社団法人 情報処理学会 研究報告
IPSJ SIG Technical Report

2006－ＦＩ－ 84 (14)
2006－ＮＬ－175 (14)
2006／9／13

類語関係抽出タスクにおけるコーパス規模拡大の影響

相澤 彰子
aizawa@nii.ac.jp

国立情報学研究所／総合研究大学院大学

本稿では、タグなしテキストから類語関係を自動抽出するタスクにおいて、コーパス規模の拡大が類似度計算に与える影響を調べる。近年ではWebに代表される大規模なテキスト集合が利用可能となり、単純な手法でもコーパス規模が十分に大きければ、LSA等の従来手法と比較し得る高い性能が得られるとの報告もある。そこで本稿では、コーパスの量が質を補うのかという問題について、実際のデータに基づき調べた結果を考察とともに報告する。特に、コーパスが大規模になると、類似度の値に対する語頻度の影響が無視できない場合があることを示し、これを回避するための単純なフィルタリング法とその効果について述べる。

On the Effect of Corpus Size in Words Similarity Calculation

Akiko Aizawa

National Institute of Informatics / Graduate School of Advanced Studies

This paper focuses on the effect of corpus size in word similarity calculation. Recently, large-scale text corpora became available for automatic synonyms extraction. And it has been reported that the performance of simple methods adapted to large-scale corpora is sometimes comparable to the one of more elaborative methods such as LDA adapted to traditional linguistic resources. In this paper, we report our experimental results as to how the quantity of the corpus complements the quality of similarity calculation. Our results show that the similarity calculation is sometimes influenced by the absolute word frequencies and that there exists a simple filtering method that can correct the bias.

1 はじめに

自然言語テキストから語の関係を自動抽出する方法として、(1) 定型表現に注目する方法と (2) 共起語に注目する方法の２つがある。(1) の定型表現に注目する方法では、たとえば「A such as B」や「A などの B」などの表現パターンを用いて、テキスト中から特定の関係にある語のペアを取り出す [1][2][3]。一方、(2) の共起語に注目する方法では、テキストの指定した範囲内で共起する語のベクトル（文脈）で各語を特徴づけ、これらの共起語ベクトルどうしの類似度によって語の類似度を数値化する [4][5]。以下、本稿では (1) を「パターン法」、(2) を「共起語ベクトル法」と呼ぶ。

パターン法では、表現パターンの選び方により、階層関係や広義には属性を含む各種の関係を扱うことができるが、抽出時の処理誤りやパターンの用法の解釈誤りが、そのまま抽出結果に含まれることになる。一方、共起語ベクトル法では、テキスト中に出現する広い範囲の語を対象にした類似度計算が可能であるが、あくまで文脈に注目した処理であるため、異なる関係の区別や細かな意味の識別は必ずしも容易ではない。近年では両者を併用して、前者におけるあいまい性解消や誤り検出のために後者を用いる方法もあり [6][7]、両者の利点をうまく組み合わせるものとして注目される。

ここで、一般に抽出数と抽出精度の間にはトレードオフの関係があり、質のよい結果を求めると得られる関係の数は少なくなり、逆に多くの語を網羅しようとすると関係の質は低下する。この問題を解決するための現実的な手段として、コーパスの規模を拡大することが考えられる。このとき、パターン法ではテキストの分量に応じて得られる関係の数が単純に増加することが予想されるが、共起語ベクトル法では大規模化の効果は必ずしも自明ではない。また、Web に代表される大量のテキストを扱う際のアプローチとして、検索エンジンのヒットカウントを用いる方法 [8][9][10]、検索エンジンが出力する検索語前後のテキストを利用する方法 [11]、コーパスから直接共起情報を抽出する方法 [6][13]、等の異なる方法が存在するが、これら相互の比較については現在のところあまり報告されていない。

以上の背景のもと、本稿では、テキスト規模の類似度計算に対する影響を調べる。具体的には新聞記事を対象として、定型表現と既存のシソーラスより評価用の類語・非類語ペアを抽出し、これを用いて語の出現頻度と類似度の計算値の関係や、共起語ベクトルの構成方法による違いを調べる。なお、コーパスの規模拡大には、同様の分野・ジャンルからのコーパスの分量を増やす場合と、多様なバリエーションをカバーする場合の 2 通りが考えられるが、本稿では前者を想定している。

