分布類似度判定における文脈の特徴量の比較と評価法に関する研究

増山篤志†　梅村恭司†　岡部正幸††

† 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

1. はじめに

文書集合から関連性の高い語を獲得するという課題は，自然言語処理の基礎的な研究課題であり，情報検索の分野において重要な要素技術の一つである．語の関連性を測る手法の一つとして「分布類似度」が提案され，広く利用されている．分布類似度とは，文脈の類似した語には関連性があると推定される「分布仮説」に基づいて計算する手法である．通常，分布類似度の計算では，文脈情報すべてを利用することは計算量的に困難であるため，定められたテキスト範囲で共起する文脈や，語に関係している有用な文脈（係り受け関係など）を素性として選択する．文脈素性の特徴量としては，文脈の出現頻度や異なり数が挙げられる．頻度は，類似した文脈が繰り返して出現するかどうかを示す特徴量であり，異なり数は，多様な文脈が出現するかどうかを示す特徴量である．特に頻度は，統計的な言語処理で頻繁に利用される特徴量であり，tf-idf など多くの指標で用いられている．

山本らによって提案され，當間らによって改良されたシソーラス構築システム[1][2]は，文書集合から分布類似度を求めることにより関連性の高い単語の対（関連語対）を抽出するシステムである．シソーラス構築システムでは，文書集合から単語の抽出，候補対の抽出，関連語対の選出という 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，共起する文脈の異なり数から分布類似度を求める異なり数モデルと，頻度から分布類似度を求める頻度モデルを比較している．しかし，當間らの研究で定義した 2 つのモデルは，文脈情報の取り方を異なるため，文脈の異なり数と頻度を厳密に比較しているとはいえない．

本研究では，関連語対の選出における分布類似度判定において，文脈の特徴量として異なり数と頻度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，當間の頻度モデルの文脈情報の取り方を，異なり数モデルと同じ手法に変更した．また，先行研究の関連語対の評価は，一部の関連語対を人手で評価していたが，関連性の有無を客観的に行うことが難しく，評価できる量も限られてしまうという問題があった．そこで，人手での評価の他に，評価辞書として日本語 WordNet 辞書，Wikipedia のキーワードリンク辞書を用意して評価を行った．結果，異なり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は，文書集合から単語の切り出し，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．これらは先行研究を踏襲している．

2.1. 単語の抽出

第 1 工程ではコーパスから単語の抽出を行う．ここでは全文検索エンジン QuickSolution[3]（以下 QS と呼ぶ）の API を使用してキーワードを抽出する．QS では，キーワードの抽出に武田のキーワード抽出アルゴリズム[4]を使用しており，辞書に登録されていない未知語や専門用語も抽出することができる．

2.2. 候補対の抽出

第 1 工程で抽出単語集合から考え得る単語対すべてについて関連性が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となる対（候補対）をざっと絞り込む．候補対の抽出では，単語の前後に出現している単語を文脈として抽出し，文脈が共通している単語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，候補対に共通して出現する文脈から分布類似度を計算して，関連語対であるかどうかを判定する．

素性としては，単語の前後に出現する 1 単語の組を周囲単語対と定義し，周囲単語対を文脈の素性として選択する．文脈の素性の取り方は多様であるが，ここでは文脈の依存情報，範囲，長さなどについては深く考慮せずに，単語同士の前後関係のみを文脈として捉える単純なモデルを用いており，本稿では文脈を周囲単語対で表現する．

分布類似度を判定するためには，候補対の周囲単語対を抽出し，周囲単語対の分布に関連性があるかどうかを調べる必要がある．ここで，周囲単語対の分布に用いる特徴量として，同じ周囲単語対の出現頻度を考慮するべきかどうかという問題がある．

3. 頻度・異なり数モデル

當間らの研究では，関連語の選出工程における文脈の特徴量として，文脈の頻度を考慮せずに異なり数から分布類似度を求める異なり数モデルと，頻度から分布類似度を求める頻度モデルを比較している．しかし，當間の頻度モデルと異なり数モデルは，文脈情報の取り方が異なるため，両モデルの比較が文脈の出現頻度と異なり数の比較とはならない．

本研究では，當間の頻度モデルの文脈情報の取り方を変更し，周囲単語対の異なり数と頻度の分布を，同じ枠踏みで比較できるモデルに変更した．異なり数モデルについては，當間の異なり数モデルを踏襲している．

3.1. 異なり数モデル

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，その前後に出現する 1 単語のペアを周囲単語対 y_j とする．x_i のすべての出現場所においての周囲単語対 y_j の集合を $Y(x_i)$，すべての x_i の周囲単語対 y_j の集合を Y とする．ここで定義する集合 Y と $Y(x_i)$ は同じ周囲単語対を一つの元としてみなし，重複を許さない．このとき，(x_1, x_2) を候補対とするとき，候補対に共通して出現する周囲単語対は $Y(x_1) \cap Y(x_2)$，x_1 には出現して x_2 に出現し

