特に語義・概念の分散表現を単語レベルのタスクに適用することを考えた場合，与えられた単語，単語列，単語集合等に対して，そこで用いられている単語の意味を適切に推定した上で，それに対応する語義・概念の分散表現を用いることが必要とされる．すなわち語義曖昧性解消のタスクに関する適用性が重要となる．

AutoExtend によって導出された語義・概念の分散表現は，前者の観点において，単語間類以度・関連度の定量化，および文脈付き単語間関連度 [13] の定量化タスクに適用したところ，その有効性が確認されている [35, 47]．また，後者の観点においては，語義曖昧性解消のタスク [17, 23] に関して有効性が確認されている [35]．

本研究で扱う 2 つのタスクに関しては，単語間意味関係分類が intrinsic evaluation，辞書の拡張タスクが extrinsic evaluation に対応する．それぞれについて評価を行うためのタスクが提案されており，評価データセットが公開されている．これらについては，関連する章で詳細を述べる．

第3章　単語間意味関係分類

単語間意味関係分類は，単語ペアが与えられたとき，その間に成立する関係性を与えられた選択肢の中から選ぶようなタスクである．このタスクを適切に解くことで，既存の辞書資源が潜在的に抱えている，本来特定の関係性を保持しているはずなのに，その記述がなされていない単語ペアに関し，適切なリンクを自動的に記述する事ができる．また，コーパスから自動的に辞書資源を構築する際 [37] にも，収集された単語群に対し，適切にリンクを付与することができる．

単語間意味関係分類の評価データはいくつか存在するが [38, 2]，ここでは広く用いられている BLESS データセット [2] を用いる．

BLESS データセット

BLESS データセットは，200 の名詞に対し，それらと以下で示す 5 つの関係のうちいずれかにある単語を収集し，ペアとしたデータセットである（計 14400 組）．

- COORD (3565 組): 兄弟関係（例： alligator - lizard）
- HYPER (1337 組): 下位 - 上位関係（例：alligator - animal）
- MERO (2943 組): 全体 - 部分関係（例：alligator - mouth）
- ATTRI (2731 組): 主体 - 形容の関係（例：alligator - aquatic）
- EVENT (3824 組): 主体 - 動作の関係（例：alligator - swim）

200 の名詞は大まかな意味（爬虫類，家具等）によって 17 のカテゴリに分けられている．

なお，ここで与えられた単語間の関係性は，WordNet で定義されたものと一致しないことに注意する．

比較評価の対称としては，単語の分散表現のみを用いて教師付き学習を行う従来手法 [29] を取り上げる．従来手法では，単語ペアが与えられたとき，それぞれの単語に対して word2vec (CBOW)[24] による単語分散表現の差分（$WECE_{BoW}^{offset}$）または連結（$WECE_{BoW}^{concat}$），及び dependency-based skip-gram[24] による単語分散表現の差分（$WECE_{Dep}^{offset}$）または連結（$WECE_{Dep}^{concat}$）のいずれかを入力に用い，教師付き学習によって多クラス分類を行っている．

3.1　提案手法

ここでは，WordNet と語義・概念の類以度を用い，教師付き学習による単語間意味関係の分類手法を提案する．その際，類似した意味（概念）を持つ単語 w_i, w_j は，別の単語 w_t に対して類似した関係性を持つ（例：類似した概念に属する語義を持つ 2 単語：car と automobile は，tire に対して似たような関係：全体-部分関係性を持つ）と仮定する．

すると，ある単語ペア（w_1, w_2）が与えられた時，そこに特定の意味関係 r が成立する可能性は，w_1 と r の関係にあることが分かっている単語 w_1^r の持つ意味（語義・概念）と，w_2 の持つ語義・概念の類似度によって表される（例：w_1: tire と w_2: car に部分-全体の関係が成立する可能性は，tire の持つ概念と部分-全体の関係にあることが分かっている w_1^r： automobile の持つ語義・概念と，w_2: car の持つ語義・概念の類似性によって表される：図 3.1）と期待できる．

この期待のもとで，単語ペア w_1, w_2 が与えられたとき，その間に成立する関係性をいくつか想定した上で，次の手順によって分類に使用する素性を集める．

1. 事前知識（WordNet）から，各単語の持つ語義･概念に加え，概念と特定の関係にある概念を収集し，関係ごとに集合（ノード集合）を構成する（3.1.1 節）

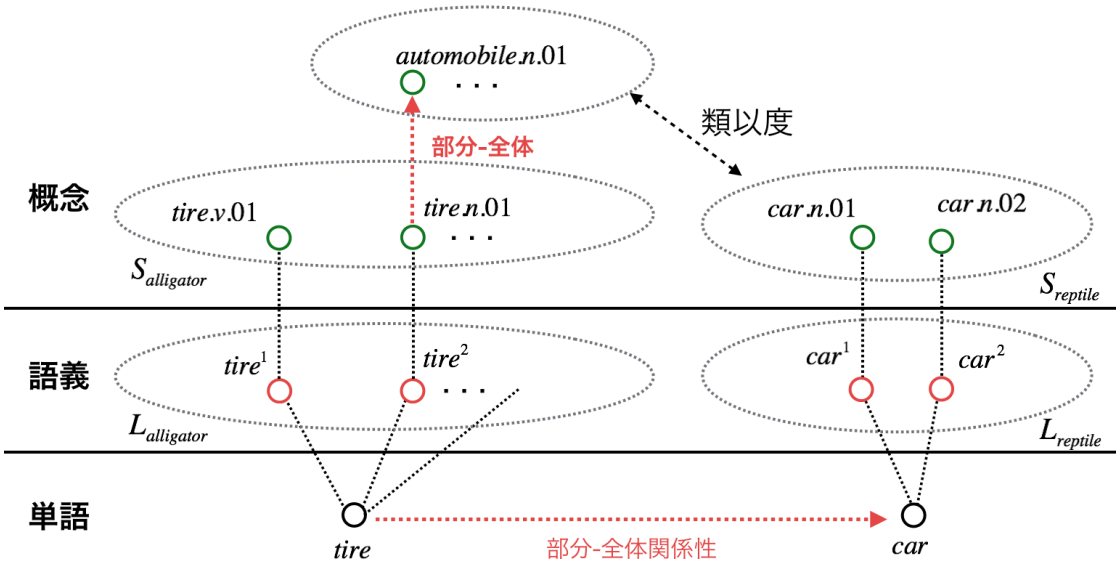


図 3.1:　語義・概念ベクトルを利用した類以度の計算（部分-全体関係時）

2. 単語間に成立しうる関係性をいくつか想定し，それぞれに応じてノード集合の組み合わせを構成する（3.1.2 節）

3. 組み合わせた語義・概念集合それぞれに対して，3 通りの手法で類以度を計算する（3.1.3 節）

最終的に 1 つの単語ペアに対して，21 種類（7 通りのノード集合 × 3 通りの計算手法）の類以度が計算される．この類以度に単語ベクトル対によるコサイン類以度を加えた 22 種類の類以度と，それぞれの類以度をもたらすベクトル対を，学習の際の素性として用いる．

3.1.1　単語に対するノード集合の構成

WordNet 中で定義された語義・概念間の意味的ネットワークを利用し，単語 w に対して以下のように 5 種類の語義・概念の集合（ノード集合）を構成する．これらのうち一部の例が図 3.1 に示されている．

