既存の言語知識を援用した想起データの分析

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

大規模コーパスに対して統計的な手法を適用することにより，単語の意味表現を多次元ベクトルとして獲得する研究 (例えば [5]) が盛んに行われており，様々な自然言語処理の応用に適用できるものとして注目を集めている．一方で，言語産出の結果であるテキストデータからは獲得できない，あるいは，獲得しにくいタイプの意味情報が存在するのも事実であり，例えば，画像から得られる視覚情報のような人間の知覚・感覚と結びついた属性情報を，テキストコーパスから得られる意味表現と統合しようとする研究 (例えば [7]) も活発化しつつある．

知覚に関する属性情報以外に，テキストデータから直接獲得することが困難なタイプの意味的な関係・情報として，心的な想起 (evocation) がある．想起は,「ある概念がどの程度，別の概念を思い浮かばせる (bring to mind) か」として定義される [1, 3]．具体的な想起データは人間の評定者から得る以外に収集の手段はなく，収集コストの高いデータである．また，想起をもたらす心的な過程の構造は必ずしも自明ではなく，結果として，想起的なデータを機械的に生成することも困難である．

本研究ではまず,「想起とは直接は観測できない概念レベルでの顕著な連想系列の始端・終端が取り出されたもの」と仮定する．この仮定のもと，すでに得られている想起データを既存の大規模言語知識に関連付けて分析することにより，想起の背景にある概念連鎖のパターンを探る [1] ことを目的とする．上記の仮定がある程度検証でき，妥当な想起をもたらす概念連鎖に関する知見が得られれば，すでに得られている想起データを拡張するための見通しが得られるものと期待する．

2 分析の概要

図 1 に本報告の分析の流れを示す．

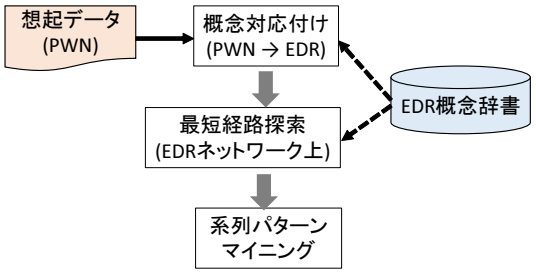

図 1: 想起データ分析の流れ

分析の対象とする想起データは英語の想起データ [1] であり，Princeton WordNet(以下，PWN) におけるある語彙概念 (起点概念と呼ぶ) がどの程度，別の語彙概念 (ターゲット概念と呼ぶ) を思い浮かばせる (bring to mind) かを被験者に評定させたデータである．PWN はネットワーク構造をなしており，起点概念からターゲット概念に至る経路を求めることが可能である．しかし，PWN における関係情報にはいわゆる格フレーム的な情報などが含まれていないため，得られる経路情報は限定的であり [2]，想起をもたらす概念連鎖のパターンを分析するには十分ではない．

そこで本研究では，より豊富な概念間の関係情報を有する言語知識として EDR 電子化辞書 (以下，EDR)[8] を援用する．このため，各想起データを構成する PWN 上の起点・ターゲット概念に対応する (あるいは，近い) EDR 概念を事前に求めておく．次に起点概念からターゲット概念に至る EDR ネットワーク上の最短経路を収集し，得られた経路情報を系列パターンマイニングの手法により分析することにより，想起をもたらす概念連鎖において特徴的なパターンを探索する．

3 想起データ

本研究の対象とする想起データは，PWN の研究グループが提供するデータ [1][3] である．このデータは，

[1] 想起の機序は，概念連鎖には限定されないと思われる．しかし本研究では，概念連鎖として定式化できる範囲を明確化したい．

[2] そもそも [1] の研究の動機も PWN により多くの関係情報を付与することにあった．

[3] http://wordnet.cs.princeton.edu/downloads.html

PWN においてコアとなる 1,000 の語彙概念 (synset) のペアから抽出された 119,652 件の語彙概念ペアの間の想起の強さを 0〜100 のスコアにより 20 名の被験者に評定させたものである．この中で，0 以上の評定値が付与されたペアは，39,309 件 (32.9%) であり，それらの評定値の平均値は 9.25 と報告されている．

