意味的係り受け情報の利用により視点を考慮した単語間類似度の評価 *

永松健司　田中英彦†
東京大学大学院 工学系研究科

アブストラクト

自然言語処理において、単語（概念）間の類似性を判定する処理は他の様々な処理への基礎となる指標を与えるものであり、多義語語義の曖昧さの解消の問題と絡めて様々な手法が提案されている。本稿では、コーパスから抽出した単語の意味的な係り受け情報を基に視点に応じた重み付けを行なう単語間類似度の評価を行なう。これにより、従来、与えられた単語対に固定であった類似度が、視点という文脈情報により柔軟に変化しうることになり、多義性解消などの応用においてもより適切な指標として利用できることが予想される。評価においては、類義語／非類義語対として与えられた大規模なデータに対する全体的な評価と、個々の類似度値が人間の判断する類似性とどの程度相似であるかの評価をまとめる。

1　はじめに

自然言語処理では、様々な場面での基礎的な指標として、二つの表現間の類似性を利用することができる。特に、意味、構造などの様々な曖昧性の解消処理では、既出の表現（文脈情報）との類似性を考慮することで、より適切な結果を得られることが期待されるが、実際、これまでの類似性に関する研究では、多義語語義の曖昧さの解消を課題として、いろいろな類似性の尺度を用いた曖昧さ解消の手法が提案されてきた[3, 1, 2]。

しかし、同じ二つの言語表現（概念）であっても、それがどのような文脈において判断されるかによって、それらの類似性も異なってくるが、これまでそのような文脈依存性は扱われていなかった。我々の研究では、そのような文脈情報を視点と呼び、視点情報によって類似の度合が変化する類似性の尺度を提案し、評価を行なっている。

本稿では、単語（概念）間の類似性判定において、コーパスから抽出した意味的な係り受け情報を用いることで、視点情報に応じて概念（単語）間の類似度が変化する類似度計算手法（以下では類似性規準と呼ぶ）を提案・評価する。本稿の内容は、最初に本手法を説明し（2節）、比較するいくつかの類似性規準を述べた後、類義語対と非類義語対のデータセットを用いた、人間の判断に依らない評価実験の結果を示す（3節）。次に心理実験を基にした人間の類似性判断と類似性規準による類似度との比較実験の結果を示す（4節）。最後に、この二つの実験結果に対して考察を行ない（5節）、まとめる（6節）。

*Evaluation of Point-of-View-based Word Similarity Measure employing Semantic Dependency Information

†Kenji Nagamatsu　Hidehiko Tanaka
{naga, tanaka}@mtl.t.u-tokyo.ac.jp
Faculty of Engineering, University of Tokyo

2　意味的な係り受け情報を基に視点変化を考慮した類似性規準

2.1　共起データへの意味的係り受けの考慮

後に挙げる結果（3節、4節）でも示されるように、これまでに提案されている類似性規準の中のシソーラスの構造に基づく手法とコーパスから抽出したデータを利用する手法では、大規模データによる評価（3節）と人間の類似性判断との比較評価（4節）で一長一短の性質を示す（5節を参照）。

これは、シソーラス構造に起因する出力可能な類似度値の少なさ、および共起データが位置の近さだけに依り、単語間の意味的な関係を取り出し得ていないことが、その主な原因である。これに対して、Resnik[4]は、シソーラス構造に情報量を導入することで出力可能な値の数の増加を図っている。

一方、本研究では、コーパスから抽出した共起データを基本データとして用い、そこに意味的な制約を加えるという逆の立場を採る。さらに、その意味的な制約の中に視点情報というパラメータを導入して、文脈情報による類似度の変化を考慮した。

2.2　共起類似度への視点情報による重み付け

本手法では、(1)式に示す共起情報に基づく類似性規準を基本とする。これは、類似度を求める単語対の双方と共起し得る単語の共起確率の和を類似度として用いるものであり、視点単語 w_p の下での単語対 (w_1, w_2) の類

似度 $Sim(w_1, w_2; w_p)$ は、

$$Sim(w_1, w_2; w_p) = \sum_{\forall w \in Co(w_1) \cap Co(w_2)} \frac{Pr(w|w_1; w_p) + Pr(w|w_2; w_p)}{2} \quad (1)$$