- 語義集合 L_w: 単語 w の持つ語義の集合

- 概念集合 S_w: 単語 w が持つ語義が指示する概念の集合
- 上位概念集合 S_w^{hyper}: S_w 中の概念と直接 hypernymy の関係（下位-上位）の関係で結びついた概念の集合
- attribute 概念集合 S_w^{attri}: S_w 中の概念と直接 attribute（主体-性質，または性質-主体）の関係で結びついた概念の集合
- meronym 概念集合 S_w^{attri}: S_w 中の概念と直接 meronymy（全体-部分）の関係で結びついた概念の集合

3.1.2　類以度計算に使用するノード集合対の構成

単語ペア w_1，w_2 が与えられたとき，以下 5 種類の意味関係について，それぞれが成立する可能性を 7 種類のノード集合の組み合わせを用いた類以度で表す．意味関係とノード集合対の対応は次の通りであり，またノード集合の組み合わせ方を以下括弧内のように呼ぶ．

- 類似性/関連性：L_{w_1} - L_{w_2} $(sense)$, S_{w_1} - S_{w_2} $(concept)$
- 下位-上位関係性：$S_{w_1}^{hyper}$ - S_{w_2} $(hyper)$
- 兄弟関係性：$S_{w_1}^{hyper}$ - $S_{w_2}^{hyper}$ $(coord)$
- 形容関係性：$S_{w_1}^{attri}$ - S_{w_2} $(attri_1)$, S_{w_1} - $S_{w_2}^{attri}$ $(attri_2)$
- 全体-部分関係性：$S_{w_1}^{mero}$ - S_{w_2} $(mero)$

類似性/関連性の有無について考えるときは，単語の持つ意味そのもの，すなわち単語の語義・概念について比較をするのが適切である．よって語義集合同士，また概念集合同士の類以度が類似性/関連性の有無の尺度となると仮定する．

w_2 が w_1 の上位語であるならば，w_1 の上位概念（w_1 の上位語の概念）と w_2 の概念が類似するはずである（図 3.1 参照）．よって $S_{w_1}^{hyper}$ と S_{w_2} の類以度が下位-上位関係性の有無の尺度になると仮定する．

w_1 と w_2 が兄弟関係にある（同じ上位語を持つ）のであれば，両者が似たような上位概念を持つはずである．よって $S_{w_1}^{hyper}$ と $S_{w_2}^{hyper}$ の類以度が兄弟関係性の有無の尺度になると仮定する．

w_1 と w_2 が形容関係にあるのであれば，w_1 の attribute 概念（w_1 を形容するような単語の概念）と w_2 の概念，または w_1 の概念と w_2 の attribute 概念（w_2 によって形容される単語の概念）が類似するはずである．よって $S_{w_1}^{attri}$ と S_{w_2}，また S_{w_1} と $S_{w_2}^{attri}$ の類以度が形容関係性の尺度になると仮定する．

w_1 と w_2 が全体-部分関係にあるのであれば，w_1 の meronym 概念（w_1 の一部分になるような単語の概念）と w_2 の概念が類似するはずである．よって $S_{w_1}^{hyper}$ と $S_{w_2}^{hyper}$ の類以度が全体-部分関係性の有無の尺度になると仮定する．

3.1.3　類以度計算手法

単語ペア w_1，w_2 に対し，c$\in \{sense, concept, hyper, attri_1, attri_2, mero\}$ という組み合わせ方でノード集合の組 X_{w_1}, X_{w_2} が与えられたとき，その類以度を以下の 3 通りの手法によって計算する．

sim^c_{max} ： sim^c_{max} は，各集合中のノードを組み合わせ，それぞれの分散表現でコサイン類以度を算出し，その最大値をノード集合間の類以度とする手法であり，以下の式で与えられる．

$$sim^c_{max}(w_1, w_2) = \max_{x_1 \in X_{w_1}, x_2 \in X_{w_2}} sim(\vec{x_1}, \vec{x_2}) \tag{3.1}$$

ただし，$sim(\vec{x_1}, \vec{x_2})$ は $\vec{x_1}, \vec{x_2}$ のコサイン類以度である．ノード集合はそこに含まれるノードの数だけ意味的曖昧性（多義性）を持つといえるが，その曖昧性を互いに最も近い意味同士と解釈することで解消した後に類以度を算出するような

手法である．類以度最大となるときのノード対の分散表現 ($\vec{x_1}$ と $\vec{x_2}$) も素性として使用する．

$\boldsymbol{sim^c_{sum}}$ ： sim^c_{sum} は，各集合に含まれるノードについて，分散表現の総和を求めた後に，総和同士のコサイン類以度を算出する手法であり，以下の式で与えられる．

$$sim^c_{sum}(w_1, w_2) = sim(\sum_{x_1 \in X_{w_1}} \vec{x_1}, \sum_{x_2 \in X_{w_2}} \vec{x_2}) \tag{3.2}$$

この手法では，それぞれのノード集合について，その全体的な意味を表すようなベクトルを作成し，それらの類以度を求めている．各ノード集合に含まれる分散表現の総和 ($\sum_{x_1 \in X_{w_1}} \vec{x_1}$ と $\sum_{x_2 \in X_{w_2}} \vec{x_2}$) も素性として利用する．

$\boldsymbol{sim^c_{med}}$ ： sim^c_{med} は，各集合中のノードを組み合わせ，それぞれの分散表現でコサイン類以度を算出し，その中央値をノード集合間の類以度とする手法であり，以下の式で与えられる．

$$sim^c_{med}(w_1, w_2) = \underset{x_1 \in X_{w_1}, x_2 \in X_{w_2}}{\text{median}} sim(\vec{x_1}, \vec{x_2}) \tag{3.3}$$

この手法においては中間的な類以度が算出されるが，この類以度に寄与するノードは各集合の中から一つづつのみである．類以度が中央値となるときのノード対の分散表現 ($\vec{x_1}$ と $\vec{x_2}$) も素性として使用する．

3.2　実験

3.2.1　実験設定

評価データには BLESS データセットを用い，分類精度を評価する．比較評価の対象としては，[29] による 4 手法（$WECE^{offset}_{BoW}$，$WECE^{concat}_{BoW}$，$WECE^{offset}_{Dep}$，$WECE^{concat}_{Dep}$）を用いる．

提案手法では RandomForest 分類器を用いて教師付き学習を行う．この際，22 種類の類以度に加え 22 種類のベクトル対の差分または連結を素性として用いる．差分を使用した場合を $Proposal_{offset}$，連結を使用した場合を $Proposal_{concat}$ と表記する．また教師データとテストデータの分割は従来手法に従い，以下の 2 通りで行う．また，評価尺度も従来手法 [29] に従い Precision (P) , Recall (R) , F-measure (F) を用いる．

In-domain (ID)： target concept のカテゴリに従いデータを 17 分割し，そのカテゴリ 1 つ 1 つに対して 5 分割の交差検定を行う．17 × 5 回の試行の平均スコアを算出する．
Out-of-domain (OoD)： target concept のカテゴリに従いデータを 17 分割し，そのうち 16 カテゴリ分のデータを学習，残り 1 カテゴリのデータをテストに用いる．17 回の試行の平均スコアを算出する．

