前後に出現する長い共通文字列を用いる関連語判定法

折原幸治　藤原大輔　梅村恭司
豊橋技術科学大学 情報工学系
{orihara,fuji}@ss.ics.tut.ac.jp, umemura@tutics.tut.ac.jp

概要

コーパスの文脈から 2 つの語に関連があるかを判定する問題において，前後に共通する長い文字列があるかどうかを判定に使用する方法について実験結果を報告する．この問題では，一般には多く出現する隣接単語や隣接修飾語のそれぞれについて，統計値を計数して総合判定することが普通であるが，本報告では共通する長い文字列という出現頻度のきわめて小さい特徴に注目し，それを用いた場合の結果について報告する．頻度の低い特徴を利用するということに加え，実際に非常に長い共通文字がとれた理由についても考察する．

1　はじめに

文章を書くときや情報検索を行うようなときなどには，似たような意味を持つ単語が役に立つことがある．同じ意味でも文章の内容に適した用語を探したり，情報検索においては，似たような言葉で検索することによって検索結果の絞り込みや漏れのない検索が期待できる．

このような場合はシソーラスと呼ばれる辞書が利用できる．シソーラスは単語を意味で分類した辞書であり，例えば，映画の同義語を引くと，キネマ，スクリーン，活動写真などの語が得られる．有名なシソーラスには，英語では WordNet [1]，日本語では日本語語彙大系 [2] がある．

これらのシソーラスは人手によって作られる．しかし，人手による方法では無数にある単語を意味によって分類しなくてはならず，非常に高価である．さらに，言語の性質にある創造性によって増加する単語に対応し続けることは難しい．そのため，シソーラスを自動的に作成することが望まれる．

シソーラスの構成要素である類似語を自動的に抽出するものとして，當間らのシステム [3] がある．このシステムは，テキストデータの統計情報のみに基づき，辞書を一切利用せずに関連語対を抽出する．當間らのシステムにおける関連語は，「文章中に同じように使われる単語」とされ，関連語の候補となる単語の前後にある単語の統計情報を元に評価している．

本報告では，このシステムの改善に単語対の前後に共通して現れる文字列を用いる手法を提案する．この手法における関連語の定義は，當間らの定義をさらに厳しくしたものであり，「文章中でまったく同じように使われる」としたものである．本報告では，この手法を用いて関連語対を抽出し，その結果を分析する．

2　本システムについて

2.1　システムの概要

本システムは當間らのシステム [3] を拡張した．當間らのシステムによって抽出された関連語対を提案手法でランキングする．以下に，システムの動作手順を示す．

1. 単語の切り出しと統計情報の取得
2. 順序対の抽出
3. 候補対の選出
4. 関連語対の判定
5. 提案手法の関連語対のランキング

1 から 4 までは當間らのシステムそのものであり，5 が本報告の提案部分である．當間らの方法では，

われわれは [**増大**] する [**知識**] を処理する方法を身につける必要にせまられているが，これは [**学校**] で教えられることがない．[**最近急速**] に [**普及**] しはじめた [**パーソナルコンピュータ**] は，[**ソフトウェア**] の [**貧弱**] さもあって [**知識**] を処理する [**道具**] として [**十分**] に [**活用**] されていない面がある．

[ ] 内はキーワード

図 1　キーワード抽出結果例

関連語の前後の単語に注目していたが，本提案では関連語の前後にある文字列に注目する．各項目について以下に示す．

2.2　當間らのシステム

2.2.1　単語の切り出しと統計情報の取得

ここでは，対象コーパスから単語の切り出しを行う．単語の切り出しには，武田ら [4] のキーワード抽出アルゴリズムを用いる．武田らのアルゴリズムは，辞書を利用せずに，コーパス中の部分文字列の出現頻度などの統計情報からキーワードらしさを判定する．このアルゴリズムを用いてテキストからキーワードを抽出した例を図 1 に示す．

2.2.2　順序対の抽出

ここでは，順序対を抽出する．順序対とは，キーワードのみを出現順に並べた時に続けて現れる傾向の高い前後のキーワード対である．順序対の抽出は χ^2 検定によって行い，帰無仮説は前後に位置するキーワード対を (s,t) としたとき，「キーワード s と t は関連がない」である．この検定を文章中で隣接したキーワード対のすべてに対し行い，合格したキーワード対が順序対となる．ただし，実験では危険率を 0.005 とした．

