Vol.2009-NL-193 No.8
2009/9/29

語の文脈の一致判定における 文脈の出現頻度と種類数の比較

増山篤志[†]　梅村恭司[†]　阿部洋丈[†]　岡部正幸[††]

関連する語であるかを判定する目的で，語の文脈の一致を利用できる．ここで二つの可能なモデルがある．語の文脈を，単語対の前後に出現する単語を結合した周囲単語対としたとき，頻度モデルでは周囲単語対の一致する頻度を，種類数モデルでは一致する種類数を計測して，周囲単語対の出現に関連があるかどうかを独立性の判定で求めた．その結果，種類数モデルの方が高い性能で関連語対を得ることが出来たことを報告する．

Comparing Variety with Frequency on Judging the correspondence of the Surrounding Contexts of Words

Atsushi Masuyama[†]　Kyoji Umemura[†]　Hirotake Abe[†]
and Masayuki Okabe[††]

It is possible to use the correspondence of the surrounding contexts of words in order to extracts the related word pairs. There are two kinds of model for the correspondence. The proposed frequency model counts numbers of correspondence of the surrounding word pairs, and the variety model counts number of kinds of the pairs. In both case, the system extracts the related word pairs by the test of independence based on these counts. Therefore, the variety model is found to be better model to get related word pairs.

1. はじめに

文書集合から関連している単語対を抽出する技術は，未知語を理解することや，関連するほかの単語を得ることに役立つ．山本らによって提案され，當間によって改良されたシソーラス構築システム[1,2]は，単語対の周囲に現れる文字列の情報から語の出現類似性を求めて関連語対を求めるもので，単語辞書を用いずに関連語対のリストを作成することができる．

シソーラス構築システムでは，文書集合から単語の切り出し，関連語候補対の抽出，関連語対の選出，という大きく 3 つの工程を経て関連語対を抽出する．當間らの研究では，関連語対の選出工程において，単語対の前後に出現する単語の出現頻度の分布から関連性を判定する頻度モデルと，単語対の前後に出現する単語が共に出現する種類数の分布から関連性を判定する種類数モデルを比較したところ，種類数モデルで関連語対を得ることが出来た．しかし，當間らの研究で扱っている 2 つのモデルは情報の取り方が異なるため，語の文脈の出現頻度と種類数の比較をしているとはいえない．

本研究では，関連語対の判定において文脈出現頻度と種類数のどちらを使用した方が高い性能で関連語対を得ることが出来るのかを比較するために，當間らの頻度モデルの情報の取り方を種類数モデルに合わせ，共通する単語対の頻度，種類数の分布から関連性を判定し，関連語対を選出するモデルに変更した．結果，種類数モデルの方が高い性能で関連語対を得ることが出来たことを報告する．

2. 本研究で扱うシステム

本研究で扱うシソーラス構築システムでは，文書集合から単語の切り出し，関連語候補対の抽出，関連語対の選出，という大きく 3 つの工程を経て関連語対を抽出する．これらは先行する研究を踏襲している．

2.1 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[3,4]使用している．このアルゴリズムは文字列の頻度情報を基にした単語らしさの尺度に従ってキーワードを抽出するもので，あらかじめ単語辞書を用意する必要がない． 図 2.1 に単語の切り出し例を示す．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している．このような単語を文脈単語と呼ぶことにする．

[†] 豊橋技術科学大学 情報工学系
[††] 豊橋技術科学大学 情報メディア基盤センター

Vol.2009-NL-193 No.8
2009/9/29

大学等での基礎的な電気回路演習を[支援]する[ＣＡＩ][ソフトウェア]とその改良について述べている．本[ＣＡＩ]はコンピュータが出題される回路を[学習]者各人のレベルに応じて[自動]的に作成すること，解答を数式で[入力]することが大きな特長である．
* []内は抽出したキーワード

図 2.1 単語の切り出し例

2.2 関連語候補対の抽出

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．関連語候補対とは，前後に同じ単語が現れる単語の対である．図 2.2 に候補対抽出の例を示す．ここの例では，「プリント」と「印刷」が候補対となる．