提案手法においては，素性となる類以度を算出する際に WordNet の情報を利用している．そのため，単にこの情報を利用して単語間意味関係分類を行ったときよりも高いスコアが得られているかどうかを確認することが重要である．よって次のようにベースラインを設定し，比較する．

ベースライン： 単語ペア w_1，w_2 が与えられたとき，S_{w_1} 中の概念と S_{w_2} の概念のうちいずれかが，WordNet 上において直接何らかの関係によって結ばれているのであれば，単語ペアに対しその関係が存在しているとする．

辞書中に EVENT の関係は定義されておらず，また BLESS の ATTRI と Wordnet における attribute は定義が異なるため，ベースラインの手法において発見できる関係は COORD, HYPER, MERO の 3 つのみである．よってこの 3 つの関係についてのみ P, R, F を算出し比較評価を行う．

3.2.2　実験結果

表 3.1 に従来手法との比較結果を示す．ベクトルの差分を利用した場合と連結を利用した場合のどちらにおいても，P, R, F 全てのスコアにおいて提案手法が従来手法を超えるスコアを獲得している．これにより，語義・概念のベクトルを利用することの有効性が確認できた．

	In-domain			Out-of-domain		
	P	R	F1	P	R	F1
$WECE^{offset}_{BoW}$	0.900	0.909	0.904	0.680	0.669	0.675
$WECE^{offset}_{Dep}$	0.853	0.865	0.859	0.687	0.623	0.654
$Proposal_{offset}$	0.913	0.907	0.906	0.766	0.762	0.753
$WECE^{concat}_{BoW}$	0.899	0.910	0.904	0.838	0.570	0.678
$WECE^{concat}_{Dep}$	0.859	0.870	0.865	0.782	0.638	0.703
$Proposal_{concat}$	**0.973**	**0.971**	**0.971**	**0.839**	**0.819**	**0.812**

表 3.1: BLESS データセットにおける従来手法との比較

	$Proposal^{OoD}_{concat}$			ベースライン		
	P	R	F1	P	R	F1
COORD	0.761	0.559	0.645	0.550	0.108	0.180
HYPER	0.767	0.654	0.706	0.746	0.199	0.314
MERO	0.625	0.809	0.705	0.934	0.034	0.065
ATTRI	0.913	0.995	0.952	-	-	-
EVENT	0.974	0.983	0.979	-	-	-

表 3.2: $Proposal^{OoD}_{concat}$ における詳細結果とベースラインの比較

OoD 分割時の $Proposal_{concat}$（$Proposal^{OoD}_{concat}$）における結果の詳細と，ベースラインとの比較結果を表 3.2 に示す．提案手法においては，異なる品詞間にある関係の ATTRI, EVENT に比べ，名詞-名詞の関係分類である COORD, HYPER, MERO の分類が低い結果となっているが，ベースラインと比較した際には Recall

スコアについて大きく上回っている．このことから，辞書情報を単に利用するよりも広い適用範囲を持っていることがわかる．

	P	R	F1	F1 変化量
$Proposal_{concat}^{OoD}$	0.839	0.819	0.812	-
−単語の分散表現	0.845	0.827	0.819	**0.008**
$-sense$	0.833	0.815	0.806	-0.006
$-concept$	0.826	0.809	0.802	-0.010
$-coord$	0.834	0.811	0.803	-0.009
$-hyper$	0.826	0.803	0.800	-0.012
$-attri_1$	0.826	0.806	0.798	-0.014
$-attri_2$	0.842	0.820	0.814	0.002
$-mero$	0.835	0.813	0.806	-0.006

表 3.3: 各ノード集合対（単語の分散表現対）に由来する素性を取り除いたときの結果

表 3.3 では，$Proposal_{concat}^{OoD}$ において各ノード対に由来する 3 つの類以度，あるいは単語の分散表現対に由来する 1 つの類以度及びベクトル対を除いて学習を行った結果を示す．ノード対を除いた場合はほぼスコアが低下していることから，各組み合わせの有効性が確認できる．一方で単語対を除いた場合はスコアが上昇していることから，語義・概念の分散表現を適切に利用した場合，単語の分散表現はむしろノイズになることが示唆される．

表 3.4 では，各計算手法に由来する 7 つの類以度及びベクトル対を除いて学習を行った結果を示す．表 3.3 に比べて多くの素性を除いているにも関わらず，どの場合においてもさほど大きくスコアが変化しないことから，1 つの手法を他の 2 つの手法が補完しあっていることが示唆される．

	P	R	F1	F1 変化量
$Proposal^{OoD}_{concat}$	0.839	0.819	0.812	-
$-sim^{c}_{max}$	0.835	0.812	0.805	-0.007
$-sim^{c}_{sum}$	0.843	0.822	0.816	0.004
$-sim^{c}_{med}$	0.838	0.811	0.805	-0.007

表 3.4: 各計算手法に由来する素性を取り除いたときの結果

3.3　議論

教師付きで単語間意味関係分類のタスクを解く際，分類器の学習において起こりうる問題として Lexical Memorization が存在する [21]．これは，文類器は単語間の意味関係を分類するように学習するのではなく，特定の関係が成立する時現れやすい「典型的な」単語を覚えているにすぎないという問題である．つまり，たとえば下位-上位関係を学習する際，(cat, animal), (dog, animal), (pig, animal),... といったように，片側に特定の単語が頻出する形で正例が与えられた場合，文類器は (X, animal) という形で与えられたデータを全て下位-上位関係に分類してしまう可能性がある．

この問題が発生しうる条件としては，学習データを文類器に与える際，

1. 単語ペアの語順を区別していること

2. 同じ入力単語に対しては，常に同じ特徴表現が与えられること

が挙げられる．

本研究で提案する手法においては，単語ペアの語順を考慮しない特徴であるコサイン類以度を用いており，また，sim^{c}_{max}, sim^{c}_{med}, において類以度を算出する際に用いるベクトルは，同じ単語が入力されてもペアとして与えられる単語が変われば変化する．一方，sim^{c}_{sum}, において類以度を算出する際に用いるベクトルは単語毎に一定であるが，この手法を用いて得られた特徴を除外しても精度に大きな変化は現れないことが，表 3.4 中の $-sim^{c}_{sum}$ と $Proposal^{OoD}_{concat}$ を比較することで

確認できる．よって，提案手法においては Lexical Memorization の影響は小さいと考えられる．

第4章　Semantic Taxonomy Enrichment

4.1　辞書の拡張

辞書の拡張，すなわち「未知語」を辞書に対して結びつける手法に関しては，様々な関連研究がある．

[28, 9] では，複数の辞書資源を統合することにより語彙サイズを増やしている．また，[45] は Wikipedia に含まれているが PWN に含まれていないような「未知語」を，Wikipedia の構造を利用することで PWN 中の概念に結びつけている．これらの手法では，ターゲットとなる辞書資源とは別の知識構造から得られる「未知語」の関連語を利用しているため，未知語が単体で提示された場合は扱うことができない．

[31] では，パターンマッチングを用いてコーパスから集めた「未知語」をフィンランド語 WordNet に結びつける手法を提案しているが，フィンランド語固有の特徴を利用しているため, 別の言語に直接適用することができる保証はない．

