

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

A trial for grounding evocation in existing onto-lexical resources

林 良彦^{*1}

Yoshihiko Hayashi

^{*1}早稲田大学理工学術院・実体情報学博士プログラム

Graduate program for embodiment informatics, Waseda University

This paper presents an on-going effort for *grounding* evocation (human-generated ratings of semantic association) in existing onto-lexical resources. Some preliminary results obtained through the application of a sequential pattern mining algorithm on semantic paths recovered from an onto-lexical resource (EDR concept dictionary) are presented, which may insist there might exist some prominent patterns of concept chaining. This paper also poses an infrastructural issue for enabling the combined exploitation of multiple semantic resources.

1. はじめに

大規模な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 [Mikolov 13, Clark 15]) 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ことも指摘されており [Bruni, et al. 14], 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 [Silberer, et al. 13, Bruni, et al. 14]) 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 [Boyd-Graber, et al. 06], 一部の想起関係は「人間の知覚や経験に基づくもの」 [Ma 13] とされている。

ある概念 (起点概念) が別の概念 (ターゲット概念) を想起させる程度 (以下、想起データ) については、人間の評定者から得る以外にその収集の手段はなく、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、想起データを機械的に生成することも困難である。

想起の機序には、もちろん様々なものがあり得るが、本研究では、「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=概念連鎖のショートカット」と仮定し、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。より具体的には、起点概念とターゲット概念を結ぶ意味ネットワークにおける経路を収集し、これらの経路群に対して系列パターンマイニングの手法を適用することにより、顕著な概念連鎖を分析する。以上の結果として、想起をもたらす概念連鎖に関する知見が得られれば、新たな概念ペアに対する想起の程度を予測するための手がかりが得られ、収集コストが高いという想起データに関する大きな問題に対する解決につながるものと期待できる。

2. 分析の概要

図 1 に本報告の分析の流れを示す。

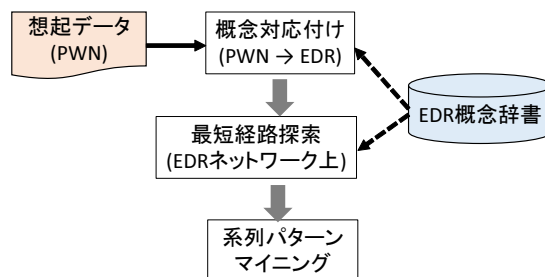


図 1: 想起データ分析の流れ

分析の対象とする想起データは英語の想起データ [Boyd-Graber, et al. 06] であり、Princeton WordNet (以下、PWN) における起点概念がどの程度、ターゲット概念を思い浮かばせるかを被験者に評定させたデータである。周知のとおり、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 (群) 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^{*1}、想起の背景に存在することを仮定する概念連鎖のパターンを分析するには十分ではないと考えられる。

そこで本研究では、より豊富な概念間関係情報を有する言語知識として EDR 電子化辞書 (以下、EDR) [Yokoi 84] における概念辞書を援用する。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き、EDR の意味ネットワーク上で最短経路を収集し、得られた経路群情報を系列パターンマイニングの手法により分析する。

なお、想起関係を適切に表す (説明できる) 概念連鎖は、最短経路によるものとは限らない。より適切には、適切にノードやエッジに重みを付与して「最短」経路探索をすべきであるが、これは今後の課題であるので、本報告では経路長による最短経路群に分析対象を限定する。

^{*1} そもそも [Boyd-Graber, et al. 06] の研究の動機もより多くの関係情報を PWN に付与することにあった。

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

3. リソースとその利用

3.1 PWN の想起データ

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 [Boyd-Graber, et al. 06]^{*2}である. 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強さを, 少なくとも 3 名の被験者に評定させたものである. 想起の強さは 0~100 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 以上の評定値が付与されたペアは, 39,309 件 (32.9%) にとどまっている.

なお, [Boyd-Graber, et al. 06] は, 想起の評定値と各種の尺度による意味的類似性の間の相関が低いことを報告しており, 想起関係が方向性のある非対称な関係であることとともに, 通常の意味的類似性とは異なるタイプの情報であることを議論している.