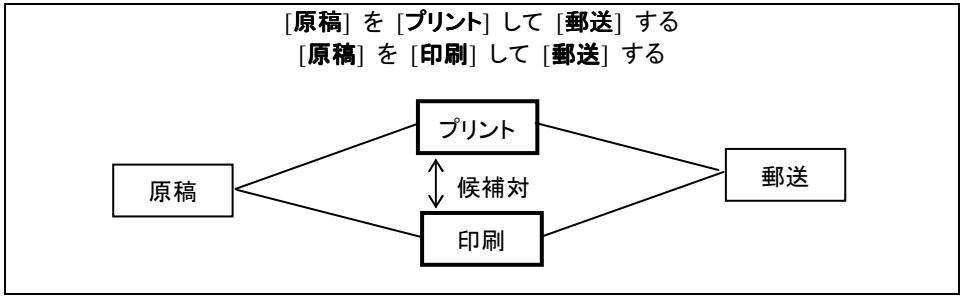

図 2.2 候補対抽出の例

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ことにする． 順序対が選出されたら，順序対を用いて関連語候補対を抽出する．この候補対の抽出は，第 3 工程である関連語の選出の前段階として，関連語になりそうなものをざっと絞り込む工程である．候補対の抽出では，図 2.2 のように，前後に同じ単語が現れる単語の対を見つけ出して抽出する．

2.3 関連語の選出

最後の工程では，候補対の周囲に現れる単語の出現類似性の尺度に従って関連度を求めて，関連語であるかどうかを判定する．候補対の周囲に現れる単語とは，第 1 工程で文書から切り出したキーワードと文脈単語を指している．當間らの研究では，関連語対の選出工程における判定方法として,頻度モデルと種類数モデルを挙げている．

3. 頻度・種類数モデル

3.1 當間の頻度・種類数モデルにおける問題点

當間らの研究[1]では関連語対の選出において，頻度モデルよりも種類数モデルの方が出力される関連語対の数が多く，精度も高くなるという結果が得られた．しかし，頻度と種類数に関して厳密に比較できておらず，頻度モデルは「前後に出現する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るモデルであるのに対し，種類数モデルは「前後に出現する単語が多くの種類で共通する」単語対を関連語対と考えるモデルとなっている．そのため分割表の形が異なっており，関連度を求める式も異なる．また，頻度モデルでは単語対の前後に出現する単語を分けて扱っているのに対し，種類数モデルでは単語対を連結させて周囲単語対として扱っている．したがって，當間の研究で扱っている頻度モデルと種類数モデルでは，情報の取り方などが異なるため，両モデルの比較が語の文脈の出現頻度と種類数の比較とはならない．

3.2 種類数モデル

x_i を関連語であるかを判定する単語とし，ある x_i の出現場所に対し，その直前に現れた単語を α，直後に現れた単語を β とするとき，ペア (α,β) を x_i の周囲単語対 y_j とする． 候補対に含まれるすべての単語 x_i の集合を $\mathbf{X}=\{x_i\}$， x_i のすべての出現場所においての x_i の周囲単語対 y_j の集合を $\mathbf{Y}(x_i)$，すべての x_i に対する周囲単語対 y_j の集合を $\mathbf{Y}$ とする．ここで定義する集合 $\mathbf{Y}$ と $\mathbf{Y}(x_i)$ は同じ周囲単語対は一つの元としてみなし，重複を許さない．このとき，$(x_1=a,x_2=b)$ を候補対とするとき，共通する周囲単語対の集合は $\mathbf{Y}(x_1)\cap\mathbf{Y}(x_2)$ となり，共通する周囲単語対の種類数は集合 $\mathbf{Y}(x_1)\cap\mathbf{Y}(x_2)$ の要素数 $\left|\mathbf{Y}(x_1)\cap\mathbf{Y}(x_2)\right|$ となる．ここで $\overline{\mathbf{Y}(x_i)}$ は $\mathbf{Y}$ を全体集合と考えたときの $\mathbf{Y}(x_i)$ の補集合である．

