2B3-OS-07a-2

単語，語義，概念: 意味タスクにおける分散表現の適用性

Investigating the applicability of word/sense/concept embeddings in semantic tasks

金田 健太郎 *1 Kentaro Kanada
小林 哲則 *1 Tetsunori Kobayashi
林 良彦 *1 Yoshihiko Hayashi

*1早稲田大学理工学術院
Faculty of Science and Engineering, Waseda University

We discuss the applicability of sense/concept embeddings generated from existing word embeddings (Word2Vec) and a lexical resource (Princeton WordNet) by the AutoExtend system to some kind of word-level semantic tasks. We found these embeddings outperformed word embeddings, especially in tasks such as estimating semantic word similarity and classifying lexical-semantic relation between words. On the other hand, there are two major problems with the AutoExtend system: the disability to generate sense/concept embeddings for words not listed in PWN, and the difficulty to assign proper embeddings for minor senses. We discuss possible solutions for these problems.

1. はじめに

近年，Word2Vec[7] や Glove[11] に代表される単語の分散表現獲得手法が注目されており，様々な自然言語処理のタスクへの適用が行われている．なかでも，言葉の意味を扱う意味タスクに対する適用性の議論が盛んである．多くの分散表現獲得手法においては，多義語の場合にも 1 つの単語に 1 つの分散表現を割り当てるため，各語義に対応した分散表現を直接得ることはできない．また，このような語義が混合 (conflate) した分散表現においては，分散表現を獲得するために用いたコーパスにおいて支配的な意味が強く現れるという特徴がある．これらの課題に対し，各単語が持つ語義に対して適切に分散表現を与えることができれば，語義曖昧性解消，概念間の意味的関連度の定量化などの意味タスクにおいて有用であると考えられる．実際，このような観点から，単語の持つ意味 (語義)，また同じ意味を持つ語義をまとめた概念に対し分散表現を与えることが最近試みられている．

本稿では，AutoExtend と呼ばれる手法 [12] により得られた語義・概念の分散表現を，単語間意味関係分類，及び単語間類以度・関連度の定量化，概念間の想起関係の予想といった意味タスクに適用した結果に基づき，単語に対する分散表現と比較した際の利点や問題点，また問題点に対する解決法について議論する．

2. 語義・概念に対する分散表現

2.1 代表的な手法

語義・概念に対し分散表現を与える手法は概ね以下の 2 種類に大別される．

一つは単語の語義を帰納的に導出し，それらに対して分散表現を与える手法である．例えば，意味の分布仮説に基づき，ある単語が出現する文脈をクラスタリングし，そのクラスタそれぞれを単語の語義と解釈する [10, 4]．このような手法においては，理論上コーパスに出現する全単語に対し，その語義に相当するベクトルを与えることができる．しかし，各クラスタが実際にどのような意味を持つかを解釈するのは難しいことが多い．また，既存の言語資源との対応を取ることができないため，その適用範囲に問題がある．

もう一方は，人手で作られた辞書資源中で定義された単語の意味 (語義) に対し，分散表現を与える手法である [12, 5]．この方法では，各語義は専門家により作成された辞書で定義されているため解釈の必要性はない．また，辞書資源と対応付けられていることから幅広いタスクに適用可能である．さらに 3.2 節で述べるように，辞書資源において定義されている語義間，概念間の関係をタスク適用において利用することもできる．

2.2 AutoExtend

我々は，語義・概念に対して分散表現を与える手法として，AutoExtend と呼ばれる手法 [12] を用いている．この手法は，既存の単語分散表現を入力とし，既存の辞書資源における単語・語義・概念のネットワーク構造を反映したオートエンコーダにより，当該の辞書資源における語義・概念に対して分散表現を与える．

より具体的には，AutoExtend におけるオートエンコーダは，単語はいくつかの語義を持ち，同一の意味を持つ語義の集合として概念が構成されるという，辞書資源の基本構造に基づき構成される．ここで，単語の分散表現はその単語の持つ語義の分散表現の総和であり，概念はそれを指示する語義の分散表現の総和であると定式化される．オートエンコーダにおける学習は，入力の単語の分散表現と再構成される単語の分散表現の間の誤差，および，中間レベルにある語義の分散表現の間の誤差を最小化するように行う．

このような学習手法をとることにより，辞書資源中の語義・概念に対して，単語の分散表現と各次元の意味合いが等しい同次元の分散表現を与えることができる．これにより，単語と語義，語義と概念といった異なるレベルの分散表現を直接比較することが可能となる．このような特徴は，辞書資源を用いる他の手法 [5] では得られない．また AutoExtend では既存の単語の分散表現を利用できることから，資源の再利用や計算量という観点からも有利である．

