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

2007－ICS－148 (9)
2007／7／4

Webコーパスを用いた語の類似度計算に関する考察

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抽出するタスクを例にとり、大規模コーパスとしてのウェブの有効性について考察する。まず、類似度の値が頻度に依存するバイアスを受けるという前提のもと、ノイズ低減のためフィルタリング法、サンプリング法の２つの方法を提案する。また、評価のための類語抽出タスクを設計し、NTCIR5のウェブ・コレクションを用いてバイアスの影響および提案手法による性能改善を確認する。

On Calculating Word Similarity using Web as Corpus

Akiko Aizawa

National Institute of Informatics / Graduate School of Advanced Studies

This paper concerns the availability of large-scale Web collection in the task of synonymous relationship identification. Assuming that the similarity calculation is affected by the word frequencies, we propose two methods for reducing the bias. The effectiveness of the proposed methods is shown using NTCIR5 web collection.

1 はじめに

本稿では、コーパスの大規模化の影響について考察し、実際にウェブ文書から抽出したデータを用いて評価を行う。我々はすでに文献[1]において、31年分の新聞記事を用いてコーパス大規模化の影響を調べ、類似度の計算値には語の出現頻度によるバイアスがかかる場合があることを指摘した。そこで本稿では、近年言語コーパスとしても注目を集めるウェブ文書を対象として、大規模化の影響と注意点を詳しく調べる。

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[2][3][4]。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[5][6]。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[7][8]、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 [9][10][11] やコーパスから直接共起情報を抽出する方法 [7][12] 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、ウェブの類語抽出タスクにおける有用性を調べる。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バイアス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／サンプリング法を提案して実験により効果を調べる。以下、まず **2.** でテキストからの共起語抽出および類似度計算の方法を述べ、ノイズを削減するための２つの方法を説明する。次に **3.** で、コーパスを利用した評価用データの構築法を述べる。さらに **4.** で実際に大規模ウェブ文書コレクション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 共起語と類似度尺度

文書、文章、句など、定められたテキスト領域内で同時に観察される語を共起語と呼ぶ。本稿では語 $w_1, w_2 \in W$ に対する典型的な類似度尺度として以下を選んで比較の対象とする。

(1) Jaccard 係数

$w_1, w_2 \in W$ に対する共起語の集合をそれぞれ V_1 、V_2 として、以下で定義される。

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

(2) Simpson 係数

$w_1, w_2 \in W$ に対する共起語の集合をそれぞれ V_1、V_2 として、以下で定義される。

$$Simpson(w_1, w_2) = \frac{|V_1 \cap V_2|}{min(|V_1|, |V_2|)} \qquad (2)$$

(3) 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 $\vec{w_1}, \vec{w_2}$ とた場合のコサイン尺度。ただし df は共起語の異なり数とする。

$$cos(\vec{w_1}, \vec{w_2}) = \frac{\vec{w_1} \cdot \vec{w_2}}{|\vec{w_1}||\vec{w_2}|} \qquad (3)$$

(4) 出現頻度による相互情報量

w_1 と w_2 が出現した文の数を $s(w_1)$, $s(w_2)$、同一文内で共起した回数を $s(w_1, w_2)$ 、文の総数 S として計算される（特殊）相互情報量。

$$PMI(w_1, w_2) = log\frac{s(w_1, w_2)S}{s(w_1)s(w_2)} \qquad (4)$$

語の類似度尺度は、共起語集合の重なりに基づく方法と、共起語の分布の類似度に基づく方法の２つに大別される。上記の類似度尺度のうち、(1)(2) は前者の、(3) は後者の典型例として選んだ。(4) は考え方が異なるが、検索エンジンを用いて語の関連度を調べる場合の典型的な計算法として選んだ。

2.2 コーパス規模拡大の影響

コーパスの規模が拡大すると、個々の共起ペアの出現回数はコーパスの大きさにほぼ比例する形で増加する。同時に、それまで観察されていなかった新たな共起ペアが出現するため、共起語の分布の幅が広がる。

