自然言語処理 134−13
(1999. 11. 26)

類似例の存在を否定的な要因として用いる重要バイグラムの収集支援方法

丹羽 芳樹, 久光 徹, 今一 修

(株) 日立製作所、中央研究所
〒350-0395 埼玉県比企郡鳩山町赤沼

{yniwa,hisamitu,imaichi}@harl.hitachi.co.jp

概要

複合語には重要な意味を担うものがあり、それらは検索支援インタフェースにおいて重要な役割を演ずる。しかし複合語にはありきたりなものも数多く、それらはむしろマイナス効果をもたらす。本報告では、複合語の中でも単語バイグラムに注目し、重要なバイグラムをそうでないものと自動的に見分ける新しい手段について述べる。ありきたりなバイグラムは顕著な類似例を持つという特徴を利用して、そのようなバイグラムを選考対象からはずすことにより、重要バイグラムの選別作業を軽減することができる。新聞データを対象とした実験では、重要複合語の損失を3〜4%程度に押えながら30%程度の作業量（選考対象数）を軽減できることが分かった。

キーワード：複合語, 類似度, バイグラム, ユーザーインタフェイス, 特徴語

Filtering of Word Bigrams by Considering Similar Bigrams as Evidence of Unimportance

Yoshiki Niwa, Toru Hisamitsu, and Osamu Imaichi

Central Research Laboratory, Hitachi, Ltd.
Hatoyama, Saitama 350-0395, Japan

{yniwa,hisamitu,imaichi}@harl.hitachi.co.jp

Abstract

There are many compound terms which play important roles in the guidance of information retrieval, whereas countless many word n-grams are unnecessary or even harmful. In this paper, we focus on word bigrams rather than general n-grams, and describe a new method of distinguishing important word bigrams from trivial ones. This method characterizes trivial bigrams as those having remarkably similar ones. By automatically discarding trivial bigrams using this method, we can reduce considerable amount of human work needed for selecting important bigrams. In an experiment in which important bigrams are selected from Japanese newspaper articles, we were able to reduce about 30% of human work with as many as 3 to 4% of loss of important bigrams.

Keywords: compound term, similarity, bigram, user-interface, topic word

1 緒言

複合語とは複数の語が組になって一語のように振舞うものであるが、コンパクトな表記で豊かな意味を担うため、発想刺激能力に優れ、ユーザーインタフェース等での活用など注目度が高まっている。しかし複合語は分野依存性が高く、かつ時とともに移ろいやすいため辞書登録では対処しきれない部分が大きく、大規模テキストデータからの自動抽出、あるいは抽出支援が強く求められている。

本研究は「重要な複合語と平凡な複合語を機械的に判別する」ことを課題とし、それに対して「類

似例の多い複合語は平凡である可能性が高いだろう」というという仮説を立て、類似例に基づいて複合語の「平凡さ」を測る尺度（＝類似例尺度）を提案する。

ところで複合語の重要度とは何であろうか。これは本質的に主観的な尺度であり、また応用目的に依存するので定義することが困難であるが、我々は情報検索のユーザーインターフェースへの利用を主な目的として重要度を判断することにした。図1は西岡ら[17]による*DualNAVI*という対話的な検索支援システムのインタフェース画面である。画面右半分は検索結果の要約を特徴語グラフ([19, 12])の形で示したものであるが、検索結果の概要を教えたり、しぼり込に適した語をアドバイスできるなどの利点がある。我々は複合語の重要度を判断する場合、このようなガイダンス画面に提示することの価値で判断することとした。

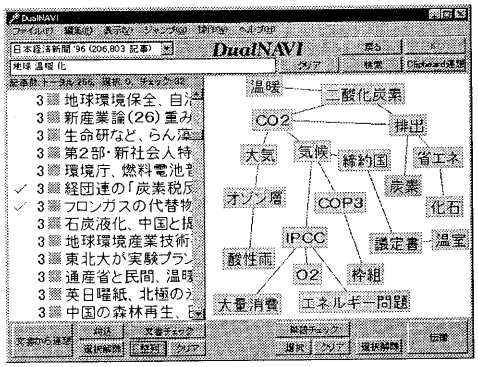

図 1. *DualNAVI*[17]のナビゲーション画面

図1は地球温暖化という検索要求で検索した場合であるが、注目して欲しいのはグラフの中ほど右寄りに現れた「炭素」という語である。これは実は「炭素税」に由来するが、「炭素税」というエントリーがないため「炭素」と「税」に別れ「炭素」だけが現れてしまった。しかし「炭素」を見て「炭素税」を想起できる人はごく稀であろう。すなわち「炭素税」は複合語のまま提示されて初めて適切なガイダンス機能を発揮できるので重要な複合語と判断される。

