単語対に対する類似性規準の心理実験による評価 *

永松健司　田中英彦†
東京大学大学院 工学系研究科

1　はじめに

自然言語処理において，二つの表現間の類似性を判定する処理は，様々な場面での基礎的な指標として利用しうる．特に，様々な曖昧性（意味，構造など）の解消処理では，既出の文脈情報との類似性を考慮することで，より適切な結果を決定できるようになることが予想され，実際に，多義語語義の曖昧さの解消に関しては，いろいろな類似性の尺度と共に，それを用いた曖昧さ解消の手法が提案されている．

本稿では，単語（概念）間の類似性判定において，いくつかの類似度計算手法（以下では類似性規準と呼ぶ）による類似度と人間の判断による類似度とを比較・考察する．本稿の内容は，まず比較する類似度の計算手法を述べ（2 節），人間の判断に依らない評価実験の結果を示す．次に，その問題を指摘し，今回行なった実験の意味・方法を述べた後に，その結果を示す（3 節）．最後に，考察を行ない（4 節），本稿をまとめる（5 節）．

2　類似性規準と類義語対集合に対する評価

2.1　評価する類似性規準

本稿で述べる実験で評価した類似性規準は以下のものである．その詳細は [2] で述べられている．

- 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）
- 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#）
- 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）
- この co に対し，共起単語の相互情報量で重み付けした生起確率合計値を類似度とするもの（パラメータを変えた pov(1.2),pov(2.0) の 2 種類）
- 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） [1]

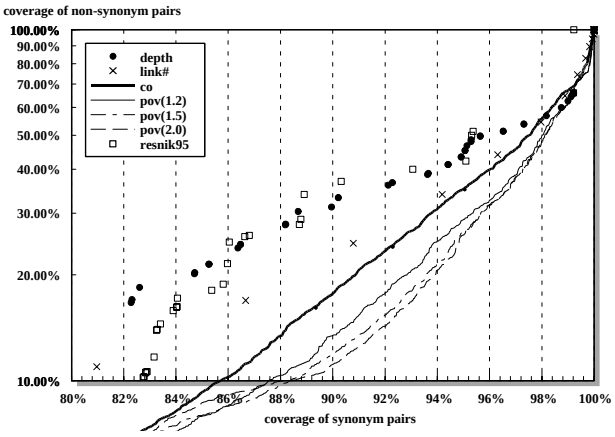

図 1: 類義語対被覆率に対する非類義語対被覆率

2.2　類義語対 − 非類義語対を用いた評価

これらの各類似性規準において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に示す．

このグラフでは，あるスレッショルド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ほど，類義語対と非類義語対の分離の度合が良いと判断できる．

つまり，共起単語情報を利用することにより（co, pov*），シソーラスのみに基づく類似性規準（depth, link#, resnik95）よりも高い分離精度が得られることが分かる．

しかし，この結果からは個々の単語対に対する類似度が妥当かどうか，すなわち人間の判断とある程度の相関を持つものかどうかを判断できない．そこで，次節では，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うことで，別の評価基準を設けて比較を行なう．

---

*Comparing Some Computational Similarity Measures with Human Beings' Judgment

†Kenji Nagamatsu　Hidehiko Tanaka
{naga, tanaka}@mtl.t.u-tokyo.ac.jp
Faculty of Engineering, University of Tokyo, 7-3-1 Hongo, Bunkyo-Ku, Tokyo, 113, Japan

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

3 人間の類似度判断に関する心理実験

3.1 実験方法

実験は，前節で述べた各手法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 辞書の類義語フィールド内から，また分類語彙表の同一最下位項目内から無作為に取り出したものをそれぞれ 50 個ずつ，計 100 個の単語対を用いている．ただし，それぞれの単語対内の単語同士は類義語関係にあるが，シソーラスである分類語彙表からの単語対は類義性が弱いことが予想される．

3.2 実験結果

実験に使用した単語対 100 個の内，2節の各手法で実際に類似度が求まったものは 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．

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す．

4 考察

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 よりも高い相関を示す．しかし，共起情報に重みを加えることで（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が示された．

	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

また，それぞれの辞書ごとの単語対に対する結果を見ると，適切なパラメータ（pov(2.0)）で，人間の判断との最も高い相関が得られていることが分かる．

次に，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を求めた結果を表 3に示す．当然ながら，シソーラスに基づくもの，共起情報に基づくもの同士では高い相関を示すが，この 2 グループ間での相関はかなり低い．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．

5 おわりに

本稿では，いくつかの類似性規準に対して，類義語対と非類義語対集合の被覆率による評価を示し，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，被覆率による評価で高い分離精度を示した類似性規準が，人間の判断結果との相関の面でも，他の類似性規準と較べて高い，または同等の精度を示すことが示された．

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，二つに分離しうることが示された．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．

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．

参考文献

[1] 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.

[2] 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.