統計情報と概念知識を用いたテキスト間の話題特定

福原 知宏・武田 英明・西田 豊明
奈良先端科学技術大学院大学 情報科学研究科
〒 630-0101 奈良県生駒市高山町 8916-5
Tel. 0743-72-5265
E-mail: tomohi-f@is.aist-nara.ac.jp

単語の出現頻度分布を用いたテキスト間の話題特定手法を提案する。テキストの示す話題はテキストを構成する単語に依存すると仮定し、単語の出現頻度分布からテキスト集合中の一般的／専門的話題を特定する。実装システムとして、検索結果をクラスタリングしてテキスト中の話題を特定するシステムを作成、話題特定の実験を行なった。提案手法とシステムの動作例を報告する。

話題特定 自動要約 複数テキスト要約 クラスタリング 統計情報

A Topic Identification Method based on Statistical and Conceptual Information of Words

Tomohiro Fukuhara, Hideaki Takeda, Toyoaki Nishida
Graduate School of Information Science,
Nara Institute of Science and Technology
8916-5, Takayama, Ikoma, Nara 630-0101 Japan

A topic identification method based on the statistical information of words is described. We identify text topics based on the generality of words. To determine the generality of words, we use the kurtosis and the skewness of the distribution of words. We developed a system which identifies general/special topics. Some example and experimental results are discussed.

Topic Identification, Text Summarization, Multiple-Text Summarization, Text Clustering, Statistical Information

1　はじめに

大量のテキスト情報から、目的とする情報を効率的に獲得するための支援が必要である。今日、WWWページや個人のメール・過去に作成したテキストなど、様々なテキスト情報が計算機上に蓄積され・検索されるようになった。一方、膨大な量の検索結果から、目的とする情報を見つけるまでに多くの時間を費すようになった。目的とする情報を獲得するためには、情報を検索できるだけでは不十分であり、検索結果から必要な情報のみを取り出し整理・組織化して提示する機能が必要である。

本研究の目標は、テキスト集合の分類と話題特定による複数テキストを対象とした自動要約の実現である。自動要約とは、文書中に記述された内容を保持したまま冗長性を低減させる処理であり [5]、情報の概要を把握する上で有効である。

本稿では、テキスト集合における一般的／専門的話題の特定手法を提案する。ある観点に沿って収集されたテキスト集合には、テキスト間に共通する一般的な話題と、特定のテキスト集合に共通する専門的な話題が存在する。ユーザにとって話題の一般性を知ることで、テキスト集合の概要把握や、必要に応じて詳細な専門的話題を参照できるようになる。本研究では、ユーザの目的に応じて可変な詳細度で要約を提供することを目的とし、テキスト間の一般的／専門的話題の特定を行なう。

本稿の構成は以下の通り。2節では要約の種類と話題の一般性について述べる。3節では、統計情報を用いた単語の一般性／専門性の計算方法を示す。4節では、実装システムとシステムの動作例を示す。5節では、本研究と関連研究との比較を行なう。6節では、今後の課題について述べる。

2　テキスト要約

テキスト要約には、要約文記述の詳細度により二つの場合分けが考えられる。この場合分けを、トップダウン的要約とボトムアップ的要約と名付け、以下に定義する。

トップダウン的要約

一つの話題に特化した要約。
情報の利用目的は明確。

ボトムアップ的要約

関連する複数の話題を示す要約。
情報の利用目的は明確ではない。

前者の例には、論文・マニュアル・専門記事などが、後者の例には、サーベイ・解説・説明記事といったテキストが相当する。

同じテキスト集合を対象としても、トップダウン的な要約とボトムアップ的な要約は各々異なる。例えば、「風力発電」について述べられたテキスト集合の要約を考える。図 1に、「風力発電」に関する周辺的な話題の関連性を示した。図 1に示されるよう、「風力発電」という話題を中心に、縦方向に一般／専門的話題、横方向に関連話題が存在する。「風力発電」には、「発電コスト」や「地熱発電」などの関連話題が存在する。トップダウン的要約では「発電コスト」や「設置状況」といった「風力発電」に関する専門的話題を展開するのに対し、ボトムアップ的要約では「地熱発電」や「環境問題」といった一般的話題を展開する。