本稿で得られた知見をまとめると以下のようになる。第一は、コーパスが大規模になると、類似度の値に対する語頻度のバイアスの影響が無視できない場合があることである。論文中ではまた、これを回避するための単純なフィルタリング法とその効果について示す。知見の第二は、共起語ベクトルの作成において、文の構文に基づく細かな言語処理が実際に性能の改善に寄与することを確認したことである。特に、実験で設定した条件のもとでは、コーパスが大規模化すると、むしろ構文解析による性能改善の効果が顕著になることがわかった。

以下、本稿では、まず **2.** テキストからの共起語の抽出および類似度計算の方法について、比較の対象とした手法を簡単にまとめる。次に **3.** で、評価のために用いた類語・非類語ペアの作成方法について述べる。さらに **4.** で評価用ペアを用いた実験結果を報告し、最後に **5.** でまとめる。

2 語の共起情報に基づく類似度の計算

2.1 テキストからの共起情報の抽出法

語 $w \in W$ として、w に対する共起語を $v \in V$、テキスト中で、w と v が共起する頻度を $freq(w,v)$ と表記する。共起情報の抽出とは、与えられたテキスト集合から、$W, V, freq(w,v)$ を定めるものである。実験では、以下の 2 通りの場合を想定して比較を行った。

(1) 係り受け関係を用いる方法（「係り受け」）

分析対象となるテキストに形態素解析 [14] および係り受け解析 [15] を適用する。次に、格助詞「を」「に」「が」「は」「で」に注目して、名詞に対する共起動詞ベクトルを作成する。

(2) 文内共起による方法（「文内共起」）

同一の文内で共起する名詞を抜き出し、これらを類語抽出のための共起ペアとする。ただし対象とする「名詞」の単位を (1) の「係り受け」法と揃えるため、名詞の抽出には、係り受け解析ツールの文節区切り機能を用いる。

表 1 に、「自転車」という語に対して、係り受け関係を用いて新聞記事から抽出した共起表現の頻度分布の例を示す。新聞記事 1 年分、31 年分のそれぞれについて、コーパス中での出現頻度を示している。表の中の順位は、31 年分のコーパスにおける 1,688 語中の頻度ランキングである。

表 1: 「自転車」に対する共起語の分布の例

順位	共起表現	f_1	f_{31}	順位	共起表現	f_1	f_{31}
1	に乗る	59	1245	1606	が持ち出される	0	1
2	で走る	10	189	1607	が似合わない	1	1
3	をこぐ	7	140	1608	が似合う	1	1
4	で逃走する	2	129	1609	が試乗できる	0	1
5	を使う	3	114	1610	が支払わない	1	1
6	で回る	2	96	1611	が思える	0	1
7	で渡る	7	84	...	...		
8	を利用する	2	70	1681	に取り組む	0	1
9	で近づく	5	65	1682	を上回る	0	1
10	で来る	11	61	1683	が進む	0	1
11	で向かう	3	61	1684	に伴う	0	1
12	を置く	3	59	1685	に向ける	0	1
13	を押す	6	53	1686	が続く	1	1
14	で横断する	1	48	1687	を進める	0	1
15	をはねる	5	46	1688	による	0	1

2.2 共起頻度行列の構成法

表 1 に例示されるように、共起語の分布は、いわゆる Zipf 型の分布であり、たとえば上位 20 語まで

の語がのべ頻度全体の約40%を占める一方で、共起語の半数以上は1度しか出現しないなど、高頻度・低頻度の語に情報が偏っている。特に注意が必要なのは、頻度が低い共起語の中に「に向ける」「が続く」などの一般的な表現が多く見られることである。これらの表現は、広範囲の語と共起するため、テキスト全体の量が大きくなると、意味的なつながりが薄いものも含めて多数の語の間で共通に観察されるようになると予想される。

本稿では、このような語は類似度計算におけるノイズになる可能性が高いという仮定のもと、共起語の選別を行うこととし、共起頻度行列 $C = (c_{i,j})$ $(i = 1, \cdots, |W|, j = 1, \cdots, |V|)$ を以下により作成した。

(1) 何もしない場合（「選択なし」）

共起頻度をそのまま共起頻度行列の要素に対応させる。すなわち、$c_{i,j} = freq(w_i, v_j)$。

(2) 高頻度語を削除する場合 （「高頻度語削除」）

共起語 $v_j \in V$ を出現頻度、すなわち $\sum_i freq(w_i, v_j)$ の順にソートし、あらかじめ定められた割合 r_1 の高頻度語を削除する。削除の対象となる高頻度語 $v_j \in V_{high}$ に対して、w_i によらず $c_{i,j} = 0$ とする。

