語の出現文脈の一致判定における 文脈出現頻度と異なり数の比較

當間 雅[†]　梅村 恭司[†]

miyabi@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

同じ文脈で出現する傾向のある語は意味上の関連があると考え，２つの単語について前後に現れる内容語を出現文脈として実験を行った．このとき，出現文脈を統計処理に使用するための統計量としては，頻度と異なり数の二つが候補になる．モデルとして赤池の情報量基準を使用して実験を行ったところ，異なり数で処理すると２つの単語についての意味の類似が観測できたことを報告する．これは頻度では観測できなかった．統計処理においては頻度を使うことが多いので，興味深い事例と考え報告する．

1. はじめに

ある語に関連する語郡を獲得するという課題は言語処理分野の基礎的な研究課題であり，シソーラスや類義語辞書など多くの研究例が報告されてきた．これらのシステムは広く情報システムへの活用が期待されている．しかし対象とする語が辞書に未登録である場合，その関連語を求めることは一般に難しいと言われている．

そのような未知語に対応した関連語の抽出システムのひとつに，山本らの提案するシソーラス構築システムがある[1]．このシステムはメリットとして辞書を一切使用せずにシソーラス構築の全工程を行なうことが出来る一方，入力した文書数に対して得られる関連語対の数が少なすぎるという問題点がある．この原因としては，システムに採用されている判定尺度が理論的な背景のないアドホックな関数であるため扱いが難しく，結果を絞り込みすぎているのではないかということが考えられる．

この問題を解決する方法として，本研究では統計モデルを用いたモデル選択手法による判定尺度を用いる．本報告では特に関連語判定の尺度について言及し，単語の頻度に着目した一般的な方法に対して，異なり数に着目した方法を提案する．単語の異なり数とは，その単語の頻度に関係なく１度でも出現したらその単語のカウントとして１を与えるということである．本報告では，関連語の判定材料として頻度ではなく異なり数を用いることの有効性を報告する．

2. 使用するモデル選択手法

本報告では，モデル選択手法として赤池情報量基準(AIC)とベイズ情報量基準(BIC)を用いた．これらは最尤法であてはめられたモデルが複数個あるとき，その中の１つを選択する基準である．詳細については文献[8][9][10]が詳しい．

一般に，最大対数尤度はモデルの自由パラメータ数が多いほど大きな値となる傾向があり，最大対数尤度の比較によってモデルを選択すると，自由パラメータ数の多いものほど選ばれやすくなる．AIC はモデルの良し悪しを評価する基準として平均対数尤度の期待値(期待平均対数尤度)したものである．期待平均対数尤度は,最大対数尤度 $l(\hat{\theta})$ とモデルの自由パラメータ数 k の差により近似的に導かれることがわかっており，歴史的経緯からそれを –2 倍した量

$$AIC = (-2) \times l(\hat{\theta}) + 2 \times k$$

がモデル選択の基準となり，AIC を最小とするモデルが最適なモデルと考えられる．

一方 BIC は,モデルの事後確率に基づく評価基準であり，あるデータ x が観測されたとき,それが i 番目のモデル M_i から生成される確率をベイズの定理より求め,事後確率が最大となるモデルを選択する方法である．BIC は最大対数尤度 $l(\hat{\theta})$,モデルの自由パラメータ数 k ,およびデータの総数 n を用いて次のように求められ，BIC を最小とするモデルが最適なモデルと考えられる．

$$BIC = (-2) \times l(\hat{\theta}) + k \times \log n$$

なお 2 つのモデルを比較する場合，AIC 又は BIC の値の差が 1 以上あれば，その差は有意な差と言える．

3. 関連語対のマイニングシステム

本報告で扱うシステムは次の 3 工程で関連語対のマイニングを行なう．

1) 単語の切り出し
2) 候補対の選出
3) 関連語の判定

以下の節ではそれぞれのステップについて順に説明する．

3.1. 単語の切り出し

第 1 工程では，コーパスから単語の切り出しを行う．これには武田のキーワード抽出アルゴリズム[2]を使用する．このアルゴリズムによって抽出されたキーワードは，以降の手順において単語として扱われる．

また本システムの第 3 工程では，キーワードほど強い意味のない一般的な単語を，文脈上の意味を把握する上で有用な情報となる単語として使用いている．このような単語を文脈単語と呼ぶことにする．この文脈単語は，武田のアルゴリズムにおいて，キーワードらしさのやや弱いものを選んでいる．