図 3.1 は $\mathbf{Y}=\{y_1,y_2,y_3\}$ ，$\mathbf{Y}(x_1)=\{y_1,y_2\}$，$\mathbf{Y}(x_2)=\{y_1\}$ であるとき，共通する周囲単語対を求めている例である． 1 列目と 1 行目は集合 $\mathbf{Y}$ の周囲単語対，2 列目と 2 行目は集合 $\mathbf{Y}(x_1)$ と $\mathbf{Y}(x_2)$ の周囲単語対，残りの対角成分は共通する周囲単語対の種類数を示している．この例では共通する周囲単語対の種類数は 1 となる．共通する周囲単語対の種類数は次のように求める．表 3.1 はこれを分割表[5]に示したものである．また，本研究では表 3.1 の分布から関連性を判定することを一致判定と定義している．

Vol.2009-NL-193 No.8
2009/9/29

$$n_{11}=\left|\mathbf{Y}(x_1)\cap\mathbf{Y}(x_2)\right|,\quad n_{12}=\left|\mathbf{Y}(x_1)\cap\overline{\mathbf{Y}(x_2)}\right|$$
$$n_{21}=\left|\overline{\mathbf{Y}(x_1)}\cap\mathbf{Y}(x_2)\right|,\quad n_{22}=\left|\overline{\mathbf{Y}(x_1)}\cap\overline{\mathbf{Y}(x_2)}\right| \qquad (3.1)$$
$$n_{i\cdot}=\sum_{j=1}^{2}n_{ij},\quad n_{\cdot j}=\sum_{i=1}^{2}n_{ij},\quad n_{\cdot\cdot}=\sum_{i=1}^{2}\sum_{j=1}^{2}n_{ij}=\left|\mathbf{Y}\right|$$

	$\mathbf{Y}$	y_1	y_2	y_3
$\mathbf{Y}$	$\mathbf{Y}(x_2)$ / $\mathbf{Y}(x_1)$	y_1	-	-
y_1	y_1	1		
y_2	y_2		0	
y_3	-			0

図 3.1 共通する周囲単語対の種類数の例

表 3.1 分割表

	$\mathbf{Y}(x_2)$	$\overline{\mathbf{Y}(x_2)}$	計
$\mathbf{Y}(x_1)$	n_{11}	n_{12}	$n_{1\cdot}$
$\overline{\mathbf{Y}(x_1)}$	n_{21}	n_{22}	$n_{2\cdot}$
計	$n_{\cdot 1}$	$n_{\cdot 2}$	$n_{\cdot\cdot}$

3.3 頻度モデル

3.2 節と同様に，x_i を関連語であるかを判定する単語とし，ある x_i の出現場所に対し，その直前に現れた単語を α，直後に現れた単語を β とするとき，ペア (α,β) を x_i の周囲単語対 y_j とする．候補対に含まれるすべての単語 x_i の集合を $\mathbf{X}=\{x_i\}$，x_i のすべての出現場所においての x_i の周囲単語対 y_j の集合を $\mathbf{Y}^+(x_i)$，すべての x_i に対する周囲単語対 y_j の集合を $\mathbf{Y}^+$ とする．ここで定義する集合 $\mathbf{Y}^+$ と $\mathbf{Y}^+(x_i)$ は同じ周囲単語対を元としていくつも含む多重集合であり，文書集合に出現する周囲単語対を，多重を許してすべて含んでいる．多重集合に同じ値の元がいくつも含まれるとき，各元の個数を重複度，重複度を求める関数を重複度関数という．ここでは，集合 $\mathbf{Y}^+$ に含まれる元 y_j の個数を重複度関数を用いて $m_{\mathbf{Y}^+}(y_j)$ と書く．同様に，集合 $\mathbf{Y}^+(x_i)$ に含まれる元 y_j の個数を重複度関数を用いて $m_{\mathbf{Y}^+(x_i)}(y_j)$ と書く．このとき，$(x_1=a,x_2=b)$ を候補対とするとき，共通する周囲単語対の集合は $\mathbf{Y}^+(x_1)\cap\mathbf{Y}^+(x_2)$ となり，共通する周囲単語対の頻度は，重複度の積 $\sum_{j=1}^{r}(m_{\mathbf{Y}^+(x_1)}(y_j)\cdot m_{\mathbf{Y}^+(x_2)}(y_j))$ となる．ここで $\overline{\mathbf{Y}^+(x_i)}$ は $\mathbf{Y}^+$ を全体集合と考えたときの $\mathbf{Y}^+(x_i)$ の補集合である．