で与えられる。ここで、$Co(w)$ は単語 w の共起単語の集合を表し、$Pr(w|w'; w_p)$ は単語 w' と共起する単語 w の共起確率に対し、次に示す (2) 式に基づく視点単語 w_p による変化を加えたものである。

二つの単語間の共起確率は、それらの関連性の強さを示すと考えられるが、その関連性の度合は、視点によって変わるものである。そこで、$Pr(w|w'; w_p)$ に対して、次式による重み付けを導入する。

$$Pr(w|w'; w_p) = \frac{\alpha^{\mu(w_p, w)} f(w, w')}{(\alpha^{\mu(w_p, w)} - 1) f(w, w') + \sum_{\forall x \in Co(w')} f(x, w')} \quad (2)$$

ここで、$f(w, w')$ は単語 w と w' の共起頻度を表し、$\mu(w_p, w)$ は視点単語 w_p と単語 w との間の意味的な係り受け関係の強さを示す値である。この式の意味は、共起する単語 w が視点単語 w_p と関連が強い程、類似度 $Sim(w_1, w_2; w_p)$ に対する寄与も大きくなるということになる。

$\mu(w_p, w)$ の値としては、本研究では意味的な係り受け情報をタグ付きコーパスから抽出し、そこでの相互情報量 (MIC) を求めて利用する。ただし、直接に相互情報量を求めることができないため、次式に示す近似を行ない、$\mu = mic(w, w')$ としている。

$$\begin{aligned} MIC(w, w') &= \log \frac{Pr(w, w')}{Pr(w) Pr(w')} \\ &\approx mic(w, w') \\ &= \log \frac{\sum_k co(w, w', r_k)}{\sum_{i,j} co(w, w_i, r_j) \sum_{i,j} co(w', w_i, r_j)} \end{aligned} \quad (3)$$

ここで意味的な係り受け情報とは、意味フレーム内での主概念と従属概念間の格属性を介した係り受け関係を指す。具体的には EDR コーパスにおける main スロット内の概念と、それに係る他の格スロット内の概念、およびその格属性の三つ組の頻度で表される。現在、EDR コーパスから 1,254,851 組の頻度データ $co(w_m, w_r, r)$ を取り出して利用している。

3 類義語対・非類義語対集合を用いた評価

3.1 比較評価する類似性規準

実験（本節、および次節の心理実験）で評価した類似性規準は以下のものである。1~3の詳細は [5] を参照。

1. 入力単語対に対し、シソーラス（EDR 概念体系辞書）中の複数の共通上位概念の深さの内、最も大きい値を類似度とするもの（depth）
2. 入力単語対に対するシソーラス（EDR 概念体系辞書）中のノード同士を結ぶ最小リンク数（の逆数）を類似度とするもの（link#）
3. 毎日新聞 94 年度から抽出した単語ごとの共起単語情報を利用し、入力単語対に対する共通共起単語の生起確率合計値を類似度とするもの（co）

   (1) 式において視点情報を考慮しない場合に一致
4. 毎日新聞 94 年度から抽出した、単語ごとの生起確率値をシソーラス（EDR 概念体系辞書）の各ノードに付与し、入力単語対に対する共通上位概念の情報量最大値を類似度とするもの（resnik95）[4]

これらに対して、前節で述べた手法 pov でパラメータ α を変えた三種類 pov(1.2)、pov(1.5)、pov(2.0) を加えて評価実験を行なった。co と pov で利用している共起データは、毎日新聞 94 年度から抽出したものである。

3.2 類義語対 − 非類義語対を用いた評価

それぞれの各類似性規準において、類似と判断されるスレッショルドを変えていった場合に、与えられた類義語対（10,297 対）と非類義語対（100,000 対）の内、どのくらいの割合が類義と判断されるか（被覆率）を評価した結果を図 1・図 2 に示す。

類義語対としては IPAL 辞書の類義語フィールドから抽出した単語対を使用し、非類義語対としては、それぞれの類似性規準で使用している単語辞書内から、無作意に選んだ単語対を（近似的に）使用している。

また、この実験では、pov に与える視点情報として、入力となる単語対中の単語を使用した。これは個々に視点情報を設定すると評価が困難になるためで、この実験のような全体的な振舞いを調べる場合には妥当だと思われる。