今回の分析においては, ある程度の強さを持つ非対称性が強い想起関係を分析対象とするため, $|evo(a, b) - evo(b, a)| > 5.0$ を満たす 1,604 件の語彙概念ペアを選んだ. ここで, $evo(a, b)$ は起点概念 a からターゲット概念 b への評定値を表す.

3.2 EDR 概念辞書

EDR 電子化辞書 [Yokoi 84]^{*3}は, 各種の辞書から構成される大規模な言語資源 (約 40 万の概念がおおよそ 900 万の概念間関係で結ばれる) であり, 概念辞書はその重要な構成要素である. 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. 今回は特に, PWN には情報が含まれていない (3) のタイプの情報が豊富であることから本言語資源を用いることとした.

なお, EDR の電子化辞書における特徴の一つとして, 概念が品詞により分類されていないことがあげられる. これは PWN と明確に異なる点である. PWN は語彙化概念を扱っているのに対し, EDR における概念は, 概念に関連する日英の単語は辞書体系中から検索できるものの, かならずしも語彙化された概念に限定されないという設計上の思想による. ただし, EDR における情報構造も PWN と同様のネットワーク構造 (概念をノード, 概念間関係をエッジとする) としてモデル化することができる [Hayashi 12b].

3.3 EDR 概念辞書による意味ネットワーク

概念辞書における各概念をノードとし, 概念間関係をラベル付き有向エッジとする意味ネットワークを構成する. 表 1 に関係の種類・説明, および, 関係の数を示す^{*4}.

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のなので, 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. このときの関係ラベルは, 上位関係を表す "upper_of" に対しては "lower_of" を与え, その他の関係については, 接頭辞 "i." を付加した. この結果, 表 1 に示した関係の数は 2 倍になっている.

構成された意味ネットワークの規模は, ノード数:392,751, エッジ数:1,794,199 であり, 無向グラフと見たときの平均クラスタ係数:0.0087, ランダムに選んだ 1,000 個のノード間の平均最短経路長:5.42 であった.

3.4 PWN 語彙概念の EDR 概念への対応付け

EDR 意味ネットワーク上で経路探索を行うために, 抽出した想起データ (1,640 語彙概念ペア) における起点・ターゲット概念 (965 の異なり PWN 語彙概念) 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に開発された手法 [Hayashi 12a] を逆方向に適用し, 対応付けの候補を抽出した. なお, [Hayashi 12a] では, PWN が提供する語義タグ付きコーパスを利用したが, これに相当する語義解消されたコーパスとして, EDR 日本語コーパスを利用した.

上記の手法により 965 個の PWN 語彙概念中の 886 個 (91.8%) に対して, 対応する EDR 概念を上位 10 件内に抽出できたので, 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た. なお, 手法により抽出した順位 1 位のものが正解であった割合は 49.5%, また, 正解の逆順位の平均 (MRR) は 0.694 であった.

EDR 概念への対応付けが行えた 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対象とするため, 分析対象の件数は, 1,350 概念ペア (歩留まり率: 84.2%) となった.

4. 主な分析結果

4.1 最短経路探索

いわゆるダイクストラのアルゴリズムにより最短経路群の探索を行った. 今回は, 特にエッジに重みを与えることをしなかったため, 最短経路は純粋にエッジの数による. 計 1,350 の概念ペアに対し, 総計 52,517 の経路が探索されたが, 各概念ペアに対して探索された最短経路の数の平均/分散は 38.9/8050.97 と非常にばらつきが大きい結果となった. 一方, 各経路の経路長の平均/分散は 2.80/0.67 であり, 上記で報告したランダムに選択した 2 つの概念ノード間の平均最短経路長より短かく, ある程度の強さの想起関係が認められる概念ペア間のネットワーク上の距離は近い傾向にあることが示唆される. また, 容易に想像できるように, 経路長と経路数との間には一定の相関関係 (Spearman の相関係数:0.51, $p < 0.001$) が認められた.