(3) 共起ペアの単位で取捨選択をする場合（「共起語フィルタリング」）

$w_i \in W$ の総頻度を $freq(w_i)$、$v_j \in V$ の総頻度を $freq(v_j)$、頻度の総数を F として、相互情報量 $log\frac{freq(w_i,v_j)F}{freq(w_i)freq(v_j)}$ の値が低い順に、定められた割合 r_2 の共起ペア (w_i, v_j) を定め、これらに対して、$c_{i,j} = 0$ とする。

(4) 語のサンプル数に上限を設ける場合（「サンプル数制限」）

各語に関するサンプル数の上限値 N を定め、コーパス中での出現順に $\sum_{w_i} freq(w_i, v_j) = N$ となった時点で w_i に関する共起語の収集を打ち切る。

(2)(3)(4) はいずれも、高頻度語との共起によるノイズを取り除く効果が期待される。(4) は、検索エンジンの結果を利用する場合に現実問題として検索語あたり高々N 件の情報しか収集できないことを意識したものである。

2.3 共起語ベクトルと類似度尺度

上記より得られた共起頻度行列を用いて、以下の類似度尺度の値を計算した。

(1) Jaccard 係数

$w_1, w_2 \in W$ に対して $c_{i,j} > 0$ となる共起語の集合をそれぞれ V_1 $(= \{v_j | c_{1,j} > 0\})$、$V_2$ $(= \{v_j | c_{2,j} > 0\})$ として、以下で定義される Jaccard 係数の値を w_1 と w_2 の類似度とする。

$$Jaccard(w_1, w_2) = \frac{|V_1 \cap V_2|}{|V_1 \cup V_2|} \quad (1)$$

(2) tf-idf コサイン尺度

共起頻度行列より tf-idf で重み付けした共起語ベクトル $\vec{w_1}, \vec{w_2}$ を求め、コサイン尺度の計算値を w_1, w_2 の類似度とする。

$$cos(\vec{w_1}, \vec{w_2}) = \frac{\vec{w_1} \cdot \vec{w_2}}{|\vec{w_1}||\vec{w_2}|} \quad (2)$$

(3) 出現頻度による相互情報量

文内共起を用いる場合に、w_1 と w_2 が出現した文の数をそれぞれ $s(w_1)$, $s(w_2)$、同一文内で共起した回数を $s(w_1, w_2)$ 、文の総数 S として、以下で類似度を求める。

$$PMI(w_1, w_2) = log\frac{s(w_1, w_2)S}{s(w_1)s(w_2)} \quad (3)$$

ここで、(1) の Jaccard 係数は、共起語の異なり数に基づく尺度の典型例、(2) の tfidf 重みによるコサイン尺度は、共起語の分布に基づく尺度の典型例として選んだものである。

3 評価・比較実験

3.1 実験に用いたコーパス

実験では、日本経済新聞 CD-ROM 版（1975 年 ～ 2005 年）[16] から、(A)1990 年 1 年分の記事（nikkei-1）、(B)31 年分の記事すべて（nikkei-31）、の 2 通りのテキスト集合を抽出した。nikkei_1 はテキスト量にして 173M バイト、名詞数は 423,500 個、nikkei_31 はテキスト量にして 3,776M バイト、名詞数は 3,738,767 個であった。

次に、**2.1** により、「係り受け」および「文内共起」に基づく 2 通りの方法で共起情報を抽出した。各名詞の共起語は、前者では動詞（＋格）、後者では名詞と

なる。さらに、**2.2**の4通りの方法で、共起頻度行列を生成した。ここで、「高頻度語削除」や「共起語フィルタリング」では、[17]の結果に基づき$r_1 = r_2 = 0.2$とし、「サンプル数制限」では、サンプル数の上限値$N = 1{,}000$とした。

3.2 タスクA：類語・非類語の判定による比較

指定されたコーパスにおける「類語」関係は、必ずしも人手により構築された体系的なシソーラスと対応がとれるわけではない。本稿では、「類語は必ずしも汎用的な体系によって定まるものではなく、類語であるかどうかの判定はコーパスが代表する語彙空間に依存して決まる」と考え、対象コーパスから特定の表現パターンにより抽出した名詞ペアを用いて評価用セットを定めた。