このグラフは、あるスレッショルドの下で、類義語対集合のある一定割合を類義と判断できる時に、類義と（間違って）判断されてしまう非類義語対の割合を示すものであり、データ系列がグラフ中で下方に位置する程、類義語対と非類義語対の分離の度合が良いと判断される。

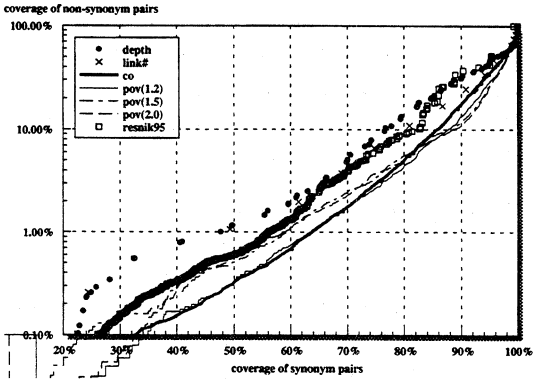

図 1: 類義語対被覆率に対する非類義語対被覆率

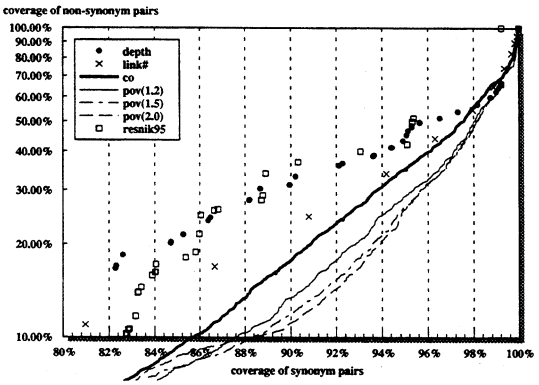

図 2: 被覆率 (図 1の拡大)

4 人間の類似度判断に関する心理実験

前節の結果からは個々の単語対に対する類似度値が妥当かどうか、すなわち人間の判断とある程度の相関を持つ値かどうかを判断できない。そこで、心理実験を通して人間が判断する類似度との比較を行なった。

4.1 実験方法

実験は、前節で述べた類似性規準それぞれにより求めた類似度と、人間の判断により与えられた類似度との比較で行なう。被験者は工学系大学院生（男性）14名であり、あらかじめ与えられた単語対100個に対して、1（類似性なし）から5（完全な同義）までの得点を付与させる。

この実験で用いた単語対は、IPAL辞書の類義語対から無作意に50個、また分類語彙表の最下位項目内から無作為に選んだ単語対から50個、計100個の単語対を用いている。双方とも単語対には類義語関係があると見做し得るが、シソーラスである分類語彙表の単語対は、IPAL類義語対よりも類義性が弱いことが予想される（表1参照）。

4.2 実験結果

実験に使用した単語対100個の内、3節の各手法すべてで実際に類似度値が求まったものは62個であった。しかし、各被験者が付与した得点は各人の主観に基づくものであり、単語対によっては大きなばらつきを示すことがあるため、以下では更に得点の標準偏差が1.0以下の単語対53個に限って検討を加えることにする。

まず、単語対を抽出した辞書の違いによる類似性の度合の違いを調べるために、IPAL辞書、分類語彙表それぞれの単語対での得点の平均値を表1に示す。

単語対のソース	IPAL	分類語彙表
有効な単語対数	29	24
得点の平均値	3.204	2.511

表 1: 元となった辞書とその単語対の類似性

手法	単語対全体	IPAL 単語対	分類語彙表
depth	0.380	0.164	0.449
link#	0.365	0.104	0.442
co	0.344	0.211	0.306
pov(1.2)	0.390	0.210	0.415
pov(2.0)	0.424	0.232	0.495
resnik95	0.426	0.235	0.420

表 2: 各手法による類似度と人間の得点との相関係数

次に、各手法により計算した類似度と人間が判断した得点との相関係数を表2に示す。ここでは、単語対全体で求めた相関係数と共に、各辞書ごとの単語対内での相関係数も併せて示している。