図 2 に実際に探索された経路群の例を示す. この例における起点概念は {arrow, pointer} であり, ターゲット概念は {legend, caption} である. これらはそれぞれ, 10c0e3, 0ec57d という概念識別子を持つ EDR 概念へと対応付けられ, 図に示すように経路長が 4 である最短経路が 2 つ探索された. ここで, 最適経路探索によって選択された逆方向のエッジは再度方向性を逆

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

*3 http://www2.nict.go.jp/out-promotion/techtransfer/EDR/J_index.html

*4 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提示されているが, 実際の辞書データには表に示した関係以外の関係は付与されていない.

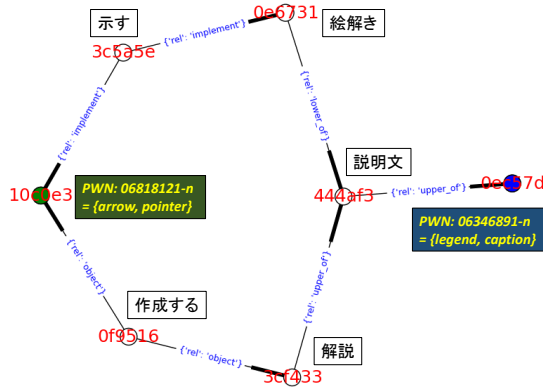


図 2: 探索された最短経路の例 (想起の評価値:5.03)

転し, EDR 概念辞書のオリジナルの関係の方向性に反していることに注意されたい. この図において, 例えば 3c5a5e (示す) と 0e6731 (絵解き) を介する経路は, ターゲット概念である legend/caption (凡例) が絵解き (illustration) と兄弟概念であり, 起点概念である arrow/pointer (矢印) と同じく示す (show) 際の手段となりうることを表している.

4.2 系列パターンマイニング

上記により得られた EDR ネットワーク上の各経路を概念連鎖の系列とみなし, Prefixspan と呼ばれる系列パターンマイニングのアルゴリズム [Pei, et al. 01] を実装したツール*5を用いて, 頻出する部分系列を抽出した.

ここで, 部分系列を抽出する単位としては様々なものが考えられるが, (1) ノードを考慮せずに, 関係種別を表すエッジのラベルのみを単位とする場合 (関係連鎖と呼ぶ), (2) ノードと, そこへ至る/そこから出て行く 2 つのエッジのラベルを組み合わせた 3 項組を単位とする場合 (3 項組連鎖と呼ぶ), の 2 通りの場合について分析を行った.

いずれの分析においても, 連鎖におけるギャップは許容せず, 連続した部分系列のみカウントした. また, サポート頻度の最小値を 250 とした.

4.3 関係連鎖の分析

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 ($\alpha \rightarrow \beta \rightarrow \gamma$) を対象とし, 概念関係の連鎖のパターン化を試みる.

図 3 は, 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) に形式的に分類している. ここでも, 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再度逆転し, EDR 概念辞書のオリジナルの関係の方向性に反していることに注意されたい. 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頻度 (すなわち部分系列が利用された頻度) を示しており, 推移的な連鎖 (p2, p4) よりも共通的な要素を介したパターン (p1, p3) の方が多いという結果となっている.

各類型において頻度が 400 以上あった概念関係の組み合わせを表 2 に示し, それぞれにおける特徴について議論する.