3. 語義・概念の分散表現のタスク適用評価

AutoExtend の原著論文 [12] においては，文脈下における単語類似度 (contextual word similarities)，および，語義曖昧性解消という 2 つの意味タスクにより，提案手法の有効性を評価している．我々は，さらに以下の 3 つの意味タスクに適用することにより，AutoExtend による語義・概念の分散表現の有効性について検討してきた．これらの適用において

は，Word2Vec による単語分散表現 [*1] (CBOW, 300 次元) と Princeton WordNet (以下では PWN) を用いて作成された語義・概念の分散表現を用いており，Word2Vec による単語の分散表現と比較，もしくは，併用することにより，AutoExtend により得られた分散表現の有効な利用法を探ってきた．その結果，多くの意味タスクにおける有用性を確認するとともに，5 節で議論するような問題点があることも明らかとなった．

3.1 単語間の意味的類以度・関連度の定量化

単語間の意味的類以度 (similarity)・関連度 (relatedness) の定量化タスクは，単語の分散表現を評価する際に最も基本的な意味タスクである．ここで，類以度は単語間の同義性の程度，関連度は同義性以外の関係性 (例: 上位・下位関係) も含めた意味的関連性の強さの程度である．このタスクに語義・概念の分散表現を適用し，単語の分散表現の適用結果と比較することでその有効性を確認した [13].

より具体的には，語義・概念の分散表現を用いて単語間の類以度・関連度を表すため，単語を，その単語が持つ語義，あるいは，その語義の属する概念の集合とみなし，それらの集合を用いて類以度・関連度を算出した．テストセットに対して得られた意味的類以度・関連度と人手により付与されたスコア (gold data) との相関により評価を行った．

その結果，単語間の意味的類以度については，概念の分散表現を用いた場合に語義の分散表現や単語の分散表現を上回る結果が得られた．一方で，単語間の意味的関連度については，語義・概念双方ともに単語の分散表現を上回る結果は確認できなかった．単語の意味的類似性 (同義性) は，単語の持つ特定の語義 (概念) の間に定義されるものであるのに対し，単語の意味的関連性は，単語の持つ語義を包括的に扱うことが必要であると考えられる．与えられた単語ペアが相互に曖昧性解消を行う可能性を考えれば，上記の結果は妥当であるといえる．

3.2 単語間の意味関係の分類

単語間の意味関係の分類は，単語間に成立しうる意味関係を分類する意味タスクである．このタスクでは，与えられた単語ペアに対してその間に存在する意味関係を分類し，正解ラベルと比較することで評価を行った [14, 6].

実験においては，使用したテストセット (BLESS [1]) において設定されている 5 つの意味関係（hypernymy (上位-下位), co-ordinate (上位概念が共通), meronymy (全体-部分），attribute (被形容-形容), event (主体-動作)) について分類を行った．その際，例えば「単語ペア w_1, w_2 について，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」というように,「単語ペアに特定の意味関係が成立する可能性は，その意味関係に応じて各単語と関連付けられた語義・概念の集合間の類似度によって表される」と仮定して計算した類以度，及び類以度計算に使用したベクトルを素性とする教師付き学習を行った．

その結果，単語の分散表現のみを素性として教師付き学習を行う従来手法 [9] を上回るスコアが得られた．また，素性として単語の分散表現による類以度を入れた場合よりも抜いた場合の方がスコアが高かったことから，単語間意味関係分類のタスクにおいて，単語の分散表現はむしろノイズになってしまっていることが示唆された．

3.3 概念間の想起関係の予想

類似性・関連性よりも広い範囲をカバーする意味関係として連想，あるいは，想起関係がある [2]．[3, 15] は，教師つき学習のアプローチにより，概念 (具体的には PWN の synset) 間の想起関係の予測を試みた．想起関係の予想においては，想起の方向性を分類問題，想起の強さの予測を回帰問題として扱った．単語間の各種の類似度・関連度に加え，荒い意味分類，品詞，PWN の意味ネットワーク構造から得た素性のほか，Word2Vec による代表単語の分散表現の差分，および，AutoExtend による語義・概念の分散表現の差分を素性とする教師つき学習のアプローチにより，従来研究を上回る精度を得ている．実験結果からは，概念の分散表現の差分が両者のタスクにおいて有用であること，特に想起の方向性の決定における概念の分散表現の差分の有効性が高いことが示された．このことから，方向性の決定というカテゴリカルな決定においては，抽象度の高い概念レベルの分散表現がよりロバストであるといえる．