SemEval 2014 Task16

SemEval 2014 Task16 [15] は，専門用語，スラングなどの PWN に登録されていない語（未知語）を, 他の辞書資源から得た定義文を用いて PWN 中の概念（synset）へと結びつけることを想定した Semantic Taxonomy Enrichiment（辞書の拡張）タスクである．前述の関連研究とは異なり,「未知語」の結びつけの際には「未知語」が別の辞書において持つ関係性ではなく,「未知語」に与えられた定義文を用

いることが期待されている．

評価データとしては，(「未知語」，品詞，定義文，正解概念）のセットが 1000 組与えられ，うち 600 組がテスト，400 組が学習用に用意されている．ここで扱う「未知語」は名詞と動詞に限っているが，その中には複数単語から成る慣用句的表現も含んでいる．

本タスクでは，手法によって得られた候補概念の正当性を，正解概念との同義性によって評価する．その際用いられる**同義性の尺度**としては, PWN で定義された概念間の上位-下位階層構造を用いて算出される, **Wu&P Similarity**[44] を用いる．2 つの概念間の Wu&P Similarity は, 次の式で与えられる.

$$wup(s_1, s_2) = \frac{2 * depth(LCS)}{depth(s_1) + depth(s_2)}$$

ここで, ある概念の深さ（depth）は, PWN で定義された概念間の上位-下位階層構造において, 最上位の概念からの最短経路長によって定義される．すなわち, depth は各概念の具体性 (specificity) の尺度となる．また, LCS（Least Common Subsumer）は, 2 つの概念に共通する上位語の中で, 最も具体的なものである.

概念間の Wu&P Similarity が高いということは, 共通する上位語の具体度が高いということ, すなわち, 2 つの概念を具体度の高いカテゴリで表現できるということであるため, Wu&P Similarity を同義性の強さの指標として用いることができる.

比較対象としては，本タスクにおいて最高精度を達成している既存手法を用意する．この手法では，定義文中の単語が持つ語義の含まれる概念を候補の集合として用意し，それらに単語分散表現や品詞タグ等で適当な特徴を与えた後，最適なものが最上位に来るような並び替え（ランキング）を行う分類器を学習することで，適切な候補を選択している．

しかし，与えられた定義文中に，必ずしも適切な概念が含まれているわけではない．また，用いている素性にどのような働きを期待しているのかが不明瞭である．

そこで本稿では，未知語に対し，その定義文中に含まれる単語の分散表現を利

用してベクトルを与え，これと辞書中の単語・語義・概念に対し与えられた分散表現を比較することによって，定義文に縛られることなく候補概念の集合を収集する．

その上で，得られた候補概念間に成り立つと仮定した関係性をもとに，辞書上における概念間類以度を適切に利用することで候補を絞り込み，教師なしでも従来手法と同程度の性能が得られるような手法を提案する．

4.2　提案手法

本稿提案する手法は，以下の 3 つのステップからなる．

候補概念の収集: 未知語定義文と概念の分散表現を用いて，候補となる概念を収集する．

クラスタリングによる候補概念の絞り込み: 互いに同義性の高い候補概念をグループ化し，そのうちで最も正解が含まれる期待が高いクラスタを選択することにより，候補選出の範囲を絞り込む．

最適な候補の選択: 選択されたクラスタから，そのクラスタに含まれる概念の意味を最も端的に表していると思われる概念を選択する．

以下の節で, それぞれの手順を説明する．

4.2.1　候補概念の収集

単語の分散表現の適用範囲は，その学習に使用するコーパスの語彙サイズによって規定される．一般に，これは辞書資源の語彙サイズに比べて遥かに大きい. 例えば, PWN に含まれる単語数は 15 万語程度（147,306 単語）であるが, Google 配布の Word2Vec[1]モデルでは, 300 万単語に対し分散表現を与えることが出来ている.

[1] https://drive.google.com/file/d/0B7XkCwpI5KDYNlNUTTlSS21pQmM/edit?usp=sharing

つまり, 本研究で対象とするような未知語であっても, その定義文が与えられれば, 分散表現を与えることが十分期待できる.

よって, 定義文中に含まれる単語の分散表現を用いてベクトルを構成し（以下, 未知語ベクトルと表記）, これと辞書中の単語・語義・概念に対し与えられた分散表現を比較することで概念を収集する. 具体的には, 以下の 3 通りの方法で概念を収集する.

word: 未知語ベクトルの m 近傍にベクトルを持つ単語が持つ語義の指示する概念のうち, 未知語と品詞が一致するもの上位 k 件を収集する.

lexeme: 未知語ベクトルの m 近傍にベクトルを持つ語義の指示する概念のうち, 未知語と品詞が一致するもの上位 k 件を収集する.

synset: 未知語ベクトルの m 近傍にベクトルを持つ概念のうち, 未知語と品詞が一致するもの上位 k 件を収集する.

ここで, 分散表現（ベクトル）間の距離尺度はコサイン類以度とする. また, k=20 とし, m はそれに合わせて適当に変化させる. これに未知語定義文中の単語に対応する概念を加えて, 候補概念集合とする.

なお, 未知語ベクトルは, 定義文中の各単語に対応する分散表現の重み付き和によって構成する. ここで, 各単語の重みは, 定義文を JoBimText[2] によって dependency parse した際に得られる, root からの深さの逆数とする. ここで深さは root を 0 とし, root の単語の分散表現に対する重みは 1 とする.

定義文だけではなく, 未知語自身に対しても単語分散表現を構成することは可能であるが, 今回は使用しない. なぜなら, 今回扱う未知語には複数語からなる慣用句的表現（未知語句）も含まれているが, それらの句が持つ意味は句中に含まれている単語と大きくかけ離れるため, 単に含まれている語の意味を単純に足

[2] http://ltmaggie.informatik.uni-hamburg.de/jobimtext/

し合わせただけではその意味をベクトル中で適切に表現できないことが想定されるためである.

4.2.2　クラスタリングによる候補概念の絞り込み

候補概念は, その分散表現と未知語ベクトルとのコサイン類以度（関連性の尺度）を基準にして収集されている. 関連性には, 同義性だけでなく様々な意味的関係性が含まれているため, 未知語ベクトルの近傍に存在する概念（未知語と関連する概念）であっても, 正解概念との同義性が低いということが起こりうる. よって,「未知語ベクトルに対し, 最近傍の概念を選ぶ」といった単純な手法で, 適切な概念を選択することは難しいことが想定される.

つまり, 適切な概念を選択するためには, 概念候補集合から「未知語ベクトルとの類似性（関連性）は高いが, 正解概念との同義性は低いような概念」を排除し, より適切な概念を選別する必要がある.