ない周囲単語対は $Y(x_1)\cap\overline{Y(x_2)}$，$x_2$ には出現して x_1 に出現しない周囲単語対は $\overline{Y(x_1)}\cap Y(x_2)$，候補対どちらにも出現しない周囲単語対は $\overline{Y(x_1)}\cap\overline{Y(x_2)}$ となる．ここで $\overline{Y(x_i)}$ は Y を全体集合と考えたときの $Y(x_i)$ の補集合である．周囲単語対の異なり数は，それぞれの集合の元の個数となるので，共通して出現する周囲単語対の異なり数の場合は $\left|Y(x_1)\cap Y(x_2)\right|$ となる． 以上から，表 3.1 の分割表が求まる．

表 3.1 分割表（異なり数モデル）

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

$$n_{11}=\left|Y(x_1)\cap Y(x_2)\right|,\quad n_{12}=\left|Y(x_1)\cap\overline{Y(x_2)}\right|$$
$$n_{21}=\left|\overline{Y(x_1)}\cap Y(x_2)\right|,\quad n_{22}=\left|\overline{Y(x_1)}\cap\overline{Y(x_2)}\right|$$

表 3.2 は $Y=\{y_1,y_2,y_3\}$ ，$Y(x_1)=\{y_1,y_2,y_3\}$，$Y(x_2)=\{y_1,y_2\}$ であるとき，共通する周囲単語対の異なり数を求めている例である． 1 列目と 1 行目は集合 $Y(x_1)$ と $Y(x_2)$ の周囲単語対，残りの対角成分は周囲単語対が一致しているかどうかを示している．この例では共通する周囲単語対の異なり数は $\left|Y(x_1)\cap Y(x_2)\right|=2$ となる．

表 3.2 共通する周囲単語対の異なり数の例

$Y(x_2)$ \ $Y(x_1)$	y_1	y_2	y_3
y_1	1		
y_2		1	
-			0

3.2. 頻度モデル

3.1 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，その前後に出現する単語のペアを周囲単語対 y_j とする． x_i のすべての出現場所においての周囲単語対 y_j の集合を $Y^+(x_i)$，すべての x_i の周囲単語対 y_j の集合を Y^+ とする．ここで定義する集合 Y^+ と $Y^+(x_i)$ 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 Y^+ に含まれる元 y_j の個数を重複度関数を用いて $m(Y^+,y_j)$ と書く．同様に，集合 $Y^+(x_i)$ に含まれる元 y_j の個数を重複度関数を用いて $m(Y^+(x_i),y_j)$ と書く．このとき，(x_1,x_2) を候補対とするとき，共通する周囲単語対の頻度は，

重複度の積 $\sum_{j=1}^{r}(m(Y^+(x_1),y_j)\cdot m(Y^+(x_2),y_j))$ となる．

ここで $\overline{Y^+(x_i)}$ は Y^+ を全体集合と考えたときの $Y^+(x_i)$ の補集合である．以上から，表 3.3 の分割表が求まる．

表 3.3 分割表（頻度モデル）

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

$$n_{11}=\sum_{j=1}^{r}(m(Y^+(x_1),y_j)\cdot m(Y^+(x_2),y_j))$$
$$n_{12}=\sum_{j=1}^{r}(m(Y^+(x_1),y_j)\cdot m(\overline{Y^+(x_2)},y_j))$$
$$n_{21}=\sum_{j=1}^{r}(m(\overline{Y^+(x_1)},y_j)\cdot m(Y^+(x_2),y_j))$$
$$n_{22}=\sum_{j=1}^{r}(m(\overline{Y^+(x_1)},y_j)\cdot m(\overline{Y^+(x_2)},y_j))$$

表 3.4 は $Y^+=\{y_1,y_1,y_1,y_1,y_2,y_2,y_3\}$ ，$Y^+(x_1)=\{y_1,y_1,y_2,y_3\}$，$Y^+(x_2)=\{y_1,y_1,y_2\}$ であるとき，共通する周囲単語対の頻度を求めている例である．1 列目と 1 行目は集合 $Y^+(x_1)$ と $Y^+(x_2)$ の周囲単語対，残りの対角成分は周囲単語対が一致しているかどうかを示している．この例では共通する周囲単語対の頻度は

$$\sum_{j=1}^{r}(m(Y^+(x_1),y_j)\cdot m(Y^+(x_2),y_j))=2\cdot 2+1\cdot 1+1\cdot 0=5$$