このように複合語には有益な効果をもたらすものがあるが、その一方で複合語を選別する必要性にも注意する必要がある。これは辞書エントリーの増加を押えたいという意味もあるが、質的な問題として、ありきたりな複合語をガイダンスとして提示することはむしろ逆効果を招く恐れがあるからである。例えば「サッカー」で検索した場面に「サッカーボール」、「サッカー選手」、「サッカー場」などありきたりな複合語がずらっと並んだことを想像すれば明らかだろう。すなわち、複合語はなんでも集めれば良いというのではなく、何らかの意味で非凡なものを選別する必要がある。

冒頭に述べたように、我々が提案する重要度を推定する指標は、類似例の有無であり、平凡なものには類似例がいくつもあるだろうという発想に基づいている。さきに掲げた「サッカー／選手」と「サッカー／くじ」の例は我々の仮説「顕著な類似例があるものは重要でない可能性が高い」を支持する例ともなっている。すなわち「サッカー／選手」には「ラグビー選手」や「野球選手」のような類似例をすぐに思い付くが「サッカー／くじ」の類似例はなかなか思い付かない。

第2節では（非）重要度を測る一つの指標としての類似例尺度を定義する。また第3節ではその類似例尺度を用いて平凡な複合語を除去することにより、いかに重要な複合語を収集する作業を効率化できるかという評価実験の結果について示す。

関連研究

大規模な文書データから自動的に収集される莫大な種類の形態素nグラムからいかにして重要な複合語を浮かび上がらせるかという課題をめぐっては自然言語処理をはじめ情報検索等々の分野で古くから研究が行われている[8, 1]。自然言語処理の分野でも８０年代末以降の大規模テキストデータの普及に伴い研究が活発化し、1990年にはCalzolari-Bindi [1]による相互情報量を用いたスコアづけが試みられている。また1993年にはDunning [5]により相互情報量に対する批判的考察がなされ、対数尤度(log likelihood)によるスコア付けの提案が発表されている。

Daille ら[3, 4](1994)はそれを受けて相互情報量、対数尤度、バイグラム頻度の比較を行い、対数尤度もしくはバイグラム頻度の優位性を示した。この結果は英語、フランス語など西欧の言語が対象であるが、日本語にも多分当てはまるのではないかと想像する。（ちなみに手元のデータで予備的に行った実験ではそれを支持する結果が得ら

れた。なお対数尤度とバイグラム頻度の優劣はやはり微細であった。）本格的な比較実験が日本語を対象になされているかどうかは筆者は未確認である。

頻度や対数尤度によるスコア付けは、最初のステップと考えられる。次に各種の理由で不要と判断される物を取り除く作業が必要となる。長い複合語の断片もその一つである。下畑ら [14] は左右隣接語の分布の乱雑さ（エントロピー）に着目し、それが低い場合には断片である可能性が高いとして除外する手法を提案し、文字列ベースでの日本語名詞句抽出と英語を対象とした（非連結を含む）連語表現抽出に適用し、有効性を示した。

中川ら [11] は影浦-海野 [8] により提唱されたtermhoodの体現として前後の接続語の多様さに着目し、それを複合語に限らず語彙の重要度スコアとして用いる手法を提案している。下畑らはある種のゴミを排除するという動機から、中川らは重要度という観点から、共通して接続語の多様さに着目したことになる。

またFrantzi-Ananiadou (1999) [6] の *C-value / NC-value* 法は統計量と言語特徴の融合への新しい試みとして注目される。

2 類似バイグラムの検出

本方法のキーとなる技術は類似バイグラムの検出である。バイグラムＡＢの類似バイグラムはＡかＢのどちらかを類似語に置き換えたＡ’Ｂあるいは Ａ Ｂ ’ という形のバイグラムの中から検出する。例えば「物価-上昇」をＡＢとすると「価格-上昇」がＡ’Ｂ型、「物価-高騰」がＡＢ’型の類似バイグラムである。