図１に「野菜」という語に対して規模の異なる Web 文書集合から抽出した共起語（係受け関係にある格＋動詞）の分布の例を示す。図 1-(a)(c) は、NTCIR-Web5 コレクション [13] の {ne, co, ac, or, go} の 5 ドメインから得られた 95,517 個の共起ペアよる分布で、図 1-(a)(c) は、{or} ドメインから得られた 11,621 個の共起ペアによる分布である。(a)(b) は共起頻度が 5 ドメインで 1~20 位の高頻度語、(c)(d) は 5,810 ~ 5,830 位の低頻度語を示している。ドメインによる多少の偏りはあるものの、共起頻度が多い上位語については、２つのコーパスの間で分布の形に大きな違いはなく（図 1 の (a)(b)）、一方で、共起頻度が 1 となる領域では語がまばらにサンプリングされるため、違いが大きく出ることがわかる（図 1 の (c) (d)）。

では、上記の場合に類似度計算に対する影響は具体的にどのようなものになるだろうか? **2.1** の tf-idf コサイン尺度のように頻度情報を重視する尺度では、共起の回数が多い共起語の影響が支配的となるため、コーパス内での語の総出現数が類似度尺度の計算値に与える影響は少ないと考えられる。一方、Jaccard 係数や Simpson 係数のように共起語集合の重なりに注目する尺度では、語の総出現数が多くなればなるほど、相対的に共起の回数が多い共起語の影響が薄れ共起の回数が少ない共起語の影響が支配的になる。

2.3 ノイズの低減

ここで注意が必要なのは、頻度が低い共起語の中には、「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「ノイズ」となる。

ここで Jaccard 係数や Simpson 係数の場合には、総出現数が多い語ほど低頻度の共起語の影響を強く受けるため、結果としてノイズの影響を強く受けることになる。すなわち、類似度の計算値に語の総出現数に依存するバイアスがかかる。さらに **2.2** では例示のため、規模の異なるコーパスを用いて同一の語に関する共起語分布の違いを比較したが、実際に問題になるのは、同一のコーパス内で類似度を計算する際の語の総出現数のばらつきであり、ばらつきはコーパスが大規模になると広いレンジにわたる。たとえば本稿の実験で用いた Web コーパスでは、各語の出現頻度は 1 から 5,000,000 となる。このような場合に、類似度計算のバイアスが問題になることが予想される。

このようなノイズに対応するための確率的なアプローチとして、たとえば共起語の分布 $P^*(v_j|w_i)$ を次式の混合分布で表して、各パラメタの値を最大エントロピー法などで推定する等が考えられる。ただし $P_0(v_1)$ を語全体に共通する共起語の分布、$P(v_j|w_i)$ を語 w_i に特徴的な共通語の分布、α を混合比とする。

$$P^*(v_j|w_i) = \alpha P(v_j|w_i) + (1-\alpha)P_0(v_j) \qquad (5)$$

しかしながら、大規模コーパスへの適用では、推定するべきパラメタ数が大きくなるため計算コストや収束の問題が予想される。そこで本稿では、大規模なコーパスにも対応できる単純なノイズ低減法として以下の 2 つの方法を提案する。

A. フィルタリング法

$w_i \in W$ 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、(特殊) 相互情報量 PMI の値がの値が閾値 β より小さい場合に、ノイズとみなして共起語から取り除く。すなわち、以下をフィルタリングの条件とする。

$$PMI = log\frac{freq(w_i, v_j)F}{freq(w_i)freq(v_j)} < \beta \qquad (6)$$

B. サンプリング法

各語 w_i について出現頻度の上限値 N を定め、コーパス中での出現頻度が N を超える場合に以下のいずれかの方針で共起ペアを選択する。

(1)乱数を用いてランダムに N 個の共起ペアを選ぶ。
(2)共起頻度が高い順に N 語の共起語を選ぶ。
(3)相互情報量の値にしたがって N 語の共起語を選ぶ。

(1) は特に、検索エンジンの結果を利用する場合に、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) や (3) では N では共起語の数となるが、(1) では N は共起頻度の合計になり、考慮される共起語の数は (2) や (3) より少ない。

