Web 検索エンジンを用いた分野連想語の自動抽出に関する研究

佃　陽平　　森田和宏　　泓田正雄　　青江順一
徳島大学　工学部　知能情報工学科
Email：{Tsukuda,kam,fuketa,aoe}@is.tokushima-u.ac.jp

1.　はじめに

現在，電子化された文書は増大の一途を辿り，それらを計算機によって自動的に処理する技術の必要性が高まってきている．例として，膨大な量の文書を話題分野に応じて自動的に分類する技術があげられる．文書の話題を特定する手段の一つとして，文書内に存在する単語から話題分野を連想する方法がある．そのため，文書の話題分野を連想する単語である分野連想語の自動抽出は重要な研究課題となっている．

従来手法[1]では，予め人手で分野分類した文書から抽出した単語の連想分野を，各分野に対する出現比率によって決定していた．しかし，結果はコーパスに依存し，分野連想語として相応しくないものが抽出される場合が多い．そこで，インターネットの検索エンジンによって収集された文書を利用し，文書中の単語から従来手法によって決定された分野連想語候補を，更に分野連想語としてふさわしいかどうかを判定する手法を提案する．

本稿では，この提案手法を用いた日本語と英語の分野連想語の抽出法について述べ，評価実験をおこなった結果から提案手法の有効性を示す．

2.　分野連想語

2.1.　分野連想語とは

分野連想語とは特定の分野を連想することのできる単語のことをいう．例えば，<baseball>分野を連想できる単語としては「home run」，「pitcher」等の普通名詞，「Alex Rodriguez」，「New York Yankees」等の人名，組織名がある．ただし，「player」，「game」等は<baseball>分野のみを連想できる単語ではないので分野連想語として扱わない．

2.2.　短単位連想語と複合連想語

本研究では，分野連想語となる可能性が高い語に名詞が多いことから，名詞のみに限定して分野連想語を抽出する．日本語の分野連想語の抽出では，文書から形態素解析によって単語を切り出す．この形態素辞書に登録されている語を短単位語と呼び，2 語以上からなる語を複合語と呼ぶ．また，英語においても同様に，1 語のみで構成され意味を持つ名詞と，複数の語で構成されて，それら全体で一つの意味を持つ名詞があることより，それぞれ短単位語，複合語と呼ぶ．例えば，前者は「ピッチャー」，「pitcher」等であり，後者は「高校野球」，「hole in one」等，更に「松井秀喜」，「New York Yankees」等の人名，組織名などにおいても，これにあたるものが存在する．本稿では，これらをそれぞれ短単位連想語，複合連想語として以降を述べる．

3.　分野連想語自動抽出システムの概要

本システムは，文書収集部，単語切り出し部，連想語候補決定部，連想語候補検証部の４つのモジュールで構成される．システムの概要を図１に示す．

3.1.　文書収集部

文書収集部では，分野連想語となる単語を抽出するために必要となる文書を，Web 検索エンジンを利用して収集する．抽出対象となる分野連想語の分野名を検索キーワードとし，Web 検索エンジンによって各分野の文書を収集する．

3.2.　単語切り出し部

単語切り出し部では，収集した文書から連想語候補となる単語を切り出し，各単語について頻度集計する．日本語

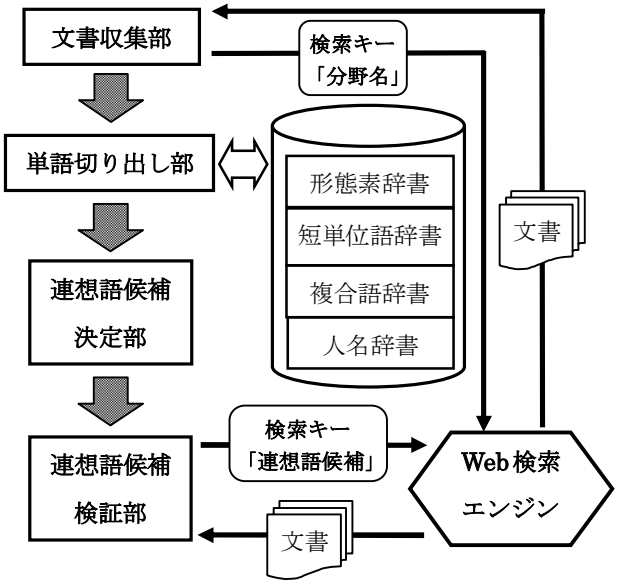

図１：システムの概要

連想語候補となる単語の切り出しには,形態素解析を用いる．また，英語連想語候補となる単語の切り出しにはEDR電子化辞書(英語単語辞書)[3],西洋人名辞書[4]を使用しておこなう.なお,これらの辞書から独自に短単位語辞書,複合語辞書，人名辞書を作成し，次に述べる規則に従い,連想語候補となる短単位語，複合語，人名を切り出す.