バイグラムはより類似した例があるほど、類似例尺度が高くなり、ユニークでないと判断される。類似例尺度の値は最も類似していると判断されたバイグラム達の類似度を上位いくつかの平均を取るなどの方法で総合判断して決定する。なおＡ’Ｂ型のバイグラムの場合の類似度はＡとＡ’の類似度、同様に、ＡＢ’型の場合にはＢとＢ’の類似度を意味するものとする。

類似バイグラムの類似度の総合判断の仕方についてはいくつもバラエティーが考えられ、最良の方法は現在模索中である。次節に示す重要複合語収集実験では、Ａ’Ｂ型の上位２位とＡＢ’型の上位２位の類似度の平均値を用いた。

以下 2.1節では本方法の技術的な基礎となる単語間の類似度計算方法（類似度の定義）を与え、2.2節ではそれを用いて類似バイグラムを検出した例を示す。

2.1 単語間の類似度

本論文のテーマは類似バイグラムの存在を重要度に関する否定的な要因として捉えようとするものである。この枠組自体は類似度の定義や計算方法からは独立したものであるが、その類似度計算がキーとなる役割を演じることもまた確かである。

単語間類似度の計算は自然言語処理において広い用途を持つ要素技術として多種多様な方法が提案され、かつ応用されている。主な手法としては共起語分布の確率・統計的距離を計算する方法 (Lin (1998) [10], Dagan et al. (1997) [2] など) とシソーラス上のリンク距離を用いる方法 (Kurohashi-Nagao (1994) [9], Fujii et al. (1997) [7]) がある。（その他関連研究はこれらの参照論文からその多くを辿ることができる。また工藤・井ノ上による1995年時におけるサーベー [16] (4.3節) がある。）

われわれの用いた類似度計算も基本的には共起語の分布による方法と位置づけられるが、確率・統計的な距離ではなく、むしろ心理言語学などで伝統的に用いられている共有属性の積み上げ式の類似度計算を用いた。(例えば Tversky [15] p.333 で *ratio-model* として紹介されているものなど。)

また共起語分布とはいっても、バイグラムのパートナーという非常に制限の強い意味での共起語を用いた。これは一つには我々の目的とするタスクとデータを共有できること、また他のより一般的な共起データよりもデータサイズが少なくて済むと言う現実的なメリットのためでもある。もちろん類似度計算ということを主目的と考えるならば、これは必ずしもベストの方法とは言えないだろう。

2.1.1 共有パートナーに基づく類似度

上記のように、我々はバイグラムのパートナーの共有度で類似度を測る、という定義を採用した。次の式はバイグラムの左側を構成する単語の類似度 sim_L であり、右側パート

ナーの共有度により定義されている。なお左右を入れ換えた sim_r も同様に定義される。

$sim_\ell(A, A') =$

$$\frac{1}{N_\ell(A) \cdot N_\ell(A')} \times \sum_{Y \in \langle A* \rangle \cap \langle A'* \rangle} weight_r(Y)$$

ここで $\langle A* \rangle$ はAの右側パートナーの集合、すなわち、

$$\langle A* \rangle = \{Y : \text{バイグラム A-Y が存在}\}$$

である。Yの重み $weight_r(Y)$ はすぐ後に定義を与える。正規化のためのAの（左側構成語としての）ノルム $N_\ell(A)$ は

$$N_\ell(A) = \sqrt{\sum_{Y \in \langle A* \rangle} weight_r(Y)}$$

で与えられ、自己類似度 $sim_\ell(A, A)$ を1に統一するような正規化である。

また $weight_r(Y)$ は単語Yが右側構成語として共有された場合の類似度へ与える貢献分であり、定性的には、多くの語とペアを組むものは低く、少数の語としかペアを組まないものは高い値を与えられるべき種類の値である。今回の実験では以下の定義を用いた。

$$weight_r(Y) = -\log\left(\frac{\#\langle *Y \rangle}{\#_\ell}\right)$$

ここで、$\#_\ell$ は考察対象のバイグラム達の左側を構成する単語の総種類数、$\#\langle *Y \rangle$ はYの左側パートナーの種類数であるから、対数の中身はある左側単語Xを与えた時にそれがXYというペアを組む確率と考えられる。属性のウェートは加算される性質の量であるから、属性毎の成立確率の対数を取るというのは、とりあえずは自然であろう（最適かどうかは別問題として）。

