自然言語処理 116－11
（1996. 11. 18）

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

永松 健司　　田中 英彦

東京大学大学院 工学系研究科

概 要

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共起情報による類似度計算に、同じくコーパスから抽出した単語間の係り受け関係に基づく重み付けを行なう手法を提案する。また、類義語対と非類義語対の集合の分離精度による評価と併せて、実際の問題への応用という面から、文書検索処理に類似度計算を利用した場合の評価を行なう。

Evaluation of a similarity measure based on co-occurrence and dependency between words

Kenji Nagamatsu　　Hidehiko Tanaka

University of Tokyo　Faculty of Engineering

ABSTRACT

In natural language processing, similarity between words can be a basis for other processings. In fact, various similarity measures have been proposed and evaluated especially in word sense disambiguation processes. In this paper, we propose a similarity measure based on co-occurrence of words and syntactic dependency of words. These informations are extracted from a corpus. Moreover, We evaluate this measure using large sets of synonym and non-synonym pairs and in a simple document retrieval task.

1 はじめに

自然言語処理において，ある表現と別の表現との類似性を判定する処理は，様々な場面で基礎的な指標としての利用が考えられる．特に，様々な曖昧性（意味，構造など）の解消処理においては，既出の文脈情報との類似性を考慮することにより，より適切な結果を決定することが可能になることが予想され，実際に，多義語の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様々な手法が提案されている．

表現の類似性を判定する処理の内，特に単語，もしくは概念間の類似度を計算する処理について，これまでの研究に見られる手法を分類すると，大きく次の3つに分類される．

1. シソーラスや概念ネットワークの構造に基づく類似度 [4, 5, 2, 9, 12, 6, 1]

   類似性を判定できる単語の範囲が広い．類似性の規準が緩やか．

2. コーパスから得た統計的情報に基づく類似度 [11, 3, 7, 14, 13, 10, 8]

   語の使用に則した様々な類似性の考慮．判定可能な単語の範囲はコーパスの大きさに依存．

3. 各概念に与えた属性値（性質）集合に基づく類似度

以上の分類，およびその特失はそれぞれ代表的な特徴を述べたものであり，これらの手法の境界が必ずしもはっきりしていることを示すものではない．例えば，[4] や [5] などは概念ネットワーク的な手法に共起情報という統計的な指標を組み合わせている他，[6] ではシソーラスと統計情報とを組み合わせることなどが研究されている．

本稿では，単語（概念）間の類似性判定処理において，統計的な共起確率情報と共に、コーパスから抽出した係り受けデータを考慮する類似度計算の手法を述べる（3 節）．また，評価の指標として類義語対と非類義語対集合の被覆率（何%カバーできるか）を用いた評価（2 節、3 節），および応用面での評価として，文書検索処理における文書のカテゴリ分け問題での評価（4 節）を示す．

2 類似性規準の被覆率による評価

2.1 類義語対と非類義語対の被覆率

これまでの概念間の類似性に関する研究における評価では，出力された類似度に対する主観的な印象や，サンプリングした標本の被験者による判断を求める心理実験によるものが多かった．これは，類似性という基準を客観的に評価することが難しかったことを示している．

これに対して本稿では，類似性規準に対してあるスレッショルドを設定した場合に，与えられた類義語対と非類義語対の集合の内，それぞれどのくらいの割合が類似であると判定されるかを示す指標（以下，被覆率と呼ぶ）の間の関係として図示することで評価する．

すなわち，ある類似性規準に対してスレッショルド t を設定した場合に，ある単語対の集合 S の内，類似と判定される単語対の集合を $S(t)$ で表記すると，

$$\begin{cases} x = \frac{|S_{\text{類義語対}}(t)|}{|S_{\text{類義語対}}|} \\ y = \frac{|S_{\text{非類義語対}}(t)|}{|S_{\text{非類義語対}}|} \end{cases} \quad (1)$$

というパラメータ表現で図示される関係となる．

2.2 代表的な既存の手法での評価例

シソーラス構造を用いた類似性規準，統計的情報を用いた類似性規準の内，代表的なものについて，2.1 節で述べた評価方法を適用した結果を図 1 と図 2 に示す．ここでは，シソーラス構造による類似性規準として，単語間の共通上位概念の深さ (depth) と単語間のリンク数 (link#) によるものを，統計的情報による類似性規準として，共通の共起単語の IDF 値 (Inverse Document Frequency) を共起確率で重み付けした値の和 (cooccur1) と共通の共起単語の共起確率値の和 (cooccur2) によるものを評価した．図中では，同じ記号で 2 つのデータ系列が示されているが，より上に位置する群は類義語対集合として分類語彙表から抽出したものを，下に位置する群は IPAL 辞書から抽出したものを使用した場合である（後述）．

