語義・概念の分散表現を利用したSemantic Taxonomy Enrichment

金田 健太郎　　小林 哲則　　林 良彦
早稲田大学 理工学術院
kanada@pcl.cs.waseda.ac.jp

1 はじめに

辞書資源のデファクトスタンダードであるPrinceton WordNet (以下，PWN) [3] は, 語義曖昧性解消 [5] や感情分析 [4] 等, 様々な自然言語処理分野のタスクに利用されている．これらのタスクの性能は, PWNの持つ語彙に大きく制限されるが, 手動で扱える語彙を増やす（辞書を拡張する）ことは, 多大なコストを要する．この問題に対し，自動的にPWNを拡張することを目的として，SemEval 2016 において Semantic Taxonomy Enrichiment というタスク (Task 14) [6] が近年提案された．このタスクでは, PWN以外の辞書から得られた，PWNに載っていない語（未知語）に対し，その品詞と定義文を手がかりに, できるだけ正解概念と同義性の高いPWN中の概念を見つけることが求められる．

このタスクで最高の精度を達成した既存手法 [8] では, 定義文中の単語が持つ語義に対応する概念を候補の集合として用意し，要素となる各概念に対して単語分散表現等を用いて表現を与えた後，最適なものが最上位に来るようなランキング学習を行うことで適切な候補を選択している．しかしこの手法には，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ことや，用いている素性の役割が明確でないなどの問題がある．

そこで本研究では，(1) 単語の持つ語義・概念に対する分散表現 [2] 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，(3) さらに辞書構造を用いた基準によりランキングを行うことにより最適な概念を選出する手法を提案する．提案する手法は教師なしの手法であるが，SemEval 2016 Task14におけるデータを用いた評価実験によれば，既存手法 [8] に匹敵する精度を達成した．これは，語義・概念の分散表現や辞書構造から得られた各種の指標を有効に利用することで，教師なし手法においても妥当な結果が得られることを示す．

2 背景

2.1 語義・概念の分散表現：AutoExtend

本研究で利用する語義・概念の分散表現の導出方法であるAutoExtendは，Word2Vec[1] などの手法により得られた単語の分散表現を入力とし，単語の持つ各語義および, それぞれの語義が指示する語彙概念に対する分散表現を導出する．この際，テキストコーパスは必要なく，そのかわりに語義・概念の関係を整理した語彙資源 (PWNなど) を利用する．

AutoExtendの根幹をなす考え方は，単語 (word) はいくつかの語義を持ち (単語と語義の組み合わせが語彙素:lexeme，以下語義と表記)，これらの語彙素の集合として概念 (synset) が構成される，というWordNetの情報構造に基づく．

AutoExtendは, 語義・概念の分散表現を単語の分散表現と同じ空間に導出する．このことから, 単語の分散表現を与えられた単語であれば, 単純な演算（コサイン類以度）によって辞書中の語義・概念との分散表現間類以度を算出することができる．

ここで, コーパス（また, コーパスから作成された単語分散表現）に由来する, 単語/語義/概念の分散表現間の**類以度**は, これらの間の同義性に限らず, 様々な意味的関係性（反義性, 全体-部分関係性…）, すなわち**関連性の強さ**を表す．

2.2 評価タスク：SemEval 2016 Task14

近年提案されたSemEval 2016 Task14[6] 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 タスクである．実際に未知語を結びつけることによって PWN の拡張手法を評価するようなデータセットは他に存在しないため, 本稿ではこのタスクにおけるデータを評価に用いる．

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**[7] 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．

3 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．

3.1 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．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, 学習済みの単語分散表現を用い, 未知語定義文に対してベクトル表現（以下，未知語ベクトル）を与え, コサイン類以度によりその k 近傍 (今回は k=20) に分散表現を持つ概念を収集する．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 [9] 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．