3.1 データ群の構成

想起の強さ，また，想起の非対称性との関連を調べるため，今回は，これらのデータから以下の 3 つのデータ群を抽出し，それぞれ分析対象とした．ただし，以下における閾値には大きな意味はなく，初期の分析に適した範囲に限定し，かつ，各データ群を同程度の規模にバランスさせるために設定した．なお以下の定義より，B 群と C 群は排他的ではなく，双方に含まれるデータが存在する．

- A 群: 想起の強さ x が $0 < x < 0.4$ と低いもの (337 件).
- B 群: 想起の強さ x が $0.4 \leq x < 20.0$ と中程度であり，かつ，想起の非対称性が比較的強い (逆方向の想起の強さとの差の絶対値が 10 以上) もの (383 件).
- C 群: 想起の強さ x が $x > 25.0$ と相応に高いもの (335 件).

3.2 PWN 語彙概念の EDR 概念への対応付け

上記のデータ群に含まれるユニークな PWN 語彙概念の総数は 745 であり，これらを EDR 概念に対応付けた．この対応付けにおいては，まず EDR 概念を PWN 語彙概念に対応付けるための手法 [2] を逆方向に適用し [4]，対応付けの候補を抽出した．具体的には，上記の手法により 745 個の PWN 語彙概念中の 683 個 (91.7%) に対して，対応する EDR 概念を上位 10 件内に抽出できた [5] ので，これらの中から人手により正解とする対応付けを決定し，対応付けが行えた PWN 語彙概念を起点・ターゲット概念とする想起データのみを以下の分析の対象とした．この結果，各データ群の件数 (歩留まり率) はそれぞれ，A 群:286 件 (84.8%)，B 群:321 件 (83.8%)，C 群:275 件 (82.1%) となった．

[4] [2] では，PWN が提供する語義タグ付きコーパスを利用したが，これに相当するものとして，EDR 日本語コーパスを利用した．

[5] 順位 1 位のものが正解であった割合は 49.5%，正解逆順位の平均 (MRR) は 0.694 であった．

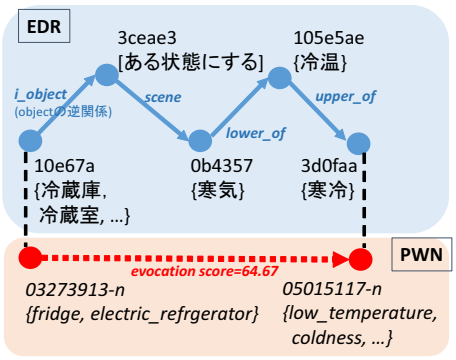

図 2: 概念連鎖の最短経路例

4 EDR 概念辞書の利用

EDR 電子化辞書 [8][6] は，概念辞書を含む各種の辞書から構成される大規模な言語資源である．概念辞書はさらに，(1) 概念見出し辞書，(2) 概念体系辞書，(3) 概念記述辞書の 3 つから構成される．(1) は概念体系におけるノードをなす各概念に識別子を与え，日本語・英語による概念見出し，および，概念説明を提示することにより概念の内容を規程・説明する．(2) は概念間の上位・下位関係を提供し，(3) は格関係中心にした上位・下位関係以外の概念間の関係を記述している．

今回の分析では，この概念辞書における各概念をノードとし，概念間関係を有向エッジとするネットワークを利用する．表 1 に関係の種別・説明，および，関係の数を示す [7]．ただし，概念辞書で与えられる概念関係の方向性はある意味恣意的なものなので，以下で述べる最短経路の探索においては，それぞれの逆関係を表す有向エッジを無条件にネットワークに加えた．

5 主な分析結果

5.1 最短経路探索

図 2 に実際に抽出した経路探索の例を示す．

最短経路探索においては，概念関係の種別にかかわらず一様の重み (=1) を各エッジに付与する場合 (α)，上位・下位関係の重みを 1 としそれ以外の関係の重みを 0.5 とする場合 (β) の 2 通りの方法を試みた．後者の意図は，上位・下位関係以外の関係によるエッジがより選ばれやすいという状況を作り出すためである．なお定義より，α の場合の重み和はパス長と等しい．