2.2.3　候補対の選出

ここでは，候補対を選出する．候補対とは，関連語対の候補となるもので，「同じように使用される」という観点から図 2 に示すような前後に同じキーワードが出現するキーワード対である．キーワード対の前後それぞれに少なくとも 1 つの共通のキーワードがある場合にそのキーワード対を候補対とする．

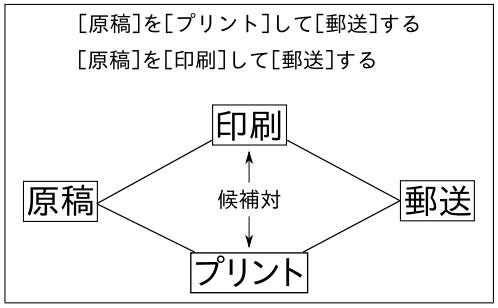

図 2　候補対の選出

2.2.4　関連語対の判定

ここでは，候補対から最終的なシステムの出力である関連語対を式 (1) の尺度を用いて判定する．この式は，候補対 (s,t) の各キーワードの前後にある単語の種類を元にして候補対 (s,t) の関連度をスコア付けする．POS は位置の集合であり，その要素 p は前または後をとる．a_p は位置 p において，s 及び t に共通する単語の種類数，b_p は s にはあって t にはない単語の種類数であり，c_p は b_p とは逆で，t にはあって s にはない単語の種類数である．

$$\mathrm{logcos}(s,t) = \log(\min(\mathrm{cf}(s),\mathrm{cf}(t)))\cos(s,t) \quad (1)$$

$$\cos(s,t) = \sum_{p \in POS} \frac{a_p}{\sqrt{(a_p+b_p)(a_p+c_p)}} \quad (2)$$

ただし，$\mathrm{cf}(x)$ は単語 x のコーパス全体での出現回数である．実験では，この尺度を用いて候補対をランキングし，上位 10,000 件を関連語対と判定した．

2.3　提案手法の関連語対のランキング

提案手法では関連語対のランキングに各キーワードの前後にある文字列を使用する．まず，関連語対の各単語の前後にある文字列を対象コーパスからすべて切り出す．そして，両単語に共通して現れる前後の文字列を抽出する．抽出した共通の文字列で最大の文字列の長さを関連語対の尺度とする．

3　実験結果

3.1　実験環境

実験として，毎日新聞の 97 年度 [5] の記事を先頭から 10,000 件を抜き出し，當間らのシステムに入力した．そして，その結果の関連語対の上位 10,000

表 1　実行結果の上位 10 件

ランク	単語対	
1	説明	指摘
2	議論	論議
3	下回	上回
4	釈放	解放
5	厚生省事務次官	厚生事務次官
6	対話	協議
7	首相	会長
8	指摘	強調
9	事故	事件
10	理由	原因

表 2　當間らの関連語対の尺度を用いない結果

ランク	単語対	
1	説明	指摘
2	ホテル	都内
3	ＮＡＴＯ	欧州
4	難航	今後
5	釈放	解放
6	被害	国民
7	前日	今年
8	ニューヨーク	ホテル
9	最大	事故
10	発表	会見

件に対して提案手法を実施した．

3.2　結果

実験から得られた関連語対の上位 10 件を表 1 に示す．(説明, 指摘) や (釈放, 解放) など類似語が多く見られる．また少数だが，(厚生省事務次官, 厚生事務次官) といった誤字・脱字や，(下回, 上回) といった反義語も現れている．

4　提案手法の特徴

この提案手法は，言い換えである可能性の高いただひとつの証拠があるかないかを判定することであるともいえる．つまり，2 つの単語の前後 Y 文字が同じである確率は $(\frac{1}{\text{文字の種類数}})^Y$ 以上であり，それが Z 文字もあるので，これほど長い文字列は異常であると判断したことになる．

そして，前後の共通文字列のうち最大長だけで評価するので，長い文字列が 1 つだけでも存在すると高い評価になる．しかし，出現頻度 1 回の情報だけで判断することは得策ではない．なぜなら，偶然が起こった場合でも正しいと判断されるため，結果に妥当性がなくなるからである．

この欠点を補うために，今回は當間らのシステムの結果を用いた．ここで，4 番目の処理である関連語対のランキングをまったく行わないで（つまり，候補対に対して提案手法を適用して）ランキングした場合の実験結果を表 2 に示す．(ホテル，都内), (難航，今後) など，誤った結果が多く現れ，本システムの動作は不良になる．

5　係り受け解析の利用の可能性

一般的な関連語対の抽出には係り受け解析で生成された構文木を用いる [6, 7] が，本システムはこれを用いていない．木の情報を使う方がノイズが少ないと考えられるため，本システムの応用として大きな構文木の一致度を判定する方法も有用であることが示唆される．ただ，構文木を用いることは言語リソースが必要なので，それなしでもテストができることは利点とも考えられ，その得失を今後，評価したい．