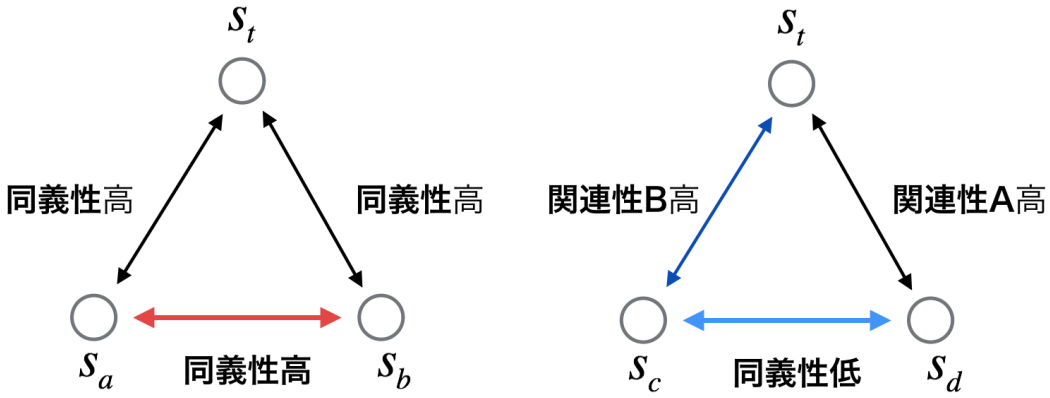


図 4.1:　概念間に成り立つと仮定される関係性

ここで候補概念集合中に混在することが仮定される,「未知語との関連性が高く, 正解概念との同義性も高い概念（S_{sim}）」と,「未知語との関連性は高いが, 正解概念との同義性は低い概念（S_{dis}）」のそれぞれについて, 次のことが成り立つと仮定できる（図 4.1）.

- S_{sim} であるような概念間の同義性は高い.
- S_{dis} は, 同義性以外の様々な関連性によって対象の未知語と結びついているため，S_{dis} であるような概念間の同義性は比較的低い.
- S_{sim} と S_{dis} は, それぞれ異なる関係性によって未知語と結びついている. よって, S_{sim} と S_{dis} 間の同義性は低い.

これらの関係性が成り立つと仮定し，次の 2 段階の手順で候補概念を絞り込む.

(1) 候補概念集合のクラスタリング　集合中の概念について，同義性（Wu&P Similarity）が高いものをグループ化することで，$Ssim$ を多く要素として持つクラスタが現れることを期待する．具体的には, Wu&P Similarity を類以度尺度として，凝集型クラスタリングを行う．ただし，凝集の打ち止めを閾値によって行うのではなく，凝集の回数によって行う．これによって，過度な凝集を防ぐことができる．今回は，10 回凝集を行った段階で，クラスタリングを打ち止めた.

(2) 最適クラスタの選択　得られた複数のクラスタと対象の未知語との間に，以下に示す尺度 :「重要度」を設定し，これが一番高くなるようなクラスタを選択することで，適切に候補が絞り込む．この尺度は，分散表現の類似性を見るだけでなく，概念間の同義性を考慮している.

$$imp(cls_i) = \max_{s_j \in\ cls_i, s_t \in\ def_{unk}} (\ \alpha\ * wup(s_j, s_t) + \ \beta\ * cos(\vec{s_j}, \vec{s_t}))$$

ここで def_{unk} は, 未知語の定義文に含まれる単語が結びついた概念の集合である. また，α，β は定数であり，ここでは $\alpha = 2$, $\beta = 1$ とする.

4.2.3　最適な候補の選択

選択されたクラスタ中から候補概念を選択する基準として，次の 2 通りを比較する.

- **dist**：関連性に注目した基準である．クラスタ中で，未知語ベクトルの最近傍に分散表現を持つ概念を選択する．

- **center**：クラスタ内での意味的代表性に注目した基準である．クラスタ中の概念をノード, 概念間の Wu&P Similarity をエッジの重みと見た無向グラフを考え, その中心に位置する概念が最も代表的な意味を持つことを期待し, 中心性の最も高い概念を選択する．ここで, クラスタ cls_i に含まれる概念 s_j の中心性 (centrality) は次の式で定義される．

$$centrality(s_j) = \frac{\sum_{s_k \in\ Cluster} wup(s_k, s_j)}{size(cls_i)}$$

4.3　実験

4.3.1　実験設定

SemEval 2016 task14 の評価データ 600 組のうち, AutoExtend によって正解の概念に分散表現を与えることが出来ているような 411 組に対して提案手法を適用し，Wu&P Similarity によって候補概念と正解概念の類似性を評価する．ここでは，提案手法におけるクラスタリングおよびクラスタ選択の有効性を確認するため，それらを行わずに得られた集合から候補を選ぶ場合と比較を行う．

4.3.2　実験結果

	word	**lexeme**	**synset**	既存手法 [39]
dist	0.467	0.498	0.500	0.523
center	0.472	0.506	**0.512**	

表 4.1: 各手法により得られた候補概念と正解概念の平均 Wu&P Similarity

主な実験結果を表 4.1 に示す．単語の分散表現のみを用いて候補概念集合を構成した場合（**word**）よりも，語義・概念の分散表現を用いて候補概念集合を構成し

	w/o clustering	w/o selection	**synset**
dist	0.469	0.493	0.500
center	0.487	0.474	**0.512**

表 4.2: **synset** 設定において各手順を省いた結果との比較

た場合（**lexeme**, **synset**）の方が高い精度が得られていることが分かる．このことから，語義・概念の分散表現の有効性が確認できる．

また，実験に用いているサンプル数が異なるため厳密な比較はできないが，今回得られた最高スコアは，教師あり学習によって最高精度を達成している既存手法 [39] に迫る性能となることが確認できた．

synset 設定において，各手順を省いた結果を図 4.2 に示す．いずれの候補選択基準を使用する場合においても，クラスタリングとクラスタ選択を行った場合において最もスコアが高くなっている．また，未知語定義文のベクトルとの類以性（関連性）を利用した選択基準 (**dist**) よりも，PWN のネットワーク上で定義される類似性（同義性）を利用した選択基準 (**center**) の方が，よい結果が得られたことから，同義性を限定して扱いたい場合に，辞書資源の構造を利用することの有効性が示唆された．

4.4　議論

本研究で提案する手法は，(1) 候補概念の収集 (2) クラスタリング (3) クラスタ選択という 3 つの手順によって選択する候補となる概念（候補概念）を絞り込んでいる．

ここでは，まず（１）の手順で適切に候補概念が収集できているかを確認した後,（１）,（２）,（３）の手順で得られた候補概念集合を比較することで，適切に候補の絞り込みが行えているかを確認する．

候補概念の収集　提案手法においてクラスタリングを行う際は，正解概念と同義性の高い概念が収集された候補中に複数含まれていることを前提としている．よって,（1）の手順で収集された候補概念集合がこの条件が満たされていることを確認する必要がある．そこで，テストデータ中の各未知語に対して収集した候補概念集合のうち，**正解概念との Wu&P Similarity が高い（0.5 以上の）概念が複数含まれているものの割合 (PossibleRatio)** を調べることで，この手順の有効性が確認できる．その結果は図 4.3 のとおりである．

	word	lexeme	synset
PossibleRatio	0.786	0.856	0.869

表 4.3: テストデータ中の各未知語に対して収集した候補概念集合のうち，正解概念との Wu&P Similarity が高い（0.5 以上の）概念が複数含まれているものの割合

lexeme, synset 設定時の割合が **word** よりも高くなっていることから，語義・概念の分散表現の有効性が確認できる．また，**synset** 設定時にはテストデータの大部分（約 87 % ）に対して提案手法が機能することが期待できる．