言語資源	語彙数
Princeton WordNet	147,306
Google news corpus	3,000,000

表 1: 各言語資源の語彙数

4. AutoExtend の問題点と解決法

タスクへの適用結果から，AutoExtend による語義・概念の分散表現は，単語の持つ特定の意味，あるいは別の単語との間に成立しうる特定の意味関係を扱う場合に有効であることが確認できた．しかしながら，それと同時に 2 つの問題点も確認された．それは，既存の辞書資源を利用することに起因する「未知語」の問題と，単語の分散表現を構成する際に用いるコーパスにおける仮想的な [*2] 語義の頻度分布における低頻出語義の問題である．

4.1 未知語に対する分散表現の割り当て

AutoExtend では，辞書に載っていない単語に対して語義・概念の分散表現を与えることが出来ない．表 1 に示すとおり，辞書資源でカバーできる語彙は大規模コーパスの 5%に満たず，単語の分散表現に比してその適用範囲が大幅に制限されてしまうという問題がある．この問題を解決するには，二通りの方法がある．1 つは，より大きな辞書資源を用いる方法である．近年，PWN を含む様々な言語資源を統合して巨大な辞書資源が作られており [8]，これらを用いることで多くの単語に対し語義・概念の分散表現を与えられることが期待できる．しかし，巨大な辞書であってもコーパス中の単語全てをカバーできるとは限らない．

もう一方で，分散表現を与えることが可能な辞書中の単語 (=「既知語」) とその語義・概念の分散表現を用いることにより，辞書に載っていない単語 (=「未知語」) を辞書中の概念に対して結びつける方法が考えられる．この手法では理論上コーパスの全単語に対し語義・概念を与えることが可能である．AutoExtend により得られる分散表現では異なるレベルの分散表現の比較であるため，未知語に対して概念を結びつける方法として，次の 3 通りの方法が考えられる (図 1).

1. 最近傍の既知語を見つけ，その単語が持つ語義が指示する概念 (以下，既知語の持つ概念と表記) を結びつける
2. 最近傍の語義を見つけ，それが指示する概念を結びつける
3. 最近傍の概念を見つけ，そこに直接結びつける

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

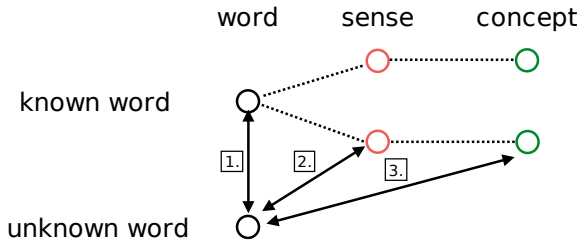

図 1: 未知語に対する概念の割り当て手法

1. は，単語と単語という同じレベルの意味表現を比較するため最も自然な手法である．ただし，この方法で適切な概念を割り当てることができるのは，対象の未知語と似た頻度分布で，1 つ以上の語義を持つ既知語が存在する場合に限られ，適用性に問題がある．

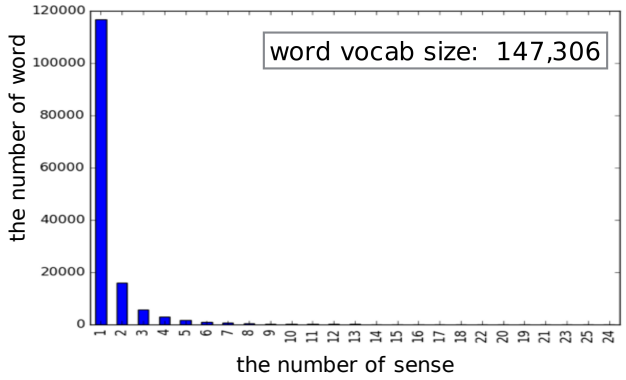

図 2: 各単語が持つ語義数のヒストグラム（上位 25 件）

2., 3. は，図 2 で示した PWN における語義数の分布（総単語の 8 割が単一の語義しか持たない）が未知語にも当てはまると仮定し，任意の未知語は単一の語義しか持たないことを前提としているが，辞書中の語義・概念一つ一つに対して類似性を評価するため，与えられた辞書においては最適な概念の割り当てが行えると考えられる．未知語に対して概念を割り当てる際に重要なのは,「各概念がどういった意味を持っているか」であり,「候補となる概念がどの単語の語義によるか」ということではない．このことから，2. よりも 3. の手法の方が，すなわち，最近傍の語義よりも最近傍の概念を見た方が適切に未知語への概念割り当てが行えると想定される．

