社団法人 情報処理学会 研究報告
IPSJ SIG Technical Report

2008－NL－188 (15)
2008／11／27

二語の共通周辺文字列の長さに着目した語文脈類似判定

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

概要

データマイニングでは，コーパス中の文脈からある二語の関係の有無を判定する問題がある．本報告では，この問題に対し対象の二語の前後に共通した文字列の長さに着目する方法を提案する．この問題では，対象の二語の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よく行われるが，本手法ではコーパス中から集めた文字列の長さが上位 n 件までの文字列のみを用いて判定を行う．実験の結果，評価が高い上位 100 件の単語対を手動で正誤判定したところ，89 件の正解を得た．

A Context Similarity for Two Words Based on The Length of Common Surrounding Strings

Koji ORIHARA and Kyoji UMEMURA
Department of Information and Computer Sciences, Toyohashi University of Technology
orihara@ss.ics.tut.ac.jp, umemura@tutics.tut.ac.jp

Abstract

This paper discusses a method to judge whether given two words have some relation by analyzing the contexts of these words. This method uses the length of common surrounding strings. Though other methods use statistics of neighbor words or modifiers, the proposed method uses strings. More precisely, the most N longest common strings are specified, and the only the length is used to judge whether the two words are used in similar contexts. This method is able to detect 89 meaningful word pairs out of 100 of the highest scored pairs.

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．文章を書くときには同じ意味で文章の内容に適した用語を探すこと，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無数にある意味によって分類しなくてはならず，非常に手間がかかる．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．そこで，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

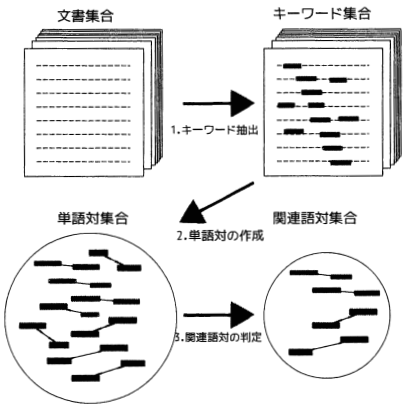

図 1　システムのデータの流れ

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．

このシステムを拡張したものに折原らのシステム [4] がある．このシステムは最長の単語対の周辺文字列のみに着目し，単語対の評価を行い効果を示した．しかしながら，最長の文字列のみを用いて評価を行うことは，偶然に発生した現象の影響を受けやすく，性能にばらつきが発生することが観測された．

そこで，最長の文字列のみを用いるという折原らのシステムの欠点を克服するために，一番長い文字列だけでなく，上位 n 件までの共通する周辺文字列を使うことで性能改善を試みた．その結果，既存のシステムでは現れなかった単語対が得られるようになった．

2　原理

2.1　システムの概要

システムのデータの遷移を図 1 に示す．このシステムは，入力に文書集合をとり，出力から関連語集合を得る．入力から出力を得るまでのシステムの一連の処理を次に示す．

1. 文書集合からキーワード集合を抽出

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 2　キーワード抽出例

2. キーワード集合から単語対集合を作成
3. 単語対集合から関連語対集合を選出

手順 1,2 は當間らが開発したシステム [3] の一部を用いる．以降で各処理の詳細について述べる．

2.2　當間らのシステム

2.2.1　キーワードの抽出

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，武田らのキーワード抽出アルゴリズム [5] 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のみからキーワードを判定する．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 2 に示す．

2.2.2　単語対の作成

単語対の作成では，まず文章からキーワード以外の文字列を除去したときに隣接する傾向が高いキーワードの対を作成する．このキーワード対を順序対と呼び，隣接する傾向が高いかどうかの判定には χ^2 検定を用いる． 帰無仮説は，キーワード対を (s,t) としたとき，「キーワード s と t は独立」である．この検定をすべての隣接するキーワード対に対して行い，順序対の集合を得る．

次に，最終的なシステムの出力である関連語対の候補となる単語対を作成する．図 3 のような関係になる単語対である．順序対集合から 4 つの順序対を取り出し，図 3 のようにひし型を構成する順序対を選び出す．図 3 では，[原稿, 印刷],[原稿, プリント],[印刷, 郵送],[プリント, 郵送] の 4 組の順序対を

