関連語対のマイニングのための評価尺度

當間 雅† 折原 幸治‡ 塩入 寛之‡ 梅村 恭司‡
{miyabi, orihara, shio}@ss.ics.tut.ac.jp, umemura@tutics.tut.ac.jp
†豊橋技術科学大学 情報工学系

概要

本報告において，「関連語対」とは，ある意味で同じような文脈で使われている語の対のことを指し，語と接続する文字やキーワードの情報を使って統計的に選出できる．本報告において，この関連語対を文書集合から獲得するためのマイニングとは，考えられうる全ての単語対の中から，関連しているという見込みのある語の対を統計的に選び出すことである．これを行う方法のひとつとして，文書集合から辞書を一切使用せずに単語を切り出し，かつ文脈情報を特定する処理について述べる．また，この処理で使用できる判定尺度を複数示す．さらに，関連語判定尺度を変更して，精度の変化を測定した結果を報告する．

1. はじめに

類義語などの知識獲得に関する研究は，シソーラス構築や情報検索などへの幅広い応用が考えられ，自然言語処理において，基本的かつ重要な課題として研究されてきた．関連語対のマイニングは，文書集合から考えられ得る全ての単語対の中から，関連しているという見込みのある対を統計的に選び出すという処理であり，知識獲得に関する研究のひとつである．本報告では，文書集合から関連していると考えられる単語対を自動的に獲得するためのマイニング処理について述べる．

本報告では，文章中で同じように扱われている単語対は，関連している単語対であると考えた．そこで，単語に接続する語の情報を用いて，辞書を一切使用せずに，統計的に関連語対の獲得を行う処理を考えた．本報告では，英語のコーパスにおいて，関連語対をマイニングするシステムについて述べる．また，このシステムで使用する判定尺度について 3 つの尺度を考えた．それぞれの尺度でシステムを実行し結果を比較し，判定尺度の変更による精度の変化について報告する．

2. 関連語対のマイニングシステム

本報告のシステムでは，大きく分けて以下の 4 工程で関連語対をマイニングする．

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）関連語の判定

それぞれのステップについて順に説明する．

2.1. 単語の切り出しと統計情報の取得

第 1 工程ではまず，単語の切り出しを行う．本報告では対象言語を英語としているので，単語はスペース区切りによって切り出している．この際，”+”や ”.” などの記号は空白として扱っている．

さらに，切り出した単語集合から，関連語として選ばれる対象となるキーワード集合を作る．これは前置詞や代名詞，関係詞などの単語を取り除いたものとする．前置詞や代名詞，関係詞などの単語は，それ単体ではあまり意味を成さないからである．しかし，これらの単語は文脈を把握する上では重要な情報となる．この工程では単語の切り出しと共に，上述のキーワードの出現頻度や，前後に出現する単語（キーワード以外の単語も含む）との共起出現頻度などの取得を行っている．図 2－1 はコーパスからキーワードとなる単語だけを抜き出したものである．

> [China] [says] it [can] [manage] [Hong] [Kong's] [transition] without [Britain] [BEIJING], [Jan] 1 ([AFP]) [China] [said] [Monday] it was [prepared] to [handle] [Hong] [Kong]'s [return] to its [fold] [July] 1, 1997 [alone], without the [cooperation] of the [territory]'s [current] [British] [masters], [if] [necessary]. [Lu] [Ping], [director] of the [Hong] [Kong] and [Macau] [affairs] [office] of the [Chinese] [state] [council], [told] [Xinhua] [news] [agency] in an [interview] that the [principle] of "[taking] ourselves as the [mainstay]" [should] be [adhered] to in [dealing] with the [issue] of [Hong] [Kong]. [Even] [if] the [British] [side] [remains] [uncooperative], he [said], "we are [confident] about a [smooth] [transition] of [Hong] [Kong]."
>
> [ ]内はキーワード