シソーラス構造による類似性規準に対しては，シソーラスとして EDR 概念体系辞書を利用し，入力の単語対に対応する概念の組み合わせの中から最も類似度が高いものを出力とした．

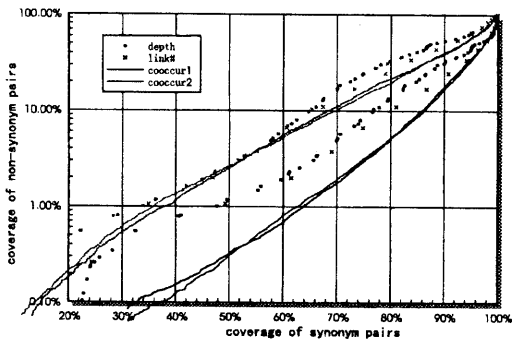

図 1: 類義語対 (x) と非類義語対 (y) の被覆率の関係

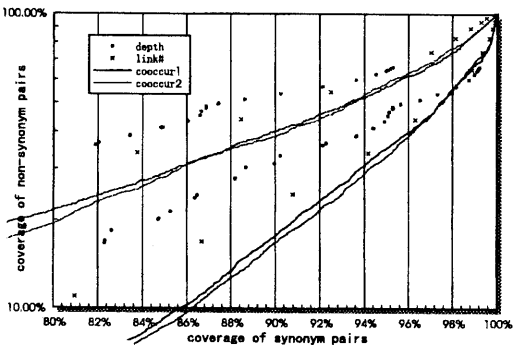

図 2: 類義語対 (x) と非類義語対 (y) の被覆率の関係 (図 1 の一部を拡大)

また，共起単語による類似性基準に対しては，毎日新聞社の CD-毎日新聞「94 年版」の記事中 74,793 記事を形態素解析プログラム JUMAN[15] で解析した結果から内容語に絞って，各語の共起単語とその頻度を抽出した共起単語辞書を作成、利用している．

この評価で使用する類義語対／非類義語対の集合として，類義語対には

1. IPA による日本語辞書 IPAL の各見出し単語と，その同義語／類義語フィールドに含まれる単語との対 (10,297 対)
2. 国立国語研究所による分類語彙表で最も下位の分類項目に含まれる単語集合から無作為に 2 つの単語を選択した単語対 (10,000 対)

の 2 つの集合を，一方，非類義語対には

1. シソーラス構造による類似度に対しては
   - EDR 日本語単語辞書から無作為に抽出した単語対 (100,000 対)
2. 共起単語による類似度，係り受け共起関係による類似度に対しては
   - それぞれの共起辞書に含まれる見出し単語から無作為に抽出した単語対 (100,000 対)

を使用した．ただし，今回は非類義語対として無作為抽出による単語対を使用したため，必ずしも類義語でないものばかりが含まれるわけではなく，一部，実際には類義だと判断されるような単語対も一定割合で入ることが考えられる．よって，ここではあくまでも，非類義語対の近似集合であり，評価の際にはそのことを考慮する必要があることを指摘しておく．

最後に，ここで評価した各手法を用いて類似性を判定することができた類義語対と非類義語対の数を表 1 に示す．それぞれの手法で用いている辞書やシソーラスに含まれない単語があるため，上述の数値よりもいくらか減少しており、これはそれぞれの手法が適用できる単語の範囲の広さを示すと見なせる。

2.3 既存の類似性規準についての考察

式 1 によると，図 1 や図 2 で下に位置する類似性規準ほど，ある一定割合の類義語対を覆うようなスレッショルドの設定に対して，より少ない割合の非類義語対を類義だと判定する，すなわち，類義語対と非類義語対の分離精度が高いことを示している．

ここから，1 節で述べたように，統計情報（ここでは共起単語）を用いた類似性規準はシソーラス構造による類似性規準よりも高い分離精度を持つことが確かめられる．

また，表 1 で実際に判定できた単語対の数を見ると，シソーラス構造による類似性規準では約 9 割以上に対して適用できるのに対して，統計情報による類似性規準では約 7 割前後となり，統計情報をそのまま使う場合の適用範囲の狭さが示されている．

