自然言語処理 115－13
情報学基礎 43－5
（1996. 9. 13）

情報検索のための複合語マッチング

山田 剛一　　森 辰則　　中川 裕志
横浜国立大学 工学部

日本語では複合語が高い頻度で用いられる。より精度の高い検索を行うためには、語が複合しているという情報を活用することが必要である。本稿では、複合語を意識したマッチング法を提案する。さらに、名詞の重要度を基本名詞の連接数により判断する手法と組み合わせ、複合語を意識した文書の重要度付与を行う方法を示す。このシステムを複合語を意識しないシステムと比較したところ、再現率、適合率とも改善されることが確認された。

Compound Matching For Information Retrieval

Koichi Yamada , Tatsunori Mori and **Hiroshi Nakagawa**
Division of Electrical and Computer Engineering, Faculty of Engineering,
Yokohama National University
{aron@naklab,mori@forest,nakagawa@naklab}.dnj.ynu.ac.jp

In Japanese, compound words are frequently used. To improve retrieval efficiency, it is necessary to use the information about the structure of compound word. In this paper, we propose a new matching method for compound words. We also describe a method of ranking texts using noun-to-noun connections. Our experimental result shows that the method we proposed improves both recall and precision than the traditional tfidf method.

1　はじめに

日本語は、複合語が多く現れる言語である。これは、複合語を作るにあたっての自由度が非常に高いからである。その場その場で作られる複合語を辞書に網羅することは不可能であるから、複合語の意味は生成的に捉えることが必要である。これには特殊な辞書が必要となり、その構築は容易ではない。しかし見方を変えて、語が複合しているという情報だけを取り出すとすると、これは、少なくとも構文解析よりははるかに簡単な作業である。本研究では、この、語が複合しているという情報を活用することを考える。

複合語は一般に複合語全体で一つの概念を表すため、語が互いに複合語を構成しているということは、語が単に共起しているのとはかなり異なった性質の関係を表しているといえる。それは単に関連があるというのではなく、係り受けのような密接な関係である。語が複合しているという情報を有効に活用できれば、検索の精度を向上させることができる

と考えられる。

大量の文献を対象としたフルテキスト検索では、検索要求文に対して文書に重要度を付与することが一般的になっている。その重要度を求める際に、複合語は複合語のままマッチングをとることにより、語と語の深いつながりを反映した、きめの細かな重要度が求められると考えられる。

我々は以前、マニュアル文を対象とした研究において複合名詞の重要度を求める方法を検討した[1]。今回はその応用として語のマッチングを複合語として行い、従来捨てられていることが多かった、語が複合しているという情報を使用することによって、検索の精度が向上することを示す。

まず第2節において語の重要度の求め方について述べ、次の第3節で、その重要度を用いた複合語マッチングの方法を示す。これを元に比較実験を行い、その結果を第4節で考察する。

2 各文書における語の重要度

文書の重要度を考える上で、まず、その文書における各語の重要度を考える必要がある。ある語がどの程度重要であるかは、文書ごとに異なるからである。

重要度の尺度として一般に用いられているのは、語の文書内出現頻度 TF である。しかし、頻度情報は複合語の重要度を捉えるのに向いていない。なぜなら、複合語は頻度が低くても重要なことが多いからである。複合語によって表現される概念は複数の語により絞り込まれたものであるため、複合語はその文書を特徴づける重要な語となっていることが多い。よって、頻度情報に依存せずに、複合語の重要度を求める必要がある。

我々は以前、マニュアル文での重要語の抽出をする際に、名詞の連接情報を用いる手法を提案した[1]。本研究では、この手法を元に複合名詞の重要度を捉える。

2.1 基本名詞の重要度

複合名詞を構成する各名詞を基本名詞と呼ぶことにする。ここではまず、複合語を構成する各基本名詞の重要度を考える。ここで、「多様な複合語を生成する基本名詞は、その文書における基本的な概念を表していることが多い」という仮説を立て、複合語の合成力が強い語ほど重要度が高いと判断することにする。

基本名詞の持つ複合語合成力は、ある文書中で、その基本名詞が前後にどれだけの種類の名詞と連接するかで判断する。ただし、「の」を伴って係る場合も含むとする。

ある基本名詞 N の、文書 D_i における複合名詞合成力 $GP^{D_i}(N)$ を、次のように定義する。

$$GP^{D_i}(N) = \left\{(Pre^{D_i}(N)+1)\times(Post^{D_i}(N)+1)\right\}^{\frac{1}{2}}$$