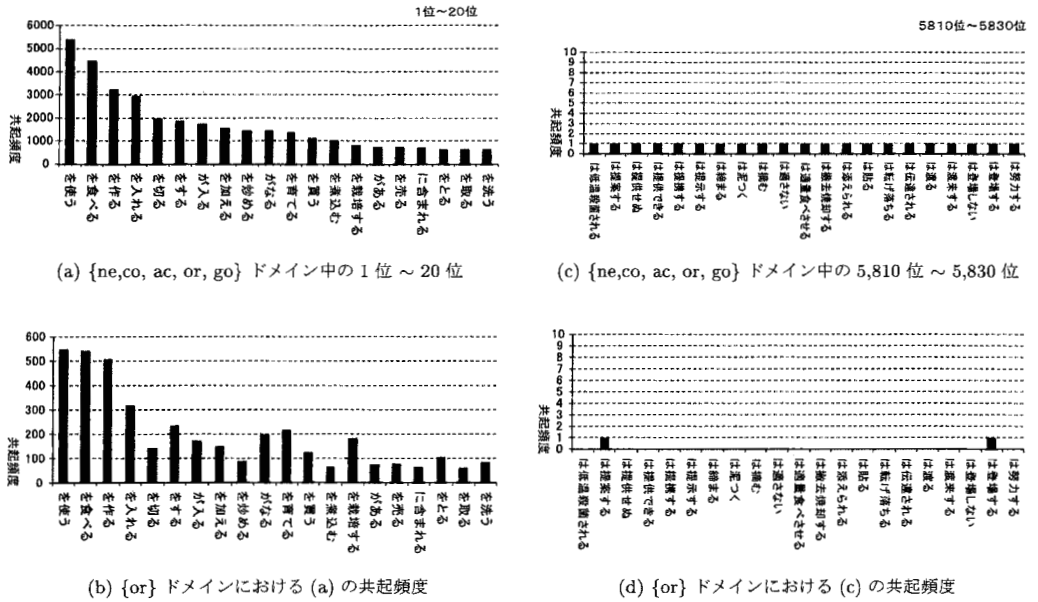
3 評価用データの作成

3.1 方針

類語関係を定義した既存の言語資源として分類語彙表 [14] 等があるが、これをそのまま評価用に用いるのは以下の点で問題がある。まず、汎用的な語彙がすべてコーパスに出現するとは考えにくい。さらに、特定のコーパスにおける類語関係は、人手により構築された体系的なシソーラスと必ずしもと対応がとれるわけではない。[1] すなわち、類語関係が存在するか否かは必ずしも絶対的ではなく、コーパスが代表する語彙空間に依存して決まると考えられる。

上記を背景に本稿の実験では、文献 [1] において筆者らが新聞記事に適用した評価方法を踏襲し、以下の 3 点に留意して評価用データの作成を行った。

[1] たとえば新聞記事コーパスの中では、「株」と「債券」は類似した文脈で出現するが、分類語彙表では異なる分類に属する。



(a) {ne,co, ac, or, go} ドメイン中の 1 位 ～ 20 位

(c) {ne,co, ac, or, go} ドメイン中の 5,810 位 ～ 5,830 位

(b) {or} ドメインにおける (a) の共起頻度

(d) {or} ドメインにおける (c) の共起頻度

図 1: 語「野菜」に対する共起語分布のコーパスによる違いの例

(i) 評価用データが対象コーパスの特徴を反映していること
(ii) 人手による判定の負担が少ないこと
(iii) 異なるレベルの類語判定タスクで評価を行うこと

具体的には、(i) を満足するため、対象コーパスから定型表現のパターン（「A-や-B-などの-C」）を使って類語の候補を抽出し、(ii) を満足するため、分類語彙表やコーパス中での出現頻度を使って評価用の候補を選別した。さらに、(iii) については、定型表現のあいまい性解消を第 2 のタスクとして設定した。以下、評価用データ作成の詳細を述べる。

3.2 タスク I：類語・非類語の判定による比較

タスク I では、まず、コーパス中の「A-や-B-などの-C」という定型表現に注目して、{A, B} を類語の候補として抽出した。次に、コーパス中での出現頻度が A、B ともに閾値 k 以上であるペアを選択し、得られたペアの中から、A、B ともに分類語彙表の見出し語であり、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ものを選び、評価用の「類語」ペアとした。

次に、関連の低い語に対する類似度の計算値と比較するために、分類語彙表上で A と第二階層でカテゴリが異なる語のうち、コーパス中での出現頻度が B にもっとも近い語 D を求め、{A,D} 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 1 に示す。たとえば、類語ペアである「株式」(=A) と「不動産」(=B) はともに分類語彙表の上で「体 / 活動 / 経済 / 資本・金銭」に分類されている、A と非類語ペアとなる「旗」(=D) は第 2 階層で「生産物」に属しており「活動」ではない、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ゼロではないが、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 1: タスク I で評価に用いた類語・非類語ペアの例