3 係り受け共起関係による類似性

本節では，本稿で評価を行なう係り受け共起関係を利用した類似性規準について述べる．

類似性の規準	IPAL 辞書の類義語対	分類語彙表の類義語対	非類義語対
シソーラス構造による類似度	9,709 / 10,297	9,126 / 10,000	98,699 / 100,000
共起単語による類似度	7,443 / 10,297	6,755 / 10,000	100,000 / 100,000
係り受け共起関係による類似度	7,443 / 10,297	—	100,000 / 100,000

表 1: 評価の際に使用した単語対の数，および実際に類似度を計算できた単語対の数

3.1 共起データへの係り受け関係の考慮

1 節，2.3 節で述べたように，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．しかし，現在のところ，各単語の共起単語として抽出する単語は，ある一定範囲内に存在する内容語などというような判断基準であり，まだ確固とした選択基準があるわけではない．

そこで，本稿ではその判断の一端として，共起単語間での係り受け関係によって重み付けを行なう類似性規準において，その評価を行なった．つまり，本研究での係り受け共起関係とは，共起データ（共起単語対）の内，元コーパスにおいて構文的な係り受け関係を持っているものを指している．

共起関係は単語間の関連性を示す尺度であり，その関連性の範囲を緩めるには抽出する共起単語の数や種類を増やすことで対応できるが，逆により強い関連性を示す方向へ類似性規準を変更しようとすると，個々の共起単語が持つ性質を考慮せざるを得ない．この理由から本研究では，意味的な関連性にも影響を持ち，ある程度，客観的な判断が行なえる，構文的な係り受けという関係を利用した．

3.2 係り受けによる共起情報の重み付け

まず，類似性規準として，単語対の共起単語の内，共通に含まれるものの共起確率値の和を利用する場合（2.2 節の cooccur2 で利用した）の式を示す．単語 w の共起単語の集合を $C(w)$ で表記すると，単語 w_1 と単語 w_2 の類似度 $Sim(w_1, w_2)$ は，

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

と表される．ここで，$Pr(w'|w)$ は単語 w に対して，単語 w' が共起する確率を示し，単語 w の共起単語 w' の共起頻度を $f(w'|w)$ で表記すると，

$$Pr(w'|w) = \frac{f(w'|w)}{\sum_{\forall x \in C(w)} f(x|w)} \quad (3)$$

で与えられる．

この式は個々の単語の性質として何も仮定しないものの場合であるが，IDF 値などのように個々の単語に付与された性質を利用する場合も，共起確率での平均操作を行なうという点ではほぼ同じである．よって，係り受けによる重み付けも，この共起確率値へ対して行なうのが自然だと考えられる．そこで，(3) 式の共起確率値 $Pr(w'|w)$ を一部変更し，次式による重み付けを導入する．

- 単語 w と w' が係り受け関係になり得る場合は

$$Pr(w'|w) = \frac{\alpha f(w'|w)}{(\alpha - 1) f(w'|w) + \sum_{\forall x \in C(w)} f(x|w)} \quad (4)$$

一方，係り受け関係にあるかどうかの判定については，類似性判定という，比較的，計算コストが小さいことが要求される処理において，構文レベルの解析処理を入れることには問題があると考え，本研究では係り受け関係の判定においてもコーパスから抽出した統計データを用いている．具体的には，解析済みコーパス（EDR コーパス）から，意味的な係り受け関係にある単語対を抽出しておき，ここに含まれる単語対に対しては係り受け関係にある（なり得る）と判定している．

3.3 評価結果

図 3 と図 4 に，この類似性規準 (ck) の評価結果を示す．ここでは，類義語対として IPAL 辞書から抽出した 10,297 対のみについて評価を行なっている．(4) 式の α には，恣意的であるが $\alpha = 2.0$ として計