図 3.2 は $\mathbf{Y}^+=\{y_1,y_1,y_1,y_2,y_3\}$, $\mathbf{Y}^+(x_1)=\{y_1,y_1,y_2\}$, $\mathbf{Y}^+(x_2)=\{y_1,y_2\}$ であるとき，共通する周囲単語対を求めている例である．1 列目と 1 行目は集合 $\mathbf{Y}^+$ の周囲単語対，2 列目と 2 行目は集合 $\mathbf{Y}^+(x_1)$ と $\mathbf{Y}^+(x_2)$ の周囲単語対，残りの白地部分は共通する周囲単語対の頻度を示している．この例では共通する周囲単語対の頻度は 3 となる．図 3.2 は図 3.1 の周囲単語対の多重を許して表現したものであり，同じ枠組みで共通する周囲単語対を捉えている．共通する周囲単語対の頻度は次のように求める．分割表の形は表 3.1 と同じになるので，同じ式で関連性を判定することが出来る．

$$n_{11}=\sum_{j=1}^{r}(m_{\mathbf{Y}^+(x_1)}(y_j)\cdot m_{\mathbf{Y}^+(x_2)}(y_j)),n_{12}=\sum_{j=1}^{r}(m_{\mathbf{Y}^+(x_1)}(y_j)\cdot m_{\overline{\mathbf{Y}^+(x_2)}}(y_j))$$
$$n_{21}=\sum_{j=1}^{r}(m_{\overline{\mathbf{Y}^+(x_1)}}(y_j)\cdot m_{\mathbf{Y}^+(x_2)}(y_j)),n_{22}=\sum_{j=1}^{r}(m_{\overline{\mathbf{Y}^+(x_1)}}(y_j)\cdot m_{\overline{\mathbf{Y}^+(x_2)}}(y_j)) \qquad (3.2)$$
$$n_{i\cdot}=\sum_{j=1}^{2}n_{ij},\quad n_{\cdot j}=\sum_{i=1}^{2}n_{ij},\quad n_{\cdot\cdot}=\sum_{i=1}^{2}\sum_{j=1}^{2}n_{ij}=\sum_{j=1}^{r}(m_{\mathbf{Y}^+}(y_j)^2)$$

	$\mathbf{Y}^+$	y_1	y_1	y_1	y_2	y_3
$\mathbf{Y}^+$	$\mathbf{Y}^+(x_1)$ / $\mathbf{Y}^+(x_2)$	y_1	-	-	y_2	-
y_1	y_1	1	0	0		
y_1	y_1	1	0	0		
y_1	-	0	0	0		
y_2	y_2				1	
y_3	-					0

図 3.2 共通する周囲単語対の頻度の例

3.4 関連語対の判定方法

表 3.1 の分割表の分布から周囲単語対の関連性を求め，候補対が関連語対であるかを判定する方法をここで述べる．本実験では，AIC(赤池情報量基準)の独立性判定，カイ二乗検定，補完類似度，AIC の独立判定＋補完類似度による判定の 4 つの判定方法を用いた．

3.4.1 AIC の独立判定による判定方法

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)としたものである．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，歴史的経緯からそれを -2 倍した量

$$AIC = (-2) \times l(\hat{\theta}) + 2 \times k \qquad (3.3)$$

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[6]. 2 つのモデルを比較する場合，AIC の値の差が 1 以上あれば優位な差と言える．