$Pre^{D_i}(N)$ ：名詞 N の前方連接数
$Post^{D_i}(N)$ ：名詞 N の後方連接数

連接には前と後があるので、その相乗平均としている。

例として「辞書」という基本名詞の複合名詞合成力 GP^{D_i}(辞書) を見てみると、図1のような前方/後方連接数から、次のような計算で求まる。

$$\begin{aligned} GP^{D_i}(\text{辞書}) &= \left\{(Pre^{D_i}(\text{辞書})+1)\times(Post^{D_i}(\text{辞書})+1)\right\}^{\frac{1}{2}} \\ &= \left\{(21+1)\times(17+1)\right\}^{\frac{1}{2}} \\ &\simeq 19.34 \end{aligned}$$

2.2 複合名詞の重要度

複合名詞の重要度は、その複合名詞を構成する各基本名詞の重要度の積とする。これは、マッチングのスコアとして使うことを想定し、複合語でマッ

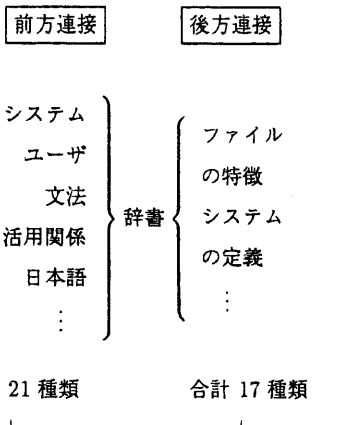

図 1: 連接数 Pre(辞書) と $Post$(辞書)

チしたということは基本名詞でマッチするよりも重要であるという意味を込めて積とするものである。

複合名詞と基本名詞をまとめると、名詞(全体)の重要度 $Imp^{D_i}(N_1, N_2, \cdots, N_m)$ は次のようになる。

$$Imp^{D_i}(N_1, N_2, \cdots, N_m) = \prod_{k=1}^{m} GI^{D_i}(N_k)$$

ただし、$N_1, N_2, \cdots, N_m$ は複合名詞を構成する基本名詞である。

基本名詞が複合語のどの部分であるかという情報は利用しない。例えば、主辞であるから重要であるかというと、語のカテゴリを決めるという意味では重要でも、検索において果たす役割という観点から見ると、重要な情報であるとは限らないからである。

3 複合語マッチング

この節では、本題である複合語マッチングについて述べる。

複合語は、検索要求文にも文書にも現れる。ここではまず、検索要求文中の一語と、文書中の一語とのマッチングに注目する。

3.1 語と語のマッチング

検索要求文中の一語を W^Q、文書 D_i 中の一語を W^{D_i} とする。どちらも複合語である可能性があるので、一般的には、N_1 の次に $N_2, \cdots$, 最後に N_n あるいは N_m という順序を持ち、次のように書ける。

$$W^Q = \langle N_1, N_2, \cdots, N_n \rangle$$
$$W^{D_i} = \langle N_1, N_2, \cdots, N_m \rangle$$

この 2 つの語から一致している部分を抽出する。その際、連続している部分は一まとまりとして取り出す。例えば次のような場合 (/ は連接を示す)、

$$W^Q = /A/B/C/D/E/$$
$$W^{D_i} = /B/C/E/$$

一致している部分は $/B/C/$ と $/E/$ である。$/B/$ や $/C/$ は $/B/C/$ に含まれているので取り出さない。なお、複合語中に同じ名詞が複数回出現する場合には、取り出そうとするパターンが重なる場合がある。このような場合は、基本語数の多いパターンを優先して取り出すこととする。

さて、取り出したそれぞれのパターンを $S_1, S_2, \cdots, S_P$ とする。上の例では、$S_1 = /B/C/, S_2 = /E/$ となる。これらもまた、基本名詞の列である。

$$S_j = \langle N_1, N_2, \cdots, N_l \rangle$$

というパターンの重みを次のように求めることとする。

$$PatternWeiyht^{D_i}(S_j) = \left\{ \prod_{k=1}^{l} GP^{D_i}(N_k) \right\} \times IDF(S_j)$$

例えばパターン $S_1 = /B/C/$ なら、その重みは次のようになる。

$$PatternWeight^{D_i}(/B/C/) = GP^{D_i}(/B/) \times GP^{D_i}(/C/) \times IDF(/B/C/)$$