[6] http://www2.nict.go.jp/out-promotion/techtransfer/EDR/J_index.html

[7] EDR 概念辞書の仕様書にはこれら以外にも多数の概念関係が示されているが，実際の辞書データには付与されていない．

表 1: EDR 概念辞書における概念関係

関係子 (ラベル)	説明	関係の数
upper_of	上位関係	413,854
agent	有意志動作を引き起こす主体	40,218
object	動作・変化の影響を受ける対象	282,733
a-object	属性をもつ対象	41,138
implement	有意志動作における道具・手段	20,038
cause	事象の原因，理由	9,620
goal	事象の主体または対象の最後の位置	46,129
place	事象の成立する場所	26,016
scene	事象の成立する場面	36,173

表 2: 各データ群におけるパス長，パス重み和の平均値

データ群	パス長	パス重み和
A	2.892	1.601
B	2.882	1.567
C	2.825	1.527

表 2 に各データ群におけるパス長，パス重み和の平均値を示す．想起の強さが強いほど，パス長・パス重み和はわずかに小さくなる傾向が確認できたが，有意差は認められなかった (p 値が最小 ($p = 0.206$) となったのは，A 群と C 群の β の場合の比較)．

一方で，比較的想起の強い C 群 (想起の強さが 20.0 より大きい) における 275 件のデータについて，想起の強さとパス長，および，パスの重み和との間の相関 (Spearman の順位相関係数; β の場合) を調べてみると，$\rho = 0.158/0.134$ 程度のごく弱い正の相関 ($p = 0.008/0.02$) がみられた．

いずれにせよ，想起の強さと概念連鎖の長さや重みの間には顕著は相関は確認できず，この結果は，想起の強さと意味的類似度の間の相関が低いことを報告した [1] の結果と符合している．以下では，比較的強い想起である C 群のデータに対して，上記 β の条件によって抽出した最短経路を検討する．

5.2 系列パターンマイニング

上記により得られた EDR ネットワーク上の経路を概念連鎖の系列とみなし，系列パターンマイニングを適用した．本報告では，エッジの両端の概念ノードは無視し，関係種別を表すエッジのラベルの連鎖のみを見た場合の結果について，サポート事例の頻度が 4 以上の部分系列についてのケーススタディを示す．なお，マイニングには，prefix span と呼ばれるアルゴリズム [6] を実装したツール [8] を用いた．

[8] http://prefixspan-rel.sourceforge.jp/

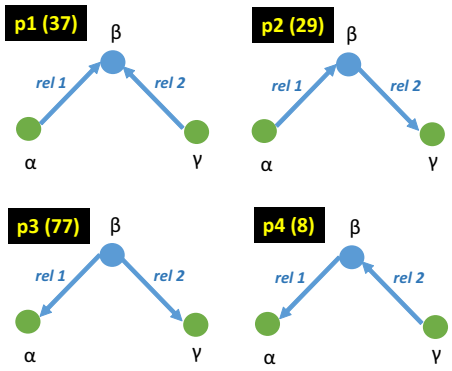

図 3: パス長=2 の概念連鎖パターンの形式的類型

5.3 概念連鎖の基本パターン

ここでは，系列マイニングにおいて頻出系列として抽出された概念連鎖系列の中から，特にパス長が 2 である部分系列 ($\alpha \rightarrow \beta \rightarrow \gamma$) のパターン化を試みる．

図 3 に 2 つのエッジと 3 つのノードから構成される部分系列を 2 つのエッジの方向性をもとに 4 つの類型 (p1〜p4) に形式的に分類する．ここでは，最適経路探索によって選択された逆方向のエッジは方向性を逆転し，EDR 概念辞書のオリジナルの関係の方向性に戻して考えている．図中の p1 などの後ろの括弧内の数字は経路における出現 (利用) 頻度を示す．推移的な連鎖 (p1, p4) よりも共通的な要素を介した連鎖 (p2, p3) の方が多いという結果となっている．

表 3 に rel-1，rel-2 の関係種別ごとにみた内訳を示す．以下では，各類型における特徴について議論する．

- p1: #1 における γ，#4, #6 における α は形容詞的概念 (以下，a 概念) にあたるので，これらは，共通の名詞的概念 (以下，n 概念) を介した a 概念と動詞的概念 (以下，v 概念) の関係を表す．それ以外のパターンは，共通の n 概念を介した v 概念どうしの関係と言える．特に，rel-1 と rel-2 の関係種別が等しい場合 (#3, #5) は，類義の n 概念