（ＡＢとＡ'Ｂの類似度のための補正） ところで右側の語を共有する二つのバイグラムＡＢとＡ'Ｂの類似度も基本的には $sim_\ell(A, A')$ で測れば良い。しかしこの場合にはＢが共通パートナーであるというのは前提条件なので、Ｂの共有分は類似度計算から除外するのが適当であると考え、以下のような補正された類似度定義 $sim_{\ell,B}$ を用いた。

$sim_{\ell,B}(A, A') =$

$$\frac{1}{N_\ell(A) \cdot N_\ell(A')} \times \sum_{\substack{Y \in \langle A* \rangle \cap \langle A'* \rangle \\ Y \neq B}} weight_r(Y)$$

2.1.2 類似度の計算例

類似度計算の一例として、「サッカー」と「テニス」左側構成語としての類似度、すなわち右側パートナー集合の類似度の計算例を示す。表1は、3つのパートからなるが、上段は両単語共通の（右）パートナーとなったもの、以下、サッカーのみ、テニスのみのパートナーと続く。各パートの中はウェートの大きさの順に上位5個ずつを掲げた。各パートの末尾には、省略されたものを含めて何種類の語が該当したかということと、そのパートのウェートの合計が記してある。

表1. サッカーとテニスの右側パートナー達とそれらの $weight_r$

サッカー、テニス共通の右パートナー	$weight_r$
準々決勝	9.08
全日（形態素解析誤り）	7.47
競技場	7.26
人生	7.03
予選	6.89
： ：	：
など 31 種, weigth 合計	179.71

サッカーの右パートナーでテニスの方にはないもの	$weight_r$
再戦	11.38
トヨタカップ	10.69
狂	10.00
春季	9.77
少年団	9.77
： ：	：
など 69 種, weigth 合計	464.00

テニスの右パートナーでサッカーの方にはないもの	$weight_r$
ウィンブルドン	10.69
日仏	9.30
全	8.68
民宿	8.68
ＵＳ	8.61
： ：	：
など 31 種, weigth 合計	210.31

この結果サッカーとテニスの類似度は、両者のノルムが $N_\ell(\text{サッカー}) = \sqrt{179.71 + 464.00} = 25.37$,

$N_\ell(\text{テニス}) = \sqrt{179.71 + 210.31} = 19.74$
と計算されるので、

$$sim_\ell(\text{サッカー}, \text{テニス}) = \frac{179.71}{25.37 \times 19.74} = 0.32$$

となる。

ここでウェート計算の例を一つ示しておく。テキストコーパスとして用いた日経新聞９７年[18]では左側構成語の種類数 ($\#_\ell$) が 87,853 であった。例えば共通右パートナーの一番上にある「準々決勝」は１０種類の左側パートナーを持つので

$$weight_r(\text{準々決勝}) = -\log\left(\frac{10}{87,853}\right) = 9.08$$

と計算される。

2.2 類似バイグラムの例

本節では、いくつかの例について類似バイグラムを示し、顕著な類似バイグラムの有無と複合語の重要度の関係を示す。

サッカー / 選手 vs. サッカー / くじ

はじめに、`平凡な'バイグラムと`非凡な'バイグラムで予想通りに類似例に差が出る例として「サッカー／選手」と「サッカー／くじ」を取り上げる。前者が平凡な組合せと思われる例であり、後者は非凡と予想した例である。
下の表２がその結果である。上は「サッカー選手」と類似したバイグラムであるが、最初にＢの方を入れ換えた場合で、サッカー／○○という複合語を構成する単語○○を「選手」との類似度が高く判定された順に５個ならべ、次にＡの方を同様に入れ換えて、「サッカー」との類似している順に５個並べたものである。右端の数字は類似度を示す。

この結果を見ると、特にＡの方の代替語としては「テニス」「ラグビー」など直観的にも「サッカー」と類似していると感じられる語が、高い類似度を伴っている。Ｂの方もそれほどでもないが、「代表」などはかなり「選手」と近い感じがするし、その他の語もそれほどかけ離れていないように感じられる。

