文章を観点とした単語間の類似性判別方式

笠原 要　　松澤 和光
NTT(株) コミュニケーション科学研究所

1 はじめに

文章から文脈を表す単語（「観点」）を決定し、辞書から構築した単語の概念知識ベース（「概念ベース」）を利用し、観点に応じて単語間の類似性を判別する方式を提案する。

人間は、単語の意味を表す「概念」間の類似性を、その単語が扱われる文脈や状況の変化に応じて柔軟に判別する．これは例えば、「馬」に対して「豚」と「自動車」のどちらが類似しているか判別する際、「動物」が話題であれば「豚」が、「乗り物」の話題では「自動車」が、各々「馬」により似ていると判別することができる。単語の類似性判別に一般的に用いられるシソーラスでは、単語間の関係が固定的に記述されており、このような柔軟な類似性判別を行なうことができない。

辞書は、単語の概念知識が記述されており、類似性判別の知識源として有望である。機械可読辞書から獲得した大規模な語彙知識を用い、状況や文脈を考慮した類似性判別方式がいくつか提案されている[1, 2]。ただし、観点をどの様に類似性判別に反映すべきか、十分な研究は進んでいない。また、機械可読辞書と言っても、あくまで人間が読むためのものであり、計算機で処理することを考えて作成されたものではない。従って、辞書から十分な品質の知識を獲得することは難しい。

我々は、日常用いられる数多くの単語に対し、観点に応じた柔軟な類似性判別を行える新しい方式を提案した[3]。この方式の特徴は、まず国語辞書の語義文から容易に獲得可能な知識だけを用いて概念知識ベース（概念ベース）を構築し、次いで「精錬」と名付けた自己参照的な手法により、概念ベースの知識を精密化することにある[4]。この二段階の処理により、日本語の辞書からでも高品質の概念知識を獲得することが出来る。そして、概念ベースに含まれる概念の中から観点に相当する概念を指定し、「変調」と名付けた処理によって、概念ベース中の任意の2概念間の類似性を観点に応じて判別することを可能としている。

上記類似性判別手法において、観点は人手で入力する必要がある。人間は、状況や文脈に敏感な類似性判別を行なえるが、状況や文脈を明確な単語としての観点に要約することは得意ではないことが、提案方式の評価の過程で明らかになった。その理由としては、観点となる単語は一語とは限らず、多数の単語について判定を下すためと考えられる。そこで、類似性判別を行なう単語を含む文章を対象とし、観点を決定する方式を提案した[5]。これは、文章を構成する自立語に対し、関連する単語を概念ベースから抽出し、その頻度に基づいて観点を決定するものである。また、文章と別の対象として、互いに類似関係にある類似語から成る集合を取り上げ、概念ベースを用いて集合を成り立たせる観点を決定する方式を提案した[6]。

本稿では、観点が文章中の単語に含まれると考え、情報検索の考え方を用いて、単語の頻度情報に基いて観点を決定する方式を提案する。また、決定された観点に基づいて、文章中の単語の類似性判別を行ない、動的に単語をクラスタリングしてシソーラスを構築する方式を提案する。

2 概念の類似性判別方式

2.1 概念ベース

概念ベースでは、人手によらない自動構築を行なうために、単語の特徴を表す属性とその重みの対のリストによる概念の単純な表現を採用している。

$$Word_i = \{(p_{i1}, q_{i1}), \cdots, (p_{in}, q_{in})\}. \qquad (1)$$

概念の属性は、辞書の語義文より獲得する。また、語義文中で出現頻度の高い属性は、その単語の概念で特徴的であると仮定し、属性の出現頻度を重みから定義した。さらに、得られた概念ベースを精錬し、現在、日常語4万語の概念知識（平均44属性/概念）を複数の辞書の参照により獲得している。

2.2 概念の類似性判別法

前記の概念ベースを用い、観点に応じた概念の類似性判別を実現した。2つの単語の意味の近さを表す類似度 S は、概念ベース中の2つの単語の概念 $Word_1$、$Word_2$ と、判別の観点となる概念（「観点」と呼ぶ）$View$ より、次の4つの手順に従って計算する（図1）。（詳しくは文献[7, 3]参照）