- p1: 後述する p3 とともに, 概念 β を介して, 概念 α から概念 γ が連想されていると見ることが出来る類型である. rel-2 に関係 a-object を持つパターンは, 動詞的概念 α

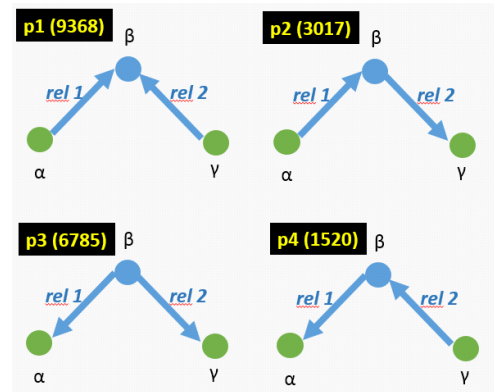


図 3: パス長=2 の概念連鎖パターンの形式的類型

表 2: 各類型において頻出する概念関係の組み合わせ

類型	rel-1	rel-2	頻度
p1	object	a-object	3,919
	scene	a-object	904
	object	place	586
	implement	object	492
	object	implement	481
	a-object	a-object	451
p2	object	upper_of	593
	scene	object	486
p3	a-object	a-object	1,129
	implement	object	991
	object	implement	824
	goal	goal	710
	scene	scene	519
	agent	agent	516
p4	cause	object	432
	object	scene	499
	scene	object	432

の格要素である名詞的概念 β を形容詞的概念 γ が限定していることを示しており, 結果として, 動詞的概念 α から形容詞的概念 γ への連想関係の可能性を示す. また, 双方の関係が a-object であるパターンは, ある共通的な名詞的概念 β を介して, 形容詞的概念である α , γ の間に連想関係があることを示す. これ以外のパターンは, 動詞的概念 α , γ が名詞的概念 β を介して連想関係にある可能性を示す.

- p2: 表には多くを記載できていないが, rel-1 として object 関係を持つものが多く含まれる. また, rel-1 として lower_of, rel-2 として upper_of を持つものが上位に来ており, 概念 β は概念階層的に上位に位置する概念であることが示唆される.
- p3: p1 同様, 概念 β を介して, 概念 α から概念 γ が連想されている類型であり, 多くのものが共通的な動詞的概念 β における格要素を占める概念 α から概念 γ への連想関係が示唆される. これに当てはまらないものが, 双方の関係が a-object である頻度 1 位のものである. このパターンは, 共通的な形容詞概念 β に対して, それにより性質などが規定される名詞的概念 α , γ 間の連想関係の可能性を示しており, 2 つの名詞概念 (例: トマト, りんご) が顕在性の高い形容詞概念 (例: 赤い) を介して連想さ

*5 <http://prefixspan-rel.sourceforge.jp/>

表 3: p3 の類型における頻出概念 (上位 10 件)

概念	単語もしくは説明	頻度
3ceae3	～になる	1,729
3ce87d	行う	781
3ce7d8	言い表す	750
1fb02f	ない/存在しない	663
1fa328	表現する	585
3d06c7	ある状態にする	535
061c7d	～と称する	383
0e8ec0	抱える	348
36c39c	～できる	344
3bc006	ある状態である	287

れている可能性が示唆される。

- p4: 逆方向に推移的なこのタイプの頻度は低く、抽出された実例も解釈が困難なものがほとんどであった。よって、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みなす。

4.4 3 項組の分析

上記の分析により、概念 α と γ の間に介在する概念 β の重要性が明らかとなった。そこで例えば、図 2 における {implement, 3c5a5e (示す), implement} のような (rel-1, 概念, rel-2) の 3 項組を単位とし、図 3 と同様の類型化を行った。上記までと同様にサポート頻度の最小値を 250 とした場合、p1:3 種、p2:4 種、p3:20 種の 3 項組が抽出された (p4 は 0 種)。

p1 に対しては、scene/place といった関係と強い結びつきから 0f17a4 (～の範囲の中に)、object の関係との結びつきから 3cf980 (意味) といった特徴的な概念が抽出された。

p2 に関しては、0f3d48 (雑然としているさま) や、3f98b4 (現象、生理的属性や生理的变化) といった抽象度の高い上位の名詞概念が抽出された。

p3 において特徴的なことは、抽出された 20 種の 3 項組の中の 16 種の 3 項組において、rel-1, rel-2 がともに関係種別 object であったことである。表 3 にこれらにおける上位 10 件の概念 β をリストする。この表に明らかなように、多くの概念の組み合わせにおいて重要な役割を果たすことが想定される、抽象度の比較的高い動詞的概念が抽出されている。

