3A2-03

医療分野における単語類似度を利用した話題語抽出方法

Extraction of Topics using Similarity Scores in the Medical Domain

日比野 哲也[*1] 山本 けい子[*2] 田村 哲嗣[*3] 速水 悟[*3]
Tetsuya Hibino Keiko Yamamoto Satoshi Tamura Satoru Hayamizu

[*1]岐阜大学大学院工学研究科
Graduate School of Engineering, Gifu University

[*2]岐阜大学産官学融合センター
Collaborative Center for Academy/Industry/Government, Gifu University

[*3]岐阜大学工学部
Faculty of Engineering, Gifu University

This paper describes how to extract the topic word in a sentence in the medical domain. Existing keyword extraction methods can extract only the keyword that appears in the text. This proposal method produces a way to extract not only topic word included in the sentence, but also topic word not included. This feature is realized by using the sum of similarity scores between words included in the sentence and words not included. The result of some experiments for comparing to TF × IDF shows the advantage of this method that extracts words not included in the sentence.

1. はじめに

近年，マルチメディア情報処理技術の進歩などにより，音声認識，手書き文字認識，かな漢字変換システムなどのような入力支援システム，あるいはパッセージ (部分テキスト) 検索 [長沼 05] のような高度な検索システムの研究が進んでいる．このようなシステムにおいては，短い入力部分から次の入力候補を推測したり，パッセージの特徴ベクトルをパッセージに含まれる単語の重要度により求めることがある [Salton 83].

現在，テキスト中の語の重要度を表現する方法として，TF × IDF や，χ^2 値 (例えば [東京 91]) を用いる方法 [松尾 02] などが用いられている．しかし，これらの手法を用いて実現できるのはテキスト文中に出現する単語のスコア付けである．一方，上に挙げたような入力支援システム，パッセージ検索システムにおいては入力文が非常に短いために，その内容を適切に表すような重要語が入力文中に含まれていないことがある．

本稿では，文中にある単語をスコア付けする重要語抽出と区別して，文中にない単語も含めて抽出できる，話題語抽出の手法を提案する．また，既存の重要語抽出手法との比較実験を行い，提案手法により入力文中に含まれていない話題語が抽出される様子を示す．なお，比較実験では提案手法が特定の話題を持つ文に対して有効であることを示すため，医療分野というドメインを設定して行った．

2. 話題語抽出システム

本研究で提案するシステムの流れは以下のようになる (図 1).

1. 入力文からの名詞抽出
2. 文中名詞によるウェブからのコーパス収集
3. コーパスからの言語モデルの構築
4. 文中名詞との類似度によるスコア付け
5. 標準病名マスターによるフィルタリング

連絡先: 日比野哲也，岐阜大学大学院 工学研究科 応用情報学専攻，〒501-1193 岐阜市柳戸 1-1，thibino@hym.info.gifu-u.ac.jp

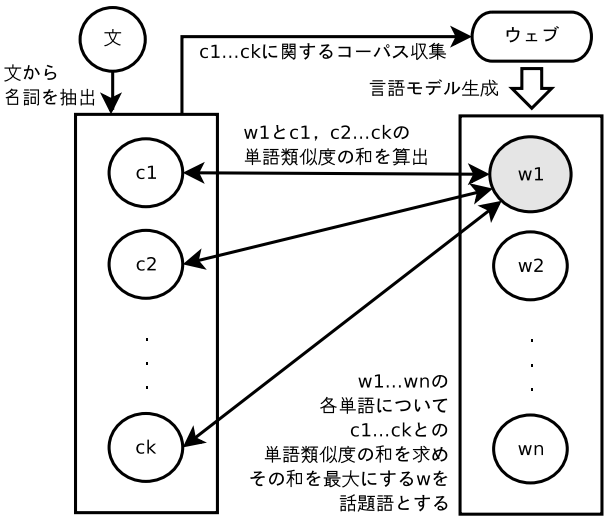

図 1: 提案手法の流れ

2.1 入力文からの名詞抽出

文中の重要語を抽出するために入力文から名詞を抽出する．形態素解析には茶筌 [松本 00] を用い，また，医療用語辞書として「ICD10 対応電子カルテ用標準病名マスター」[財団 06] を用いる．

2.2 文中名詞によるウェブからのコーパス収集

文中の名詞に関連するコーパスを収集するために，文中の各単語をキーにしてウェブ検索を行い，ウェブページを収集する．ウェブ検索エンジンには Google[*1] を用いる．また，単語 w について収集する文書数 m_w を文中名詞の出現文書数に反比例させるように (1) 式を用いて決定する．

$$m_w = M \times \frac{f(n_w)}{\sum_u f(n_u)}, \quad f(n_w) = \frac{N}{\frac{N}{g} + n_w} \qquad (1)$$