具体的には、まず、「A-や-B」という定型表現に注目して、{A, B}を類語関係の候補として抽出した。次に、定型的な表現を除外するため、順番を逆にした「B-や-A」の出現頻度が極端に少ないペアを削除した。さらに、コーパス中での出現頻度がA、Bともに閾値kより大きいペアを評価用の類語セットとして選択した。実験では$k = 5$としてnikkei-31に適用した。さらに、得られた類語ペアの中から、A、Bともに分類語彙表[18]の見出し語であり、かつ分類語彙表の第4階層のレベルで同一のカテゴリに登録されているようなペアを選び候補とした。最後に、各Aについて複数の候補が存在する場合には、コーパス中での出現頻度がもっとも多い候補ペアだけを選択し、最終的に685ペアの類語候補を得た。

さらに、関連のない語に対する類似度の計算値と比較するために、分類語彙表の上でAと第二階層以下が異なるカテゴリであるもののうち、コーパス中での出現頻度がBにもっとも近い語Cを求め、{A,C}を「非類語」ペアとした。出現頻度が近い語を選ぶのは、類似度計算の条件をなるべく揃えるためである。評価用に選んだペアの例を表2に示す。たとえば、A.「自転車」とB.「バイク」はともに分類語彙表の上で「体-生産物-機械-乗り物（陸上）」に分類されており、非類語のC.「無条件」は第2階層で「生産物」には属していない、等である。

上記の構築法は人手チェックを含んでいないため、適切でない関係が含まれる可能性は排除できないが、コーパスが大規模になった場合にも、その分野特性を反映する評価用データが容易に得られるという利点がある。

表2: タスクAで評価に用いた類語・非類語ペアの例

語 (A)	類語 (B)	非類語 (C)
琴	尺八	高級官僚
心臓病	脳卒中	委任状
南部	西部	取り込み
アイスクリーム	ゼリー	主将
内容	仕組み	ベストセラー
金融機関	証券会社	目
紅茶	緑茶	変わり
二酸化炭素	メタン	兼業農家
テレビ	携帯電話	講座
自転車	バイク	無条件

3.3 タスクB：定型表現の用法判定による比較

表2からも明らかな通り、タスクAで選んだ類語・非類語のペアには意味の上で大きな隔たりがあり、これらの区別は比較的容易であることが予想される。そこで、より細かな語義の区別について調べるために、「パターン法」において定型表現から語の上位－下位関係を抽出するタスクを想定し、以下の手順で評価用のセットの作成を行った。

具体的には、まず、コーパスから「A-や-B-などの-C」という定型表現を抽出した。次に、A,B,Cともコーパス中での出現頻度が$N = 10$以上で分類語彙表の見出し語となっているものを選び、、2名の判定者で「CがA-や-Bの上位概念になっているもの」「そうでないもの」を分類した。ただし、「カーテンやカーペットなどのスポーツ用品」のように整合性がとれないものや「インドや中国などの市場」のように判断がむずかしいと思われる用法については、評価セットからは除外した。このようにして選んだ定型表現の例を表3に示す。最後に、{A, C}を評価用の語ペアとして、上位－下位関係にあるペア308件、そうではないペア463件を選択した。

表3: タスクBで評価に用いた定型表現の例

上位・下位関係を示すもの	上位・下位関係を示さないもの
紳士服や婦人服などの衣料品	台湾や日本などのアーチスト
コーヒーや紅茶などの飲料	フランスや英国などのクラブ
クラシックやポップスなどの音楽	マーケティングや開発などのコスト
高卒や大卒などの学歴	結婚式や葬式などのしきたり
スーパーや商店などの業者	クラシックやジャズなどのショー
円や三角形などの形	ゴルフや野球などのスイング
音楽や美術などの芸術	関東や北陸などのスーパー
クレジットやリースなどの債権	韓国や中国などのスタッフ
タオルや歯ブラシなどの雑貨	ドラマや映画などのソフト

4 実験結果

4.1 タスクAのF値による比較

まず、**3.2**で述べたタスクAについて、Jaccard尺度およびtf-idfコサイン尺度による類語・非類語ペアの類似度を計算した。そして、閾値δに対して、類語ペアの数をa、非類語ペアの数をb、類語と判定されたペアの数をc、正しく類語と判定されたペアの数をdとして、$p = d/a$, $r = d/c$, $F = 2pr/(p+r)$によりF値を計算し、δを変化させて最大値を求めた。表4に、各条件によるF値の最大値を示す。