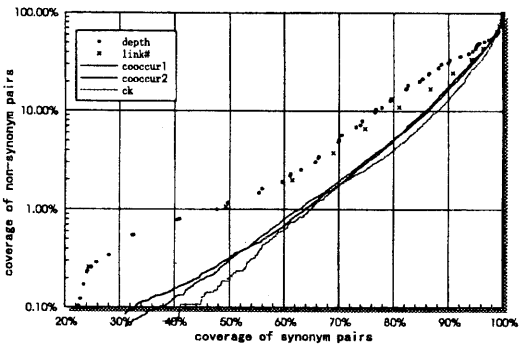

図 3: 類義語対 (x) と非類義語対 (y) の被覆率の関係

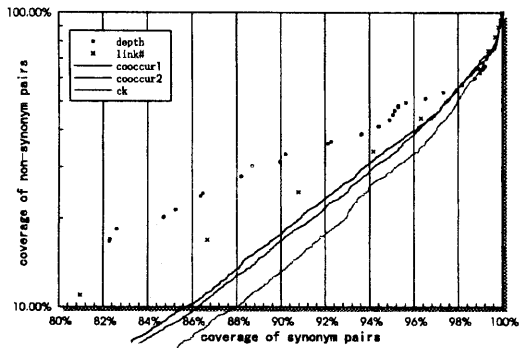

図 4: 類義語対 (x) と非類義語対 (y) の被覆率の関係 (図 3 の一部を拡大)

算した．また，2.2 節に示したシソーラス構造と共起情報を用いた類似性規準（それぞれ 2 つずつ）の結果も比較のために示している．

これらの図から分かるように，係り受け関係による重み付けを採用することで類義語対／非類義語対の分離の精度が向上している．具体的に，類義語対の被覆率のいくつかに対する非類義語対の被覆率で見ると，

類義語対	link#	cooccur1	cooccur2	ck
80%	10.9	5.0	5.2	4.0
90%	24.0	17.8	16.8	13.1
95%	33.9	35.2	33.0	28.9

のようになり，係り受け関係を判定することで，従来の手法よりも非類義語対を過剰に類似と判定してしまう割合を低くすることが可能となっている．

4 文書検索実験での評価

本節では，類似性規準を実際の問題に応用する場合の一例を通して，類似性規準の評価を行なう．

ここで行なった実験は，毎日新聞 94 年 1 月分から無作為抽出した 464 記事について，そのカテゴリ推定の正答率の評価，および代表的なカテゴリに属する記事ついて検索を行なった場合の適合率・再現率による評価である．この新聞記事 DB では各記事に，「1 面」，「社説」などのカテゴリ名が付与されており，ここではその内，内容との関連が高いと思われる，「文化」，「経済」，「芸能」，「国際」，「家庭」，「社会」，「スポーツ」の 7 カテゴリを選び，サンプルとなる記事もこれらのカテゴリに含まれるもののみに限定してある．

それぞれの記事に対しては以下の処理を行なう．

1. 上記の 7 カテゴリのラベルの共起単語の内，生起頻度が大きいものからそれぞれ 5 単語を選択する．
2. 記事中の内容語の内， IDF 値が大きいものから 5 単語を選択する．
3. 各カテゴリの 5 単語と記事に対する 5 単語の，すべての組み合わせにおける類似度の和を求め，この値をその記事がそのカテゴリに属する度合を示す値とする．
4. カテゴリ推定では，この値の最も大きいカテゴリを出力する．
5. 検索では，与えられたカテゴリに対して，この値へのスレッショルドを次第に低くしていった場合に，判定される記事が正しく選択されたかどうかで，その適合率と再現率を求める．

以下に、カテゴリ推定での正答率の結果を示す．

手法	depth	link#	
正答率	16.8%	17.3%	
手法	cooccur1	cooccur2	ck
正答率	34.3%	30.9%	34.7%

図 1（図 2）で下にある手法（類義語対と非類義語対の分離精度が良い）ほど，ここでの正答率も高くなっていることが分かる．

一方，検索処理での再現率と適合率の結果を図 5 に示す．これはカテゴリ「経済」に対する記事を検索した場合での結果である．

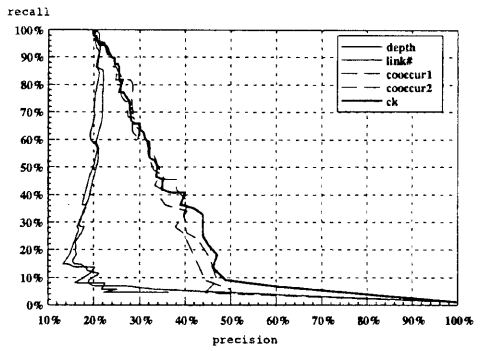

図 5: 新聞記事のカテゴリ検索での再現率と適合率

全体的な精度はあまり高くないが、結果について言えば，カテゴリ推定での順番，および図 1（図 2）での順番と一致していることが示されている．

本節での評価結果は，2 節や 3 節で示した大量のデータに対する類似性規準の統計的な性質が，具体的な個々の単語間での類似性判定でも同様に期待できることを示していると考えられる．