要約文の生成において、話題の一般性を重視するか専門性を重視するかは、状況に応じて異なる。ユーザが、求める対象について背景知識を持たない場合、対象自体と対象に関連する話題を説明する必要がある。逆にユーザが十分な背景知識を持つ場合、関連する話題は冗長となり、対象に関する話題を深く掘り下げて説明しなければならない。

以上から、システムはユーザの要求に応じてトップダウン／ボトムアップ二種類の要約を提供しなければならない。このため、テキスト中の話題の一般性を特定する必要がある。以下では、テキスト中の話題の一般性を特定する手法について述べる。

3　話題特定手法

話題の一般性を特定する手法について述べる。まず、話題の一般性を以下に定義する。

一般的話題

テキスト集合中で全体的に出現する記述。

専門的話題

テキスト集合中で部分的に出現する記述。

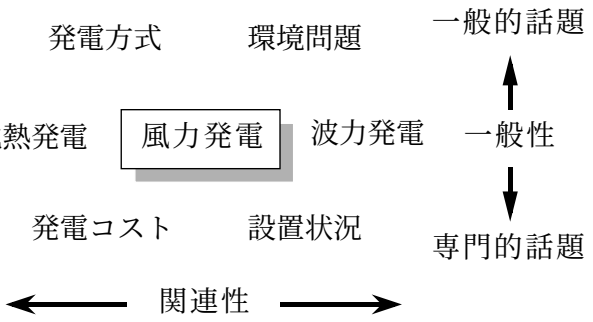

図 1: 話題の一般性／関連性

一般的話題は、テキスト集合全体に共通する話題であり、図 1の例では「発電方式」、「環境問題」が相当する。専門的話題は、テキスト集合中、特定の部分集合に共通する話題であり、図 1の例では「発電コスト」、「設置状況」が相当する。

一般的話題はテキスト集合全般に出現する話題だが、専門的話題はテキストの部分的集合として現れる話題である。このため、専門的話題を特定するためには、テキスト集合を部分集合として分類しなければならない。本稿では、テキスト集合を部分集合に分割するため、クラスタリングを行なう。クラスタリング結果から、各クラスタ固有の話題を専門的話題、クラスタ間に共通する話題を一般的話題として特定する。

話題の一般性は、クラスタを構成する単語の出現頻度を用いて計算する。単語出現頻度の分布形状から、単語の一般性／専門性を判定する。以下、各々の過程について述べる。

3.1 テキスト数量化

テキストの数量化にはVSM(Vector Space Model)[3]を用いた。VSM では、テキストの特徴は $tfidf$ 値により表される。$tfidf$ 値は、あるテキストにおける単語の特徴量を示す値であり、あるテキストに多く出現する単語は大きな値を、逆に多くのテキストに出現する単語は低い値を取る。テキスト i における j 番目の単語 $term_{ij}$ の $tfidf$ 値 w_{ij} は以下の式より求まる。

$$w_{ij} = tf_{ij} \cdot \log(\frac{N}{df_j}) \quad (1)$$

tf_{ij} はテキスト t_i における単語 $term_j$ の出現数を表す。N は単語総数を、df_j は $term_{ij}$ の出現するテキスト数を表す。

テキストはベクトルとして表現される。テキスト $\boldsymbol{D}_i$ は、各単語の重み w_{ij} からなるベクトル

$$\boldsymbol{D}_i = (w_{i1}, w_{i2}, \cdots, w_{iN}) \quad (2)$$

として表される。

テキスト間の類似度はベクトル間の内積として計算される。テキスト p と q の類似度 $sim(\boldsymbol{D}_p, \boldsymbol{D}_q)$ は、

$$sim(\boldsymbol{D}_p, \boldsymbol{D}_q) = \sum_{i=1}^{N} w_{pi} \cdot w_{qi} \quad (3)$$

と表される。

3.2 クラスタリング

テキスト集合に含まれる個々の専門的話題を特定するため、テキスト集合を分類する。本稿では階層クラスタリングを用いてテキストを分類する。クラスタリングの結果、テキスト集合は各クラスタ固有の専門的話題を含むクラスタに分類される。

クラスタリング手法には最遠距離法を用いた [4]。最遠距離法はクラスタ間の距離を最も遠い要素間の距離に取る。これにより、クラスタが一つのクラスタだけにまとまる問題 (鎖効果) を避けられる。また、最近距離法に比べ、より多くのクラスタを生成できる。