表4: F値による比較（タスクA）

		nikkei-1		nikkei-31	
		Jaccard	コサイン	Jaccard	コサイン
係り受け	選択なし	0.841	0.788	0.932	0.876
	高頻度語削除	0.856	0.853	0.928	0.952
	共起語フィルタリング	0.879	0.865	0.982	0.951
	サンプル数制限	0.843	0.787	0.970	0.874
文内共起	選択なし	0.834	0.758	0.908	0.771
	高頻度語削除	0.875	0.905	0.895	0.933
	共起語フィルタリング	0.920	0.929	0.948	0.951
	サンプル数制限	0.897	0.741	0.902	0.671

表4の結果から、まず、nikkei-1では「文内共起」の方がスコアが高いが、nikkei-31では「係り受け」の方がスコアが高いことがわかる。これは、文内共起の方が共起語数が平均で10倍程度多く、テキストの分量が十分でない場合に、より多くの手がかりが得られるためであると考えられる。表4からは、また、高頻度語の削除や情報量を使った共起語フィルタリングによってスコア値が改善すること、特に後者について改善の度合いが高いことがわかる。語あたりのサンプル数を制限する方法では、Jaccard尺度に関してスコア値の改善がみられるが、コサイン尺度に対してはむしろスコア値が下がる傾向がみられる。スコアの分布については**4.3**でさらに細かく調べる。

なお実験では、同一の単語に対して類語・非類語のスコアを比較した場合の誤り率についてもあわせて調べた。具体的には、「琴」に対して「尺八」と「高級官僚」のいずれが近いと判定されるか、等である。結果の詳細は省略するが、表4のスコアとほぼ対応しており、誤り率はnikkei-1の場合で10 ～ 20 %、nikkei-31の場合で0 ～ 数%であった。もっともスコア値が高かったnikkei-31の「係り受け＋共起語フィルタリング」の条件では、判別誤りはゼロであった。

また、参考までに、google APIにより各語のヒットカウント$n(w_1)$, $n(w_2)$、両者を含むクエリのヒットカウント$n(w_1, w_2)$を求め、文献[8]の最も単純なベースラインにしたがって$log\frac{n(w_1,w_2)}{n(w_1)n(w_2)}$により類似度を計算した場合のF値の最大値は0.785であった。

4.2 タスクBのF値による比較

次に、**3.3**で述べたタスクBについて、同様にして類似度を計算し、F値の最大値を求めて比較を行った。表5に結果をまとめる。全体的な傾向は、表4のタスクAの結果とほぼ一致している。上位下位関係と誤って判定してしまった例としては、「消費や投資などの内需」「憲法や民法などの条文」、上位下位関係を正しく判定できなかった例としては、「カステラや和菓子などの製品」「ニューヨークやパリなどの国際都市」などがあった。

また、文内共起において、式(3)を用いて求めた類似度によるF値の最大値は、nikkei-1で0.662、nikkei-31で0.688であった。google APIを使った場合の性能は0.654であった。

表5: F値による比較（タスクB）

		nikkei-1		nikkei-31	
		Jaccard	コサイン	Jaccard	コサイン
係り受け	選択なし	0.575	0.589	0.585	0.649
	高頻度語削除	0.589	0.613	0.595	0.679
	共起語フィルタリング	0.656	0.672	0.752	0.751
	サンプル数制限	0.574	0.592	0.714	0.669
文内共起	選択なし	0.578	0.572	0.580	0.571
	高頻度語削除	0.583	0.594	0.581	0.634
	共起語フィルタリング	0.674	0.679	0.681	0.682
	サンプル数制限	0.611	0.574	0.618	0.572

4.3 タスクAにおける類似度の分布による比較

最後に、タスクAでnikkei-31を用いる場合の類語・非類語ペアに対する類似度の分布を図1および図2に示す。グラフの各プロットは異なる語ペア（w_1, w_2）に対応しており、縦軸は類似度の値、横軸は出現頻度の積（$freq(w_1) \times freq(w_2)$、対数目盛り）である。グラフ中では、類語を網掛け、非類語を黒として区別している。

まず、図1-(a)(e)より、Jaccard係数では類似度の値は語の出現頻度の影響を受けており、類語・非類語を問わず出現頻度が高いほど値が大きくなる傾向があることが確認できる。これは、**2.2**で予想したように、汎用的な動詞との共起あるいは解析誤りが

ノイズとして影響を与えるためだと考えられる。また、図2-(a)(e)から、tf-idfコサイン尺度で文内共起を利用する場合には、Jaccard係数の場合と同様に、頻度によるバイアスが観察されることがわかる。ただし、係り受け関係を利用する場合には、語頻度による影響はあまりみられなかった。