ここで，M は収集するウェブページの合計数，n_w は単語 w の出現文書数である．(1) 式を用いて求められた収集するウェ

∗1 http://www.google.co.jp/

ブページ数は，出現文書数が少なければ多く， 出現文書数が多ければ少なくなり，その影響は重み g により決定される．

このような操作を行うのは，出現文書数が多い単語ほど一般的な単語であるといえるので，文意に関係が浅い可能性が高く，逆に，出現文書数が少ない単語は専門性の高い単語である可能性が高いと考えられるからである．

2.3 コーパスからの言語モデルの構築

言語モデルには単語-文書間のベクトルモデルを用いた．単語に $w_1 \cdots w_n$ について文書 $d_1 \cdots d_l$ から作成した特徴ベクトルを並べて作った行列は (2) 式のようになる．a_{ij} は文書 d_i における単語 w_j の TF × IDF 値である．

$$\begin{array}{c} \\ d_1 \\ d_2 \\ \vdots \\ d_l \end{array} \begin{array}{c} \begin{array}{cccc} w_1 & w_2 & \dots & w_n \end{array} \\ \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{l1} & a_{l2} & \dots & a_{ln} \end{pmatrix} \end{array} \quad a_{ij} = TF_{ij} \times \log\left(\frac{N}{DF_j}\right) \tag{2}$$

2.4 文中名詞との類似度によるスコア付け

単語のスコアは，文中単語との類似度の和により付与する．単語 w_i の特徴ベクトルを $\boldsymbol{w_i}$，単語 w_j の特徴ベクトルを $\boldsymbol{w_j}$ とした時の w_i と w_j の類似度 $\mathrm{sim}(w_i, w_j)$ を (3) 式で定義する．

$$\begin{aligned} \mathrm{sim}(w_i, w_j) &= \cos(\boldsymbol{w_i}, \boldsymbol{w_j}) \\ &= \frac{\boldsymbol{w_i} \cdot \boldsymbol{w_j}}{|\boldsymbol{w_i}||\boldsymbol{w_j}|} \end{aligned} \tag{3}$$

この類似度を用いて，単語 w のスコア $s(w)$ を (4) 式で得ることができる．なお，$c_1 \cdots c_k$ は文中から抽出した名詞群である．

$$s(w) = \sum_{i=1}^{k} \mathrm{sim}(w, c_i) \tag{4}$$

2.5 標準病名マスターによるフィルタリング

本稿の実験では特定の話題領域についての抽出精度の評価を行うために，対象領域を医療分野に限定して行った．そのため，医療分野というドメインにおいて抽出精度を向上させるために，辞書として用いた標準病名マスターの中に存在する単語以外を除去する処理を行った．

3. 評価実験

3.1 実験対象と実験方法

本研究では，提案手法と TF × IDF で比較実験を行った．病名マスターに含まれているものから無作為に選んだ 8 病名についての文を対象とし，各病名について病名を含む文と含まない文の合計 16 文について実験を行った．例として脳梗塞についての文で，病名を含むものを図 2 に，病名を含まないものを図 3 に示す．

> 脳梗塞が起こった場所によって症状は様々ですが、よく知られた症状は麻痺、しびれ、構語障害、などです。

図 2: 病名「脳梗塞」を含む脳梗塞についての文

本稿では，これらの各病名についてのスコア上位 5 単語について,「関連性がない」,「関連性がある」,「病名そのもの」で

> 心臓や心臓を出てから脳に至る前の血管の中で血液が固まった血栓が出来て、これが血液の流れに乗って脳の血管に入り込んで脳の血管をつめてしまう状態です。

図 3: 病名「脳梗塞」を含まない脳梗塞についての文

表 1: 実験結果の F 値

	病名を含む文		病名を含まない文	
	F_1	F_2	F_1	F_2
提案手法	0.375	0.375	0.2	0.25
TF × IDF	0.7	1.0	0.375	0.0

ある単語の数をそれぞれ n_0，n_1，n_2，抽出された総単語数を n，実験対象とした文の数を m としたときの F 値，

$$F_1 = \frac{n_1}{n} \tag{5}$$

$$F_2 = \frac{n_2}{m} \tag{6}$$

を評価基準として用いる．F_1 は抽出された単語が文の内容に関連するものである確率を示し，F_2 は文で話題にしている病名が上位 5 単語の中に抽出される確率である．