文書内でのパターンの重みを、パターンを構成する基本名詞の複合名詞合成力 GP の積としている。パターン間のスコアは加算するため、複合しているという情報にはより大きい重みを与えるために積とした。

また、パターン全体の IDF を用いている。ただし、IDF(inverse document frequency) は次の定義に従う。

$$IDF(S_j) = \left(\log_2 \frac{DBsize(DB)}{freq(S_j, DB)}\right) + 1$$

DBsize: DB 内の総文書数

freq: DB 内で S_j が出現する文書数

複合語全体の重要度や IDF は用いない。どんな複合語の中で出現したのかということをスコアに加味することも考えられなくはないのだが、複合語が全体として重要だからといって、マッチした部分が重要な情報を担っているとは限らない。

3.2 検索語に対する文書のスコア

検索要求文内のある 1 つの語 (検索語) に対して、一般には何種類もの複合語が (部分) マッチする。そこで、マッチしたパターンのスコアを合計し、その文書のスコアとする。ただし、文書内での複合語の多様性についてはすでに基本語の重要度に折り込み済みなので、同一パターンのスコアは一回しか加算しない。

$$Weight^{D_i}(W_j^Q) = \sum_{k=1}^{P} PatternWeight^{D_i}(S_k)$$

S_k: マッチしたパターン

P: マッチしたパターンの種類の数

3.3 検索要求文に対する文書のスコア

最終的に求めるべき文書のスコアは、検索要求文にあるすべての語に対するスコアの総和として定義する。

$$Weight^{D_i}(Q) = \sum_{j=1}^{M} Weight^{D_i}(W_j^Q)$$

$$Q = \{W_k^Q | k = 1 \ldots M\}$$

4 評価

本稿で提案した複合語を意識したマッチング法の有効性を検証するため、以下で述べる基本的なシステムとの比較実験を行った。

4.1 比較するシステムの構成

単語の重みには $TF \times IDF$ を用い、文書の重み付けにはベクトル空間モデルを用いた。

ただし、IDF には以下の定義を用いた。

$$IDF(N) = \left(\log_2 \frac{DBsize(DB)}{freq(N, DB)}\right) + 1$$

DBsize: DB 内の総文書数

freq: DB 内で名詞 N が出現する文書数

文書の長さによる正規化処理は行っていない。

4.2 評価条件

4.2.1 評価用データ

評価には、情報検索評価用データベースである BMIR-J1 を利用[1]した。これは文書 600 記事と検索要求文 60 文、およびその正解からなるものである。この正解には A,B の 2 ランクがあるが、今回の評価では同一に扱った。

[1] 株式会社 日本経済新聞社の協力により、社団法人 情報処理学会・データベースシステム研究会・情報検索システム評価用データベース構築ワーキンググループが、1993 年 9 月 1 日から 12 月 31 日の日本経済新聞記事を基に構築した情報検索評価用データベース (テスト版) を利用