3.2.1. 短単位語切り出し規則

短単位語の切り出しは，形態素辞書，EDR 電子化辞書から品詞が名詞である短単位語のみを抜粋した短単位語辞書を用いておこなう．形態素辞書，短単位語辞書に登録されている，名詞である単語を文書から切り出し，切り出された各単語について頻度集計する.

3.2.2. 複合語切り出し規則

複合語の切り出しは，形態素辞書，EDR 電子化辞書から品詞が名詞である複合語のみを抜粋した複合語辞書を用いておこなう.前項で述べた短単位語の切り出しと同様に,形態素辞書，複合語辞書に登録されている，名詞である単語を文書から切り出し,切り出された各単語について頻度集計する.

また,英語には複数の名詞を修飾語として連続させて複合語とする「名詞の形容詞用法」がある.これは,「infield hit」のように名詞が連続する場合に，前の名詞が後の名詞を修飾する形となる文法である.このような形の複合語は複合語辞書には登録されていない場合が多い．そこで,短単位語辞書に登録されている名詞が文中で連続して出現する場合,それら複数の単語を一つの複合語として切り出す.

3.2.3. 人名切り出し規則

人名の切り出しは，形態素辞書，西洋人名辞書を加工した人名辞書を用いておこなう．人名辞書は，西洋人名辞書に登録されている人物名を姓と名で分割して,各自個別に辞書登録して作成している.これは，西洋人名辞書に登録された人名だけでなく,未登録の姓名の組合せからなる人名を抽出するためである.

人名の切り出しは,前項で述べた短単位語辞書による複合語の切り出しと同様に，形態素辞書，人名辞書に登録されている姓，または名が文中で連続して出現する場合，姓名の組合せを一つの人名として切り出す.

また，図 2 に示す「Sharapova」のように未登録である人名が出現する場合,上で述べた手法のみではフルネームである「Maria Sharapova」を取得できない．このような

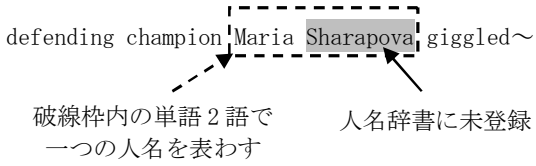

図 2：人名辞書に未登録の人名の例

場合,英文において人名は文中でもキャピタライズされていることから，人名である単語をおおよそ見当できる.例の場合，名である「Maria」に続く「Sharapova」がキャピタライズされていることから「Maria Sharapova」を一つの人名であるとし，切り出す.これにより連想語となる人名が取得できる可能性が高くなる.

3.3. 連想語候補決定部

連想語候補決定部では,単語切り出し部で分野ごとに頻度集計された単語から，各分野の連想語候補を決定する.連想語候補の決定方法として辻の手法[1]を用いる．この手法では,文書から切り出した単語の各文書に出現する頻度をもとに連想語候補を決定する.以下に連想語候補決定手法を示す.

各分野の文書を均一に収集するのは難しい.そこで分野$<C>$に出現する全ての単語の合計頻度を$T(<C>)$とし，単語 w の分野$<C>$の頻度を$F(w,<C>)$で表わす．以後，分野に$<C>$おける単語 w の頻度には次のように正規化した頻度$R(w,<C>)$を使用する.

$$R(w,<C>)=\frac{F(w,<C>)}{T(<C>)}$$

ここで，各分野$<C>$を総称して全分野$<S>$とする.全分野$<S>$における単語 w の頻度$R(w,<S>)$は，全分野$<S>$の$R(w,<S>)$を合計した頻度とする．分野$<C>$における単語 w の集中率$P(w,<C>)$を以下に定義する.

$$P(w,<C>)=\frac{R(w,<C>)}{R(w,<S>)}\geq\alpha$$

ある分野$<C>$において，単語 w が集中するか否かの条件を上の式で判定し，この条件式を満たせば単語 w を分野$<C>$の連想語候補に決定する.条件式を満たさない場合は，単語 w は分野$<C>$の連想語候補でないとする.