表 3: 概念連鎖のパターン分類

#	類型	rel-1	rel-2	頻度
1	p1	object	a_object	12
2	p1	agent	object	7
3	p1	scene	scene	6
4	p1	a_object	scene	4
5	p1	agent	agent	4
6	p1	scene	a_object	4
7	p2	object	object	18
8	p2	object	upper_of	7
9	p2	implement	object	4
10	p3	goal	object	14
11	p3	object	scene	12
12	p3	scene	goal	12
13	p3	agent	object	10
14	p3	scene	object	10
15	p3	a_object	a_obj	8
16	p3	object	implement	6
17	p3	object	agent	5
18	p4	agent	object	4
19	p4	object	scene	4

の間の関係となっていることが推定される．

- p2: #8 は上位概念を利用した格関係を表す．それ以外には，2 つの格フレームが関与しており，例えば，「予約:α を 済ませること:β を 求める:γ」における α と γ (すなわち，予約と求める) のような高階な関係を表す．このような類型の頻度が高くないことは，ここでの α と γ は想起の基盤を形成する可能性が高くないことを示唆する．逆にいうと，このパターンに現れる α と γ のペアは，特別な個別的な事由で結ばれている可能性がある．
- p3: この類型が全体の約半数を占めた．#15 は，β にあたる単一の a 概念が 2 つの n 概念の属性となっている場合である．すなわち，a 概念の属性が共通する n 概念が想起されているパターン (例: リンゴもトマトも赤い) である．その他は，全てある v 概念に対して格要素となっている n 概念間の想起を表すといってよい．すなわち，同一の格フレームにおける別の構成要素が想起の基盤を形成することが最も多いといえる．
- p4: 逆方向に推移的なこの類型の頻度は非常に低く，抽出された実例も適切とは考えにくいものがほとんどであるため，想起の概念連鎖のパターン化において考慮する必要は低いと思われる．

6 おわりに

想起を概念連鎖のショートカットと仮定し，観測できない概念連鎖のパターンを言語知識のネットワークにおける最短経路から探る試みとその可能性について議論した．

最短経路が妥当な概念連鎖 (連想) を与えうるためには，ネットワークのノードやエッジに適正な重みが付与されていることが必要となる．本報告ではアドホックな重みを付与したに過ぎないので，中心性 (centrality) のようなノード，エッジの重要性を表す指標 [4] を導入するとともに，得られる経路の妥当性の評価を進める必要がある．ある程度の量の評価データが収集できれば，機械学習的な手法の適用も考えられる．

本研究では，PWN ベースの想起を分析するために，全く別の言語資源である EDR を援用するという「荒技」を用いたが，本来ならば，この 2 つに限らず様々な言語資源を組み合わせ的に利用できることが望まれる．このためには，語義・概念のレベルで言語資源の構成要素を対応付けること，さらには，そのような統合的な利用を可能とする枠組みの研究が必要と考える．

一方で，想起のような心的な関係が実際の言語処理の応用 (例:多義解消，意味的類似度，含意認識) においてどの程度有用であるかも評価していく必要がある．

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] Hayashi, Y. 2012. Computing cross-lingual synonym set similarity by using Princeton annotated corpus. *Proc. of the 6-th International Global WordNet Conference (GWC2012)*. pp.134–141.

[3] 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.

[4] Mihalcea, R., and Radev, D. 2011. *Graph-Based Natural Language Processing and Information Retrieval*. Cambridge University Press.

[5] Mikolov, T., et al. 2013. Distributed Representations of Words and Phrases and their Compositionality. *Proc. of NIPS 2013*.

[6] Pei, J. et al. 2001. PrefixSpan: mining sequential patterns efficiently by prefix projected pattern growth. *Proc. of ICDE'01*, pp.215–224.

[7] Silberer, C., et al. 2013. Models of semantic representation with visual attributes. *Proc. of ACL 2013*, pp.572–582.

[8] Yokoi, T. 1995. The EDR electronic dictionary. *Communications of the ACM*, Volume 38, Issue 11, pp.42–44.