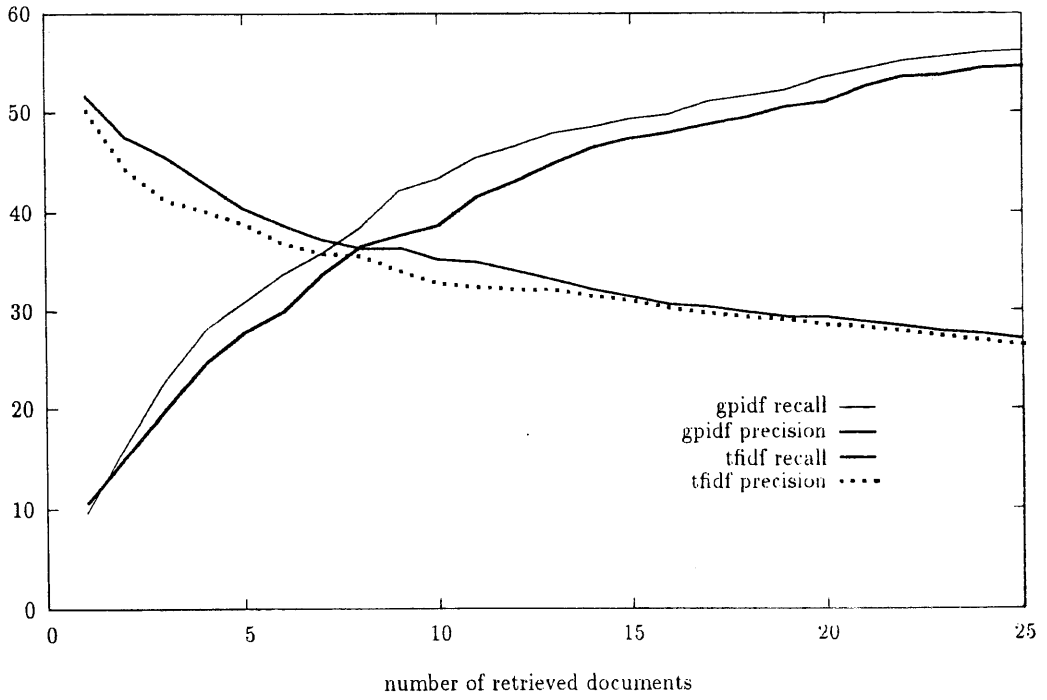

図 2: 出力数による再現率、適合率の変化

検索要求文 60 文の中には複合名詞を含まないものも多いのであるが、一般的な傾向を知るため、総ての検索要求文を利用して評価した。

なお、記事の情報は本文のみを利用し、タイトルや付与されているキーワード、記事の重要度等は使用していない。

4.2.2 評価の方法

検索要求文は JUMAN(3.0beta) を用いて形態素解析し、その中の名詞を抽出してシステムに渡すようになっている。もちろん複合語マッチングを行うシステムには、複合語の構成がわかるようになっている。なお、形式名詞、時相名詞、数詞は名詞から除外している。逆に、未定義語は名詞として扱っている。記事本文の形態素解析にも JUMAN を用いているが、どちらの場合も JUMAN の誤解析は修正していない。また、JUMAN の辞書の見出し語には若干の複合語が存在するが、これをさらに分割することはしていない。

各検索要求文に対して、システムは文書のスコアを出力する。この上位何位までを正解出力とするかをパラメータとして再現率、適合率を算出し、これを各検索要求文に対して平均した。

4.3 評価結果と考察

出力文書数によらず全般的に、再現率、適合率ともに若干の向上が認められた (図 2; gpidf が本手法)。

なお、どちらのシステムにおいても再現率、適合率とも低めであるが、これは今回の評価が複合名詞を意識するか否かを調査するのが目的で、検索語として名詞しか対象にしていないことが主な原因であると考えられる。

また、BMIR-J1 には言語知識や世界知識などの高いレベルのリソースを必要とする検索要求文が多く含まれている。そのような検索要求文では、今回比較システムではどちらも太刀打ちできず、よって両者の差もあまり出なかった。

5 関連研究との比較

複合語を扱った研究は数多くある [2] [3] [4] [5] [6] [7] [8]。ここでは本研究と近いアプローチである以下の研究について比較検討を行なう。

本研究と最も近いのは、小川らの研究 [6] である。しかし、複合語でマッチングをとるという点以外はすべて異なっている。

彼らは、複合語において基本語がどのような役割を果たすのかをタイプ分けし、この情報を辞書とは別に持つようにしている。このことは我々よりも複合語の構造を細かく捉えていることを意味するが、同時に特別なリソースを要求されることになる。彼らの基本語の重みづけはこの複合語の構造に依存するものであり、我々のように文書ごとに異なる値を持たない。よって、その文書でどれだけ多様に使われている語であるかということはスコアに反映されない。

複合語のマッチングでは、連接に着目する点は同じであるが、彼らは基本語の重みと連接数を別々に扱っている。我々の手法では、連接している場合には重要度の積を取るため、重要度の高い名詞が連接していると特にスコアが高くなるようになっている。ただ、そもそも語の重みの付け方が異なるため、マッチング法が違ってくるのも当然で、どちらが良いかを議論することはできない。

さらに違うのは、彼らは文書の重要度を、その文書に含まれる各キーワードに対するスコアの最大値としている点である。我々は、マッチしたパターンすべての和をとっている (同一パターンはとらない)。

われわれの手法は、名詞の重要度を連接数から求めることも含め、出現形態の多様性を重視した考え方をとっている。一方、彼らの手法は、語の構造を基礎としたマッチングによるスコアを素直に反映させるようにしている。これはあくまで予想であるが、両システムは基本姿勢が根本的に異なるので、得意不得意の領域に差が出るものと思われる。