3.2. 候補対の選出

第 1 工程で切り出された単語集合から考え得る単語対すべてについて関連関係にあるか調査する場合，計算量が問題となる．そこで第 2 工程では関連語の候補対を求める．関連語の候補対とは，前後に同じ単語が現れる単語のついである．図 3.1の例では，「プリント」と「印刷」が候補

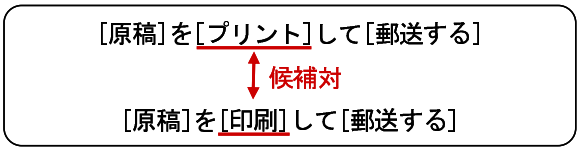

図 3.1 候補対選出の例

対にあたることになる．ここでは効率的に候補対を求めるため，単語を出現する順序に並べたとき，続けて現れる傾向の高い単語の対をあらかじめ求めておく．このような単語対を順序対と呼ぶ．

山本らのシステムでは，連語の共起頻度を用いたアドホックな判定方法をとっているが，ここでは BIC を用いた順序対選出の方法について説明する．単語 a,b を順序対であるか判定される単語対とする．また単語 x の後に単語 y が現れることを単語列 xy と表現すし，文書中に現れた全ての単語列 xy の数，つまり順序対であるかどうか判定される対象となる単語対の数を N とする．さらに，$cf(z)$ を単語列 z の出現頻度とする．2 つの単語の並びを考えたとき，前方に単語 a が出る事象を A_1，前方に単語 a 以外の単語が出る事象を A_2，後方に単語 b が出る事象を B_1，後方に単語 b 以外の単語が出る事象を B_2 とする．このとき，$A_i(i=1,2)$ かつ $B_j(j=1,2)$ にあたる単語列の数を次のように求める．表 3.1はこれを表に表したものである．

$$n_{11}=cf(ab),\qquad n_{12}=\sum_{y\neq b}cf(ay)$$
$$n_{21}=\sum_{x\neq a}cf(xb),\quad n_{22}=\sum_{x\neq a,y\neq b}cf(xy)$$
$$n_{i\cdot}=\sum_{j}n_{ij},\qquad n_{\cdot j}=\sum_{i}n_{ij},\qquad n_{\cdot\cdot}=\sum_{i,j}n_{ij}$$

A_i と B_j の同時確率を $p(A_i,B_j)$ と表すと，表 3.1のような結果が得られる確率の対数尤度は次のようになる．

$$l\left(\{p(A_i,B_j)\}\right)=K_1+\sum_{i,j}n_{ij}\log p(A_i,B_j)$$

ただし，$K_1=\log\left(n_{\cdot\cdot}/\prod_{i,j}n_{ij}\right)$である．ここでは「単語 a と単語 b の出現は独立である」とするモデル M1 と，「単語 a と単語 b の出現は独立でない」とするモデル M2 を考え，どちらのモデルが実際のデータへの当てはまりが良いかを評価する．

まずモデル M1 では，a,b の出現が独立なので次式が成り立つ．

$$p(A_i,B_j)=\theta(A_i,\cdot)\theta(\cdot,B_j)$$

ただし，制約条件として以下が成り立つ．

$$\theta(A_i,\cdot)=\sum_j p(A_i,B_j),\quad \sum_i\theta(A_i,\cdot)=1$$
$$\theta(\cdot,B_j)=\sum_i p(A_i,B_j),\quad \sum_j\theta(\cdot,B_j)=1$$

これらより，最尤推定量 $\hat{\theta}(A_i,\cdot)=n_{i\cdot}/n_{\cdot\cdot}$，$\hat{\theta}(\cdot,B_j)=n_{\cdot j}/n_{\cdot\cdot}$ が得られる．また，モデル 1 のパラメータ数は 4 であるが，制約条件から自由度は 2 となる．したがってモデル M1 の BIC は次のようになる．

表 3.1 単語列の出現頻度

	B_1	B_2	計
A_1	n_{11}	n_{12}	$n_{1\cdot}$
A_2	n_{21}	n_{22}	$n_{2\cdot}$
計	$n_{\cdot 1}$	$n_{\cdot 2}$	$n_{\cdot\cdot}$

$$BIC_{M1}=(-2)\left[K_1+\sum_{i,j}n_{ij}\log\frac{n_{i\cdot}n_{\cdot j}}{n_{\cdot\cdot}}\right]+2\times\log N$$