語 (A)	類語 (B)	非類語 (D)
航空券	特急券	かわ
パソコン	ワークステーション	今晩
アメリカ	韓国	スタイル
株式	不動産	旗
テレビ	ビデオ	差
地名	人名	狂牛病

3.3 タスクII：定型表現の用法判定による比較

表1からも明らかな通り、タスクI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、タスクAで用いた「A-や-B-などの-C」という定型表現について、コーパス中での出現頻度がA、B、Cともに閾値k以上で、かつA、B、Cともに分類語彙表の見出し語であるようなものを選択し、「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」を人手により判定した。

判断がむずかしい場合には、まずウェブ上の文書を参照して一般的な用法を調べ、それでも判定できない場合には評価セットから除外した。[2]また、今回は予備的に判定者1名（筆者）で判定を行ったため、「ニュースや天気予報などの生活」のように前処理段階での誤りの影響が予想されるものや「高校生や大学生などのユーザ」のように人間の判断がゆれることが予想される用法については、あらかじめ評価セットからは除外した。得られた定型表現の例を表2に示す。

表2: タスクIIで評価に用いた定型表現の例

◎上位・下位関係を示すもの
エイズや肝炎などの病気
スタッカートやテヌートなどの表現
アイスクリームやシャーベットなどの氷菓
雑誌や書籍などの著作物
国債や地方債などの債権
タオルや歯ブラシなどの消耗品
◎上位・下位関係を示さないもの
エイズやペストなどの病原菌
新聞や放送などの表現
南極や北極などの氷
雑誌や書籍などのガイドブック
国債や地方債などの利子
プリンターや複写機などの消耗品

4 実験

4.1 実験の条件

実験の対象としたのは、NTCIR5-Webテストコレクション[13]に含まれるうち、{ ne, co, ac, or, go }の5つのドメインのいずれかに属する約430Gバイトの Web 文書である。前処理として、まずタグを除いたテキストに形態素解析[15]および係受け解析[16]を適用し、次に格助詞「を」「に」「が」「は」「で」に注目して、{名詞,「格＋動詞」}の共起ペアを抽出した。結果として、名詞98,638,646個、格＋動詞2,625,231個を含むのべ数で675,344,497、異なり数で98,638,646の共起ペアが得られた。設定した実験の条件は以下の通りである。

- 格情報の有無

  比較のため、{名詞,「格＋動詞」}の共起ペアに加えて格を考慮しない{名詞、動詞}ペアもあわせて抽出した。

- 類似度の計算法

  2.1で述べたJaccard係数、Simpson係数、*tf-idf*コサイン尺度の3つを適用して類似度を求めた。また google API 経由で検索エンジンのヒット数を用いる式(4)の方法も参考のため行った。

- ノイズ低減法の効果

  得られた共起ペアをすべて類似度計算に用いる場合、**2.3**で述べたフィルタリング法およびサンプリング法の適用により一部の共起ペアを取り除く場合について比較を行った。

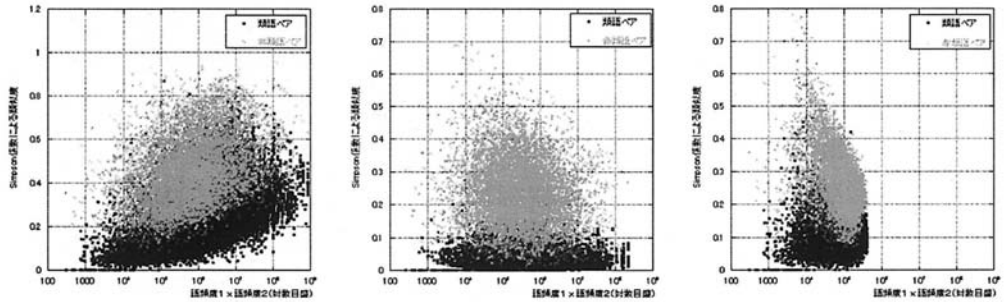
また、評価用セットの構築では、コーパス中での出現頻度が$k = 10$以上である名詞を対象として、タスクIについては25,740個ずつの類語・非類語ペアを、タスクIIについては正解1,265個、不正解754個を含む合計2,019個のフレーズを人手判定により選んだ。

4.2 頻度の類似度計算に対する影響

まず、フィルタリングやサンプリングの効果を調べるため、タスクIの類語・非類語ペアに対する類似度の値の分布を調べた結果を図2-(a)に示す。グラフの各プロットは個々の共起ペアに対応しており、縦軸はSimpson係数による類似度の値、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、類語ペアを黒、非類語ペアを灰色として区別している。[3]