3.2 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．

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 1）．

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

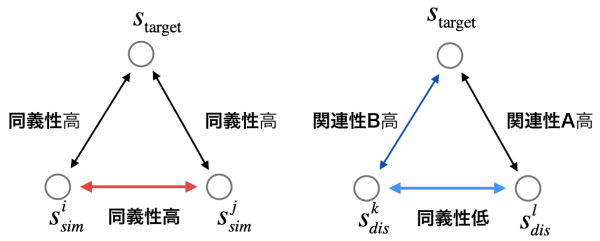

図 1: 概念間に成り立つと仮定される関係性

- 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．

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

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．

3.3　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．

	w/o clustering	w/o selection	提案
dist	0.469	0.493	0.500
center	0.487	0.474	**0.511**

表 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

- **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　実験

4.1　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．

4.2　実験結果

主な実験結果を表 1 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．

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従来手法 [8] による最高スコア（0.52）に迫る性能となることが確認できた．

4.3 議論；Wu&P Similarity の分布

選択基準に **center** を利用した場合の提案手法によって得られた選択概念と正解概念との間の Wu&P Similarity の分布を図 2 に示す．

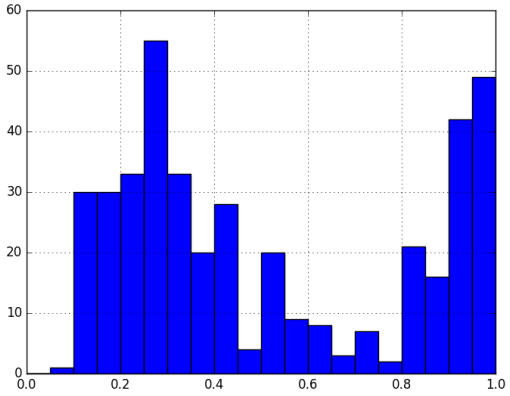

図 2: Wu&P Similarity の分布

与えられた全ての未知語に対し高い Wu&P Similarity が得られるような概念を選択することが理想であるが，本手法においては低い値を示す未知語群が存在している．これらに対し適切な未知語を選択するためには，よりよい候補集合を収集しておく必要がある．

5 おわりに

本稿では, 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 [2] 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，(3) さらに辞書構造を用いた基準によりランキングを行うことにより，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 [8] に迫る精度が得られた．

本手法では，候補集合中に，２つ以上候補として選ぶのに適切な概念が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．手順 (1) で候補概念を収集する際に，適切な概念をより多く集めることができれば，さらなる精度の向上が期待される．

参考文献

[1] Mikolov,Tomas, et al. Linguistic Regularities in Continuous Space Word Representations. In Proceedings of NAACL HLT, 2013.

[2] Rothe, Sascha, and Hinrich Schutze. ”AutoExtend: Extending word embeddings to embeddings for synsets and lexemes,” arXiv, 1507.01127, pp.1-11, 2015.

[3] G. A. Miller, R. Beckwith, C. D. Fellbaum, D. Gross, K. Miller, ”WordNet: An online lexical database,” Int. J. Lexicograph. Vol.3, No.4, pp.235-244, 1990.

[4] Esuli, Andrea, and Fabrizio Sebastiani. ”SentiWordNet: a high-coverage lexical resource for opinion mining.” Evaluation (2007): 1-26.

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

[6] Jurgens, David, and Mohammad Taher Pilehvar. ”SemEval-2016 Task 14: Semantic Taxonomy Enrichment.” SemEval@ NAACL-HLT. 2016.

[7] Wu, Zhibiao, and Martha Palmer. ”Verbs semantics and lexical selection.” Proceedings of the 32nd annual meeting on Association for Computational Linguistics. Association for Computational Linguistics, 1994.

[8] Schlichtkrull, Michael Sejr, and Héctor Martínez Alonso. ”MSejrKu at SemEval-2016 Task 14: Taxonomy Enrichment by Evidence Ranking.” SemEval@ NAACL-HLT. 2016.

[9] Biemann, Chris, et al. ”JoBimText visualizer: a graph-based approach to contextualizing distributional similarity.” Proceedings of TextGraphs-8 Graph-based Methods for Natural Language Processing (2013): 6-10. APA