重要複合語の自動判別方法

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。

1 緒言

1.1 動機と目的

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」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。

複合語を選択する際の価値基準は応用目的に依存し、かつ本質的に主観的な尺度である。本研究では情報検索へのガイダンスとして提示するのに適した語という観点から重要度を考えることにした。図1は西岡らによる *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[16]。画面右半分は検索結果の要約を特徴語グラフ([18])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。

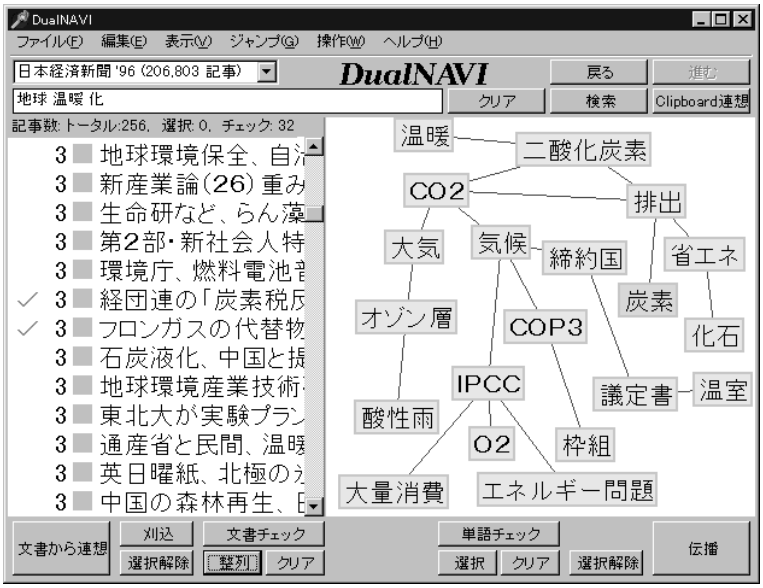

図 1. *DualNAVI* [16] のインタフェース画面

この図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のだ。しかし「炭素」を見て「炭素税」を想起できる人はいないだろう。複合語「炭素税」の効果が分かる。

このように複合語は有益な効果をもたらすものであるが、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、逆効果を招くからである。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??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の評価実験を行った。

なお本研究の関連研究として久光による語彙の重要度を測る統計的な新指標が提案され

ている。

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 (影浦-海野 [8] Calzolari-Bindi [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 [3, 4]。この結果は英語、フランス語などヨーロパ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。実際にそのような定量比較実験が日本語を対象になされたかどうか、筆者にはまだ調べがついていない。

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。次に各種の不要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 [14, 12]。

中川らは影浦-海野により提唱された 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 [19, 11]。下畑らは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 [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 ＡＢ’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」の場合には「価格-上昇」がＡ’Ｂ型、「物価-高騰」がＡＢ’型である。

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。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、最良の方法が決まった訳では

ないが、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。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’の類似度の意味である。

以下 2.2.1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、2.2.2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。一つはバイグラムの直観的なユニークさと類似例の少なさが符合する場合、もう一つはそれが符合しない（すなわち期待に反する）場合である。

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計算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ける有用な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など (Dekang Lin [10], Dagan et al. [2]) とシソーラス上のリンク距離を用いる方法 (黒橋 [9], 藤井 [7]) などがある。（その他関連研究は多い。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イ [15] (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、むしろ伝統的な共有属性の積み上げ式の類似度計算を用いた。(例えば Tversky [13] 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、またより一般的な共起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。

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

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 $weight_r(Y)$ は後に定義を与える。正規化のための A のノルム $N(A)$ は

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

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。

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

ここで、N_ℓ 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、対数の中身は任意の左側単語Ｘが与えられた時にそれがＹを右側パートナーに取ってＸＹというペアを構成するかどうかの確率と考えられる。さらに対数を取っている理由については、理論的な根拠があるというよりもむしろ大きさのレンジを適度に押えるためなのであるが、属性のウェートというのは加算される性質の量になって欲しいという要請からすれば、属性毎の成立確率の対数を取るというのは、自然であると考えられる（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］ところで右側の語としてＢ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。