次にモデル M2 では，a,b の出現が独立ではないので次式が成り立つ．

$$p(A_i,B_j)=\theta(A_i,B_j)$$

ただし，$\sum_{i,j}\theta(A_i,B_j)=1$ という制約条件が成り立つことから，最尤推定量 $\hat{\theta}(A_i,B_j)=n_{ij}/n_{\cdot\cdot}$ が得られる．モデル 2 のパラメータ数は 4 であるが，制約条件から自由度は 3 となるので，モデル M2 の AIC は次のようになる．

$$BIC_{M2}=(-2)\left[K_1+\sum_{i,j}n_{ij}\log\frac{n_{ij}}{n_{\cdot\cdot}}\right]+3\times\log N$$

出現に関連のある単語対を順序対として選択するので，モデル M2 の BIC 値が小さい単語対を選べばよい．したがって次のような順序対集合 *Linked* が得られる．

$$Linked=\{ab\,|\,BIC_{M1}-BIC_{M2}>1\}$$

3.3. 関連語の 2 次判定

最後の工程では，候補対となった単語対について出現類似性の尺度に従って関連語であるかどうかを判定する．山本らの方法は周囲の文字列の情報を用いたアドホックな判定方法を用いていたが，ここでは「周囲の単語が同じような頻度分布をしている」単語対を関連語対として考え，頻度を基にしたモデルを考えた．この方法は AIC によるスタンダードな方法を使った試みであるが，山本らのシステムより精度が落ちてしまうという重大な問題がある．本報告ではこれに対して新たに異なり数によるモデルを考えたが，それについては第 4 章で述べ，ここでは頻度に着目した方法について述べる．

単語を出現する順に並べたとき，単語 x の後に単語 y が現れることを単語列 xy と表現し，$cf(z)$ を単語列 z の出現頻度とする．また，単語 x の直前に現れる単語(キーワードまたは文脈単語)の集合を $BCI(x)$，直後に現れる同様の単語集合を $FCI(x)$ とする．簡単のため前出単語の頻度分布の同一性判定について説明する．候補対 x_1,x_2 が与えられたとき，次のように前出単語の集合を求める．

$$\{w_1,w_2,\cdots,w_m\}=BCI(x_1)\cup FCI(x_2)$$

単語列 $w_ix_j(i=1,2,\cdots,m,\ j=1,2)$ \$について，次のように頻度を定義する．表 3.2はこれを表に示したものである．

$$n_{ij}=cf(w_ix_j),\quad n_{i\cdot}=\sum_j n_{ij}$$
$$n_{\cdot j}=\sum_i n_{ij},\quad n_{\cdot\cdot}=\sum_{i,j}n_{ij}$$

表 3.2 前出単語の頻度分布

	B_1	B_2	計
w_1	n_{11}	n_{12}	$n_{1\cdot}$
⋮	⋮	⋮	⋮
w_m	n_{m1}	n_{m2}	$n_{m\cdot}$
計	$n_{\cdot 1}$	$n_{\cdot 2}$	$n_{\cdot\cdot}$

表 3.3 周囲単語対の異なり数

	B_1	B_2	計
A_1	n_{11}	n_{12}	$n_{1\cdot}$
A_2	n_{21}	n_{22}	$n_{2\cdot}$
計	$n_{\cdot 1}$	$n_{\cdot 2}$	$n_{\cdot\cdot}$

表 5.4 評価基準

正誤	評価	詳細
正	同義語	表記ゆれ、略語、その他の同義語
正	関連が強い対	その他主観により関連があると認めた単語対
誤	関連がわからない対	上記に当てはまらない単語対
誤	非単語対	単語対でない対

ただし，頻度が 0 となる単語列を考慮して，そのような単語の頻度を 0.001 としてスムージングを行なった．単語 x_j が現れたときの前出単語 w_i の出現確率 $P(w_i \mid x_j)$ を用いると，表 3.2のような結果が得られる確率の対数尤度 $l(\{p(w_i \mid x_j)\})$ は次のようになる．

$$l(\{p(w_i \mid x_j)\}) = K_2 + \sum_{i,j} n_{ij} \log p(w_i \mid x_j)$$

ただし，$\sum_{j=1}^{2} p(w_i \mid x_j) = 1$，$K_1 = \log\left(n_{i\cdot}! / \prod_{i,j} n_{ij}\right)$ である．