[2]たとえば、「心臓病や腎臓病などの合併症」は、表現の上ではあいまい性があるが、高血圧の合併症としてあげる文書が多いので正解とした。

[3]横軸を語頻度の積としている理由は、語w_1,w_2について語vが共起する確率をそれぞれ$P(v|w_1)$, $P(v|w_2)$とするとき、f_1, f_2回のサンプリングでvが実際に共通の動詞として観察される確率が近似的に$\left(1-(1-P(v|w_1))^{f_1}\right) \times \left(1-(1-P(v|w_2))^{f_2}\right) \sim f_1 f_2 P(v|w_1)P(v|w_2)$となり、$f_1 \times f_2$の項を含むためである。（ただし、確率が十分に小さいとして1次の項のみで近似。）



(a) 頻度によるバイアス　(b) フィルタリングの効果　(b) サンプリングの効果

図 2: 頻度による類似度計算値（タスク I、Simpson 係数）

図 2-(a) は、テキストから抽出した共起ペアをそのまま類似度計算に用いる場合である。Simpson 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。一方、図 2-(b) はフィルタリング法、図 2-(c) はサンプリング法を適用する場合で、頻度によるバイアスが低減されていることがわかる。

4.3 F 値による性能

前節で定義したタスク I、タスク II について、Jaccard 係数、Simpson 係数および 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 3 に、各条件による F 値の最大値を示す。

表 3 より、タスク I、タスク II とも、フィルタリング法と Simpson 係数を組み合わせる場合にもっとも性能値が高くなった。Jaccard 係数と Simpson 係数では、Simpson 係数の方が高い性能値が得られた。これは Simpson 係数では頻度が少ない語にあわせた正規化が行われるので、類似度計算の対象とする 2 つの語の頻度の違いに頑強であるためと考えられる。

また、格情報を考慮し「格+動詞」を共起語とする場合、Simpson 係数についてタスク I で 0.851 から 0.971、タスク II で 0.710 から 0.862 の性能改善がみられた。一方、格情報を考慮せず「動詞」だけを共起語とする場合、Simpson 係数についてタスク I で 0.810 から 0.962、タスク II で 0.674 から 0.852 とより大幅な性能改善がみられた。これは格情報を考慮しない場合の方が「ノイズ」にあたるデータが多く、フィルタリングによって効果的にノイズが低減されるためであると考えられる。

サンプリング法の中では、単純なランダム選択でもっとも高い性能値が得られた。ランダムに共起語を選択する場合の方が、高頻度語や相互情報量が高い共起語を優先的に選ぶ場合よりも利用する情報が少ないことを考慮すると注目に値する。特に Jaccard 係数については、ランダム選択（$N = 1000$）はフィルタリング法よりも高い性能を示した。

tf-idf コサイン尺度についても、フィルタリングやサンプリングの効果が観察されたが、性能は全般に Jaccard 係数/Simpson 係数の方が高かった。ただし、tf-idf における「文書頻度」(df) が何であるかは、検索エンジン経由の利用などでは定義があいまいである。実験ではそれぞれの共起語の名詞頻度（その動詞が何種類の名詞と共起したか）を「文書頻度」として用いたが、解釈は他にも考えられるので注意を要する。

なお、参考までに google API を用いて各語をクエリとした場合の検索エンジンヒット数 $n(w_1)$、$n(w_2)$、および両者を含むクエリのヒット数 $n(w_1, w_2)$ を求め、文献 [9] 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。この場合の F 値の最大値は、タスク I で 0.757、タスク II で 0.707 であった。ヒット数を用いる方法は手軽ではあるが、意味的な類似性ではなく共起の度合を測定することになる。これより、「国民年金や厚生年金など

