

語彙知識と分散表現を用いた概念間の想起関係の予測

林 良彦 (早稲田大学理工学術院・実体情報学博士プログラム)

yshk.hayashi@aoni.waseda.jp

1 はじめに

近年、言語産出の結果であるコーパスからは得ることができない、あるいは、得ることが難しい意味的な属性を単語などの基本的な言語単位の意味表現に組み入れる必要性が主張され、特に視覚に代表される人間の知覚情報を取り込んだ意味表現の研究 [9, 2] が活発化している。本研究で扱う想起 (evocation) [1] も人間の知覚や経験に根ざした意味関係として、一部で注目されている [6]。

想起関係は基本的には心理実験の形式でしか収集することができないため、言語コーパスに比べて収集コストの高い情報である。このため、任意の概念間の想起関係を精度よく予測する機械的な手段が求められている。次節で詳しく述べるように、想起関係は関係の強さの程度を表す重みを有する方向性を持つ関係であるから、与えられた概念ペアに対し、それらの間の想起関係の強さと方向性を予測できることが望まれるが、一般的にはこれは困難な問題と認識されている。

本稿では、先見的な言語知識を整理した語彙資源 (WordNet) とコーパスから学習された分散表現から導いた各種の属性を用いることにより、回帰問題として想起関係強度を予測する手法を提案し、その評価結果について報告する。また、想起関係の方向性の予測も困難を伴う問題であることを示し、同じ属性を用いる分類問題として扱う可能性を検討する。

今回の結果により、一定の精度で想起関係の強さと方向性を予測する可能性が示されたが、さらなる精度の向上のためには、既存の語彙資源などに暗黙的に含まれる情報をフルに活用することが必要と考えられる。

2 想起関係とそのデータ

2.1 想起関係

想起とは、ある概念 s (起点概念) が別の概念 t (ターゲット概念) を思い浮かばせる (bring to mind) こと、また、その程度である [1]。

想起関係についてコンセンサスの取れている特徴は以下のように列挙できる。

- 基本的には概念間の関係であり、
- 同一の品詞の間には限らない。
- 方向性のある非対称な関係であり、
- 想起の程度を表す重み付きの関係である。

想起関係の分類やその機序はいまだ明らかではないが、意味的な同義性を示す類似度 (similarity) や、上位・下位関係などの比較的な定義が明らかな意味関係による意味的関連度 (relatedness) に比べ、より広く、かつ、曖昧な関係を含むものと考えられている。

一方で、想起関係は「人間の知覚や経験に基づくもの」であることから、言語コーパスから抽出される言語由来の情報を補完する情報として有用であることが期待されている [6]。

2.2 Princeton WordNet の想起データ

想起関係の背後にある概念の連鎖を特徴付けることを試みた本報告者による先行研究 [10] と同様、本研究においても Princeton WordNet (以下、PWN) の研究グループによる想起データ¹[1] を用いる。

このデータは、PWNにおいてコアとなる1,000の語彙概念 (core synsets) からランダムに抽出された119,652件の語彙概念ペアの間の想起の程度 (以下、想起強度) を、少なくとも3名の被験者に評定させたものである。想起強度は、0から100の間の評定値により表現されるが、何らかの心的な関連性があることを意味する0より大きい評定値が付与されたペアは、39,309件 (全体の約33%) にとどまっている。

[1]では、想起強度と各種の尺度による意味的類似度・関連度との間の相関が低いことを報告しており、想起関係が通常の意味的類似性・関連性とは異なるタイプの情報であり、その予測は困難を伴うことを議論している。

¹<http://wordnet.cs.princeton.edu/downloads.html>

3 回帰問題としての想起強度の予測

想起関係には、多くの要因が関連していることが想定されるため、寄与の可能性が考えられる様々な属性を用いる回帰問題として想起強度の予測を行うことが妥当である。回帰によって予測される想起強度と gold data との相関により、予測の良さを評価する。

Deyne ら [3] も指摘しているように、想起強度には、 s と t の間の基本的な関連度による対称的成分と想起の特徴を付加する非対称的成分が関わっていると考えられる。そこで、それぞれの成分への寄与が想定される以下の属性群を用いる。