The 【Yankees】 had a better first baseman in Moose Skowron. The Red Sox gave up the chance to sign future Hall of Famer Willis Mays, who would go on to hit more career home run.

※網掛け部分の単語が<baseball>の分野連想語

図 3：<baseball>の連想語「Yankees」を含む文書の例

3.4. 連想語候補検証部

連想語候補検証部では，連想語候補決定部で獲得した各連想語候補に対し，Web 検索エンジンを利用して分野連想語として相応しいかどうかを検証する．

一般的に，文書中には話題に関する分野連想語が 1 語しか含まれていないことは珍しく，複数の分野連想語が含まれている可能性が高い．例えば，図 3 のような<baseball>について書かれている文書であれば，文書中に存在する分野連想語が「Yankees」だけでなく「hit」，「home run」などの分野連想語が複数含まれている可能性が高い．以上のことから，連想語候補が分野連想語として相応しいかどうかは，他の分野連想語とのつながりから判定することができる．よって，分野連想語としての相応しさを示す尺度として，連想語候補について収集された文書内に分野連想語が含まれる割合を用い，これを連想語候補の分野連想語としての相応しさを検証する手法とした．

ここで，連想語候補検証処理の詳細について述べる．分野＜C＞における各連想語候補を

$$w_1,\ w_2,\ w_3,\ \cdots\ w_n$$

とし，検証する連想語候補を w_i としたとき，他の連想語候補の集合 X を以下のように表わす．

$$X = \{\ w_j \mid 1 \le j \le n,\ j \ne i\ \}$$

また，w_i を検索キーワードとして Web から収集する文書数を N，それら収集した文書の集合 Y を

$$Y = \{\ y_k \mid 1 \le k \le N\ \},$$

各文書 y_k に含まれる単語の集合を $S(y_k)$ で表わし，$(X \cap S(y_k)) \ne \phi$ となる y_k の個数を I としたとき，連想語候補 w_i の連想語信頼度 $B(w_i)$ を以下に定義する．

$$B(w_i) = \frac{I}{N} \ge \beta$$

ある分野＜C＞において，連想語候補 w_i が連想語であるか否かを上の条件式で判定し，条件式を満たせば連想語候補 w_i を分野＜C＞の連想語とし，満たさない場合は連想語でないとする．

4. 実験

4.1. 評価方法

提案手法による分野連想語の抽出精度を確かめるために，評価実験をおこなった．実験は日本語，英語の分野連想語の抽出に対しておこない，3.3 節で述べた連想語候補決定部における条件式 α の値は，辻[1]による実験をもとに日本語，英語ともに 0.9 とした．また，3.4 節で述べた連想語候補検証部における連想語候補 1 個につき収集する文書数は収集する時間も考慮した上で 20 とし，β の値は日本語，英語ともに閾値決定のための予備実験をおこない決定した．そして，本実験として提案手法の有効性を確かめるために，予備実験と異なるデータを用いておこなった．予備実験，本実験ともに検索エンジンを利用して収集される文書を使用し，予備実験では 5 分野各 50 ファイル，本実験では 10 分野各 100 ファイルを用いた．抽出した分野連想語が分野連想語として相応しいかどうかの判定は人手でおこない，適合率での評価のみをおこなった．適合率を求める式を以下に示す．

$$\text{適合率}(\%) = \frac{\text{システムが抽出した正解分野連想語数}}{\text{システムが抽出した全分野連想語数}} \times 100$$

4.2. 実験結果

4.2.1. 予備実験結果

表 1 にβ値別の日本語，英語の分野連想語の抽出結果を示す．また，表上の「正解」列はシステムが抽出した正解分野連想語数，「不正解」列はシステムが抽出した不正解分野連想語数，「適合率」列は抽出結果の適合率をそれぞれ示している．結果が示すように，日本語，英語共にβ値を 0.2 から 0.3 に上げる際にそれ以降と比べ，システムが抽出した分野連想語として相応しくない単語の数が減っており，適合率も大きく向上している．また，0.3 から 0.4 にβ値を上げる際には，システムが抽出した分野連想語と

表 1：予備実験における抽出結果

	日本語			英語		
	正解	不正解	適合率	正解	不正解	適合率
従来手法	411	579	41.51	338	248	57.63
提案手法						
β=0.1	406	267	60.32	329	220	59.93
β=0.2	398	165	70.69	319	146	68.60
β=0.3	387	116	76.93	294	84	77.78
β=0.4	361	93	79.51	265	57	82.30
β=0.5	329	64	83.71	248	40	86.11