6　非常に長い文字列が存在する場合

新聞記事に依存する現象だと考えられるが，関連語対の前後に非常に長く共通した文字列が存在する場合がある．実行結果（表 1）の 1 位である (説明, 指摘) に実際に観測された例を図 3 に示す．$\langle \mathbf{X} \rangle$ の部分に”説明”と”指摘”が位置する．図 3 の記事の掲載日は，”説明”が 97 年 1 月 27 日，”指摘”が 97 年 2 月 4 日である．これほど長い文字列の一致は偶然で起こることはまずなく，”説明”と書かれていた部分がなんらかの理由によって，”指摘”に置き換え

同営業本部は「電子メールの利用だけで本当にパソコンネットワークを導入した効果を出せるか」との疑問を動機に，一昨年から本格的なイントラネットの構築に着手．昨年４月にはシステムの基礎となる社員４００人全員のホームページを社内ネット上に開設する「マイホーム作戦」を完了した．構築にあたっては，マウスのクリックだけで簡単に情報が引き出せるように工夫．潮田邦夫営業本部長は「団塊の世代には機械音痴が多いが，仕事の主力選手の彼らが操作に煩わされていては，メリットは半減する」と導入のポイントを ⟨**X**⟩ する．

図 3　(説明, 指摘) の前後に共通してある文 (243 文字)

られたと考えるのが妥当である．

この例のように文の一部が変化した場合，その部分に入る語は同じように使われる語である可能性が高い．また，これほどの長さの文であると，置き換えることができる語は限られていると考える．これらのことは，前後に共通して存在する文字列が一定以上の長さを持ったとき，変化した部分に入る単語はより関連が高いことを示している．新聞記事のような文章においては，過去の記事を利用し，変形して新しい記事を作成することは自然である．それならば，編集の過程の情報の痕跡を利用するという着想は，十分に合理的なものと考えられる．また，編集の過程の情報が実際上は利用できない状況であるので，長い共通する文字列を検出するという手続きは実際上意味のある手続きとなる．編集作業によっては長い文字列が残らないこともあるし，長い文字列が必ずしも編集の履歴を捉えているということは言えないが，情報源として有効に利用できることは示唆される．

一般の文章において，新聞記事と同様に長い文字での共通部分が観測されるかどうかは興味深い問題であり，常に，このように長い文字列のケースがあるとは限らないと思われる．ただし，そのようなケースを検出できたとしたら，それは 2 つの単語が編集の課程で変化しうるという意味での関連性について，強い情報になっていることが示唆されるため，そのような共通の文字列を検出しようとする作業は無駄にはならないことが示唆される．

7　まとめ

本報告では関連語対の評価尺度に関連語の前後に共通する長い文字列を利用する手法を提案した．共通する長い文字列は文章の一部の単語を置き換えたことによって発生し，その部分に入る単語が同じ意味である可能性が高いことを利用した手法である．

今後は提案手法の効果を定量的に計測し，また新聞記事以外の文章の適用可能性を検討したい．

謝辞

この研究は，住友電工情報システム株式会社との共同研究の成果である．また，この成果を分析するときに使用したシステムには，平成 19 年度科学研究費課題（課題番号 19500120）の研究成果を使用した．

参考文献

[1] Christiane Fellbaum. *WordNet: an electronic lexical database*. MIT Press, 1998.

[2] 池原悟ほか. 日本語語彙大系. 岩波書店, 1997.

[3] 當間雅, 折原幸治, 塩入寛之, 梅村恭司. 関連語対のマニインングのための評価尺度. 言語処理学会第 13 回年次大会予稿集, pp. 526–529(B3–7), 2007.

[4] 武田善行, 梅村恭司. キーワード抽出を実現する文書頻度分析. 軽量国語学第 23 巻第 2 号, 2001.

[5] 毎日新聞社. 毎日新聞コーパス. 97 年.

[6] 萩原正人, 小川泰弘, 外山勝彦. perplexity を用いた類義語獲得の自動評価. 言語処理学会第 12 回年次大会予稿集, pp. 767–770(B4–4), 2006.

[7] Masato Hagiwara, Yasuhiro Ogawa, and Katsuhiko Toyama. PLSI Utilization for Automatic Thesaurus Construction. *The Second International Joint Conference on Natural Language Processing(IJCNLP-05)*, pp. 334–345, 2005.