階層クラスタリングのアルゴリズムを以下に示す。

0. 各テキストをクラスタとする
1. クラスタ間の距離を計算する
2. 最も距離の小さな組を新たなクラスタとして統合する
3. クラスタ数が 1 になるまで 1. に戻り繰り返す

クラスタ間の距離を次式で示す。

$$d_{XC} = \frac{1}{2}d_{XA} + \frac{1}{2}d_{XB} - \frac{1}{2}|d_{XA} - d_{XB}| \quad (4)$$

上式は、クラスタ A, B が統合され、クラスタ C が生成される時の任意のクラスタ X とクラスタ C との距離 d_{XC} を表す。

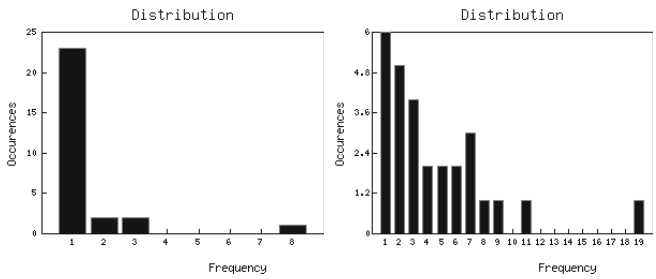

図 2: 単語の出現頻度分布

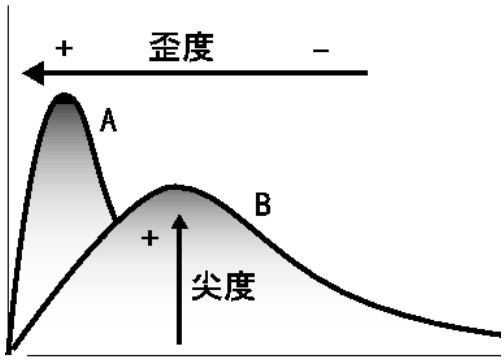

図 3: 尖度と歪度

3.3 単語の一般性特定

テキストの全体集合、及び各クラスタに出現する単語の頻度分布から単語の一般性を判定する。一般的／専門的話題を示す単語を各々、一般語／専門語とする。本稿では、話題の一般性に関して以下の仮定を置く。

一般的話題

一般語を多く含む文または文の集合

専門的話題

専門語を多く含む文または文の集合

単語の出現頻度分布が話題の一般性を判定する指標となる例を示す。

図 2 は、テキスト集合中の二つの単語の出現頻度分布である。図中、x 軸 は単語の出現回数を、y 軸は単語を含むテキストの出現頻度を表す。右の図は、テキスト集合に広範に出現する単語の分布を、左の図は、特定のテキスト集合に瀕出する単語の分布を表す。右の図では、分布が x 軸に従って緩やかな形を示すのに対し、左の図では、x 軸の低い値で突出した形を示す。このように、広範なテキストに出現する単語と、特定のテキスト集合に出現する単語では、出現頻度分布の形状に違いが生じる。本稿では、分布形状の比較により単語の持つ話題の一般性を特定できると考える。

以上の仮説に基づき、単語の一般性を分布の尖度と歪度を用いて特定する。尖度は分布の尖り具合を、歪度は分布の偏りを表す指標である。尖度と歪度の関係を図 3 に示す。図 3は、分布形状の異なる A, B 二つの分布を示す。分布 A は専門語の分布を、分布 B は一般語の分布を表す。専門的話題 (A) では尖度・歪度とも大きな値を取ることがわかる。一方、一般的話題 (B) は A に比べ尖度・歪度とも低い値を取る。以上から、単語の示す話題が専門的になるのに従って尖度・歪度とも高い値を取る。

尖度 α_4 、歪度 α_3 は以下の式で表される。

$$\alpha_3 = \frac{1}{n}\sum_{i=1}^{n}\frac{(x_i-\bar{x})^3}{s^3}$$

$$\alpha_4 = \frac{1}{n}\sum_{i=1}^{n}\frac{(x_i-\bar{x})^4}{s^4} - 3$$

n はクラスタ数を、x_i は各クラスタ中の単語の出現頻度を、$\bar{x}$ はクラスタ中の単語出現頻度の平均値を表す。s は標準偏差 $s = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2}$ である。