[原稿]を[プリント]して[郵送]する
[原稿]を[印刷]して[郵送]した

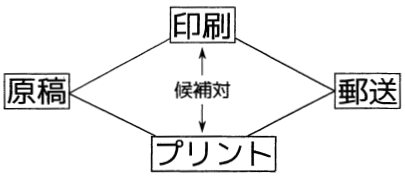

図 3　関連語対の候補となる対

図 4　単語対 (プリント, 印刷) の両側にある文字列の例

用いて，ひし型を構成し，その真ん中にできる (印刷，プリント) を最終的な関連語対の候補となる単語対として選び出す．

2.3　提案方法

単語対 (x, y) が与えられたとき，以下の条件を満たす文字列 α, β を定義する．

「文字列 $\alpha x \beta$ がコーパス中に存在し，かつ文字列 $\alpha y \beta$ もコーパス中に存在する．また α, β ともにこれを満たす中で最大長のものとする」

この周辺文字列の組 $< \alpha, \beta >$ を c で表し，c を単語対 (x, y) の**共通コンテキスト**と呼ぶ．また，α を c の**左コンテキスト**，β を c の**右コンテキスト**と呼ぶ．共通コンテキストの例を図 4 に示す．この図において単語対 (印刷, プリント) に対する共通コンテキストは (原稿を, して郵送) である．共通コンテキストの条件にある最大長とは，この例では (を, して) 等の文字列の対が共通コンテキストではないことを意味する．

共通コンテキスト c に関する式 (1)，(2)，(3) を定義する．これらは，それぞれ c の左右のコンテキストの長さに対して，算術平均，幾何平均，調和平均をとる式である．

$$\text{算術平均: } \mathrm{AA}(c) = \frac{\mathrm{length}(\alpha) + \mathrm{length}(\beta)}{2} \quad (1)$$

$$\text{幾何平均: } \mathrm{GM}(c) = \sqrt{\mathrm{length}(\alpha)\,\mathrm{length}(\beta)} \quad (2)$$

$$\text{調和平均: } \mathrm{HA}(c) = \frac{\mathrm{length}(\alpha)\,\mathrm{length}(\beta)}{\mathrm{length}(\alpha) + \mathrm{length}(\beta)} \quad (3)$$

$\mathrm{length}(s)$ は文字列 s の長さを求める関数である．$\mathrm{AA}(c)$，$\mathrm{GM}(c)$，$\mathrm{HA}(c)$ を総称して $\phi(c)$ と書くことにする．

共通コンテキストは 1 つの単語対に対し，コーパス中に複数個存在する．単語対 (x, y) に対する全共通コンテキスト c を $C_{(x,y)}$ と表す．$C_{(x,y)}$ に関して，(4) 式を導入する．

$$\mathrm{maxN}(x, y, n) = \underset{c \in C_{(x,y)}}{\mathrm{rank}}\ (n, \phi(c)) \quad (4)$$

$\mathrm{rank}_{c \in C_{(x,y)}}(n, \phi(c))$ は集合 $C_{(x,y)}$ のすべての要素に ϕ 関数を適用し，その中から n 番目に大きな値を返す関数とする．ただし $|C_{(x,y)}| < n$ なら，$\mathrm{rank}_{c \in C_{(x,y)}}(|C_{(x,y)}|, \phi(c))$ を返す．この (4) 式を単語対の関連度とする．

本報告の先行研究である [4] は，n が 1 で ϕ 関数に算術平均を用いた (4) 式で単語対を評価したときと同等の結果が得られる．本報告は，[4] を特別な場合に含む方法であり，max 関数と ϕ 関数を一般化した関数を関連度にする方法を提案するものである．

3　実験

実験には毎日新聞の 97 年度 [6] を用いた．先頭から 30,000 記事をシステムに入力し，関連語対の候補となる単語対を 198,642 件得た．得られたすべての単語対に対して，次の操作を n と ϕ 関数のすべての組み合わせに大して実施する．ここで，n の範囲は 1 から 9 であり，ϕ 関数は算術平均，幾何平均，調和平均のいずれかである．

1. 式 (4) を用いて単語対の関連度を算出
2. 関連度の高い単語対の上位 100 件に対して，正解判定を実施

正解判定は著者の主観で行った．コーパスや関連度を算出する関数などを考慮せずに 2 つの単語に関連があるかどうかを判断した．関連があるなら 1 点，

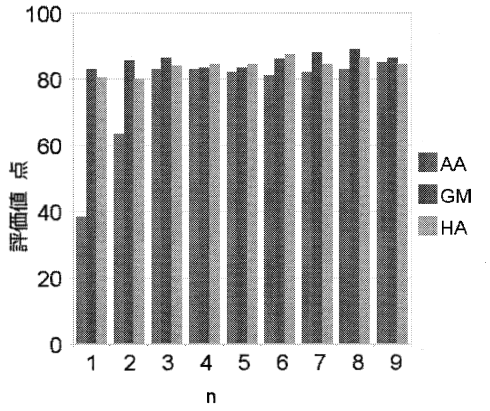

図 5　n 番目の平均値を評価値とした場合の結果

不正解なら 0 点，どちらとも言えない場合 (例えば，女性と子供，最近と当時) は 0.5 点とした．(4) 式で算出した関連度が高い単語対の上位 100 件に対し採点し，その合計を評価値とする．また，(4) 式で用いる単語対の共通コンテキスト集合は毎日新聞の 97 年度全体から収集した．