高木らの研究 [7] は、文書重要度の付与に単語の文書内での共起関係を用いるものだが、検索語間の共起関連性を判断する際に、複合語を構成しているか否かという情報を利用している。しかし、文書内での共起関係を判断する際には、その語が複合語を形成しているかを見ておらず、複合語としてマッチさせているわけではない。彼らの手法による精度向上は、我々の手法とは別の要因による部分が大きいと考えられる。

亀田の研究 [8] では、複合語内の部分文字列に注目し、単語頻度を修正する形で複合語に対する重み付けをしている。この修正単語頻度は、各複合語の文字数と、異なる複合語間で共通する文字の数を考慮することにより得られる。この研究では複合語を基本語に分解せずに文字として扱うので、基本語の連接という概念自体が存在しない。基本語の重要度を基盤とする我々の研究とは根本的に発想が異なるといえる。この研究は、[8] の段階では重要キーワードと重要文の抽出を目的としているので、マッチング法については述べられていない。ただし、文書検索での文書の評価に利用する予定とのことである。

6 おわりに

本研究では、シソーラスを用いなくても、複合語の部分マッチを考慮に入れれば、かなり柔軟な重要度算出が可能であることを示している。しかし、シソーラスが本質的に必要である場合もあるため、シソーラスを用いたより柔軟な重要度付与の手法を構築することが望まれる。また、文書内で共起しているという情報も利用できればなおよいが、語の複合とは異なる次元の話をうまく反映させる枠組みが必要となる。

本研究では、語が複合しているという情報を使用するか否かで、検索の精度に差が認められるかどうかを検証することが目的であったので、名詞に絞ったシステムで比較評価をした。しかし、複合語

を扱う一般的なシステムとしては、複合名詞だけを扱うのでは情報の活用の度合いが足りない。そこで名詞以外の語、接頭語、接尾辞の類や数詞と助数辞などを扱うことになるのだが、名詞が単独で概念を表しているのにくらべ、これらはあきらかに性質が異なるので、名詞と同列に扱うことはできない。名詞のように連接数で重要度を判断できるかは不明であるし、マッチングにおいてもパターンがフラットな列であるとするのは乱暴かもしれない。

本研究では「の」による連体修飾を複合語の形成の一部として捉えたが、逆に「の」で接続されているという情報は、多様な接続を許す複合語形成の中では、構造を捉える有力な手がかりになる。それは、複合語内の接続とはレベルの違う現象だからである。もちろん「の」による連体修飾は意味的に非常に多様であるから、意味的な構造までは捉えようがないが。

「の」による接続や接頭語、接尾辞などを考慮し、ある程度の階層構造を持たせた上でのマッチングを考えてみたいと思っている。

謝辞 BMIR-J1 を提供してくださった方々、特にリコー 小川さん、富士通 松井さんに感謝いたします。また、JUMAN を公開、発展させ続けている方々に感謝いたします。

参考文献

[1] 松崎知美, 雨宮秀文, 森辰則, 中川裕志. 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重要語の抽出. 情報処理学会研究報告 96-NL-113-16, 自然言語処理研究会, 情報処理学会, May 1996.

[2] David A. Evans and Chengxiang Zhai. Noun-phrase analysis in unrestricted text for information retrieval. In *Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics*, June 1996.

[3] 大井耕三, 隅田英一郎, 飯田仁. 単語間の意味的類似度に基づく文書検索手法. 言語処理学会第2回年次大会発表論文集, pp. 109-112, March 1996.

[4] 小林義行, 徳永健伸, 田中穂積. 複合名詞の構成要素間の関係推定の一手法. 言語処理学会第1回年次大会発表論文集, pp. 85-88, March 1995.

[5] 宮崎正弘. 係り受け解析を用いた複合語の自動分割法. 情報処理学会論文誌, Vol. 25, No. 6, pp. 970-979, Nov 1984.

[6] Yasushi Ogawa, Ayako Bessho, and Masako Hirose. Simple word strings as compound keywords: An indexing and ranking method for japanese texts. In *ACM-SIGIR'93*, pp. 227-236, June 1993.

[7] 高木徹, 木谷強. 単語出現共起関係を用いた文書重要度付与の検討. 情報処理学会研究報告 96-FI-41-8, 情報学基礎研究会, 情報処理学会, April 1996.

[8] 亀田雅之. キーワード相関法による重要キーワードと重要文の抽出. 言語処理学会第2回年次大会発表論文集, pp. 97-100, March 1996.