となる．

表 3.4 は表 3.2 の周囲単語対の多重を許して表現したものであり，異なり数もモデルと同じ枠組みで周囲単語対の一致を捉えている．

表 3.4 共通する周囲単語対の頻度の例

$Y^+(x_2)$ \ $Y^+(x_1)$	y_1	y_1	y_2	y_3
y_1	1	1		
y_1	1	1		
y_2			1	
-				0

3.3. 関連語対の判定方法

表 3.1，表 3.3 の分割表から分布類似度を求め，候補対が関連語対であるかを判定する．分割表の分布から関連性を求める手法はいくつかあるが，本実験では過去の実験で特に良い結果が得られた AIC（赤池情報量基準）の独立判定によって判定を行う．

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

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

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[5]. 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．

候補対を (x_1, x_2) としたとき，「(x_1, x_2) の周囲単語対は独立である」とするモデル M1 と，「(x_1, x_2) 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表 3.1，表 3.3 の分割表に AIC を適用して変形すると次のようになる．

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

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

モデルの AIC の差 $AIC_{M1} - AIC_{M2}$（以後この値を関連度 AIC と呼ぶ）が 1 より大きいとき独立でないと判定され，システムの出力となる．AIC の値は，周囲単語対の関連性の強さを示しているため，AIC が大きいほど関連語対の関連性が高いことになる．

4. 実験

分布類似度判定において，周囲単語対の異なり数と頻度のどちらを使用した方が，高い精度で関連語対を得ることが出来るのかを確かめる為に，比較実験を行った．

4.1. 実験条件

関連語の抽出には 2 節で述べたシソーラス構築システムを使用し，関連語対の選出工程では 3.1 節，3.2 節で述べた異なり数モデルと頻度モデルを用いて周囲単語対の特徴量を求め，3.3 節で述べた AIC の独立判定によって関連性を判定した．対象コーパスとしては NTCIR1 の学術論文記事 16 万件（テキスト分量で 150M バイト）を使用した．

4.2. 評価法

関連性の評価法としては，人手での評価の他に，評価辞書として日本語 WordNet 辞書，Wikipedia のキーワードリンク辞書を用意して評価を行った．

人手による評価

人手によって抽出した関連語対に関連性があるかどうか主観的に評価をした．関連があるものは 1 点，関連がないものは 0 点, 関連が低い, あるいはどちらともいえないものは 0.5 点の評価点をつけた．

日本語 WordNet 辞書

日本語 WordNet は NICT が開発した WordNet の日本語版である．日本語 WordNet は synset と呼ばれる類義関係のセットでグループ化され，簡単な定義や他の同義語グループとの関係が記述されている．本実験では，単語対が同じ synset に含まれていた場合，関連があると評価した．

Wikipedia キーワードリンク辞書

オンライン百科辞書 Wikipedia では，記事中のキーワードから他の Wikipedia 記事へのリンクが張られており，これをキーワードリンクと呼ぶ．キーワードリンクは記事の記者が指定するリンクであり，記者が主題と関わりがあると判断したキーワードがリンクとして張られている．そこで，キーワードリンクをしている，あるいはされているキーワードには関連性があると想定し，これを評価辞書として利用した．本実験では，単語対同士が片方向でもリンクしているとき関連があると評価した．

4.3. 実験結果

システムの出力した関連語対のうち, AIC の値が上位 1000 位までの関連語対から無作為に 100 件を人手で評価したときの評価値のグラフを図 4.1，AIC の値が上位 1000 位までの関連語対を日本語 WordNet 辞書評価したときの正解数のグラフを図 4.2，Wikipedia 辞書評価したときの正解数のグラフを図 4.3 に示す．

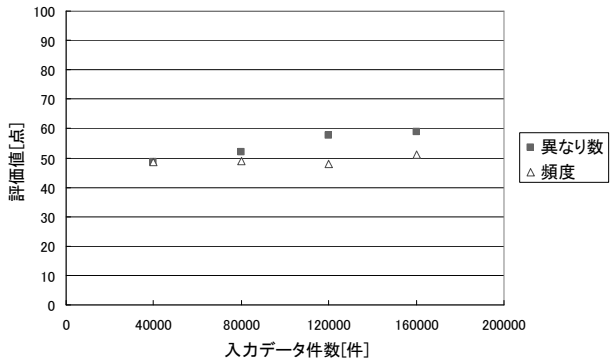

図 4.1 人手評価 (AIC 上位 1000 からランダム 100 件)

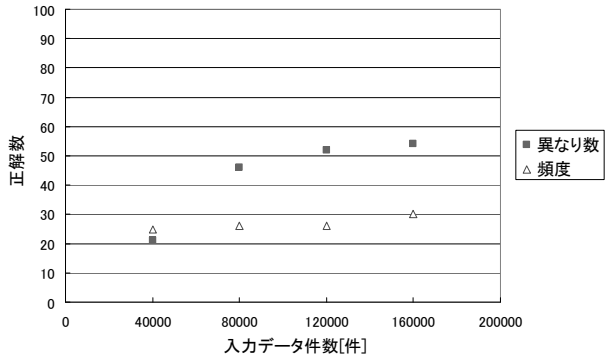

図 4.2 日本語 WordNet 辞書評価 (AIC 上位 1000 件)