尖度と歪度を用いた単語の一般性／専門性の判定基準を以下に示す。単語の一般性／専門性を α_3, α_4 を用いて定義する。

$$\mathcal{G} = \alpha_3 + \alpha_4 \tag{5}$$

$\mathcal{G}$ は単語の一般性を示す。$\mathcal{G}$ は、尖度および歪度とも同符合の時、高い値を取る。ここで、閾値を τ_a, τ_b に取ると、単語の一般性は

$$\mathcal{G} = \begin{cases} \text{一般的} & (\alpha_3 < \tau_a, \quad \alpha_4 < \tau_b) \\ \text{専門的} & (\alpha_3 > \tau_a, \quad \alpha_4 > \tau_b) \end{cases} \tag{6}$$

と特定できる。$\alpha_3 - \tau_a, \alpha_4 - \tau_b$ が共に正の時、単語の分布形状は図 3の A(専門的話題) に近付く。反対に、$\alpha_3 - \tau_a, \alpha_4 - \tau_b$ が共に負の時、分布形状は図 3 の B(一般的話題) に近付く。

以下では、これまでに述べた考えを実装したシステムについて述べ、実装システムでの実験結果について述べる。

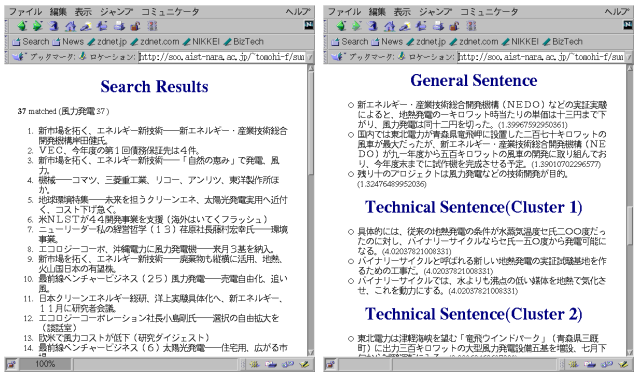

図 4: 実装システム

No.	タイトル
1	風力・地熱発電 - 容量の確保が課題
	風力・地熱発電 - 地球環境を汚さず
2	新市場を拓く、エネルギー新技術 - 「自然の恵み」で発電、風力。
	NEDO、風力発電機 3 基、宮古島に増設完了。
3	ユナイテッドケミカル、風力発電機を輸入販売。- デンマークから。
	最前線ベンチャービジネス - 風力発電売電自由化、追い風。

表 1: クラスタ構成

4 実験

本節では実装システムを用いた実験結果について述べる。システムは新聞記事から検索を行ない、検索結果をクラスタリングする。次に、クラスタリング結果から、クラスタ全体を代表する一般語、および各クラスタを代表する専門語を特定する。最後に、特定された単語を用いて一般的／専門的話題を代表する文を抽出する。

4.1 実装システム

実装システムの概要について述べる。システムは、与えられた検索式中の単語を元に検索し、検索結果をクラスタリング、クラスタ全体および各クラスタを代表する文を抽出する。

図 4 に実装システムの画面を示す。左の図はクラスタリングの結果を、右の図は抽出された代表文を表す。

システムは以下のモジュールにより構成される。

検索モジュール

検索式中の単語をタイトル／本文に含むテキストを検索する。検索データには日本経済新聞 95 年度版を用いた。

クラスタリングモジュール

検索結果を最遠距離法に基づき階層クラスタリングする。

一般性特定モジュール

クラスタリング結果からクラスタ全体を代表する一般語／各クラスタ固有の専門語を特定する。

代表文抽出モジュール

特定された一般語／専門語から、一般的／専門的話題を示す代表文を抽出する。

4.2 クラスタリング

テキストを検索し、クラスタリングした例について述べる。「風力発電」という単語で検索、クラスタリングした結果を表 1に示す。

表 1 は、37 記事中 28 記事を含む 8 クラスタを生成した時点での上位 3 クラスタのタイトルを示す。表中の番号は、生成されたクラスタの番号である。各クラスタのタイトルは、クラスタ中、最も類似度の高い組を示す。表中の各クラスタを構成するテキスト間の類似度は各々、0.67, 0.56, 0.54 である。なお、類似度の計算式は、式 (3), (4) による。