一方下の方は「サッカー／くじ」に関する`類似例'である。これらは計算上最も類似していると判定されたものではあるが、直観的にも全く類似性が感じられず、また類似度の計算値もそれに符合して低い値になっている。すなわちこの場合にも直観と計算が合っていることになる。

表 2. 「サッカー／選手」と「サッカー／くじ」の類似例

	サッカー	選手	類似度
1		代表	0.12
2		選手権	0.11
3		競技	0.11
4		日本	0.10
5		記者	0.09
	- -		
1	テニス		0.35
2	ラグビー		0.32
3	バスケットボール		0.29
4	ボクシング		0.26
5	スキー		0.26

	サッカー	くじ	類似度
1		一次	0.04
2		全日 (形態素解析誤り)	0.03
3		ど (形態素解析誤り)	0.03
4		選手	0.03
5		型	0.02
	- -		
1	振興		0.08
2	話		0.06
3	スピード		0.06
4	シート		0.05
5	プレゼント		0.05

炭素 / 原子 vs. 炭素 / 税

次に、予期に反した結果になる場合もある例として「炭素-原子」と「炭素-税」の対比を示すことにする。

予想したところでは、炭素-税の方が炭素-原子よりも非凡な感じがするので、後者の方により類似度が高いと判定される例が現れると予想した。

しかし実際には次の表３が示すようにわずかな差ではあるが、炭素-税の方の類似例の類似度の方が高い結果となった。確かに炭素-原子の方には炭素の類似代替語として、塩素や酸素など直観的に類似度の高いと感じられる語が現れているが、計算された類似度はそれほど高い値とはなっていな

表 3. 「炭素 - 原子」と「炭素 - 税」の類似例

	炭素	原子	類似度
1		含有量	0.11
2		化合物	0.11
3		粉体	0.09
4		濃度	0.07
5		粉末	0.07
1	塩素		0.12
2	金属		0.10
3	酸素		0.09
4	水素		0.08
5	ビスマス		0.07

	炭素	税	類似度
1		税制	0.13
2		税収	0.13
3		基金	0.10
4		事業	0.08
5		法	0.08
1	二酸化炭素		0.13
2	バルブ		0.09
3	ガス		0.08
4	関連		0.07
5	促進		0.07

い。これは化学に関する分野が新聞ではマイナーな話題であり、従って統計的な類似度計算がうまく働かなかった結果と考えられる。

3 重要複合語収集への応用

本節では類似例尺度を重要複合語収集というタスクに応用し、その有用性を検証する。類似例尺度がある閾値を越えるものを探索対象から除くというヒューリスティクスを導入することにより、作業がどの程度効率化され、また副作用としてどの程度の重要複合語のロスが見込まれるかを小規模ではあるが実験で評価した。今後より大規模な実験が必要である。

3.1 実験方法

今回の実験には日経新聞(CD-ROM 版) 9 7 年[18] の約 2 0 万記事のタイトルと本文を利用し、形態素解析により単語＋品詞の列に分解し、品詞が名詞または未知語で構成されるすべてのバイグラムとその頻度をカウントした。形態素総数は 44 メガ個、名詞および未知語で構成されるバイグラムの延べ総数は 6.9 メガ個、種類数は 965,923 種類であった。なお形態素解析器として用いた ANIMA[13] は約 7 万文字／秒 (DEC Alpha 300MHz 上) の解析性能を持ち、単語分割に要した時間は 5 0 分程度であった。

(第 1 段階：対数尤度による順位付け) 最初に対象バイグラムを対数尤度をスコアとして順位付けを行った。その定義は以下で与えられる。

$$LL(AB) = fr(AB) \cdot \log \frac{fr(AB) \cdot N}{fr(A) \cdot fr(B)}$$

主尺度として対数尤度（の簡略版）を用いることにした理由は Daille ら [3, 4](1994) による比較実験で相互情報量に対する優位性が示されていたからである。上位 1 0 万個を取り、それらに対して類似例尺度を計算した。計算に要した時間は 3 時間弱であった。

(ヒューリスティクスによるふるい落し) 次に以下に示すようなヒューリスティクスによりふるい落しを行い、さらに両方の構成語が 2 文字以上という条件を付けて残ったものの上位 1 0 0 0 個（対数尤度による順位）を実験対象として用いた。
構成語が 2 文字以上という条件を付けた理由は、1 文字語を含む場合には形態素解析誤りの場合が多数含まれていて今回の実験の目的には適さないと判断したためである。