3.1 対称性を表す属性群

対象となる概念、あるいは、その代表単語の間の意味的関連度を対称性を表す属性として用いる。

単語間の意味的関連度に基づく属性としては、Word2Vec [7]²、LSA、LDA 手法により作成した³ベクトル間のコサイン類似度、計 4 種類を用いた。この中で特に Word2Vec に基づく属性 (w2v-sim) をベースラインの一つとする。

語義・概念間の意味的関連度としては、WordNet に基づく Wu and Palmer による関連度 (wup-sim) に加え、AutoExtend と呼ばれる手法 [8] に基づく分散表現ベクトル間のコサイン類似度 3 種類 (語義ベクトル、概念ベクトル ae-syb-sim, 単語, 語義, 概念ベクトルの和ベクトル) を用いる。特に、wup-sim と ae-syn-sim をベースラインとして比較する。

AutoExtend: 任意の単語の分散表現を入力とし、単語の持つ各語義、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手法である。WordNet などの語義・概念の関係を整理した語彙資源を利用することが特徴であり、単語ベクトルと同じ空間に語義・概念ベクトルを導出するという利点がある。今回は、[8] の著者が公開しているベクトルデータ⁴を利用した。

3.2 非対称性を表す属性群

3.2.1 概念ノードの特徴の順序付き組合せ

まず、 s , t それぞれの概念ノードの特徴を表す属性を、この順序で並べることにより非対称性の特徴

²<https://code.google.com/p/word2vec/>

³Wikipedia コーパス enwik9 (<http://mattmahoney.net/dc/textdata>) を利用。

⁴<http://www.cis.lmu.de/~sascha/AutoExtend/embeddings.zip>

付けに寄与することを期待する。ノードの特徴を表す属性としては、品詞、おおまかな意味分類 (WordNet における lexicographer file を分類体系として利用)、WordNet グラフにおける当該ノードの重要度を表す中心性指標、上記の各種分散表現ベクトルの分散値を利用した。品詞と意味分類に関しては、1-of-k 表現をとっている。これらの属性群をまとめたもの (nodes; 116 次元) もベースラインの一つとする。

3.2.2 属性の包含関係

事例数は多くはないものの非対称的な類似度・関連度に関する研究がいくつか行われている [5, 4]。これらにおける基本的な考えは、属性の包含関係 (feature inclusion) である。概念的には、より狭い範囲の属性を持つ要素 x 、より広い範囲の属性を持つ要素 y があるとき、 $rel(x \rightarrow y) > rel(y \rightarrow x)$ という関係が成り立つことを表す。既存研究 [5, 4] では、言語コーパスを構文解析するなどして獲得した属性間の包含関係を用いて方向性のある類似度・関連度を定義しているが、本研究においては、この考え方を以下のリソースから得る情報に対して適用し、非対称性を表すであろう属性として利用する。

WordNet における k -近傍ノード集合: WordNet をグラフとしてみたとき、 s , t ノードの k -近傍には共通するノードがいくつか存在することが期待される。これはまた、両者の間にパスが存在することを表す。 $nb(s, k)$ を s の k -近傍にあるノード集合とするとき、以下の指標 $drel$ を計算する。今回は予備実験の結果から $k = 3$ とした。

$$drel(s, t, k) = \frac{|nb(s, k) \cap nb(t, k)|}{|nb(s, k)|}$$

分散表現ベクトルにおける要素: 分散表現ベクトルの各次元要素はそれ自体が一つの属性であると考えることができる。そこで、前述の 4 種の分散表現ベクトルに対して、以下の指標 $feat_inc$ を計算する。ここで、 $vec(s, f)$ は、 s に対する分散表現ベクトルの f 番目の次元要素を表す。ただし、事前に一定の閾値以下の要素は 0 に置き換えている。

$$feat_inc(s, t) = \frac{\sum_f \min(vec(s, f), vec(t, f))}{\sum_f vec(s, f)}$$

また、両ベクトルからの属性 (次元) を重要度順に並べた時の重なり具合を情報検索で用いられる平均適合度により測り、属性とすることが提案されている [5] ので、この指標 (simpAP) も属性に加える。

表 1: 順方向の想起強度の予測結果

feature group	# dims.	RF	NN
w2v-sim	1	0.1905	0.2477
wup-sim	1	0.1236	0.0832
ae-syn-sim	1	0.1790	0.2400
nodes	116	0.2922	0.2424
non similarities	13	0.2629	0.2219
similarities	7	0.3231	0.3149
all	136	0.3773	0.3932