- **STEP 1: 属性のシソーラス圧縮**
  類似度の計算は同じ属性同士の重みの比較に基づくため、表記が異なるが意味の似た属性は等しく扱う必要がある。そこで、属性をシソーラスのカテゴリーに変換する。
- **STEP 2: 重みの正規化**
  概念毎に、重みをノルム1になるように正規化を行なう。
- **STEP 3: 観点に応じた重みの変調**
  観点に応じた類似度の計算とは、観点に含ま

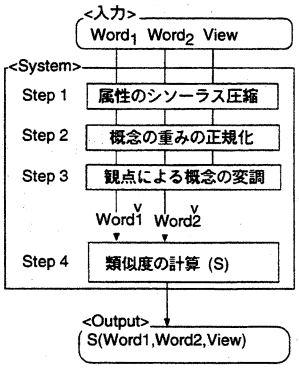

図 1: 類似性判別方式の全体図

れる特徴的な属性を重視して行なうものと仮定する。つまり、比較する概念中で観点と共通する属性の重みを大きくする。

- **STEP 4: 類似度の計算 (S)**
  情報検索で一般に行なわれる、ベクトル空間での検索手法の考え [8] を用い、類似度 $S(0 \geq S \geq 1)$ を、属性空間上の 2 つの変調概念がなすベクトルの余弦から計算する。

表 1に、以上の方式を用いて類似性判別を行なった例を示す。「馬」に対して「豚」と「自動車」のどちらが類似しているかを判別する際に、観点「動物」では、「豚」の方が類似していることを示すが、観点「乗る」では、反対に「自動車」の方が類似していることを示している。これは、人間の感覚に合った判別結果と言える。

表 1: 類似性判別の実例

観点	S('馬', '豚', V)	S('馬', '自動車', V)
動物	0.84	0.29
乗る	0.23	0.60

3 文章を観点とした単語間の類似性判別方式

3.1 観点決定方式

文書中で、その文脈を良く表現する単語を観点として抽出する方式を提案する。観点となる単語を決定するためには、文章の構文・意味理解が必要であるが、それらは研究の途上にある。そこで、ここでは最も簡単な方式として、情報検索におけるキーワードの重み付けの考えを用いた 3 種類の方式を提案する。

【方式 1】
情報検索において、ある文書 d における単語 t のキーワードとしての重要度を、t の term frequency $tf(d,t)$ により定義する方式がある [9]。これは、文書内で頻出する単語程重要であるという考え方に基づく。この考えを文書 d 中の単語から観点 k を決定する方式として用いる。k は、以下の式を満たす単語である。

$$tf(d,k) > tf(d,t_i). \tag{2}$$

t_i とは、文書 d に含まれる任意の単語を表す。

【方式 2】
文書内出現頻度のみで重要度を付与した場合、形式名詞や頻出する言い回し等の無意味であるが頻出する単語の重要度が高まる。そこで、単語数 N からなる全文書中に対し、単語 t の出現頻度 $df(t)$ を用いて、その単語の「全体として少数しか出現しない」程度 inversed document frequncy idf(t) [10] を求め、tf の値を補正した $tf * idf$ をキーワードの重要度とする考え方があり、これを観点 k の決定方式に用いる。

$$tf(d,k) * idf(k) > tf(d,t_i) * idf(t_i)$$
$$idf(t) = log\frac{N}{df(t)}. \tag{3}$$

【方式 3】
方式 2 で単語の重要度を補正した idf を、特定の分野の文書から計算する場合、その分野に共通して出現する重要な単語の重要度も下げてしまうおそれがある。例えば、新聞の経済分野の記事を対象とした場合、文書には平均的に単語「経済」が頻出する。「経済」は観点として重要と考えられるが、idf(「経済」) は、その計算方法ゆえに観点「経済」の重要度を相対的に下げる働きを行なう。

そこで、分野依存が少なく定義的な文書からなる辞書から idf を計算し、単語の出現頻度を補正して観点の重要度とする。

$$tf(d,k) * idf(D,k) > tf(d,t_i) * idf(D,t_i)$$
$$idf(D,t) = log\frac{N}{df(t)}. \tag{4}$$

$idf(D,t)$ とは、辞書 D 中での単語 t の idf を表す。

3.2 文書中の単語からの動的シソーラスの構築

上記の文書から観点決定により、文書中の単語を類似性に基づいて分類することが可能となる。この分類は観点を考慮しているので、固定的に記述されたシソーラスに比べ、人間の意図に応じた分類と考えられる。また、観点に応じて分類を行なうことにより、シソーラスを動的に構築することができる。その方法は、以下の通りである。