4　結果

実験の結果を図 5 に示す．横軸は rank 関数に与える順位 n，縦軸は評価値，棒グラフは n と ϕ 関数の組み合わせをそれぞれ示す．図 5 から以下の点が分かる．

- 実験を行った範囲で最大の評価値は，n が 8 で幾何平均を用いた場合である．このとき正解と判定した対が 88 組，不正解と判定した対が 10 組，判定不能とした対が 2 組で評価値は 89 点である．この結果は，本報告の先行研究の [4] の方法である n が 1 で算術平均を用いた場合の結果よりも優れている．
- n が 9 のときを除き，常に算術平均の評価値よりも幾何平均および調和平均の評価値が高い．
- n を 1 から 9 の範囲で変化させても，幾何平均と調和平均の評価値はほとんど変わらない．算術平均の評価値は n が 1 から 3 までは増加し，以降は他の平均方法と同じく 80 点代に位置し，ほとんど変わらない．

⟨**X**⟩ に盛り込まれた周辺有事の対米協力の項目で，憲法との関係で問題点があると思われるものを質問（複数回答）... 以下，略 (全 139 文字)
(⟨**X**⟩ に報告または具体例が入る)

図 6　(報告, 具体的) の共有コンテキストの実例

5　考察

5.1　平均方法の比較

実験結果から算術平均よりも他の平均が優れているという結果を得た．この理由は，図 6 に示すような左右どちらか一方に非常に長い共有コンテキストを持つ単語対が算術平均の結果の上位に現れるためである．図 6 は，コーパス中に"中間報告に盛り込まれた"で始まる文と"中間報告に具体的に盛り込まれた"で始まる 2 つの文が存在することによって得られる共通コンテキストである．このような文節に形容詞節を挿入する場合が日本語ではよくあるので，片側のみの共通コンテキストは単語対を関連付ける情報には役立たないと考える．幾何平均および調和平均は片側のみの共通コンテキストを無視し，両側が非常に長い共通コンテキストのみを関連度の算出に用いるため算術平均よりも高い評価を得たのだと推察する推察する．

5.2　n が増加した場合の単語対の分析

rank 関数は，極めて出現が少ない非常に長い共通コンテキストの影響を大きくする目的で定義した．目的に沿へば，n が 1 に近い場合は rank 関数が非常に長い共通コンテキストを選択して関連度が大きくなり，n が大きくになるにつれて rank 関数が非常に長い共通コンテキストを選択しなくなり，関連度が小さくなることを予想する．しかし，図 5 の結果は，n が 9 でも，算術平均，幾何平均，調和平均すべての評価値が大きいままで，予想とは異なる．もし 1 つの単語対に 9 種類の非常に長い共通コンテキストが存在するならば，n が 9 でもその単語対は大きな関連度を得るが，非常に長い共通コンテキストはほとんど出現しないのでこのようなことは

表1　nが1と9の上位100件の重複

	正解	不正解	判別不能	総計
両方	23	1	0	24
1のみ	55	11	10	76
9のみ	62	11	3	76

表2　nが1または9の場合の関連語対の一部

1のみ出現	9のみ出現
(調査, 検査)	**(巨人, ロッテ)**
(会社, メーカー)	(清水宏保, 島崎京子)
(逮捕, 拘束)	(関東, 東北)
(製造, 製作)	(ワシントン, ロンドン)

起こらない．したがってnが大きい場合，非常に長い共通コンテキスト以外の共通コンテキストが評価に良い影響を与えていると考える．

原因の共通コンテキストを特定するために，nのみが変化した場合の上位100件の単語対の変化を分析した．nが1と9で幾何平均を用いたときを調べた．その結果を表1に示す．表1から，選ばれた単語対が7割程度異なることが分かる．

次に，nが1または9のどちらか一方の上位100件に現れた単語対の実例を表2に示す．また，太字で示した (調査, 検査) と (巨人, ロッテ) について，実際に出現した共通コンテキストを図7と図8にそれぞれ示す．各行は左から順に左コンテキストの長さ，右コンテキストの長さ，共通コンテキストから成る．幾何平均の値が大きい順に上位9件までを示す．太字で記された共通コンテキストはrank関数が選択し関連度に用いたコンテキストである．また，共通コンテキスト内にある記号Sは空白文字が2文字以上続く文字列を置き換えたものである．