表 2: 順方向・逆方向の想起強度の同時予測結果

feature group	NN2 (f)	NN2 (b)	NN
w2v-sim	0.2384	0.2383	0.2380
wup-sim	0.0257	0.0291	0.0327
ae-syn-sim	0.2310	0.2307	0.2302
nodes	0.2002	0.1978	0.2030
non similarities	0.1814	0.1883	0.1817
similarities	0.2960	0.2959	0.2952
all	0.3487	0.3503	0.3500

3.3 その他の属性

Word2Vec が注目を集めた理由の一つとして、意味関係がベクトルの差分に現れるという特性がある。本研究の設定では、 $vec(t) - vec(s)$ に各種の想起関係の特徴が埋め込まれる可能性があるが、今回はこの差分ベクトルの分散値が想起強度と関連する可能性を想定し、これを属性として加えた (rel_var)。

4 評価実験

4.1 主な結果: 想起強度の予測

PWN 想起データの全データ (119,652 件) 全てを対象とし、5 分割交差検定により想起強度の予測に関する評価を行った。この際、gold data g についてはその値域が広いので、対数を用いて $g' = \log(g+1)$ なる変換を行っている。回帰の手段としては、Random Forest (RF)、および、3 層のフィードフォワード・ニューラルネットワーク (NN) を試した。前者には scikit-learn⁵、後者には Chainer⁶ を用いた。

表 1 に、結果 (想起強度の予測値と gold data との Pearson 相関係数) を示す。属性群のうち、最初の 4 つはベースラインとするものであり、non similarities は、3.2, 3.3 節で導入した非対称性を総合した属性群である。また参考のため、表には各手法における属性ベクトルの次元を示している。この結果からは以下のようなことが言える。

- 全ての属性を使う (all) のが良く、NN によって 0.4 に迫る相関を得た。これは、[1] に報告されている予備調査結果 (Spearman 相関係数で 0.131 程度。LSA を利用) を大きく上回るだけでなく、今回の各ベースラインの結果も大きく上回る。また、提案する属性群はある程度、相補的であることも確認できた。

- 意味的類似度系の属性を統合的 (similarities) を用いた場合、非類似度系の属性を用いる場合 (non similarities) よりも高い相関を示している。ただし、前者で得られる予測値はあくまで s, t に関して対称的であることに注意しなければならない。
- 意味的類似度系の属性では、分散表現ベースの属性 (w2v-sim, ae-syn-sum) の有用性が明らかである。ベースラインとしては、w2v に基づく類似度 (w2v-sim) がもっとも良い結果であるが、AutoExtend による概念ベクトル (ae-syn-sim) もこれに近い結果を出しており、特定の単語に限定されずとも概念間の想起関係を捉えることができていることが確認できた。
- s, t の双方の概念ノードに関する属性の組合せ (nodes) は当初の予想以上に有用であることが確認できた。コーパス分析のカバー性の問題から必要な分散表現ベクトルが得られない場合のボトムラインを与えることが期待できる。
- 機械学習の手法としては、総合的には NN の方が RF より良いが、その差はさほど大きくない。

4.2 順方向・逆方向の想起強度の同時予測

PWN の想起データにおいては、両方向の想起強度が記録された概念ペアが存在する。これらのデータは、想起強度がゼロのものも含めて 7,164 ペアに過ぎないが、 s から t への順方向、 t から s への逆方向の想起強度の同時予測を試みた。結果を表 2 に示す。この際、回帰の枠組みとしては、出力層のノードを 2 つ持つフィードフォワード・ニューラルネットワークを用いた (NN2)⁷。表には、同じデータを前節に示した順方向の回帰として扱った場合の結果 (NN) を比較している。

表に示すように、順方向 (f)、逆方向 (b) の相関係数は同等であり、両方向の想起強度を同時予測することは可能である。しかし、それぞれを順方向の想起関係

⁵<http://scikit-learn.org/>

⁶<http://chainer.org/>

⁷正準相関分析も試みたが、NN2 を大きく下回った。

と考へて単純に予測した場合の結果を上回ることではなく、同時予測を行うメリットは観察されなかった。このような結果は、順方向、逆方向の想起強度の間の相関が高くないことを予想させるが、実際は、双方の想起強度がゼロのものを含めた場合に 0.5315、含めない場合でも 0.4539 と、ある程度強い相関を示した。このため、今後、両方向の想起強度の同時予測の精度を高めるためには、より多くの学習データを用い、同時予測に適した属性を探索する必要があると思われる。