ヒューリスティクスとしては重要度に関する否定的な要因として、構成語の（少なくとも）一方が極めて多種類の語をパートナーとするもの、また、より長い複合語の部分と考えるられるもの、格助詞などを補って分解表現へ換言できるもの（コーパス中に分解表現が発見されたもの）などをある閾値を設定してふるい落した。詳細は本稿の主題と直接には関係しないので省略する。

3.2 作業量削減効果と重要語損失率の関係

類似例尺度の閾値を小さく設定すれば、それだけ類似例尺度の値が閾値を越え選出対象から除かれるバイグラムの個数が大きくなる。またその分副

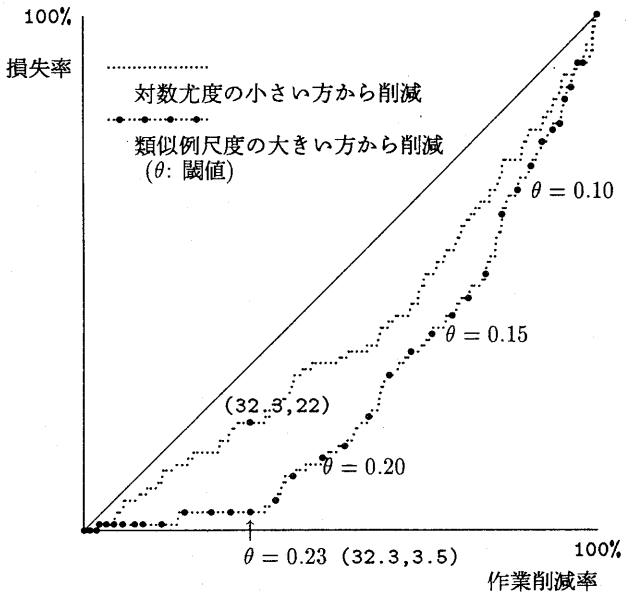

図 2. 類似例尺度の閾値 (θ) を変えた時の作業量削減効果と重要複合語損失率の推移

作用として多くの重要な複合語が失われる可能性もある。

実験対象とした１０００個のバイグラム中から、あらかじめ重要と判断されるものを選んでおき、それらがどの程度取り漏らされるかで損失率を計算した。重要かどうかの判断は筆者らの主観判断によるが、図１のような検索ナビゲーションの画面で一語として表示することの価値、言い替えれば、構成語に分けると著しく意味が損なわれるかどうかで判断した。結果として８６個のバイグラムが選ばれた。

図２は類似例尺度の閾値θを変化させた場合の作業量削減率と重要複合語損失率の推移を示したグラフである。
また比較のため、主尺度として用いた対数尤度の小さい方から順に削減した場合の推移も示した。グラフ中細かい点のみの方がそれであり、所々太い点が混じっているのが類似例尺度を用いた場合である。

このグラフで重要なことは類似例尺度を用いた方のグラフの最初の部分の立上りが遅いことである。すなわちある程度までは作業量削減しても損失が軽微に押えられることを示している。例えば類似例尺度の閾値を0.23に設定した場合、32%の削減率で損失が３～4%程度であることを示している。対数尤度による場合、同じ削減率での損失が22%に及ぶのに比べれば６分の１程度に押えられていることになる。

4　結言

コーパスから抽出した多数の日本語単語バイグラムを対象とし、統計的なスコアによる順位付けをした後、さらに重要な複合語の濃度を高める手段として、顕著な類似例のある複合語は平凡である可能性が高いので除くという方法を提案した。

新聞から抽出した単語バイグラムのセットを対象とした実験によりその効果を確認した。統計的スコア上位１０００位を対象とした場合、重要複合語の損失を３～４%程度に押えながら３０%程度の作業量を削減できることが分かった。

今後の課題

本提案手法は単語類似度の測り方に大きく依存する。従ってその最適化が課題となる。本実験ではバイグラムの右側にくる語としての類似度を計算

する場合には左側に伴う語の分布のみを用いた。
しかしながら純粋に単語間の類似度を測る目的であれば、左右片方より両方用いた方が（さらには遠距離の共起関係なども用いた方が）良い結果がでるはずであろう。
ただし、本研究の目的である重要な複合語を抽出するという目的に用いる場合には、どちらが良い結果になるかは明らかとは思われず、実験的に確認する必要がある。

共起分布とシソーラスベースの類似度を用いた場合の比較も興味深い。後者はネットワークタイプの語彙データベース上での単語間距離に基づく類似度であるが、その場合未登録語の影響が不利に働くのではないかと予想される。

参考文献

[1] Nicoletta Calzolari and Remo Bindi. Acquision of lexical information from a large textual Italian corpus. In *Proceedings of COLING'90*, pp. 54–59, Helsinki, 1990.