関連語候補対を(a,b)としたとき，「(a,b)の周囲単語対は独立である」とするモデル M1 と，「(a,b)の周囲単語対は独立でない」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．(3.3)を表 3.1 の分割表に適用して変形すると次のようになる．

$$AIC_{M1} = -2\sum_{i,j} n_{ij} \log \frac{n_{i.} n_{.j}}{n_{..}} + 2 \times 2, \quad AIC_{M2} = -2\sum_{i,j} n_{ij} \log \frac{n_{ij}}{n_{..}} + 2 \times 3 \qquad (3.4)$$

(3.4)の差 $AIC_{M1} - AIC_{M2}$ が大きいほど (a,b) の周囲単語対に関連があることになるので, この値を関連度とし, 次のような関連語対集合 $Relevants_{aic}$ が得られる．

$$Relevants_{aic} = \{(a,b) \mid AIC_{M1} - AIC_{M2} > 1\} \qquad (3.5)$$

3.4.2 カイ二乗検定による判定方法

適切な仮定のもと，下記の確率変数 Z は自由度 1 のカイ二乗分布に従う．

$$Z = \sum_{i=1}^{2} \sum_{j=1}^{2} \frac{(n_{ij} - \hat{n}_{ij})^2}{\hat{n}_{ij}} \qquad (3.6)$$

確率変数 Z がある水準値 α より大きければ，関連語候補対 (a,b) は関連語対と判定される．したがって， Z をカイ二乗検定の関連度とし，次のような関連語対集合 $Relevants_{chi}$ が得られる．本実験では，検定危険率は 0.005 と設定し，水準値 α は 7.88 とする．

$$Relevants_{chi} = \{(a,b) \mid Z > \alpha\} \qquad (3.7)$$

3.4.3 補完類似度による判定方法

二つのパターンの類似度を求める場合，補完類似度は包含関係をもつような一対多関係を推定する問題において，高い性能で類似度を得ることが出来る[7].

$$S_c = \frac{n_{11} n_{22} - n_{12} n_{21}}{\sqrt{(n_{11} + n_{21})(n_{12} + n_{22})}} \qquad (3.8)$$

補完類似度 S_c が閾値 β より大きければ，関連語対であると判定される．したがって， S_c を関連度とし，次のような関連語対集合 $Relevants_{Sc}$ が得られる．本実験では，閾値 β を 0 と設定し，補完類似度が正の値のとき関連語対であると判断した．

$$Relevants_{Sc} = \{(a,b) \mid S_c > \beta\} \qquad (3.9)$$

3.4.4 AIC の独立判定＋補完類似度による判定方法

AIC の独立判定と補完類似度による判定で関連性があると判定された候補対を関連語対とする．したがって，次のような関連語対集合 $Relevants_{aic+}$ が得られる．また，関連度は AIC の独立判定の関連度である $AIC_{M1} - AIC_{M2}$ を用いる．

$$Relevants_{aic+} = Relevants_{aic} \wedge Relevants_{Sc} \qquad (3.10)$$

3.5 実験内容

種類数モデルと頻度モデルを使用して関連語対を選出した．実験には 2 節で説明したシソーラス構築システムを使用してキーワード，文脈単語，候補対を抽出し，各モデルの一致判定により周囲単語対の関連性を判定して，関連語対を選出した．対象コーパスには NTCIR の学術論文記事を使用した．実験は 2 種類行った．一つ目は同記事 20 万件を使用し，3.4 節で説明した 4 つの判定方法を用いて，種類数モデルと頻度モデルを判定方法別に比較した．二つ目は同記事を 2,4,8,12,16,20 万件ずつ使用し，AIC の独立判定+補完類似度による判定を用いて，種類数モデルと頻度モデルを入力データ件数別に比較した．得られた関連語対の評価は，人手で正誤評価を行う．モデルの比較方法としては,選出した関連語対すべてを人手で評価することは困難であるため,選出した関連語対を関連度のスコア順にソートしたときのスコア上位 100 位の評価値の合計で比較する．