タスクA,Bにおいて、相互情報量による共起語フィルタリングのスコアが高かった理由は、図1,2-(c)(g)から説明できる。すなわち、類似度の値を用いて類語・非類語の判別を行う場合には、頻度によるバイアスが性能低下の要因となるが、共起語フィルタリングは、このような状況でバイアスを取り除く効果がある。図1,2-(b)(f)に見られるように、単純に高頻度の共起語を取り除くだけでは、頻度の影響を除くことはできないため、スコア値の大幅な改善は期待できない。また、図1,2-(d)(h)に見られるように、各語についてサンプル数1000の上限を設定する場合、語頻度のばらつき度合いは大幅に減少するが、やはり分離の程度には大きな変化がない。

5 まとめ

本稿では、類大規模テキスト活用の1つの鍵は、パターン法において必要となる様々なあいまい性解消タスクにおける共起語ベクトル法の活用であると考え、語抽出タスクにおけるコーパス規模拡大の効果を実際の新聞記事コーパスを使って調べた。上記の結果を踏まえ現在、相互情報量による共起語フィルタリング適用後のデータに同時クラスタリングを組み合わせることで、大規模な類語・例文辞書の自動構築を試みており[17]、Web文書にも適用する予定である。

参考文献

[1] Marti A. Hearst: "Automatic Acquisition of Hyponyms from Large Text Corpora," in Proc. of the 14th International Conference on Computational Linguistics, 539-545 (1992).

[2] 安藤まや、関根聡、石崎俊:「定型表現を利用した新聞記事からの下位概念単語の自動抽出」情報処理学会研究報告、FI-72/NL-157, 77-82 (2003).

[3] Emmanuel Morin and Christian Jacquemin: "Automatic Acquisition and Expansion of Hypernym Links," Computer and the Humanities, 38(4), 343-362 (2004).

[4] Dekang Lin: "Automatic Retrieval and Clustering of Similar Words," in Proc. of the 36th Annual Meeting of the Association for Computational Linguistics and the 17th International Conference on Computational Linguistics, 768-774 (1998).

[5] Lillian Lee: "Measures of Distributional Similarity," in Proc. of the 37th Annual Meeting of the Association for Computational Linguistics, pp.25-32 (1999).

[6] 新里圭司、鳥澤健太郎:「HTML文書からの単語間の上位下位関係の自動獲得」自然言語処理, vol.12, No,1, 125-151 (2005).

[7] Rion Snow, Daniel Jurafsky, Andrew Y. Ng: "Semantic Taxonomy Induction from Heterogenous Evidence," in Proc. of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the the Association for Computational Linguistics, 801-808 (2006).

[8] Peter D. Turney: "Mining the Web for Synonyms: PMI-IR versus LSA on TOEFL," Proceedings of the Twelfth European Conference on Machine Learning (ECML-2001), 491-502 (2001).

[9] M. Baroni and S. Bisi: "Using cooccurrence statistics and the web to discover synonyms in a technical language," in proceedings of the 4th International Conference on Language Resources and Evaluation (LREC 2004), 1725-1728 (2004).

[10] Vinci Liu and James Curran. Words and Word Usage: "Newspaper Text versus the Web," In Proceedings of the Australasian Language Technology Workshop, Sydney, Australia (2005).

[11] Mehran Sahami and Timothy D. Heilman: "A Web-based Kernel Function for Measuring the Similarity of Short Text Snippets," in Proc. of World Wide Web Conference (2006).

[12] M. Baroni and M. Ueyama: "Building general- and special-purpose corpora by Web crawling," in Proceedings of the 13th NIJL International Symposium, Language Corpora: Their Compilation and Application, 31-40 (2006).

[13] 河原大輔, 黒橋禎夫:「Webから獲得した大規模格フレームに基づく構文・格解析の統合的確率モデル」言語処理学会 第12回年次大会, 1111-1114 (2006).

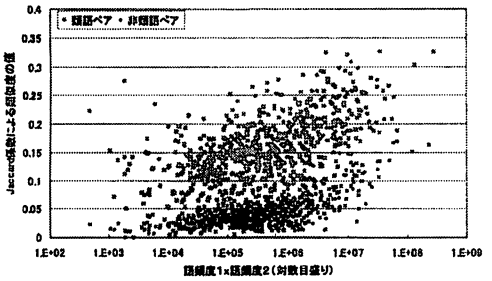
[14] 松本裕治、北内啓, 山下達雄, 平野善隆, 松田寛, 高岡一馬, 浅原正幸:「日本語形態素解析システム『茶筌』」version 2.2.1 使用説明書 (2000).