5 考察

図1・図2から、単語の共起情報を用いた類似性規準は、判定可能な単語（概念）が利用したコーパスに依存するという問題があるとは言え、シソーラス構造を用いた類似性規準よりも分離精度は高いと言える。そして、co に対して意味的係り受け情報と視点情報による重み

	link#	co	p(1.2)	p(2.0)	res95
depth	0.970	0.132	0.175	0.211	0.910
link#		0.198	0.247	0.268	0.883
co			0.938	0.809	0.125
p(1.2)				0.942	0.200
p(2.0)					0.249

表 3: 各手法間での相関係数

付けを行なった pov は、オリジナルよりも高い分離精度を示すことが示された。

図1では、パラメータ α の選び方により（$\alpha = 1.5$ or 2.0）、co よりも分離精度が落ちる場合があるが、率としてはわずかであり、また類似性規準が実際に利用される場合では、類義語の 80% を覆うなどの設定がなされるため、実質的な問題はない。

次に、表2の結果より、単語対全体で見た場合、シソーラスに基づく類似性規準（depth, link#, resnik95）は共起情報のみに基づく co よりも、人間の判断との高い相関を示すことが分かる。これは、シソーラス自体が人間の判断を基に構成されたものであり、その影響が直接に現れているからだと考えられる。

しかし、共起情報に2節で述べた重み付けを行なうことで（pov(1.2), pov(2.0)）、人間の判断との相関はそれよりも高くなり、適切なパラメータ（pov(2.0)）を選択することで、評価した手法の中では最も高い相関を示すことも示された。また、それぞれの辞書ごとの単語対に対する結果を見ても、pov(2.0) では、人間の判断との相関が最も高くなっていることが分かる。

表3に、それぞれの類似性規準による類似度同士での相関係数を求めた結果を示す。当然ながら、シソーラス構造に基づくもの、共起情報に基づくもの同士では高い相関値を示すが、一方、この2グループ間での相関はかなり低くなる。つまり、それぞれのグループごとに妥当な類似度を出力しうる単語対のある範囲があることが予想される。

6 おわりに

本稿では、共起データを用いる類似性規準に対して、コーパスから抽出した意味的な係り受け情報を重み付けという形で導入する手法を提示した。この手法ではさらに、パラメータとして視点情報を考慮し、視点に応じて類似度値が変化しうるようにしている。

次に、この手法と他のいくつかの類似性規準と共に、類義語対と非類義語対集合の被覆率による評価を行ない、その問題点を補う目的で、心理実験を通して人間の判断による類似度との比較を行なった。その結果、本手法は、被覆率による評価で高い分離精度を示し、人間の類似性判断との相関の面でも、他の類似性規準と較べて高い精度を示した。

さらに、シソーラスに基づく類似性規準と共起情報に基づく類似性規準は、その類似度値の相関において、二つに分かれていることが示された。このことは、それぞれを相補的に用いることで、更に高い精度および、人間の判断と高い相関を持つ類似性規準が得られる可能性を示すと思われる。今後は、それらをどのように組み合わせることで精度が向上するかを調べる必要がある。

本研究には、情報処理振興事業協会の計算機用日本語辞書 IPAL、国立国語研究所の分類語彙表、毎日新聞社の CD-毎日新聞「94 年版」を利用させていただいたことを感謝します。

参考文献

[1] E. Agirre and G. Rigau. A proposal for word sense disambiguation using conceptual distance. In *Proceedings of 1st International Conference on Recent Advances in Natural Language Processing*, 1995.

[2] Y. Karov and S. Edelman. Learning similarity-based word sense disambiguation. In *Proceedings of the Fourth Workshop on Very Large Corpora*, 1996.

[3] Y. Niwa and Y. Nitta. Co-occurrence vectors from corpora vs. distance vectors from dictionaries. In *Proc. COLING 94*, Vol. 1, pp. 304–309, 1994.

[4] P. Resnik. Using information content to evaluate semantic similarity in a taxonomy. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence*, Vol. 1, pp. 448–453, 1995.

[5] 永松, 田中. コーパスから抽出した係り受け共起情報に基づく類似度と文書検索における評価. 情報処理学会研究報告 自然言語処理 研究会, 96-NL-116, 96(114):73–78, Nov. 1996.