[2] Ido Dagan, Lillian Lee, and Fernando Pereira. Similarity-based methods for word sense disambiguation. In *Proceedings of the 35th Annual Meeting of the ACL*, pp. 56–63, Madrid, Spain, 1997.

[3] Béatrice Daille. Study and implementation of combined techniques for automatic extraction of terminology. In *The Balancing Act (Workshop at the 32nd Annual Meeting of the ACL)*, pp. 29–36, New Mexico, 1994.

[4] Béatrice Daille, Éric Gaussier, and Jean-Marc Langé. Towards automatic extraction of monolingual and bilingual terminology. In *Proceedings of COLING'94*, pp. 515–521, Kyoto, 1994.

[5] Ted Dunning. Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, Vol. 19, No. 1, pp. 61–74, 1993.

[6] Katerina T. Frantzi and Sophia Ananiadou. The *C-value/NC-value* domain-independent method for multi-word term extraction. *Journal of Natural Language Processing*, Vol. 6, No. 3, pp. 145–179, 4 1999.

[7] Atsushi Fujii, Kentaro Inui, Takenobu Tokunaga, and Hozumi Tanaka. Case contribution in example-based verb sense disambiguation. *Journal of Natural Language Processing*, Vol. 4, No. 2, pp. 111–123, Apr 1997.

[8] Kyo Kageura and Bin Umino. Methods of automatic term recognition. *TERMINOLOGY*, Vol. 3, No. 2, pp. 259–289, 1996.

[9] Sadao Kurohashi and Makoto Nagao. A method of case structure analysis for Japanese sentences based on examples in case frame dictionary. *IEICE Transactions on Information and Systems*, Vol. E77-D, No. 2, pp. 227–239, 1994.

[10] Dekang Lin. Automatic retrieval and clustering of similar words. In *Proceedings of COLING-ACL'98*, pp. 768–774, Aug. 1998.

[11] Hiroshi Nakagawa and Tatsunori Mori. Nested collocation and compound noun for term extraction. In *COMPUTERM '98 (Coling-ACL '98 workshop)*, 1998.

[12] Yoshiki Niwa, Shingo Nishioka, Makoto Iwayama, and Akihiko Takano. Topic graph generation for query navigation: Use of frequency classes for topic extraction. In *Proceedings of NLPRS'97*, pp. 95–100, 1997.

[13] Hirofumi Sakurai and Toru Hisamitsu. A data structure for fast lookup of grammatically connectable word pairs in Japanese morphological analysis. In *Proceedings of ICCPOL'99*, pp. 467–471, 1999.

[14] Sayori Shimohata, Toshiyuki Sugio, and Junji Nagata. Retrieving collocations by co-occurrence and word order constraints. In *Proceedings of ACL-EACL '97*, pp. 476–481, Madrid, Spain, 1997.

[15] Amos Tversky. Features of similarity. *Psychological Review*, Vol. 84, No. 4, pp. 327–352, 7 1977.

[16] 工藤育男, 井ノ上直己. コーパスに基づく共起知識の獲得とその応用. 人工知能学会誌, Vol. 10, No. 2, pp. 205–212, 3 1995.

[17] 西岡慎吾, 丹羽芳樹, 岩山真, 高野明彦. 文献検索支援インタフェース *DualNAVI*. In *Proceedings of WISS'97*, pp. 43–48. 日本ソフトウェア科学会, 1997.

[18] 日本経済新聞社. 日本経済新聞ＣＤ－ＲＯＭ１９９７年版, 1997.

[19] 丹羽芳樹. 動的な共起解析を用いた対話的文書検索支援. 情報処理学会・自然言語処理研究会報告, Vol. 96-NL-115, pp. 99–106, 9 1996.