[15] 工藤拓, 松本裕治:「チャンキングの段階適用による日本語係り受け解析」情報処理学会論文誌, vol.43, no.6, 63-69 (2002).

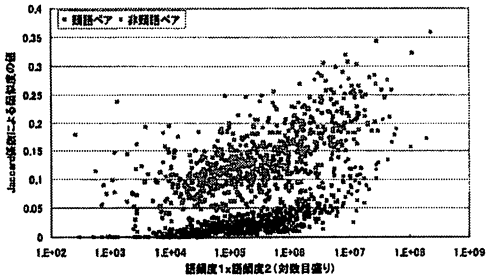
[16] 日本経済新聞社「日経全文記事データベース 1975〜2005年版日本経済新聞」.

[17] 相澤、中渡瀬「係り受け関係を利用した類語・例文辞書構築法と大規模コーパスへの適用」第20回人工知能学会全国大会, 2E1-5.

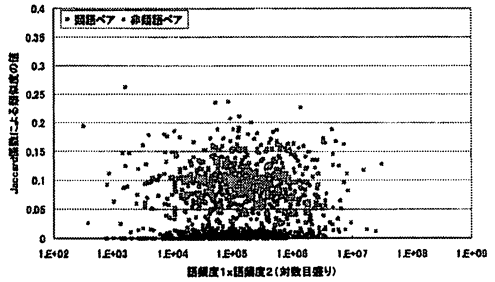
[18] 国立国語研究所編:『分類語彙表 増補改訂版』、大日本図書 (2004).