表 2：本実験における抽出結果

	日本語			英語		
	正解	不正解	適合率	正解	不正解	適合率
従来手法	985	1066	48.02	607	506	54.54
提案手法	915	324	73.84	561	193	74.40

して相応しい数が，それまでに比べて大きく減少しており，従来手法と比べてもかなり減少している．以上の 2 点を考慮した上で，今回の実験において $\beta = 0.3$ が最も適していると判断し，本実験に用いる．

4.2.2. 本実験結果

表 2 に本実験での従来手法と提案手法による結果をそれぞれ示す．実験結果から，提案手法が従来手法に比べ分野連想語の抽出結果の適合率が向上し，提案手法の有効性が確認できた．

4.3. 考察

実験をおこなった結果，従来手法に比べ，提案手法による分野連想語の抽出が日本語，英語共に良好な結果が得られた．予備実験の抽出結果に比べ，本実験の抽出結果が日本語，英語ともに精度が下がったことに関しては，予備実験では主にスポーツの分野を対象として実験をおこなったためであると考えられる．スポーツ分野は，Web 検索エンジンの検索上位となる文書が，そのスポーツに関する，リアルタイムのニュース記事が多くを占めるために，内容もその分野に通じるものがほとんどであった．一方，本実験で使用した〈天気〉，〈音楽〉などの分野では，一部関連する語を含むが，文書全体としては対象分野と全く関連性のない文書があった．このことから，Web 検索エンジンで文書を収集する場合，分野名を検索キーワードとするだけでは，分野によって結果が異なることが考えられる．

また，連想語信頼度の定義づけによる問題点から誤判定する場合があった．今回，検証する連想語候補に対する連想語信頼度が条件を満たす場合に分野連想語と判定したが，その中でも互いに連想語信頼度が等しいにも関わらず，分野連想語として相応しい語，相応しくない語がそれぞれ存在した．例として，〈野球〉の連想語である「代打」と「契約更改」をあげて説明する．これらは互いに連想語信頼度は等しくなったが，「契約更改」は〈野球〉の分野連想語とはいえない．これら 2 つの連想語候補の大きな違いとして，互いに検索キーワードとして収集した文書に存在する分野連想語の頻度と種類数にある．「代打」について収集した文書は分野連想語の出現頻度も多く，種類数も豊富であったのに対し，「契約更改」について収集された文書は，収集時期において「契約更改」をおこなった人物名のみが出現し，その他の〈野球〉の分野連想語を含むことはなかった．このことから，連想語を含む文書数のみを考慮した判定方法でなく，収集した文書における分野連想語の出現頻度，種類数を考慮した判定法を用いることで更なる精度向上が望めると考えられる．

英語の特徴の一つである，同表記の単語において複数の品詞を持つ語が存在することが結果に影響することも見受けられた．今回使用した辞書は名詞として機能する単語を，EDR 電子化辞書から全て抜粋して作成されたため，文中では他の品詞として機能している語も名詞として抽出する場合があった．例として，「great MLB」のように後に続く名詞に対して，前の単語が形容詞として機能している語を抽出した．これは，3.2.2 項で述べた名詞の形容詞用法とは異なり，「great」は本来形容詞として名詞を修飾する．このような，複合語を分野連想語として認めてしまうと，形容詞の数だけ「MLB」を含む分野連想語を抽出するが，分野連想語の必要性を考慮した場合あまり意味がない．よって名詞辞書の見直し，もしくは文脈から単語の品詞を判断することが必要となる．

5. まとめ

本稿では，Web 検索エンジンを利用した分野連想語の抽出方法を提案し，提案手法の有効性を評価するための実験をおこなった．

今後の課題は 4.3 節で述べた問題点を解決すること，また，検索エンジンを利用することは多大な時間と労力を要するため，連想語候補検証部における処理に変わる精度向上法の考案が必要となる．

参考文献

[1]辻孝子:“分野連想語の効率的構築法と早期分野決定への応用”，徳島大学博士論文，2000.

[2]辻孝子:“複合語の分野連想語の効率的決定法”，自然言語処理，Vol.7，No.2，pp.3-26，2000.

[3]“EDR 電子化辞書”，(株)日本電子化辞書研究所，1996.

[4]“西洋人名辞書”，日外アソシエーツ株式会社，2001.