4.3 順方向・逆方向の想起強度の差分の予測

想起は方向性を持つ非対称な関係であるため、2つの概念が与えられた時、どちらの方向の想起が強いかを予測できることが望ましい。そこで、双方の想起強度を記録した 7,164 ペアのデータに対し、順方向と逆方向の想起強度の差分を予測する実験を行った。全ての属性群を利用した場合に得られた Pearson 相関係数は、NN の場合 0.1209、RF の場合 0.0522 とかなり低い値となった。この結果は、想起強度の方向性の違いによる差分、あるいは、想起の方向性の判定は、単純に加減算的には扱うことはできないことを示唆する。

4.4 分類問題としての想起の方向性の予測

そこで、Ma[6] により公開されているデータを用い、想起関係の方向性の予測を試みた。このデータは、WordNet における synset と単語の組のペア (348,447 件) に対して、想起の方向性を (a) s から t (outbound), (b) t から s (inbound), (c) 双方向 (bidirectional), (d) 不明の 4 クラスで評定したものである。

実験では想起強度の予測に用いたものと同じ全属性を用い、RF によりマルチクラス分類問題として評価を行った。5 分割交差検定の結果、全体の分類精度 (accuracy) は 0.785、outbound の精度/再現率は 0.751/0.869、inbound の精度/再現率は 0.774/0.592、双方向性 (bidirection) の精度/再現率は 0.961/1.000 であった。全般的にある程度の精度が得られたが、inbound は再現率が低い結果となった。

なお、この想起方向性の予測においては、想起強度の予測においては貢献度が低かった非対称性を表す属性群の貢献が高い傾向にあることが分かった。これより、非対称性をより良く表す属性を開発できれば、方向性予測精度の向上が見込めると考えられる。

Ma のデータの性質は PWN 想起データとは大きく異なるので、現時点で確固たることは言えないが、前

節の結果から予想されるよりも良好な結果を得ていることから、方向性の予測は強度の予測とは切り離して考えるのが良いことが示唆される。

5 おわりに

語彙資源と分散表現を用いることにより、想起の強さと方向性のある程度の精度で予測できることを示した。今後は、既存の語彙資源などに暗黙的に含まれる情報を活用する方向性を検討する。特に、想起を概念連鎖と考えたときの意味ネットワーク上での経路情報の利用 [10] について探求したい。

謝辞

本研究は JSPS 科研費 #26540144 の助成を受けた。

参考文献

- [1] Boyd-Graber, J., et al. 2006. Adding dense, weighted, connections to WordNet. *Proc. of the Third International WordNet Conference*, pp.29–36.
- [2] Bruni, E., et al. 2014. Multimodal distributional semantics. *Journal of Artificial Intelligence Research* 49, pp.1-47.
- [3] De Deyne, S., et al. 2015. Associative strength and semantic activation in the mental lexicon: evidence from continued word associations. *Proc. of the 35th Annual Conference of the Cognitive Science Society*, pp.2142–2147.
- [4] Gawron, J. M. 2014. Improving sparse word similarity method with asymmetric measures. *Proc. of ACL 2014 (Short Papers)*, pp.296–301.
- [5] Kotlerman, L., et al. 2010. Directional distributional similarity for lexical inference. *Natural Language Engineering*, 16 (4): 359–389.
- [6] Ma, X. 2013. Evocation: analyzing and propagating a semantic link based on free word association. *Language Resources and Evaluation*, Volume 47 Issue 3, pp.819–837.
- [7] Mikolov, T., et al. 2013. Distributed Representations of Words and Phrases and their Compositionality. *Proc. of NIPS 2013*.
- [8] Rothe, S., and Schütze, H. 2015. AutoExtend: Extending word embeddings to embeddings from synsets and lezemes. *Proc. of ACL 2015*, pp.1793–1803.
- [9] Silberer, C., et al. 2013. Models of semantic representation with visual attributes. *Proc. of ACL 2013*, pp.572–582.
- [10] 林 良彦. 2015. 意味ネットワークの探索と系列パターンマイニングによる想起の類型化. 2015 年度人工知能学会全国大会 2F5-OS-01b-3.