表中の各クラスタを要約すると以下のようになる。

1. 風力・地熱発電の解説記事。
   具体的には、発電容量、発電コスト。
2. 風力発電の実証試験の報告。
   具体的には、青森県竜飛岬、沖縄県宮古島の発電機の設置状況。
3. 風力発電に取り組む企業の紹介。
   具体的には、発電機の輸入販売、ユナイテッド社の納入事例。

要約の作成は、実際にテキストを読み、テキスト間に共通して述べられている記述を手作業で抽出した。

例では、クラスタリング結果の最大類似度と最小類似度は各々、0.67, 0.23 である。最大類似度は、

一般性	単語	尖度 α_4	歪度 α_3	$\mathcal{G}$
一般	国内	-1.84	0.00	-1.84
	機構	-1.58	-0.20	-1.79
	実験	-1.61	-0.17	-1.78
	全国	-1.58	0.08	-1.50
	開発	-1.44	0.09	-1.35
専門 (1)	バイナリ	3.14	2.27	5.41
	水蒸気	3.14	2.27	5.41
	サイクル	3.14	2.27	5.41
	容量	3.03	2.23	5.25
	風車	2.72	2.11	4.82
専門 (2)	増設	2.90	2.18	5.08
	予測	2.20	1.90	4.10
	制御	1.51	1.75	3.26
	インド	1.23	1.54	2.77
	設備	1.06	1.42	2.48
専門 (3)	日本食	3.14	2.27	5.41
	ユナイテッド	3.14	2.27	5.41
	納入	2.90	2.18	5.08
	エコロジー	2.84	2.15	4.98
	小島	2.84	2.15	4.98

表 2: 尖度・歪度・$\mathcal{G}$

最初のクラスタを生成した時点でのテキスト間の類似度を、最小類似度は、最後にクラスタに生成・統合した時点でのクラスタ・テキスト間の類似度を表す。

クラスタリングの結果、部分的に共通する話題が含まれるテキストがクラスタとして統合された。クラスタリング結果の類似度は、テキスト数に反比例する。テキストを多く含むほどクラスタリング結果の類似度は減少する。この結果、本来専門的であるはずのクラスタ中の話題は一般的になる。例では、クラスタ中のテキスト間には部分的に共通する記述が存在したが、クラスタを構成するテキストの増加に伴い、クラスタの専門性は減少する。

4.3 一般語／専門語の特定

本提案手法を用いて特定された一般語／専門語の例を示す。対象とするテキスト集合は、「風力発電」の検索結果をクラスタリングした集合である。この集合から、クラスタ間に共通する一般語と各クラスタ固有の専門語を抽出した。

各クラスタから抽出した単語と、尖度・歪度・$\mathcal{G}$ の各数値を表 2 に示す。 単語の一般性には (5) 式を、式 (6) の閾値には $\tau_a = 1, \tau_b = 0$ を用いた。「専門」の左の括弧内は、クラスタの番号を示す。それぞれ、$\mathcal{G}$ 値の上位 5 単語を抜き出した。

$\mathcal{G}$ 値を用いた単語の話題特定能力の結論は以下の通り。$\mathcal{G}$ 値は専門語に対して識別力を持つが、一般語の識別力に欠ける。

専門語の抽出では、$\mathcal{G}$ 値は識別力を示す。$\mathcal{G}$ 値により特定された専門語はクラスタ固有の話題を示している。例としてクラスタ (1) を取り上げる。クラスタ (1) 中の専門語である「バイナリ」、「水蒸気」、「サイクル」の各単語は「地熱発電」に関連する単語である。また、クラスタ 1 の話題には、「風力発電」とともに「地熱発電」が含まれるため、これらの単語はクラスタ固有の単語と考えられる。この他、「容量」は「発電容量」に、「風車」は「風力発電」に関連する単語である。以上から、$\mathcal{G}$ 値によって特定された専門語はクラスタ固有の話題を識別する能力を持つ。

一方、一般語の特定では、$\mathcal{G}$ 値を用いた方法では十分な識別力を示さない。表 2中の一般語は、各クラスタ間に共通する話題である「風力発電」に直接的に関連する単語ではない。「キロワット」、「風力」といった「風力発電」に関連する単語は、表 2中の単語より下位に出現する。出現頻度順では、「風力」、「キロワット」という単語は多くのクラスタに含まれるため、一般的な単語として上位に出現する。しかし、$\mathcal{G}$ 値順では下位になる。このことは、単語の出現頻度分布の形状だけでは単語の持つ一般性を判定できないことを意味する。