- 文書より観点を決定
- 文書中の単語同士の全組合せについて、観点に基づく類似度を計算
- 類似度を距離に変換
- 単語同士の距離行列を元に分類

多次元空間上の概念はノルム1のベクトルであり、類似度は2つのベクトルの余弦であるので、類似度から距離 $L(0 \leq L \leq 1)$ を計算できる。この距離を元にして分類、階層的クラスタリングなどを行ない、動的なシソーラスを獲得する。

4 実験結果

文書の観点の決定法は、確立された方式が存在しない。そこで、新聞記事の見出し中の単語を観点とみなして、方式の定量的評価を行なった。また、決定された観点を用い、文書内の単語を分類し、観点の効果を調べた。

4.1 観点決定法の評価

日本経済新聞の記事600件[1]を形態素解析し[11]、概念ベースに含まれる4万語の日常語を含む595件の記事対象として、記事の本文から観点を決定する実験を行なった。記事の見出しは、記事を的確に要約したものであり、見出し中には高い確率で記事の観点が含まれていることが期待される。そこで、見出しの単語を観点と見なし、提案した観点決定方式で重み付けした観点のリストを評価した。本文中に、見出しの単語が全く含まれない記事が13件存在し、これを除いた582件の記事について評価を行なった。その結果を表2に示す。平均順位とは、観点リスト

表 2: 観点決定方式の評価結果

方式	平均順位	平均適合率
1	32.54	0.1800
2	20.52	0.2194
3	21.05	0.2201

中での正しく検索された観点の順位の平均を、全記事について平均したものである。また、平均適合率は、観点リストの1位から n 位までの観点と見なして適合率と再現率を求め、n を1からリストを構成する検索単語数まで変化させた時の適合率の平均値を表す。平均順位と平均適合度ともに、出現頻度 tf のみから観点を決定する方式1に比べて、idf を考慮した方式2・3が良い評価を与えている。

[1]「株式会社 日本経済新聞の協力によって、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」

また、方式2と3は、単語の特殊性を表す idf をそれぞれ新聞記事と辞書から算出している点が異なるにもかかわらず、ほぼ同じ評価結果を与えている。辞書から計算した idf は単語の、分野依存性の少ない一般的な特殊性を表現している可能性が考えられる。

4.2 動的シソーラスの構築

上記で評価した観点決定方式を用い、文書中の単語の動的シソーラスを構築した。以下に、対象とした新聞記事の本文を取り上げる。

> **三井物産、子会社の宝飾事業縮小――経営再建に専念**
> 三井物産は宝飾・貴金属製品などの販売子会社、物産ジー・アンド・エー（東京・千代田、三神広臣社長）の宝飾事業を大幅に縮小、経営再建に専念する。同社は宝飾品の売り上げ不振で過剰在庫を抱え、…

この記事から決定された観点は、表3の通りであるここでは、方式2・3によって選択された観点「物産」を用いる。クラスタ分析は、S言語[12]によ

表 3: 決定された観点一覧

方式1		方式2		方式2	
観点	重み	観点	重み	観点	重み
事業	6	物産	60.1	物産	30.2
撤退	4	撤退	53.8	撤退	28.7
販売	4	事業	46.7	事業	27.6
物産	4	在庫	39.1	在庫	23.6
在庫	3	販売	37.5	販売	20.4
…		…		…	

る最長距離法を用いた。クラスタ数を10とした場合の、類似語集合を表4にあげる。また、観点「物産」における記事本文中の階層的クラスター分析結果を図2に示す。これは、一種のシソーラスであり、観点に応じて動的に変動することが特徴である。

5 おわりに

文書から、その文書の文脈を良く表す単語である観点を、統計情報に基づいて決定する3方式を提案した。新聞記事を文書とした評価実験を行ない、文書中の単語の出現頻度 tf だけではなく、文書全体から単語の特殊性を表す指数 idf により出現頻度を補正した値から、観点を決定することが良い結果を