図 2-1 キーワードの抜き出し

2.2. 順序対の抽出

第 2 工程では，キーワードを出現する順序に並べたとき，続けて現れる傾向の高いキーワードの対を抽出する．このようなキーワード対を順序対と呼ぶことにする．順序対の抽出は，χ^2 検定によって行う．

s と t を異なるキーワードとし，帰無仮説として「キーワード s と t は関連がない（独立である）」を考える．続いて現れる 2 つのキーワードを考えたとき，事象 A を，前者が s である場合，$\overline{A}$ を前者が s 以外のキーワードである場合，事象 B を後者が t である場合，$\overline{B}$ を後者が t 以外のキーワードである場合であるとする．このとき，表 2-1 にように、それぞれの場合の出現回数を計数し，以下のような確率変数を定義する．

$$X = \sum_{i=1}^{2}\sum_{j=1}^{2}\frac{(x_{ij} - a_i b_j / N)^2}{a_i b_j / N} \qquad (2-1)$$

この確率変数 X は，自由度 1 の χ^2 分布に従うことがわ

かっている．したがって，X の値がある閾値より大きい場合，帰無仮説は棄却され，s と t は順序対として抽出される．ただし，式（2－1）のままでは頻度の小さいものの推定ができないか，不安定な値となってしまう．そこで，x_{ij} の値にあらかじめ 1 を加え，次のように期待度数に補正を行った．

$$\tilde{x}_{ij} = \frac{a_i + 2}{N + 4} \cdot \frac{b_j + 2}{N + 4} \cdot N \qquad (4-2)$$

表 2－1　キーワード対の出現回数

	B	$\overline{B}$	**計**
A	x_{11}	x_{12}	a_1
$\overline{A}$	x_{21}	x_{22}	a_2
計	b_1	b_2	N

2.3. 候補対の選出

第 3 工程では，キーワード集合から，関連語対の候補となる単語対を選出する．この工程では，前後に同じキーワードが出現しているキーワード対を，「同じように使用されている単語対」とみなし，候補対を選出している（図 2－2 参照）．

x，c をキーワードとし，x のあとに c が続く順序対を xc と表現し，順序対の集合を $Combined$ とする．さらに，x について，その直前・直後に現れるキーワードの集合を次のように定義する．

$$pre(x) = \{c \mid cx \in Combined\}$$
$$post(x) = \{c \mid xc \in Combined\}$$
$$Fpre(x) = \{c' \mid c' \in post(c), c \in pre(x)\}$$
$$Bpost(x) = \{c' \mid c' \in pre(c), c \in post(x)\}$$

このとき，次のように候補対の集合を得ることができる．

$$Candidates = \{(a,b) \mid b \in Fpre(a) \wedge Bpost(a)\}$$

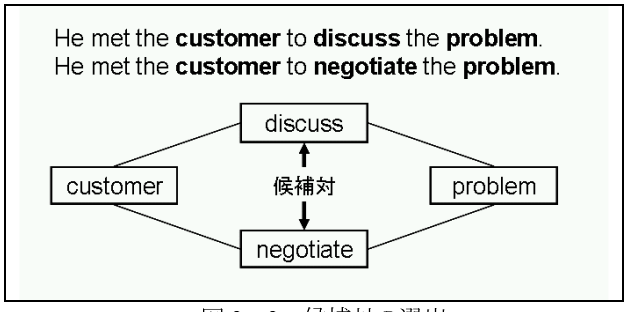

図 2－2　候補対の選出

2.4. 関連語の判定

最後の第 4 工程では，候補対となった単語対について，ある尺度をもって関連語であるかどうかを判定し，最終的なシステムの出力である関連語対の集合を求める．関連語判定には，キーワードの前後に出現する単語の情報を用いる．候補対選出のときとは違い，キーワードではない単語の情報も用いる．

まず，候補対 s - t について，前後に出現する単語を考え，s の前(後)にも t の前(後)にも出現する単語の種類数を a_{pre} (a_{post})，s の前(後)には現れるが t の前(後)には現れない単語の種類数を b_{pre} (b_{post})，t の前(後)には現れるが s の前(後)には現れない単語の種類数を c_{pre} (c_{post}) とする．つまり，前後単語の共通数や異なり数を使って関連語判定を行う．これらを使った尺度はさまざまなものが考えられ，有用なものは多数ある．ここでは以下の 3 式を検討した．

$$score_1 = a_{pre} + a_{post}$$

$$score_2 = \sum_{POS} \frac{2a_{POS}}{\sqrt{(a_{POS} + b_{POS}) + (a_{POS} + c_{POS})}}$$

$$score_3 = \log(\min(cf(s), cf(t)))score_{\cos}$$

ただし，$cf(x)$ を単語 x のコーパス全体での出現回数であるとする．スコア 1 では，単語 s と t の前後に出現した単語のうち，共通のものの数になっている．スコア 2 では，代表的な類似尺度のコサイン尺度を使って，s と t の前後単語の出現類似度をそれぞれ計算し加算している．スコア 3 は，スコア 2 に s と t の出現頻度で重みを与えたものである．スコア 2 とスコア 3 のでは，スコア 1 と違って，共通する前後単語だけでなく，異なっている前後単語の情報も使っている．