表2を見ると，nが1と9では，得られる関連語対の特徴が異なることが分かる．nが1の場合は同義語に分類される単語対を選び，nが9の場合は固有名詞の単語対を多く選ぶ．

選択する単語対の違いは，実際にrank関数が選択した共通コンテキストの違いによる．図7の (調査, 検査) の場合，1番目に両辺が非常に長い共通コンテキストが存在するが，2番目は左コンテキストの長さが1と急激に短くなる．この単語対は，nが1のときは非常に高い関連度を得るものの，nが大きくなるにつれて極端に関連度を下げる．この単語対は先行研究 [4] に記された稀に出現する共通コンテキストを重視することで得られる単語対である．

同義語に分類される単語対が選ばれた理由として，非常に長い共通コンテキストは文全体の意味が

34 62 [S昨年１１月以降の平均単価で購入，修理費を計算し比較すると，立ち入り ⟨X⟩ 前の９４年度は約１８億６０００万円，９５年度が計約１６億円，今年度は約７億９０００万円も高く落札されていたことになるという．]
1 37 [り ⟨X⟩ を受け，今年度平均は４７０９円に，昨年１１月以降だと３８１１円まで落ちた．]
1 15 [り ⟨X⟩ を実施することを明らかにした．]
4 3 [抜き打ち ⟨X⟩ を実施]
2 5 [ング ⟨X⟩ を実施する]
2 5 [ング ⟨X⟩ を実施して]
1 8 [の ⟨X⟩ を行っているが，]
1 7 [S⟨X⟩ 結果によると，]
1 7 [の ⟨X⟩ 結果によると，]

図7　(調査, 検査) の共通コンテキストの上位9件

19 1 [［キャンプだより］プロ野球＜１４日＞S⟨X⟩S]
19 1 [［キャンプだより］プロ野球＜２２日＞S⟨X⟩S]
9 2 [［焦点］プロ野球S⟨X⟩ ９－]
9 2 [［焦点］プロ野球S⟨X⟩ ５－]
17 1 [［ドラフト情報］プロ野球＜１日＞S⟨X⟩，]
17 1 [［キャンプだより］プロ野球S ３日S⟨X⟩S]
17 1 [［１２球団キャンプルポ］プロ野球S⟨X⟩S]
17 1 [［キャンプだより］プロ野球S ２日S⟨X⟩S]
5 3 [プロ野球S⟨X⟩ ２－１]

図8　(ロッテ, 巨人) の共通コンテキストの上位9件

大きく変わるような単語対を選ばない可能性が高いことを考える．図7の1番目にある共通コンテキストを見ると分かるように，非常に長い共通コンテキスト内にある $\langle \mathbf{X} \rangle$ に挿入できる単語はほとんどない．このことは非常に長い共有コンテキストを含む文の意味がほぼ確定していることを意味し，その結果，意味が同義語を選出する可能性が高くなる．

一方，n が9の場合の一例である (ロッテ, 巨人) の共通コンテキストについて考える．(ロッテ, 巨人) の場合は，(調査, 検査) とは異なり，上位9件の共通コンテキストの左右の長さの積はほとんど変化がない．そのため，このような単語対は n が1から9に変わっても関連度をほとんど変えないが，(調査, 検査) のような単語対が n を大きくするにつれて関連度を著しく下げるので，上位に現れる．

図8には”[キャンプだより]”や”プロ野球”を含む共通コンテキストが存在する．このような共通コンテキストは新聞のスポーツ面等に現れる定型句に含まれる．新聞で使用する定型句は固有名詞を伴うことが多いので，この共通コンテキストが得る単語対は固有名詞を多く獲得する可能性が高い．

6 まとめ

本報告では，折原らの関連語抽出方法 [4] を特別な場合とする一般化した方法を提案した．提案した一般式により，[4] よりも効果が高い引数の組み合わせを発見した．共通コンテキストの長さの幾何平均の集合から8番目に大きい値を関連度とすることが，調査した範囲ではもっとも良い評価をえることが分かった．さらに，一番大きな値よりも n 番目の長さの平均値を関連度としたときに結果が向上する理由を考察した．

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] 折原幸治, 藤原大輔, 梅村恭司. 前後に出現する長い共通文字列を用いる関連語判定法. 自然言語処理学会第14次年次大会, 2008.

[5] 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.

[6] 毎日新聞社. 毎日新聞コーパス. 97年.