表 4: 類似語集合

No.	類似語集合
1	落ち込む, 貴金属, 急激, 強化, 金属, 継続, 子会社, 撤退, 長引く, 物産
2	業績, 結果
3	大きい, 大幅, 過剰, 軽減, 減少, 半減
4	神, 考え付く, 期, 近い, 二
5	英, 営業, 会社, 業界, 今期, 今後, 三月, 社長, 臣, 新規, 倒産, 当面, 柱, 場合, 不況
6	売り上げ, 卸, 小売り, 在庫, 仕入れ, 製品, 販売, 負債
7	経営, 先, 事業, 対す
8	悪化, 言う, 関する, 決断, 検討, 三つ, 事実, 専念, 努める, 見方
9	広い, 部門, 見通し
10	既存, 縮小, 民間

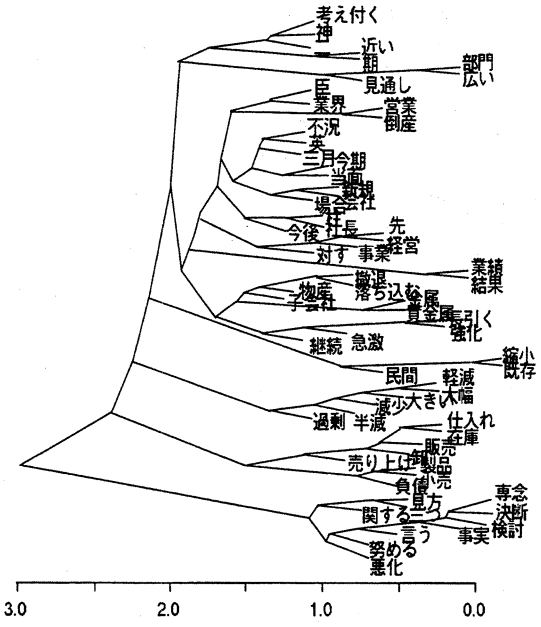

図 2: 観点「物産」に基づくシソーラス

示すことを明らかにした。また、*idf* として、辞書から獲得した結果が、分野に依存せずに利用できる可能性を示した。さらに、観点に応じた類似性判別方式を用い、文書中の単語から動的シソーラスを構築する手法を提案した。今後は、応用を通して動的シソーラスの評価を行なう予定である。

References

[1] 北川, 清木, 人見: 意味の数学モデルとその実現方式について, 信学技報, Vol. DE93-4, pp. 25–31 (1993).

[2] 小嶋, 伊藤: 意味空間のスケール変換による動的シソーラスの実現, 信学技報, Vol. NL95, No. 19, pp. 1–8 (1995).

[3] Kaname, K., Matsuzawa, K., Ishikawa, T. and Kawaoka, T.: *Viewpoint-Based Measurement of Semantic Similarity between Words*, Lecture Notes in Statistics: Learning From Data, Vol. 112, Springer-Verlag, chapter 41, pp. 433–442 (1996).

[4] Kasahara, K., Matsuzawa, K. and Ishikawa, T.: Refinement Method for a Large-Scale Knowledge Base of Words, *the Third Symp.on Logical Formalizations of Commonsense Reasoning*, pp. 73–82 (1996).

[5] 笠原, 松澤: 類似語検索における観点の自動生成法, 情報学基礎, Vol. 42-5, pp. 29–36 (1996).

[6] 笠原, 松澤: 辞書を用いた類似語集合の観点決定法, 自然言語処理シンポジウム'96 (1996).

[7] 笠原, 松澤, 石川, 河岡: 観点に基づく概念間の類似性判別, 情報処理学会論文誌, Vol. 35, No. 3, pp. 505–509 (1994).

[8] Salton, G. and McGill, M.: *Introduction to modern information retrieval*, McGraw-Hill (1983).

[9] Luhn, H. P.: A statistical approach to mechanized encoding and searching of literary information, *IBM Journal*, pp. 309–317 (1957).

[10] Sparck Jones, K.: A statistical interpretation of term specificity and its application in retrieval, *Journal of Documentation*, Vol. 28, No. 1, p. 11.21 (1972).

[11] 池原, 宮崎, 横尾: 日英機械翻訳のための意味解析辞書, 情報処理学会自然言語処理研究会, Vol. 84-13, pp. 95–102 (1991).

[12] Becker, R. A., Chambers, J. M. and Wilks, A. R.: *The New S Language: A Programming Environment for Data Analysis and Graphics*, Wadsworth & Brooks (1988).