ところで，各手法の妥当性を検証するためには，実際に未知語に対して手法を適用する前に，辞書中からある既知語を取り出し，これを未知語とみなした上で各手法を適用し，割り当てられた概念と実際の概念を比較することを繰り返すことが必要である．ここで，ある既知語を取り出す際には，その語義・概念の情報も未知とすることが望まれる．しかし，この場合は正解となる概念の情報が利用できないため，3. の手法に対する評価が行えない．また，1.，2. の手法についても評価が困難である．なぜなら，図 2 が示すように，PWN 上の概念の半数以上は単一の単語から成り立っているが，これは抜き出した既知語以外に正解が存在しないケースが全体の半数以上であることを示しており，事前に評価を行える概念が限定されてしまうためである．

*2 語義タグ付きコーパスを用いていないため，実際の語義の分布は不明である．

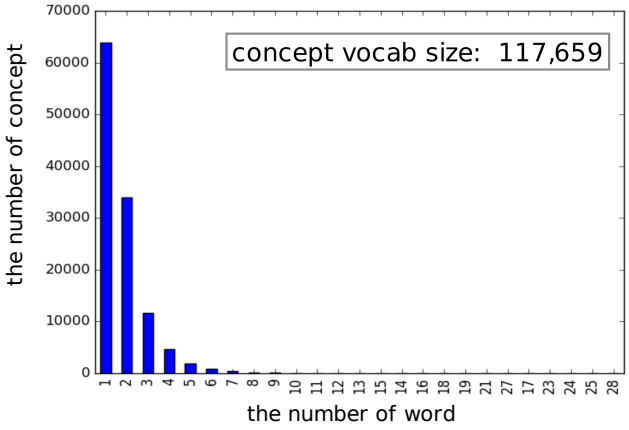

図 3: 各概念が持つ単語数のヒストグラム

また，ここで挙げた 3 手法は，未知語が任意に語義を持つことが想定されていない．すなわち，1. の手法では対象の未知語と同様の語義群を持つ既知語の存在が前提となっており，2.，3. の手法では，そもそも未知語が持つ語義が 1 つのみであると限定している．そこで，例えば，単語の分散表現を用いた類似度において，n 近傍に存在する単語が共通して持っている概念を割り当てることが考えられるが，PWN 上の概念の半数以上は単一の単語から成り立っている（図 3）ために，類似単語中に共通する概念が存在しない可能性が高い．

これらの問題を解決するためには，辞書 (PWN) 中の概念を適当に集約 (クラスタリング) し，これらの概念グループを利用することが考えられる．その際には，辞書に定義された上位-下位関係などによる階層構造を用いる方法と，概念の分散表現を用いる方法が考えられる．

4.2 低頻出の語義に対する分散表現の割り当て

AutoExtend において，単語の分散表現は，その単語の持つ語義の分散表現の和である．この考え方は，単語の分散表現に全ての語義の分散表現が含まれている場合，すなわち，コーパス中に辞書で定義された全ての語義が（仮想的に）出現している状況を仮定している．また，概念に関しても，その概念を指す語義が 1 つでも出現していることを仮定している．このことから，概念間の関係性を利用して表現を補正してはいるものの，ある多義語の全く出現しない/他の語義に比べて出現頻度の低い語義，あるいはそのような語義のみからなる概念は，適切な表現が得られないことが想定される．

例えば，長さの単位を指す概念である `hand.n.09` の近傍にある単語は図 4 に示すようになっている．これらの単語は，`hand.n.09` の指す意味を表しておらず，hand という単語において支配的に出現していると思われる，体の部位に関係している．こうした問題は，辞書中の語義が全て出現しているようなコーパスを利用することで解決できる．しかし，全ての語義が均等に出現するような大規模コーパスを作成するのは現実的に難しい．

このような問題を解決する一案として，概念を定義説明する gloss を利用することが考えられる．すなわち，gloss 全

```
synset : 'hand.n.09 '
gloss :
 a unit of length equal to 4 inches;
 used in measuring horses

neighbor words :
[(u'hand', 0.67089716707504254),
 (u'hands', 0.40105846911705301),
 (u'finger', 0.29894136419969936),
 (u'fist', 0.2890361243661238),
 (u'handed', 0.28110865410315478),
 (u'arm', 0.27435457173860178),
 (u'ear', 0.26022667456320531),
 (u'other', 0.25479011172359678),
 (u'clenched_fist', 0.24753984224938555),
 (u'door', 0.24679898049815663)]
```

図 4: 概念の n 近傍の単語群 (適切でない例)

体の意味を表す分散表現が適切に得られれば，これを用いて AutoExtend により得た概念の分散表現を補正すること，あるいは，概念の分散表現の適切さを定量化することが考えられる．