3.6 実験結果

各判定方法別に，候補対数と 3.4 節で述べた関連語対集合の判断基準を満たす関連語対数と関連度のスコア上位 100 件の評価値を図 3.3 に示す．入力データ件数別に，候補対数と関連語対集合の判断基準を満たす関連語対数と関連度のスコア上位 100 件の評価値を図 3.4 に示す．また，得られた関連語対の一部を

表 3.2 に示す.評価値は人手で主観的に判断し,選出した関連語対に関連があれば○,関連がないか関連が弱い対は×，どちらともいえない対は△とし，それぞれ 1,0,0.5 点としている．ただし，単語の切り出し過程におけるミス(非単語や助詞が加わっているだけの単語など)は除いた結果の評価をしているため，実際の性能はもう少し低い．

表 3.2 関連語対の一部

単語 a	単語 b	評価	単語 a	単語 b	評価
遺伝アルゴリズム	遺伝的アルゴリズム	○	符号化	圧縮	○
ニューラルネットワーク	神経回路網	○	データ	情報	○
ひびわれ	ひび割れ	○	ポリマー	高分子	○
タンパク質	蛋白質	○	動画	画像	○
識別	認識	○	演算	処理	△
制御系	制御器	○	電極	モデル	×
ヒドロゲル	ハイドロゲル	○	試料	方式	×
線形	非線形	○	制御	分布	×
誘導電動機	誘導機	○	顔画像	文字	×