表 3: F 値によるタスク性能の比較

		タスク I			タスク II		
		Jaccard 係数	Simpson 係数	tf-idf コサイン	Jaccard 係数	Simpson 係数	tf-idf コサイン
選別なし	格＋動詞	0.800	0.851	0.820	0.697	0.710	0.758
	動詞（格情報なし）	0.757	0.810	0.759	0.670	0.674	0.696
フィルタリング	格＋動詞	0.936	0.971	0.919	0.806	0.862	0.826
	動詞（格情報なし）	0.921	0.962	0.896	0.790	0.852	0.795
サンプリング	ランダム (N=100)	0.887	0.881	0.818	0.811	0.814	0.755
	ランダム (N=1000)	0.943	0.954	0.822	0.827	0.853	0.763
	頻度 (N=100)	0.667	0.667	0.667	0.662	0.662	0.662
	頻度 (N=1000)	0.735	0.731	0.722	0.770	0.775	0.766
	PMI (N=100)	0.929	0.939	0.818	0.819	0.824	0.757
	PMI (N=1000)	0.862	0.944	0.820	0.735	0.814	0.758
google ヒット数		0.757			0.707		

の保険料」などの例で判定誤りが生じることが観察された。

なお、上記に限らず、上位下位関係を正しく判定できなかった例としては、「テレビやラジカセなどの商品」（正解）や「経済学や政治学などの理論」（不正解）のように上位語の範囲が広いものや、「ベルベッドやベロアなどの衣料」（正解）や「書道や水墨画などの稽古事」（不正解）のように頻度が少ないものなどがあった。これらは共起語に基づく自動判定の限界を示すものと考えられる。

4.4 パラメタ値の影響

フィルタリング法では共起ペアごとの相互情報量の閾値 β が、サンプリング法ではサンプル数上限値 N がパラメタとなる。タスク I およびタスク II で Simpson 係数を用いる場合について、各パラメタの値に対する F 値性能の変化を図 3 および図 4 にそれぞれ示す。フィルタリングでは $\beta = 3$ 付近、サンプリングでは $N = 1000$ 付近に最適値があることがわかる。図には示していないが、他の尺度についても同様の傾向が確認された。

最後に実験を通して、タスク I とタスク II に対する結果は整合性がとれており、パラメタの最適値を含め傾向が一致することが確認された。

5 考察

本稿では、ウェブ文書に代表される大規模コーパス登場を背景として、言語的な利用の立場から類似度計算における規模拡大の影響を調べた。まず、大規模コーパス活用の 1 つの鍵は、言語処理の様々なあいまい性解消タスクにおける共起情報の活用であると考え、評価を行うための類語抽出タスクを設計

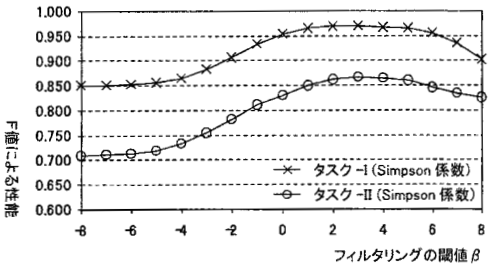


図 3: フィルタリングの閾値 β の性能値に対する影響

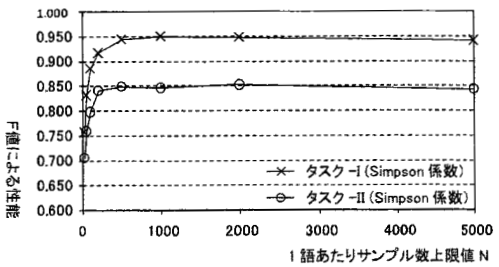


図 4: サンプリング上限値 N の性能値に対する影響

した。また、類似度の値が頻度に依存するバイアスを受けることを実際のコーパスによって確認し、ノイズ低減のためフィルタリング法、サンプリング法と呼ぶ２つの方法を提案して効果を確認した。上記の結果を踏まえ現在、フィルタリング適用後のデータを用いた大規模な類語・例文辞書の自動構築を試みている。

参考文献

[1] 相澤, 彰子「類語関係抽出タスクにおけるコーパス規模拡大の影響 (言語モデル・単語)」, 情報処理学会研究報告. 自然言語処理研究会報告 NL-94, pp. 91-98 (2006).

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

[3] 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).

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

[5] 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).

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

[7] 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).

[8] 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).

[9] 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).

[10] 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).

[11] 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).

[12] 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).

[13] Masao Takaku, Keizo Oyama, Akiko Aizawa, Haruko Ishikawa, Kengo Minamide, Shin Kato, Hayato Yamana, Junya Hayashi: "Building a Terabyte-scale Web Data Collection "NW1000G-04" in the NTCIR-5 WEB Task," NII Technical Report; NII-2006-012E; National Institute of Informatics; 8p. (2006-09).

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

[15] 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).

[16] 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).