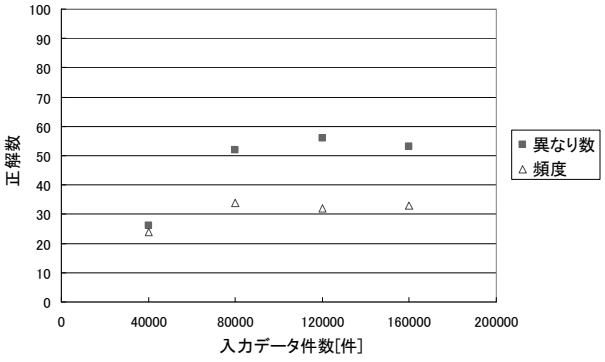

図 4.3 Wikipedia 辞書評価 (AIC 上位 1000 件)

図 4.1, 図 4.2, 図 4.3 より，モデルの比較結果としては，3 つの評価法すべてで異なり数モデルのほうが，頻度モデルよりも高い精度で関連語対を抽出できていることがわかる．また，入力データ件数少ないうちはモデル間の精度に差が見られないが，入力データ件数が増えるにつれて関連語対の精度が増加していることが確認できる．人手の評価と辞書による評価では，精度に大きな差があるが，以上の評価の傾向は一致している．よって，辞書による評価が妥当な評価法として使用できることが確認できる．

5. 考察

異なり数モデルは，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．

異なり数モデルの方が高い精度であった理由を考察する．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．そのため，出現するかしないかのみを考慮する異なり数モデルの方が，有用な文脈の素性を利用できたのではないかと示唆される．

また，抽出した関連語対の周囲単語対に着目したところ，頻度モデルでは周囲単語対が複合語による影響を強く受けている様子が見られた．例えば，異なり数モデルで得られた関連語対の周囲単語対は，関連語対の単語に近接している単語や，係り受け関係にある名詞や動詞などの様々な種類が見られた．しかし，頻度モデルでは，“離散/(サイン,フーリエ)/変換”や“リニア(直流,交流)/モータ”などのように，単語のすぐ隣に連接している周囲単語対が高頻度で出現していた．これは形態素解析による問題であり，複合語として抽出できなかった複合語は複数の単語列に区切って抽出されてしまうために，複合語の構成語を周囲単語対として捉えてしまうことが原因であると考えられる．複合語の構成語も有用な素性ではあるはずだが，それ以外の周囲単語対よりも高頻度で出現しやすいため，偏った関連語対が得られてしまう．また，複合語を構成する名詞は専門用語であることが多いことから[6]，専門用語が辞書に載っておらずに評価されなかったのではないかと考えられる．仮に辞書に載っていたとしても，専門用語同士の関連性は判断が難しいため，関連語として評価されない可能性が高く，一般的な辞書による評価は難しい．

複合語による問題を解決するためには，単語の抽出段階で，複合語をきちんと認識して，分割せずに切り出す必要がある．複合語を抽出する方法はいくつかあるが[6]，完全に取り出すことは難しい．また，コーパスによっては，複合語以外にも，特有の定型文として使用される文字列が，高頻度で出現する[7]．そのため，頻度モデルは，偏った文脈の出現の影響を受けやすいモデルであると考えられる．

6. まとめ

本研究では，分布類似度判定によって関連語対を求める工程において，文脈の出現頻度と異なり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できるか比較するために，當間の頻度モデルの文脈情報の取り方を，異なり数モデルと同じ手法に変更した関連語対の評価方法としては，人手の評価に加え，評価辞書として日本語 WordNet 辞書，Wikipedia のキーワードリンク辞書を用意して評価を行った．結果，異なり数モデルの方が高い性能で関連語対を得ることができ，人手と辞書の評価が一致していることを確かめたことを報告する．

今後の課題としては，複合語の抽出精度を上げることで，複合語による影響を検証することや，他のコーパスや判定手法でも同様の比較実験を行うことで，本実験の結果が一般化できているか確認することが挙げられる．

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あり，サポートに感謝します．

参考文献

[1] 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.

[2] 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.

[3] 住友電工情報システム株式会社．全文検索エンジン QuickSolution.

[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] 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.

[6] 中川裕志，森辰則，湯本紘彰: 出現頻度と連接頻度に基づく専門用語抽出，自然言語処理，19(1)，pp.27-45, 2003.

[7] 増山篤志，梅村恭司，阿部洋丈，岡部正幸．隣接単語で表現した文脈における 高頻度文脈の傾向分析．言語処理学会第 16 年次大会, 2010.