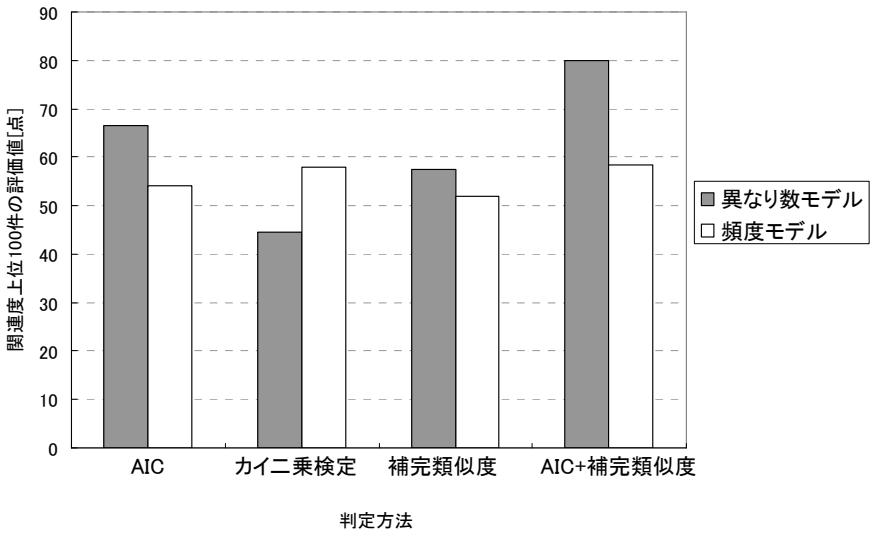

図 3.3 判定方法別 上位 100 件の評価値

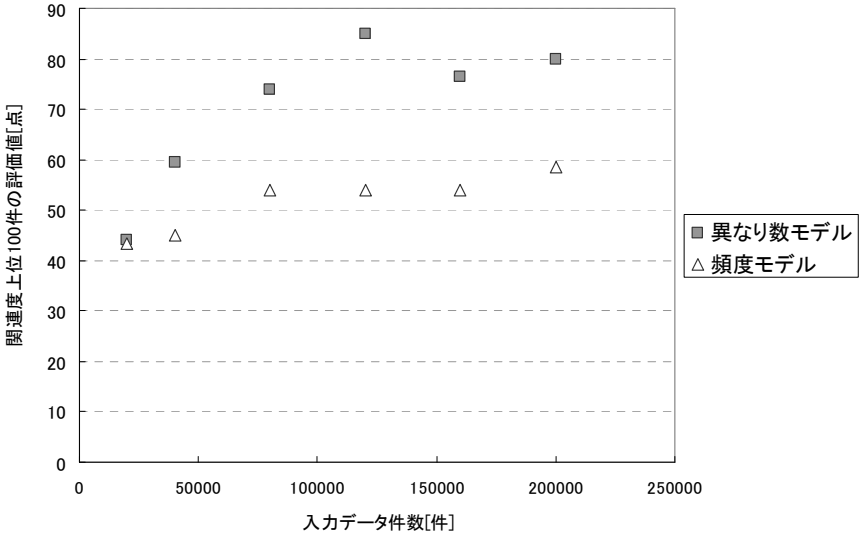

図 3.4 入力データ件数別 上位 100 件の評価値

3.7 考察

図 3.3 より，判定方法別に関連度上位 100 件の評価値を比較すると，カイ二乗検定以外の 3 つの判定方法では種類数モデルの方が評価値は高い．各モデルで一番評価値が高い判定方法は AIC の独立性判定＋補完類似度による判定であったが，それを比べても種類数モデルの評価値の方が高いことが分かる．図 3.4 より，入力データ件数別に関連度上位 100 件の評価値を比較すると，入力データが 2 万件のときは各モデルの差はあまりないが，それ以降では種類数モデルの評価値の方が高いということが分かる．また，入力データ件数が増えるにつれて，どちらのモデルも評価値は増えているが，種類数モデルの方が増える割合が高いということも分かる．

以上より，判定方法を変更してみても，種類数モデルの方が頻度モデルより多くの関連語対を得ることができた．種類数モデル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，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．統計的な言語処理では，重み付けとして頻度が良く使われることが多いため，興味深い事例である．

これについて，著者らは今のところ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．これは，同じ種類の文書中では同じ文字列を繰り返して使う可能性が高いということからも推測できる．また，頻度モデルの関連度と種類数モデルの関連度の差が大きい関連語対を選んだところ，共通して出現する周囲単語対の種類数は 1～3 種類ほどであるのに対し，頻度は 90 回以上であった．つまり，頻度モデルでは共通して出現する周囲単語対が，少ない種類で重複していることになる．本研究は文脈の一致判定で関連語対を選出しているが，頻度を一致判定に用いると，このような過剰に出現する単語の一致によって関連度が高くなってしまう可能性がある．そのため，頻度を使用するよりも種類数のみを用いたほうが良い結果になったのではないかと思われる．

4. まとめ

本研究では，関連語対の抽出において，語の文脈の一致判定における文脈の出現頻度と種類数のどちらを使用した方が高い性能で関連語対を得ることが出来るのかどうかを比較した．比較方法としては，共通する周囲単語対の頻度分布と種類数分布の関連性を判定して，関連性が高かった単語対を関連語対として抽出した．結果，種類数モデルの方が高い性能で関連語対を得ることが出来たことを報告する．

今後の課題としては，判定方法や実験対象を変更して，実験の数を増やし，結果に差があることをさらに確実にしたい．また，差の理由として考えられる説明を別の実験で確認したい．

5. 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あり，サポートに感謝します．

参考文献

[1] 當間 雅, 梅村 恭司: 語の出現類似性のための統計的モデルとシソーラス構築への適用, 言語処理学会第 13 年次大会, 2007.

[2] Eiko Yamamoto and Kyoji Umemura: Related word-pairs extraction without dictionaries, In *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC-2004)*, pp. 1309–1312, 2004.

[3] 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, 計量国語学会, Vol. 23, No. 2, pp. 65–90, 2001.

[4] Yoshiyuki Takeda and Kyoji Umemura: Selecting indexing strings using adaptation, In *Proceedings of The 25th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 42–43, 2004.

[5] 坂元 慶行, 石黒真木夫, 北川源四郎: 情報量統計学, 共立出版, 1983.

[6] 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.

[7] 山本 英子, 梅村 恭司: コーパス中の一対多関係を推定する問題における類似尺度, 自然言語処理, Vol.9, No.2, pp.45-75, 2002.