5 おわりに

本稿では，コーパスから抽出した係り受け関係を利用した類似性の判定手法について述べた．また，類義語対と非類義語対の被覆率による評価を示し，他の既存の類似性判定手法との比較を行ない，本手法の有効性に関して考察を行なった．さらに，応用面での評価の一例として，文書検索処理に利用した場合の実験を行なって，ある場合には，その他の手法よりも有用であることを示した．

本稿で示した様々な類似性判定の手法にはそれぞれ適する場合が存在する．今後は，それらをどのように組み合わせることで全体的な判定の精度が向上するかを調べる必要があると考える．

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] Agirre, E. and Rigau, G.: A Proposal for Word Sense Disambiguation using Conceptual Distance, *Proceedings of 1st International Conference on Recent Advances in Natural Language Processing* (1995).

[2] Dagan, I. and Pereira, F.: Similarity-Based Estimation of Word Cooccurrence Probabilities, *Proceedings of ACL-94* (1994).

[3] Iwayama, M. and Tokunaga, T.: Hierarchical Bayesian Clustering for Automatic Text Classification, 人工知能学会研究会資料, pp. 127–134 (1994). SIG-J-9401-17.

[4] Kozima, H. and Furugori, T.: Similarity between Words Computed by Spreading Activation on an English Dictionary, *Proceedings of the 6th Conference of the European Chapter of the Association for Computational Linguistics (EACL-93)*, ACL, pp. 232–239 (1993).

[5] Niwa, Y. and Nitta, Y.: CO-OCCURRENCE VECTORS FROM CORPORA VS. DISTANCE VECTORS FROM DICTIONARIES, *Proc. COLING 94*, Vol. 1, pp. 304–309 (1994).

[6] Resnik, P.: Using Information Content to Evaluate Semantic Similarity in a Taxonomy, *Proceedings of the 14th International Joint Conference on Artificial Intelligence*, Vol. 1, pp. 448–453 (1995).

[7] Yang, Y. and Chute, C. G.: An Example-Based Mapping Method for Text Categorization and Retrieval, *ACM Transactions on Information Systems*, Vol. 12, No. 3, pp. 252–277 (1994).

[8] 新谷研, 角田達彦, 大石巧, 長尾眞: 形態素の共起頻度と出現位置による新聞関連記事の検索手 法, 信学技報言語理解とコミュニケーション, Vol. 96, pp. 1–8 (1996). NLC96-1.

[9] 池野篤司: 分野依存性を考慮した単語間類似度の獲得と利用, 人工知能学会研究会資料, pp. 143–150 (1994). SIG-J-9401-19.

[10] 宇津呂武仁: 類似度計算テンプレートを用いた検索質問生成による最近隣検索法の最適化, 人工知能学会研究会資料, pp. 1–8 (1995). SIG-KBS-9502-1.

[11] 芥子育雄, 乾隆夫, 石鞍謙一郎: 大規模文書データベースからの連想検索, 信学技報人工知能と知識処理, Vol. 92, pp. 73–80 (1993). AI 92-99.

[12] 佐々木寛, 羽生田博美, 木下哲男: 単語表記情報に基づく情報検索向き単語関係の抽出, 信学技報 人工知能と知識処理, pp. 9–16 (1995). AI 95-7.

[13] 城風敏彦, 羽生田博美, 木下哲男: 統計的シソーラスを用いた分散型ネットワークニュース検 索システム, 信学技報人工知能と知識処理, pp. 15–22 (1995). AI 95-24.

[14] 徳永健伸, 岩山真: 重み付き IDF を用いた文書の自動分類について, 情報処理学会研究報告, Vol. 94, No. 28, pp. 33–40 (1994). 94-NL-100.

[15] 松本裕治, 黒橋禎夫, 宇津呂武仁, 妙木裕, 長尾真: 日本語形態素解析システム JUMAN 使用説明書 version 2.0 (1994).