候補概念の絞り込み　候補概念を絞り込む際には，選択するのに適切な，すなわち正解概念と Wu&P Similarity の高い概念をできるだけ多く残しながら，候補数を削減することが望まれる．すなわち，**PossibleRatio** を維持しながら，**正解概念との Wu& P Similarity が高い（0.5 以上の）概念の割合（Propotion)** を増やしていくことが望まれる．

また，最終的に得られた集合中から候補概念を選択するため，集合中に含まれる概念と正解概念との Wu&P Similarity の最大値によって，提案手法の性能上限が決定される．すなわち，絞り込みの際には**正解概念との Wu& P Similarity の最大値（UpperBound)** が高い値を維持していることが望まれる．

よって，(1) 候補概念の収集 (2) クラスタリング (3) クラスタ選択の各段階で得られた集合中の概念に対し，

- **PossibleRatio**
- **Propotion**
- **UpperBound**

を算出，比較することで各手順の有効性を確認する．ここでは **synset** 設定時を取り上げ，それぞれの平均を 4.4 表に示す．

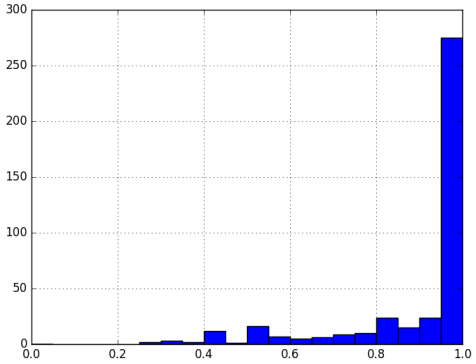
	UpperBound	Propotion	PossibleRatio
候補収集	0.901	0.244	0.869
クラスタリング	0.730	0.346	0.720
クラスタ選択	0.541	0.416	0.426

表 4.4: **synset** 設定時の各手順における平均結果

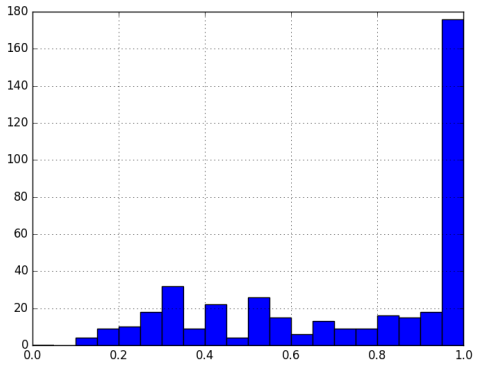
Propotion に関しては，手順を経るにつれて期待通り増加している事がわかる．一方で **UpperBound, PossibleRatio** に関しては, 手順を経るにつれて明らかな減少が見られる．特にクラスタ選択の段階で，**PossibleRatio** の減少が顕著である．このことは，得られたクラスタの中から不適切なもの（正解概念との同義性が低い概念の集合）を選択してしまったことを意味する．

以下では，各手順における **UpperBound, Propotion** に関して，その分布を図 4.2, 4.3 に示す．

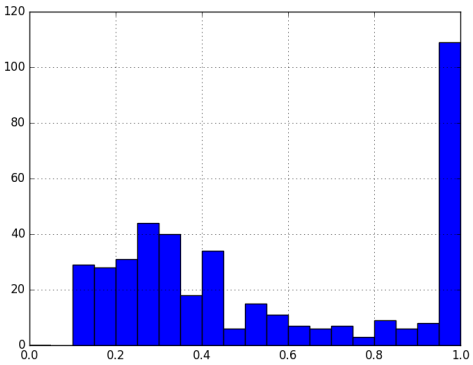
Propotion に関しては，(1) 概念収集時には低い値を取っていたサンプルが (2) クラスタリングによって高い値を取るようになっていることが分かる（図 4.3，(1) と (2) を比較）．すなわち，クラスタリングによって概ね適切に候補概念を絞り込めていることが分かる．しかし (2) の手順においては (1) の手順に比べて候補中に正解概念との Wu&P Similarity の高い概念を殆ど含まないようなサンプル（**Propotion** が 0 に近いサンプル）も増えており，(3) の手順では **Propotion** が 0 に近いものと 1 に近いもので 2 極化している（図 4.3, (3)）．クラスタ選択を行った段階で，全てのサンプルにおいて **Propotion** が 1 となるのが理



(1) 概念収集



（２）クラスタリング



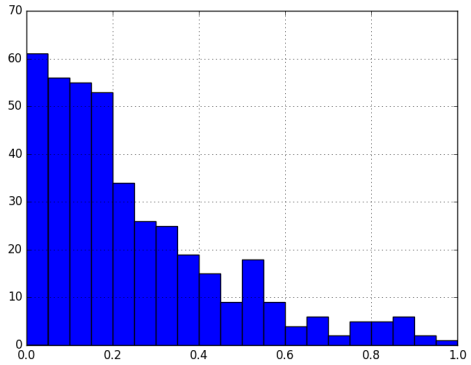
（３）クラスタ選択

図 4.2: テストデータに **synset** の各手順を適用した際の **UpperBound** の分布

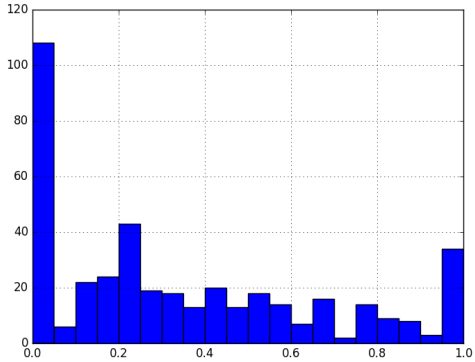
想であるが，提案手法は一部のサンプルにおいて理想的に機能しているものの，そうでないサンプルも多く存在していることが分かる．

UpperBound に関しては，手順 (1), (2) でさほど大きな変化はない（図 4.2, (１) と (2) を比較）．しかし，手順（３）においては，UpperBound の低い部分に山ができていることがはっきりと分かる（図 4.2, (3) ）．

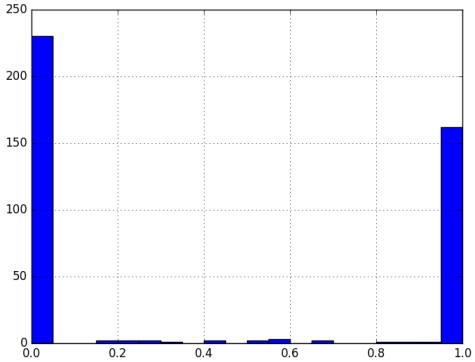
提案手法は最終的に教師ありの既存手法に迫る精度を達成することが出来たが，特にクラスタ選択の手順においては，「重要度」の設定に改善の余地があることが確認できた．



(1) 概念収集



（２）クラスタリング



（３）クラスタ選択

図 4.3: テストデータに **synset** の各手順を適用した際の **Propotion** の分布

第5章　結論

本研究では，AutoExtend により作成された語義・概念の分散表現を単語間意味関係分類，および Semantic Taxonomy Enrichment のタスクに適用する手法を提案した．それぞれの結果については次のとおりである．

単語間意味関係分類：　与えられた単語ペア中の各単語に対して PWN 中で定義されたネットワークを利用して語義・概念を結びつけ，それらの類以度，および類以度算出時に使用したベクトルを素性とした教師付き学習による単語間意味関係の分類手法を提案した．提案手法を BLESS 評価データ [2] に適用したところ，単語の分散表現のみを利用した既存手法を上回る精度が確認できた．