ここでは「x_1, x_2 の前出単語は同一分布である」とするモデル M1 と，「x_1, x_2 の前出単語は異なる分布である」とするモデル M2 を考える．つまり，次の 2 つのモデルの比較を行うことになる．

モデル M1　$p(w_i \mid x_j) = \theta(w_i)$

モデル M2　$p(w_i \mid x_j) = \theta(w_i \mid x_j)$

モデル M1 の最尤推定量は $\hat{\theta}(w_i) = n_{i\cdot} / n_{\cdot\cdot}$ となり，パラメータ数は m となるが，制約条件より自由度は $m-1$ となる．したがって，モデル M1 の AIC は次のようになる．

$$AIC_{M1} = (-2)\left[K_2 + \sum_i n_{i\cdot} \log \frac{n_{i\cdot}}{n_{\cdot\cdot}} \right] + 2 \times (m-1)$$

同様にモデル M2 の最尤推定量は $\hat{\theta}(w_i \mid x_j) = n_{ij} / n_{\cdot j}$ となり，制約条件より自由度は $2(m-1)$ となる．したがって，モデル M2 の AIC は次のようになる．

$$AIC_{M2} = (-2)\left[K_2 + \sum_{i,j} n_{ij} \log \frac{n_{ij}}{n_{\cdot j}} \right] + 2 \times 2 \times (m-1)$$

以上から，前出単語の頻度分布が同じ単語対の集合を次のように求めることが出来る．

$$RelB = \{ab \mid AIC_{M2} - AIC_{M1} > 1\}$$

後出単語についても同様に集合 *RelF* を考え，最終的に次のように関連語対集合 *Relevants* を定義する．

$$Relevants = \{(a,b) \mid RelB \cap RelF\}$$

4. 異なり数を基にしたモデルによる関連語の 2 次判定

3.3 節で説明した関連語の 2 次判定の方法では，周囲の単語の頻度に着目したが，ここでは単語の異なり数に着目し，候補対として a,b を考えたとき，その周囲の単語について a,b に共通する単語の種類が多いものほど関連が高いと考える．

x を単語とし，その直前に現れたキーワード又は文脈単語を α，直後に現れたキーワード又は文脈単語を β とするとき，ペア (α,β) を x の周囲単語対とする．また，a,b を関連語であるか判定される候補対とする．このとき，入力文書から考えられる全ての単語ついについて，それが単語 a の周囲単語対である(ない)という事象を A_1 (A_2)，単語 b の周囲単語対である(ない)という事象を B_1 (B_2) とし，$A_i (i = 1,2)$ かつ $B_j (j = 1,2)$ にあたる周囲単語対の数を次のように求める．表 3.3はこれを表に示したものである．

$$n_{11} = |A_1 \cap B_1|, \quad n_{12} = |A_1 \cap B_2|$$
$$n_{21} = |A_2 \cap B_1|, \quad n_{12} = |A_2 \cap B_2|$$
$$n_{i\cdot} = \sum_{j=1}^{2} n_{ij}, \quad n_{\cdot j} = \sum_{i=1}^{2} n_{ij}, \quad n_{\cdot\cdot} = \sum_{i=1}^{2}\sum_{j=1}^{2} n_{ij}$$

ここではモデル M1 として「a,b の周囲単語対は独立である」というモデル，モデル M2「a,b の周囲単語対は独立でない」というモデルを考え，独立性の否定されたものを関連語対と刷ることを考える．これには AIC による独立性検定を用いた．これは 3.2 節と同様の方法で，BIC ではなく AIC を使用したものである．各モデルの AIC の値は，共通項を無視すると次のようになる．

$$AIC_{M1} = (-2)\sum_{i,j} n_{ij} \log \frac{n_{i\cdot} n_{\cdot j}}{n_{\cdot\cdot}} + 2 \times 2$$

$$AIC_{M2} = (-2)\sum_{i,j} n_{ij} \log \frac{n_{ij}}{n_{\cdot\cdot}} + 2 \times 3$$

したがって次のような関連語対集合 *Relevants′* が得られる．

$$Relevants' = \{ab \mid AIC_{M1} - AIC_{M2} > 1\}$$

5. 実験

5.1. 実験内容

ここでは関連語の 2 次判定として，山本らの方法，3.3 節で示した頻度モデルによる方法，4節で示した異なり数モデルによる方法をそれぞれ適用し，その結果を比較する実験を行った．

関連語かどうか検査される候補対は3.2節の方法で得られた候補対 11,838 対を使用した．また，結果に対して表 5.4のように人手で正誤評価を行なった．評価の対象は AIC 値の差が大きかったもの上位 100 件である．