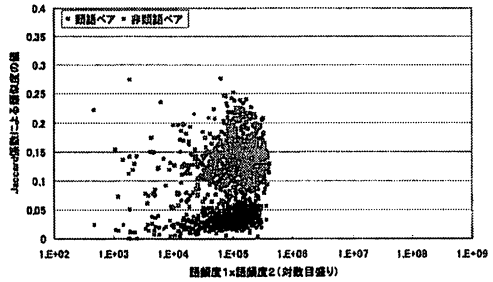
(a) 係り受け-選択なし



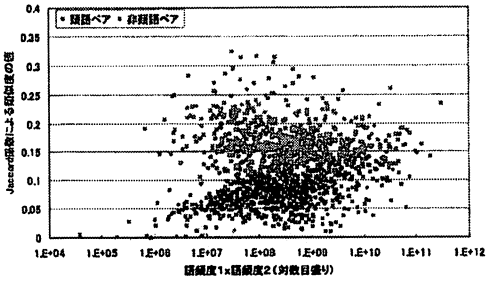
(b) 係り受け-高頻度語削除



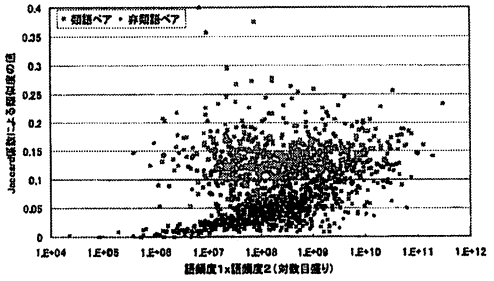
(c) 係り受け-共起語フィルタリング



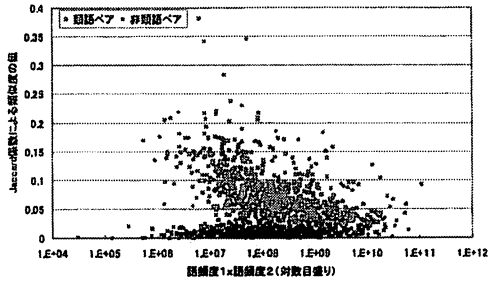
(d) 係り受け-サンプル数制限



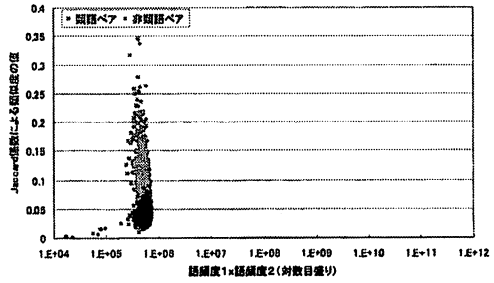
(e) 文内共起-選択なし



(f) 文内共起-高頻度語削除

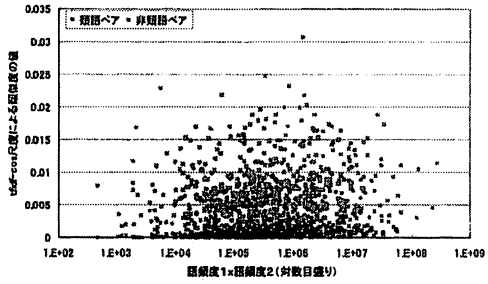


(g) 文内共起-共起語フィルタリング

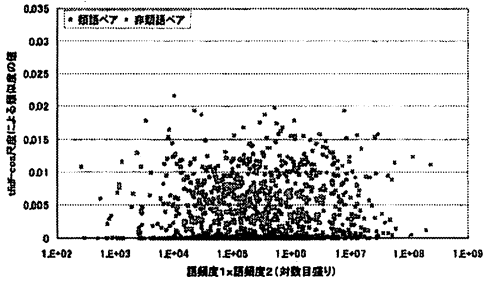


(h) 文内共起-サンプル数制限

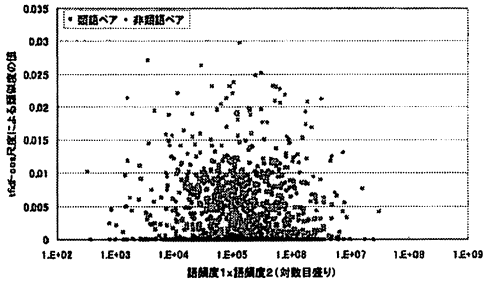
図 1: タスク A による類語・非類語ペアの判定結果（Jaccard 尺度）



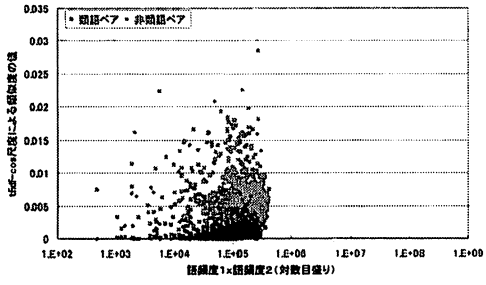
(a) 係り受け-選択なし



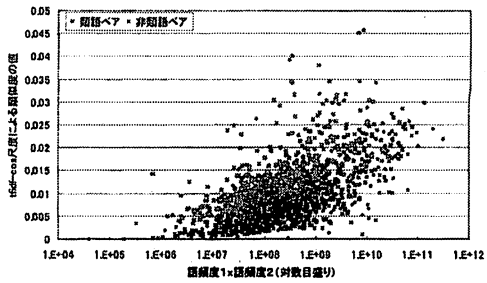
(b) 係り受け-高頻度語削除



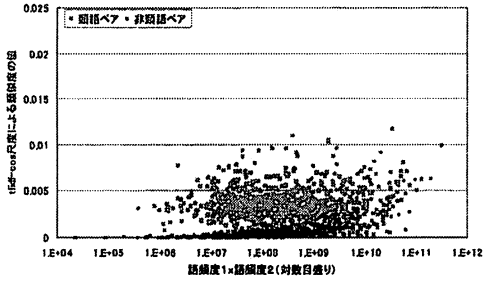
(c) 係り受け-共起語フィルタリング



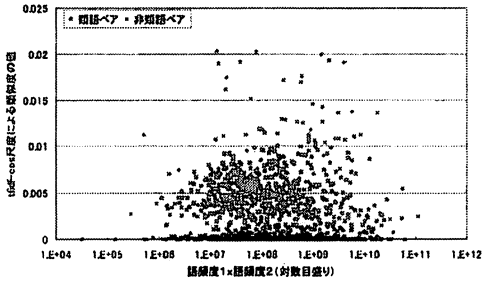
(d) 係り受け-サンプル数制限



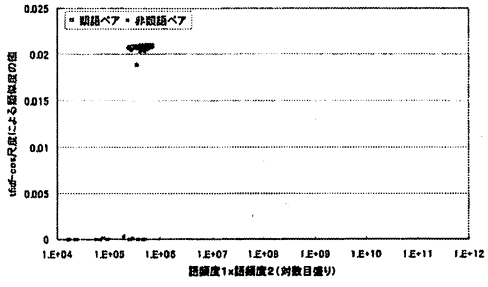
(e) 文内共起-選択なし



(f) 文内共起-高頻度語削除



(g) 文内共起-共起語フィルタリング



(h) 文内共起-サンプル数制限

図 2: タスク A による類語・非類語ペアの判定結果（コサイン尺度）