5. 議論

本研究では、PWN ベースの想起を分析するために、PWN とは本来別の言語資源である EDR を援用した。これは、PWN には格関係に相当する関係が付与されていないことによるが、一方で EDR には、全体部分関係のような概念間関係が欠落している。また、概念間関係ではなく語彙的な関係ではあるが、反意語のような関係も含まれていない。すなわち、可能ならば、PWN, EDR 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。

これを可能とするためには、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が必要となる。UBY (A Large-Scale Unified Lexical-Semantic Resource) プロジェクト [Gurevych 12] においては、ドイツ語、英語を対象として 12 種類の言語資源を意味レベルで統合しており、EDR 電子化辞書のような日本発の大規模言語資源もこのような枠組みに組み入れていくことが望まれる。このためにも、語義・概念レベルでの辞書エンタリの対応付けは重要な要素技術である。

6. おわりに

想起を概念連鎖のショートカットと仮定し、大規模言語資源が提供する意味ネットワークにおける最短経路に対する系列パターンマイニングにより、概念連鎖のパターンを探る試みについて報告した。その結果、共通する第 3 の概念を介して 2 つの概念が結び付けられるという局所的な連鎖構造に関して、いくつかの顕著なパターンを抽出することができた。

しかし、想起関係の背後にあると仮定する妥当な概念連鎖を与えうするためには、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た状況での最短経路を探索することが必要になる。この点については、グラフ理論的な手法 [Mihalcea and Radev 11] を導入していく予定である。また、与えられた概念ペアに対して想起の程度を予測する手法を確立することも必要である。これについては、現在、機械学習に基づく手法の検討と評価を進めている。さらには、想起のような心的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実証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

- [Boyd-Graber, et al. 06] Boyd-Graber, J., et al.: Adding dense, weighted, connections to WordNet. *Proc. of the Third International WordNet Conference*, pp.29–36 (2006).
- [Bruni, et al. 14] Bruni, E., et al.: Multimodal distributional semantics. *Journal of Artificial Intelligence Research* 49, pp.1–47 (2014).
- [Clark 15] Clark, S.: Vector space models of lexical meaning. In: Lappin, C. and Fox, C. (Eds.), *Handbook of Contemporary Semantics, 2nd edition*. Wiley (In press) (2015).
- [Gurevych 12] Gurevych, I. et al.: UBY — A large-scale unified lexical-semantic resource based on LMF. *Proc. of EACL 2012*, pp.580–590 (2012).
- [Hayashi 12a] Hayashi, Y.: Computing cross-lingual synonym set similarity by using Princeton annotated corpus. *Proc. of the 6-th International Global WordNet Conference*, pp.134–141 (2012).
- [Hayashi 12b] Hayashi, Y., et al.: LMF-aware Web services for accessing semantic lexicons. *Language Resources and Evaluation*, Volume 46, Issue 2, pp.253–264 (2012).
- [Ma 13] Ma, X.: Evocation: analyzing and propagating a semantic link based on free word association. *Language Resources and Evaluation*, Volume 47, Issue 3, pp.819–837 (2013).
- [Mihalcea and Radev 11] Mihalcea, R., and Radev, D.: *Graph-Based Natural Language Processing and Information Retrieval*. Cambridge University Press (2011).
- [Mikolov 13] Mikolov, T., et al.: Distributed representations of words and phrases and their compositionality. *Proc. of NIPS 2013* (2013).
- [Pei, et al. 01] Pei, J. et al.: PrefixSpan: mining sequential patterns efficiently by prefix projected pattern growth. *Proc. of ICDE'01*, pp.215–224 (2001).
- [Silberer, et al. 13] Silberer, C., et al.: Models of semantic representation with visual attributes. *Proc. of ACL 2013*, pp.572–582 (2013).
- [Yokoi 84] Yokoi, T.: The EDR electronic dictionary. *Communications of the ACM*, Volume 38, Issue 11, pp.42–44 (1995).