一般語の判定には分布形状の他、単語の出現頻度といった数値を用いる必要がある。$\mathcal{G}$ 値は単語の分布形状のみから一般語を判定するが、分布形状に加え出現頻度を用いることで、一般語の識別精度の改善を期待できる。

4.4 代表文の抽出

特定された一般語／専門語から、一般的／専門的話題を代表する文の抽出を行なった。表 3に、一般的／専門的話題を代表する文を示す。「専門」の左の括弧内はクラスタの番号を、各文に付された括弧内の数字は代表文の得点を表す。文の得点は、各文に含まれる一般語／専門語の $\mathcal{G}$ 値の平均を用いた。以下に文 S_i の得点 $score(S_i)$ の計算式を示す。

$$score(S_i) = \frac{1}{N_i} \sum_{term_i \in S_i}^{N_i} |\mathcal{G}(term_i)|$$

話題	代表文
一般	新エネルギー・産業技術総合開発機構（ＮＥＤＯ）などの実証実験によると、地熱発電の一キロワット時当たりの単価は十三円まで下がり、風力発電は同十二円を切った。(1.40)
専門 (1)	具体的には、従来の地熱発電の条件が水蒸気温度セ氏二〇〇度だったのに対し、バイナリーサイクルならセ氏一五〇度から発電可能になる。(4.02)
専門 (2)	東北電力は津軽海峡を望む「竜飛ウインドパーク」（青森県三厩町）に出力三百キロワットの大型風力発電設備五基を増設、七月下旬から試験運転に入る。(2.21)
専門 (3)	風力発電機を輸入販売するエコロジーコーポレーション（東京、小島社長）は十月二十六日から沖縄・宮古島で開催する「風のサミット」に、風車模型を出展する。(3.56)

表 3: 一般的／専門的話題の代表文

N_i は S_i に出現する単語総数を、$term_i$ は S_i に含まれる単語を表す。$\mathcal{G}(term_i)$ は $term_i$ の $\mathcal{G}$ 値である。

抽出された文は、各クラスタ／クラスタ全体を代表する話題である。本稿では、文に含まれる一般語／専門語の量と文の示す話題に相関関係を仮定している。このため各文には、表 2中の語が含まれる。

手作業で作成した要約と、表 2中の代表文とを比べると、クラスタ (1), (2) で特定された代表文と手作業で特定した要約文とでは、大きく異ならない結果となった。一方、クラスタ (3) は「輸入販売」の単語のみが該当する結果となった。この例では、手作業の要約と統計情報から得た代表文に大きな違いは見られなかった。しかし、今回、比較したデータは 3 つのクラスタについてのみであり、他のクラスタおよび検索結果を扱っていない。今後、より多くのデータに対し評価を行なう必要がある。

5 関連研究

本節では、提案手法と関連研究との比較を行なう。これまで、テキスト間の話題特定について、多くの研究がなされてきた。一例を挙げると、テキストの分析に基づき単語やフレーズの特徴的な表現から共通的話題を特定する手法 [6] や、単語間の関連をグラフで表し、グラフの共通ノード・相違ノードの比較から共通的・相違的話題を特定する手法 [1] などがある。

以下では、話題特定と自動要約に関する先行研究との比較を行なう。

5.1 テキストの話題特定

テキストの話題特定に関する先行研究として、Mani らの研究がある。Mani らは、複数テキストからの共通的／相違的話題の特定手法を提案した [1]。Mani らは、テキスト間に共通する話題をグラフの比較により特定する。単語間の概念関係を意味ネットワークを用いて特定、単語間にリンクを生成しグラフを構築する。テキスト間の共通点／相違点は、同じ話題で生成したグラフを比較することで特定する。

これに対し本研究では、単語の分布情報を用いて一般的／専門的話題を特定する。Mani らの特定する共通性／相違性は、本研究の一般性／専門性とは異なるものである。「共通性」と言った場合、内容的に同質か否かが議論の対象となるが、「一般性」と言った場合、テキスト集合中、どれだけ話題が分布しているかが問題となる。