Semantic Taxonomy Enrichment：　与えられた未知語に対し，(1) 単語の持つ語義・概念に対する分散表現 [35] と未知語に与えられた定義文を利用して候補となる概念集合を導出し，(2) PWN の辞書構造を利用したクラスタリングを行い候補を絞り込み，(3) 残った候補を辞書構造を用いた基準によりランキングすることにより，教師データを必要とすることなく，候補として適切な概念を選択する手法を提案した．

評価データ [15] に適用した実験結果から，クラスタリングにより候補概念を絞り込むことの有効性を確認した．また，最良の条件において教師あり学習を行う従来手法 [39] に迫る精度が得られた．

今回扱った 2 つのタスクそれぞれにおいて，語義・概念の分散表現の有効性が確認できたといえる．残された課題とそれに対するアプローチには，次のことが

考えられる．

単語間意味関係分類：　提案手法においては，与えられた単語ペア間に成立する関係性をいくつか想定した上で，事前知識（WordNet）を用いて分類に用いる素性を構成しているが，今回の評価実験においては，事前に成立を想定した関係性と実際に分類を行う関係性の粒度が大きく変わることはなかった．

より細分化された関係性を分類するようなデータセットに対しても適用・評価を行うことで，提案手法の更なる有効性が検証できると考えられる．

Semantic Taxonomy Enrichment：　提案手法では，候補集合中に候補として選ぶのに適切な概念が複数含まれていることを想定した上で，凝集型クラスリングを行い，選択候補を絞り込んでいる．

収集された候補概念の大半については，その中に適切な（正解概念と同義性の高い）概念が複数含まれていることが確認されたが，そうでないものも存在した．句にも対応可能な形で作成した未知語の分散表現や，辞書中の概念に与えられた定義文をもとに作成したベクトルなどの，より多くの手がかりをもとに候補概念を探索することで，より適切な候補概念を収集することが期待できる．

また，候補を絞り込む段階では，特にクラスタリング後の最適なクラスタを絞り込む際に改善の余地が見られた．分散表現・辞書の構造から得られる類以度だけでなく，分散表現から得られるクラスタのセントロイド等を素性として既存手法 [39] と同様に教師ありのランキング学習を行うことで，精度の改善が期待できる．

Semantic Taxonomy Enrichment のタスクに関しては，既存の辞書の構造を利用して評価することになるが，そもそも手動で作られた辞書は完全でないという問題がある．すなわち，与えられた未知語に対して適切な概念を結びつける際，実際は正解と同義性の高い概念を収集・選択できているのに，それらと正解概念間の経路が辞書上で定義されていないために不当に精度が下がってしまうというこ

とが起こり得る.

このことから今後は,

1. 概念に対し与えられた分散表現を利用して「関連性は高いが関係性は結ばれていない」ような概念ペアを収集し,
2. それらに成立しうる関係性を自動的に分類･アノテーションして関係性のネットワークをを充実させた辞書を作った上で,
3. Semantic Taxonomy Enrichment に取り組む

ことが重要であると考える.

謝辞

研究に際し，様々な議論にあたって多くの時間を頂き，綿密な御指導，御助言を頂いた林良彦教授に，心より感謝申し上げます.

また，研究方針やプレゼンテーション方法について御指導，御助言を頂いた小林哲則教授に，心より感謝申し上げます.

参考文献

[1] Timothy Baldwin, Marie-Catherine de Marneffe, Bo Han, Young-Bum Kim, Alan Ritter, and Wei Xu. Shared tasks of the 2015 workshop on noisy user-generated text: Twitter lexical normalization and named entity recognition. In *Proceedings of the Workshop on Noisy User-generated Text*, pp. 126–135, 2015.

[2] Marco Baroni and Alessandro Lenci. How we blessed distributional semantic evaluation. In *Proceedings of the GEMS 2011 Workshop on GEometrical Models of Natural Language Semantics*, pp. 1–10. Association for Computational Linguistics, 2011.

[3] Sergey Bartunov, Dmitry Kondrashkin, Anton Osokin, and Dmitry Vetrov. Breaking sticks and ambiguities with adaptive skip-gram. In *Artificial Intelligence and Statistics*, pp. 130–138, 2016.

[4] Antoine Bordes and Jason Weston. Embedding methods for natural language processing, 2014.

[5] Danqi Chen and Christopher Manning. A fast and accurate dependency parser using neural networks. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pp. 740–750, 2014.

[6] Xinxiong Chen, Zhiyuan Liu, and Maosong Sun. A unified model for word sense representation and disambiguation. In *Proceedings of the 2014 Confer-*

ence on Empirical Methods in Natural Language Processing (EMNLP), pp. 1025–1035, Doha, Qatar, October 2014. Association for Computational Linguistics.

[7] Cicero dos Santos and Maira Gatti. Deep convolutional neural networks for sentiment analysis of short texts. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pp. 69–78, Dublin, Ireland, August 2014. Dublin City University and Association for Computational Linguistics.

[8] Andrea Esuli and Fabrizio Sebastiani. Sentiwordnet: a high-coverage lexical resource for opinion mining. *Evaluation*, pp. 1–26, 2007.

[9] MS Fabian, K Gjergji, WEIKUM Gerhard, et al. Yago: A core of semantic knowledge unifying wordnet and wikipedia. In *16th International World Wide Web Conference, WWW*, pp. 697–706, 2007.

[10] Lev Finkelstein, Evgeniy Gabrilovich, Yossi Matias, Ehud Rivlin, Zach Solan, Gadi Wolfman, and Eytan Ruppin. Placing search in context: The concept revisited. In *Proceedings of the 10th international conference on World Wide Web*, pp. 406–414. ACM, 2001.

[11] John R Firth. A synopsis of linguistic theory, 1930-1955. *Studies in linguistic analysis*, 1957.

[12] Felix Hill, Roi Reichart, and Anna Korhonen. Simlex-999: Evaluating semantic models with (genuine) similarity estimation. *Computational Linguistics*, Vol. 41, No. 4, pp. 665–695, 2015.

[13] Eric Huang, Richard Socher, Christopher Manning, and Andrew Ng. Improving word representations via global context and multiple word prototypes. In

Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 873–882, Jeju Island, Korea, July 2012. Association for Computational Linguistics.

[14] Ignacio Iacobacci, Mohammad Taher Pilehvar, and Roberto Navigli. Sensembed: Learning sense embeddings for word and relational similarity. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 95–105, Beijing, China, July 2015. Association for Computational Linguistics.

[15] David Jurgens and Mohammad Taher Pilehvar. Semeval-2016 task 14: Semantic taxonomy enrichment. In *SemEval@ NAACL-HLT*, pp. 1092–1102, 2016.

[16] Kentaro Kanada, Tetsunori Kobayashi, and Yoshihiko Hayashi. Classifying lexical-semantic relationships by exploiting sense/concept representations. In *Proceedings of the 1st Workshop on Sense, Concept and Entity Representations and their Applications*, pp. 37–46, Valencia, Spain, April 2017. Association for Computational Linguistics.