3. 判定尺度の比較実験

実際に関連語対の抽出を行ない，結果を比較する実験を行った．実験では英字新聞コーパスである English Gigaword Corpora[8]を使用した．1994～1997，2001，2002 年度のデータからそれぞれ 1000 記事を抜き出し，各関連語判定尺度を適用したシステムに入力し実行した．評価方法は，萩原らの方法[3]と同じく，既存のシソーラスである WordNet を用いて正解となる基準類似度を計算し，この正解判定をもってシステムの適合率を求めた．

WordNet は語の関係を木構造により具現化している(図 3-1)．WordNet 中の語義 w_i に対応する節点の深さを d_i，v_j に対応する節点の深さを d_j とする．また，これら 2 接点の共通祖先の深さの最大値を d_{dca} とする．このとき，基準類似度を以下の式で求める．

$$sim(w_i, v_j) = \frac{2 \cdot d_{dca}}{d_i + d_j}$$

本実験では，$sim(w_i, v_j) \geq 0.6$ となった単語対を正解と判定した．また，WordNet の辞書に含まれない単語については，評価対象外とした．

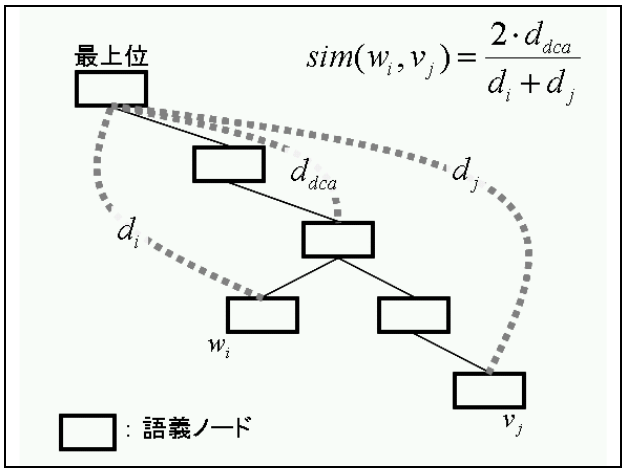

図 3-1　WordNet による基準類似度

4. 実験結果と考察

表 5-1～3 は各尺度での出力結果の例である．正解と評価されたものは○，不正解には×，WordNet 辞書に含まれておらず，評価対象外となったものは？という評価を付与している．

スコア 1 の結果では，同じ動詞や助動詞の変化形の対が多く見られた．しかし，said - Friday のように，品詞の異なる誤った対も含まれていた．また，この尺度は単純な共通部分の加算であり，出現頻度の高い単語対に大きな値を与える傾向がある．

スコア 2 の結果では，人名や地名などの対が非常に多く見られた．この尺度では，コーパス中での出現頻度が極端に少ない固有名詞を含む対に，大きなスコアを与える傾向がある．また，表記ゆれやスペルミスの対も含まれていた．

スコア 3 の結果では，月日や単位など，似た概念の単語対や negotiate - discuss などのような言い換え可能な語や同義語対，また反意語や表記ゆれやスペルミスも見られた．

表 5-4 は，各評価尺度での適合率の表である．()内は実際の正解数である．全体的に，スコア 1 よりスコア 2 やスコア 3 の方が適合率が高くなった．しかし，スコア 2 では得られる関連語対の数が他の 2 つに比べて極端に少ないという問題がある．これは，コーパス中での出現頻度が極端に少ない固有名詞を含む対に，大きなスコアを与える傾向があるため，WordNet のシソーラス辞書に登録されていない単語が多かったためであると考えられる．スコア 3 では，単語の出現頻度を考慮に入れているので，このようなことはなかった．

これらの結果から，関連語の判定尺度の要素として，前後単語の共通数だけでなく，異なり数も加えることが効果的であることがわかった．さらに，キーワードの出現頻度による重みを加えることで，頻度の極端に少ないキーワードによる雑音を取り除くことができたといえる．