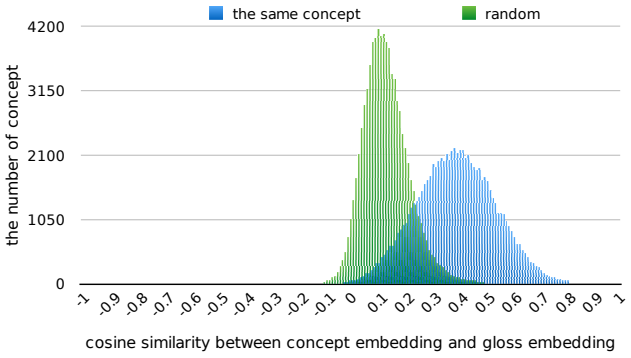

図 5: 概念の分散表現と gloss の分散表現の比較

実際に，gloss 中の各単語の分散表現の平均を求めるという最も単純な手法によって得られた gloss の分散表現を AutoExtend による概念の分散表現と比較した結果を図 5 に示す．この図に示されているように，概念の分散表現と当該概念の gloss の分散表現の類似度は，ランダムに選定した gloss の分散表現との類似度に比べ高い．このことは，全く異なる情報源，過程により得られる両者の分散表現を用いることにより，概念の分散表現を改善できることを示唆している．なお，上記で試した gloss の分散表現を求める手法は最も単純なものであり，単語の分散表現から文の分散表現を求めるより高度な手法を適用することにより，さらに有用な結果が得られることが期待できる．

5. おわりに

本稿では，AutoExtend と呼ばれる手法によって PWN 中の語義・概念に対して分散表現を作成し，単語間類以度・関連度の定量化，単語間意味関係分類，単語間想起関係の予想タスクへ適用し，有効性を確認した，このとき，語義・概念の分散表現は，単語の特定の意味や，単語間に成立しうる特定のみ関係を扱うタスクに対して特に有効であることが確認できた，一方で，語義・概念の分散表現は辞書中の単語にしか割り当てられないため適用範囲が限られてしまう問題，また低頻出の語義に対しては適切に表現を作成できないという問題が確認されたが，それぞれに対する解決方法を検討した，今後は，解決法を実装し，定量的に評価を行う予定である，

謝辞

本研究は JSPS 科研費#26540144, #25280117 の助成を受けた．

参考文献

[1] 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*, GEMS '11, pages 1–10, 2011.

[2] Jordan Boyd-Graber et al. Adding dense, weighted connections to wordnet. In *Proceedings of the third international WordNet conference*, pages 29–36, 2006.

[3] Yoshihiko Hayashi. Predicting the evocation relation between lexicalized concepts. In *Proceedings of COLING 2016*, pages 1657–1668, December 2016.

[4] Eric H Huang et al. Improving word representations via global context and multiple word prototypes. In *Proceedings of ACL 2012*, pages 873–882, 2012.

[5] Ignacio Iacobacci et al. Sensembed: Learning sense embeddings for word and relational similarity. In *Proceedings of ACL 2015 and IJCNLP 2015*, pages 95–105, July 2015.

[6] Kentaro Kanada, Tetsunori Kobayashi, and Yoshihiko Hayashi. Classifying lexical-semantic relationships by exploiting sense/concept representations. *Proceedings of EACL 2017 Workshop on Sense, Concept and Entity Representations and their Applications*, 2017.

[7] Tomas Mikolov et al. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013.

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

[9] Silvia Necsuleşcu et al. Reading between the lines: Overcoming data sparsity for accurate classification of lexical relationships. In *Proceedings of *SEM 2015*, pages 182–192, June 2015.

[10] Arvind Neelakantan et al. Efficient non-parametric estimation of multiple embeddings per word in vector space. In *Proceedings of EMNLP 2014*, pages 1059–1069, October 2014.

[11] Jeffrey Pennington et al. Glove: Global vectors for word representation. In *Proceedings of EMNLP 2014*, pages 1532–1543, October 2014.

[12] Sascha Rothe and Hinrich Schütze. Autoextend: Extending word embeddings to embeddings for synsets and lexemes. In *Proceedings of ACL 2015 and IJCNLP 2015*, pages 1793–1803, July 2015.

[13] 金田 健太郎, 小林 哲則, and 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)*, pages 1069–1072, 2016.

[14] 金田 健太郎, 小林 哲則, and 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)*, pages 214–217, 2017.

[15] 林 良彦. 概念間の想起の強さと方向性の予測. *2016* **年度 人工知能学会全国大会** *(JSAI2016)*, 2016.