5.2. 結果と考察

表 5.7は頻度モデルと異なり数モデルそれぞれ 2 つの手法を適用して得られた関連語対の数と，AIC 値の差上位 100 位の正解率を示したものである．表 5.5は頻度モデルの結果例，表 5.6は異なり数モデルを適用した場合の関連語対の例である．

頻度モデルによる方法は，出力される関連語対の数については異なり数モデルに比べて多いが，正解率が著しく悪くなっている．一方，異なり数モデルによる方法は，出力される関連語対の数，正解率共にやや増加した．異なり数

表 5.5 頻度モデルでの結果例

Word1	Word2	正誤
測定器	測定	正
時制論理	設計	誤
構造	共通部	誤
並行開発	プログラム	誤
ベース・システム	記述	誤
画像	平坦	誤
用例検索	画像	誤
地図記号	画像	誤
分光診断	生成	誤
SIH	生成	誤

表 5.6 異なり数モデルでの結果例

Word1	Word2	正誤
スーパー	スーパ	正
デジタル	ディジタル	正
FDT	LOTOS	正
吊橋	架橋	正
PAM	PWM	正
短絡	ショート	正
彩色する	辺彩色	誤
がいし	がいしの	誤
教育	演習	誤
法律文	翻訳	誤

表 5.7 実験結果

判定手法	関連語対数	正解率
山本らの手法	824	0.22
異なり数モデル	1099	0.31
頻度モデル	4207	0.01

モデルをとることで,確率での意味づけを持つこととシステムとして動作することが両立したことは興味深い結果と考えられる.

異なりモデルの立場は,ある事象が出現するか出現しないかは考慮する一方で,一度出現した事象であれば,それが何度出現しようとも無視するというモデルである.いいかえれば,このモデルは統計的な言語処理では重み付けとして重要と認められている頻度を無視するという異端のモデルである.この報告の処理では,ほぼ同じ枠組みの処理を行ったにもかかわらず,頻度を利用するとシステムとして動作しなかったことから考えると,異なりモデルもなんらかの関連する言語事象があることが示唆される.

これについて,著者らは以下のように解釈している.単語の文書中の出現については,1度単語が出現することを前提条件としたとき,その単語が2度以上出現する確率は大きいことが観測される.この確率が1に近ければ,文書内の頻度には情報がなく,単語が出現するかどうかだけが意味をもつことになる.本稿のシステムが動作したことから考えて,語の接続関係において,その接続関係が生じたことを前提条件としたとき,一度生じた接続関係がほかの場所で出現する条件確率は,最初に出現する確率とは無関係であることを示唆される.

6. まとめ

本報告では,文書中から関連語対を統計的に選び出すシステムについて説明し,従来まで採用されてきたアドホックな関数による判定尺度を統計モデルに基づいた方法に置き換えてシステムを再構築した.また,関連語の判定方法として,語の頻度分布に着目したモデルと語の異なり数に着目したモデルの2つを考え比較をし,関連語の判定材料としては頻度より異なり数の方が有効であることを報告した.

7. 謝辞

この研究は、住友電工情報システムとの共同研究の成果です.また,この成果を分析するときに使用したシステムは,平成19年度科学研究費補助金(課題番号19500120)の研究成果を使いました.

8. 参考文献

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

[2] 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. 計量国語学, Vol.23, No.2, pp.65-90, September 2001

[3] 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

[4] 當間雅, 折原幸治, 塩入寛之, 梅村恭司. 関連語対のマイニ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, 2007

[5] 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, 2006

[6] 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

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

[8] 坂本慶行, 石黒真木夫, 北川源四郎. 情報量統計学. 共立出版, 1983

[9] 赤池弘次, 甘利俊一,北川源四郎,樺島祥介,下平英寿. 赤池情報量基準 AIC. 共立出版, 2007

[10] 小西貞則. 情報量基準. 朝倉出版. 2004

[11] 薩摩順吉. 確率・統計. 岩波書店, 1989

[12] 北研二. 確率的言語モデル. 東京大学出版会, 1999

[13] 北研二, 津田和彦, 獅々堀正幹. 情報検索アルゴリズム. 共立出版, 2002

[14] Patrick Pantel and Marco Pennacchiotti. Espresso: leveraging generic patterns for automatically harvesting semantic relations. ACL '06: Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the ACL, pp. 113-120, 2006