表 5-1　スコア 1 の結果

No.	単語 1	単語 2	評価
1	has	had	○
2	would	will	○
3	said	told	○
4	Monday	said	×
5	people	government	×
6	four	two	○
7	announced	said	○
8	police	said	×
9	Wednesday	here	×
10	people	troops	×
11	party	Government	○
12	also	now	×

表 5-2　スコア 2 の結果

No.	単語 1	単語 2	評価
1	Phoenix	Tucson	○
2	WELLINGTON	JOHANNESBURG	○
3	Chart	Map	○
4	neighbour	neighbor	○
5	Hart	Houghton	○
6	necessarily	neccessarily	?
7	Exchequer	Exchecquer	?
8	Zinc	Nickel	○
9	dong	peso	○
10	pound	peso	○
11	unwise	Willing	×
12	dissatisfied	demented	×

表 5-3　スコア 3 の結果

No.	単語 1	単語 2	評価
1	year	month	○
2	September	December	○
3	third	second	○
4	visit	move	○
5	give	provide	○
6	Iran	Israel	○
7	unable	able	×
8	failing	trying	×
9	build	Make	○
10	authorities	government	○
11	understand	do	×
12	win	victory	○

表 5-4　適合率と正解数

尺度 / 年度	スコア 1	スコア 2	スコア 3
1994	0.21(270)	0.35(21)	0.42(428)
1995	0.21(266)	0.32(19)	0.37(408)
1996	0.22(231)	0.46(29)	0.40(441)
1997	0.25(226)	0.32(26)	0.42(393)
2001	0.21(235)	0.37(27)	0.37(445)
2002	0.18(266)	0.37(20)	0.34(520)

5. 他の研究との比較

シソーラス構築に関する萩原らの研究[3,4]では係り受け解析を行っているが，本システムでは行なっていない．その点において精度比較をすべきであるが，システムを改良中なのでまだ行っていない．また，日本語を対象にしてこのマイニング処理を行った研究[2]もある．これは山本らのシステム[1]を改良したものであるが，本報告で示したように，関連語判定尺度に a_{pre} (a_{post})，b_{pre} (b_{post})，を c_{pre} (c_{post})を使用することを提案したものではない．これらの使用は本報告の独自性のあるところである．

6. まとめ

本報告では，文章中で同じように扱われている単語対は，関連している単語対であると考え，前後に出現する単語の情報を用いて，辞書を一切使用せずに，統計的に関連語対の獲得を行うマイニングシステムについて述べた．また，このシステムで使用する判定尺度について 3 つの尺度を考えた．スコア 1 は前後単語の共通単語数を加算する単純な尺度，スコア 2 はコサイン尺度を試用して前後単語の類似度を加算した尺度，スコア 3 はスコア 2 に判定キーワードの出現回数を重みに加えたものである．これらを検証するため，各尺度を適用したシステムを用いて，英語の文書集合に対して関連語対を抽出し，結果を比較する実験を行った．その結果，関連語の判定尺度として，前後単語の異なり数やキーワードの出現頻度による重みが効果的であることを報告した．

7. 謝辞

本研究は，株式会社住友電工情報システムとの共同研究の成果である．また，本研究は文部科学省 21 世紀 COE プログラム「インテリジェント　ヒューマンセンシング」の援助により行われた．さらに，名古屋大学情報科学研究科外山グループの皆さんとのディスカッションは，今後の研究に大変参考になりました．

8. 参考文献

[1] Eiko Yamamoto and Kyoji Umemura. Related Word-pairs Extraction without Dictionaries, LREC-2004 pp.1309-1312, 2004

[2] 當間 雅，梅村 恭司．語の接続情報によるシソーラス自動構築システムの実装と評価．第 48 回プログラミング・シンポジウム報告書，2007

[3] 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，B4-4 (pp. 767-770), 2006

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

[5] Christopher D, Manning and Hinrich Schutze. Foundations of Statistical Natural Language Processing. The MIT Press, 1999

[6] 薩摩 順吉．確率・統計．岩波書店，1989

[7] 林 周二．基礎過程 統計および統計学．東京大学出版，1988

[8] LDC(2004). “http://www.ldc.upenn.edu/.”