[17] Adam Kilgarriff. English lexical sample task description. In *The Proceedings of the Second International Workshop on Evaluating Word Sense Disambiguation Systems*, pp. 17–20. Association for Computational Linguistics, 2001.

[18] Angeliki Lazaridou, Elia Bruni, and Marco Baroni. Is this a wampimuk? cross-modal mapping between distributional semantics and the visual world. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1403–1414, Baltimore, Maryland, June 2014. Association for Computational Linguistics.

[19] Omer Levy and Yoav Goldberg. Dependency-based word embeddings. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 302–308, Baltimore, Maryland, June 2014. Association for Computational Linguistics.

[20] Omer Levy and Yoav Goldberg. Neural word embedding as implicit matrix factorization. In *Advances in neural information processing systems*, pp. 2177–2185, 2014.

[21] Omer Levy, Steffen Remus, Chris Biemann, and Ido Dagan. Do supervised distributional methods really learn lexical inference relations? In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 970–976, Denver, Colorado, May–June 2015. Association for Computational Linguistics.

[22] Wang Ling, Chris Dyer, Alan W Black, and Isabel Trancoso. Two/too simple adaptations of word2vec for syntax problems. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 1299–1304, Denver, Colorado, May–June 2015. Association for Computational Linguistics.

[23] Rada Mihalcea, Timothy Chklovski, and Adam Kilgarriff. The senseval-3 english lexical sample task. In *Proceedings of SENSEVAL-3, the third international workshop on the evaluation of systems for the semantic analysis of text*, 2004.

[24] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

[25] George A Miller. Wordnet: a lexical database for english. *Communications of the ACM*, Vol. 38, No. 11, pp. 39–41, 1995.

[26] Andrea Moro, Alessandro Raganato, and Roberto Navigli. Entity linking meets word sense disambiguation: a unified approach. *Transactions of the Association for Computational Linguistics*, Vol. 2, pp. 231–244, 2014.

[27] Roberto Navigli. Word sense disambiguation: A survey. *ACM Computing Surveys (CSUR)*, Vol. 41, No. 2, p. 10, 2009.

[28] Roberto Navigli and Simone Paolo Ponzetto. Babelnet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, Vol. 193, pp. 217 – 250, 2012.

[29] Silvia Necsulescu, Sara Mendes, David Jurgens, Núria Bel, and Roberto Navigli. Reading between the lines: Overcoming data sparsity for accurate classification of lexical relationships. In ** SEM@ NAACL-HLT*, pp. 182–192, 2015.

[30] Arvind Neelakantan, Jeevan Shankar, Alexandre Passos, and Andrew McCallum. 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, Doha, Qatar, October 2014. Association for Computational Linguistics.

[31] Paula Pääkkö, Krister Lindén, et al. Finding a location for a new word in wordnet. In *Proceedings of the Global Wordnet Conference*, 2012.

[32] Maria Pelevina, Nikolay Arefiev, Chris Biemann, and Alexander Panchenko. Making sense of word embeddings. In *Proceedings of the 1st Workshop on Rep-*

resentation Learning for NLP, pp. 174–183, Berlin, Germany, August 2016. Association for Computational Linguistics.

[33] Jeffrey Pennington, Richard Socher, and Christopher Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1532–1543, Doha, Qatar, October 2014. Association for Computational Linguistics.

[34] Mohammad Taher Pilehvar and Nigel Collier. De-conflated semantic representations. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 1680–1690, Austin, Texas, November 2016. Association for Computational Linguistics.

[35] Sascha Rothe and Hinrich Schütze. Autoextend: Extending word embeddings to embeddings for synsets and lexemes. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 1793–1803, Beijing, China, July 2015. Association for Computational Linguistics.

[36] Dwaipayan Roy, Debjyoti Paul, Mandar Mitra, and Utpal Garain. Using word embeddings for automatic query expansion. *arXiv preprint arXiv:1606.07608*, 2016.

[37] Pavel Rychlý and Adam Kilgarriff. An efficient algorithm for building a distributional thesaurus (and other sketch engine developments). In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pp. 41–44, Prague, Czech Republic, June 2007. Association for Computational Linguistics.

[38] Enrico Santus, Frances Yung, Alessandro Lenci, and Chu-Ren Huang. Evalution 1.0: an evolving semantic dataset for training and evaluation of distributional semantic models. In *Proceedings of the 4th Workshop on Linked Data in Linguistics (LDL-2015)*, pp. 64–69, 2015.

[39] Michael Sejr Schlichtkrull and Héctor Martínez Alonso. Msejrku at semeval-2016 task 14: Taxonomy enrichment by evidence ranking. In *SemEval@ NAACL-HLT*, pp. 1337–1341, 2016.

[40] Fei Tian, Hanjun Dai, Jiang Bian, Bin Gao, Rui Zhang, Enhong Chen, and Tie-Yan Liu. A probabilistic model for learning multi-prototype word embeddings. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pp. 151–160, Dublin, Ireland, August 2014. Dublin City University and Association for Computational Linguistics.

[41] Peter D Turney and Patrick Pantel. From frequency to meaning: Vector space models of semantics. *Journal of artificial intelligence research*, Vol. 37, pp. 141–188, 2010.

[42] Giannis Varelas, Epimenidis Voutsakis, Paraskevi Raftopoulou, Euripides GM Petrakis, and Evangelos E Milios. Semantic similarity methods in wordnet and their application to information retrieval on the web. In *Proceedings of the 7th annual ACM international workshop on Web information and data management*, pp. 10–16. ACM, 2005.

[43] Ivan Vulić and Marie-Francine Moens. Monolingual and cross-lingual information retrieval models based on (bilingual) word embeddings. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 363–372. ACM, 2015.

[44] Zhibiao Wu and Martha Palmer. Verbs semantics and lexical selection. In *Proceedings of the 32nd annual meeting on Association for Computational Linguistics*, pp. 133–138. Association for Computational Linguistics, 1994.

[45] Ichiro Yamada, Jong-Hoon Oh, Chikara Hashimoto, Kentaro Torisawa, Jun'ichi Kazama, Stijn De Saeger, and Takuya Kawada. Extending wordnet with hypernyms and siblings acquired from wikipedia. In *IJCNLP*, pp. 874–882, 2011.

[46] Zhi Zhong and Hwee Tou Ng. It makes sense: A wide-coverage word sense disambiguation system for free text. In *Proceedings of the ACL 2010 system demonstrations*, pp. 78–83. Association for Computational Linguistics, 2010.

[47] 金田健太郎, 小林哲則, 林良彦. 語義・概念ベクトルによる意味タスクの精度向上. 言語処理学会第 22 回年次大会 (NLP2016), pp. 1069–1072, 2016.

[48] 金田健太郎, 小林哲則, 林良彦. 語義・概念の分散表現を利用した単語間の意味関係分類. 言語処理学会第 23 回年次大会 (NLP2017), 2017.

[49] 金田健太郎, 小林哲則, 林良彦. 単語, 語義, 概念: 意味タスクにおける分散表現の適用性. 人工知能学会全国大会（第 31 回） (JSAI2017), 2017.

[50] 金田健太郎, 小林哲則, 林良彦. 語義・概念の分散表現を利用した semantic taxonomy enrichment. 言語処理学会第 24 回年次大会 (NLP2018), 2018.