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

ここで分母のノルム $N(A)$ や $N(A')$ には同様の補正を行っていないのは不統一感を免れないが、本方法では類似度が小さめに測られるよりも大きめに測られることを警戒しなければならないので、その意味から分母を小さくする恐れのあるノルム値の補正は行わなかった。特にＡやＡ’がごく少数のパートナーとしか組まない場合、分母からＢのウェート分を除くと、類似度が極端に大きく測られてしまう恐れがある。

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類似度計算

サッカー、テニス共通の右パートナー	重み
準々決勝	9.08
全日	7.47
競技場	7.26
人生	7.03
予選	6.89
: :	:
など 31 種, 重み 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	重み
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, 重み 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	重み
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, 重み 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
$N(\text{サッカー}) = \sqrt{179.71 + 464.00} = 25.37,\ \ N(\text{テニス}) = \sqrt{179.71 + 210.31} = 19.74$
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。

次にウェート計算の例を示す。テキストコーパスとして用いた日経新聞９７年 [17] では左側構成語の種類 (N_ℓ) 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以下のように計算される。

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

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探る。

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できるような○○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。

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

表 2.　「サッカー／選手」と「サッカー／くじ」の類似例

	サッカー *soccer*	選手 *player*	類似度
1		代表 (*representative*)	0.12
2		選手権 (championship)	0.11
3		競技 (*game*)	0.11
4		日本 (*Japan*)	0.10
5		記者 (*reporter*)	0.09
	- -		
1	テニス (*tennis*)		0.35
2	ラグビー (*Rugby*)		0.32
3	バスケットボール (*basketball*)		0.29
4	ボクシング (*boxing*)		0.26
5	スキー (*ski*)		0.26

	サッカー *soccer*	くじ *lottery*	類似度
1		一次 (*first*)	0.04
2		全日 (* JMA error)	0.03
3		ど (* JMA error)	0.03
4		選手 (*player*)	0.03
5		型 (*style*)	0.02
	- -		
1	振興 (*promotion*)		0.08
2	話 (*talk*)		0.06
3	スピード (*speed*)		0.06
4	シート (*sheet*)		0.05
5	プレゼント (*present (gift)*)		0.05

表 3.　「炭素-原子」と「炭素-税」の類似例

	炭素 *carbon*	原子 *atom*	類似度
1		含有量 (*content*)	0.11
2		化合物 (*compound*)	0.11
3		粉体 (*powder*)	0.09
4		濃度 (*density*)	0.07
5		粉末 (*powder*)	0.07
	- -		
1	塩素 (*chlorine*)		0.12
2	金属 (*metal*)		0.10
3	酸素 (*oxygen*)		0.09
4	水素 (*hydrogen*)		0.08
5	ビスマス (*bismuth*)		0.07

	炭素 *carbon*	税 *tax*	類似度
1		税制 (*tax system*)	0.13
2		税収 (*tax revenues*)	0.13
3		基金 (*foundataion*)	0.10
4		事業 (*enterprise*)	0.08
5		法 (*method*)	0.08
	- -		
1	二酸化炭素 (*carbon dioxide*)		0.13
2	パルプ (*pulp*)		0.09
3	ガス (*gas*)		0.08
4	関連 (*related*)		0.07
5	促進 (*promotion*)		0.07

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 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. Extraction of index words from manuals. In *Proceedings of RIAO'97*, pp. 598–611, 1997.

[12] 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.

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

[14] 下畑さより, 杉尾俊之. 隣接文字情報を用いたn-gram抽出文字列からの名詞句の自動抽出. 情報処理学会・自然言語処理研究会報告, Vol. 96-NL-114, pp. 13–18, 7 1996.
[15] 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.
[16] 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.
[17] 日本経済新聞社. 日本経済新聞ＣＤ－ＲＯＭ１９９７年版, 1998.
[18] 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.
[19] 中川裕志, 森辰則, 松崎知美. 日本語マニュアル文における名詞間の連接情報を用いた索引語の抽出. 情報処理学会・自然言語処理研究会報告, Vol. 96-NL-116, pp. 65–72, 11 1996.