5.2 複数テキストの自動要約

複数テキストの自動要約に関する先行研究として、難波らの研究、McKeown らの研究を挙げる。

難波らは学術論文の参照関係を用いた要約手法を示した [7]。難波らは、論文を参照する際出現する手がかり語から参照目的を特定、複数テキストの要約を行なう。

これに対し本研究では、明示的に参照情報が示されないテキストを扱う。学術論文では、テキストの関係が参照関係として明示的に出現するが、通常のテキストにテキスト間の関係を示す情報が現れるのは稀である。このため、テキスト間に共通するテキ

ストを明らかにする必要がある。本研究ではテキストのクラスタリングから、類似テキストを共通集合として分類する。

McKeown らは、情報抽出に基づく要約システムを示した [2]。システムは、テロ事件について述べられた複数の新聞記事から、事件の場所や日時、被害者数などを抽出する。システムは、抽出した情報を比較することで、記事間に共通／相違する情報を示す。

これに対し、本研究では統計情報を用いた話題特定を行なう。McKeown らの情報抽出に基づくアプローチでは、文中に出現する手がかり語に依存するため、扱うテキストの分野は限定される。本研究では、クラスタ中に出現する単語の分布形状を元に話題の一般性を判定する。テキスト中に出現する単語の統計情報を用いることで、扱うテキスト分野を限定しない。今後、WWW に代表される雑多な情報環境で要約を行なう際、分野を限定しない要約が必要となる。テキストの統計情報を用いることで、分野の限定を避けられる。

6 今後の課題

今後の課題として、単語の概念知識を用いた話題特定がある。本稿では、テキストの統計情報のみを用いた話題特定手法を示した。統計情報を用いることで、テキスト間に共通する話題の概要を把握できる。しかし、統計情報はテキスト間の話題を把握する上で近似でしかない。標本数が少ない場合や、単語や文の意味といった統計情報だけでは対応できない場合も存在する。

ここでの概念知識とは、単語間の包含・同値・共起・般化／特化関係などである。「風力発電」の例では、上位概念に「発電方式」、下位概念に「風車」、共起概念に「波力発電」などが存在する。こうした概念間の関係を用いることで、テキスト中の話題が具体的なのか抽象的なのか、専門的なのか一般的なのか特定できる。

また、単に概念知識を利用するだけでなく、新たに概念知識を構築することも考えられる。実装システムでは、検索式として与えられた単語からテキストを検索し、検索結果をクラスタリング、テキストを複数のクラスタに分割した。ここで、検索式中の各単語とクラスタ中の専門語との間には関係が生じる。この関係から単語間の関係を概念知識として自動的に構築することも考えられる。

7 おわりに

本稿では、単語の統計情報を用いたテキスト間の話題特定手法を示した。要約の種類には、専門的話題を説明するトップダウン的要約と、一般的話題を説明するボトムアップ的要約の二種類が存在する。これらの要約を提示するため、話題の一般性を特定するための手法を提示した。話題の一般性はテキストに出現する単語に現れると仮定し、単語の一般性の特定手法を示した。

参考文献

[1] Mani, I., Bloedorn, E. *Multi-document Summarization by Graph Search and Matching.* Proceedings of the 14th National Conference on Artificial Intelligence, pp.622-628(1997)

[2] McKeown, K., Radev, D.R. *Generating Summaries of Multiple News Articles.* In Proceedings of ACM-SIGIR'95, pp. 74-82(1995)

[3] Salton, G. Automatic text processing: the transformation, analysis, and retrieval of information by computer. Addison-Wesley(1989)

[4] 奥村 忠一, 久米 均, 芳賀 敏朗, 吉沢 正. 多変量解析法. 日科技連出版社, pp.391-410(1973)

[5] 奥村 学, 難波 英嗣. テキスト自動要約の現状と課題. 北陸先端科学技術大学院大学 情報科学研究科 Research Report, IS-RR-98-0010I(1998)

[6] 船坂 貴浩, 山本 和秀, 増山 繁. 冗長度削減による関連新聞記事の要約. 電子情報通信学会 自然言語処理研究会, NLC96-15(1996-07)

[7] 難波 英嗣, 奥村 学. 論文間の参照情報を考慮した学術論文要約システムの開発. 言語処理学会第 4 回年次大会. pp.638–641(1998)