3.2 実験結果と考察

実験結果の F 値を表 3.2 に示す．提案手法は，抽出した単語が対象とする疾患に関連がある単語である割合を示す F_1 値において TF × IDF よりも劣った結果であった．

しかし，病名を抽出できた割合である F_2 値においては，文中に病名を含まない文で TF × IDF が一つも病名を抽出できていないことに対し，8 病名中，2 つの本文中に出現しなかった正しい病名を抽出できた．この結果は，本文中に話題語を含まない文からも適切な話題語を抽出できるという提案手法の特長を示している．

抽出結果の例として，図 2 の解析結果を表 3.2 に，図 3 の場合の解析結果を表 3.2 に示す．この場合には，スコアのもっとも高い単語に「よう」を抽出しており，また，表 3 の結果においては 2 番目にスコアの高い単語に「心筋梗塞」という誤った単語を抽出している．しかし，表 3.2 の結果においては図 3 の文中に登場しなかった「脳梗塞」という単語を抽出できており，文中に登場しない単語を抽出するという点では一定の成果を得ることができた．

図 3 の解析における重み g の変化に対する病名のスコアの変化を図 4 に示す．$g = 1$ の時に比べて $g = 500$ や $g = 1,000$ の時の方が平均と比較しても若干良い結果が得られていること

表 2: 図 2 の文の解析結果

TF × IDF		$g = 500$	
スコア	単語	スコア	単語
6.98	構語障害	2.60	よう
5.64	症状	2.21	構語障害
3.69	麻痺	2.16	腫瘍
3.57	脳梗塞	2.04	脳梗塞
3.46	しびれ	2.01	難聴

表 3: 図 3 の文の解析結果

TF × IDF		$g = 500$	
スコア	単語	スコア	単語
9.48	血管	3.74	よう
7.77	脳	2.23	心筋梗塞
6.72	心臓	2.17	脳梗塞
6.01	血液	2.11	外傷
3.78	血栓	1.85	出血

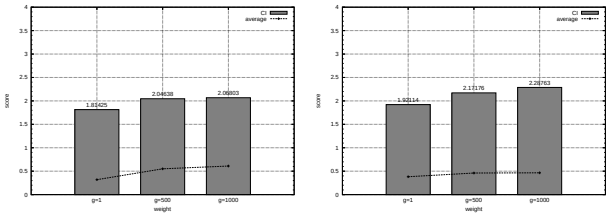

図 4: 重み g と病名のスコアの関係 (左: 図 2 の文，右: 図 3 の文)

が分かる．提案手法における出現文書数に応じてコーパス収集数を増減させる方法が一定の成果を挙げていることが分かる．

なお，提案手法が TF × IDF に対して抽出精度が低い原因としては以下のような原因が考えられる．

- コサイン尺度を用いたためにベクトルの大きさが類似度に与える影響が小さくなった．そのため，頻度の面で極端に出現傾向が異なる単語が類似語としてスコア付けされることがあった．ユークリッド距離などの他の尺度を用いた類似度算出についても検討する必要がある．
- 文中から抽出した名詞群の中に文意と関係のない単語が含まれていた．提案手法は文中から抽出した名詞群全てとの類似度が最も大きい単語を抽出する手法である．そのため，文中に文意と関係のない名詞が含まれていると結果がその名詞に影響を受ける可能性がある．提案手法においては文中から名詞を抽出した後にコーパス収集文書数を出現文書数により増減することによってこのような重要でない単語の影響を小さくすることを試みた．その改良案として，単語の重要度として評価実験でも良い結果を得た TF × IDF を用いたり，文中単語の重用度をコーパス収集文書数の増減ではなくスコア算出時の係数として利用するなどの方法が考えられる．
- 提案手法は文中にない単語も対象とするために，高々数十単語を対象とした TF × IDF に比べて，精度が下がりやすい傾向がある．評価実験で用いた文は対象疾患に関する情報が十分に得られるようなものであり，文中の重要単語を抜き出すだけでその話題を認識することができるものであった．文中の単語のみからでは十分に意味が得られないような文に対しては提案手法と TF × IDF の差が少なくなると思われる．

4. まとめと今後の課題

本研究では，文中に出現しない語を話題語として抽出する方法として，ウェブを用いた話題語抽出方法を提案し，病名マスターを用いて医療分野の文の解析実験を行った．実験の結果，抽出精度は既存手法の TF × IDF に対して劣るものの，文中にない単語を抽出するという点では一定の成果を得ることができた．

A00-B99 感染症および寄生虫症

⋮

G00-G99 神経系の疾患

G00-G09 中枢神経系の炎症性疾患

G00 細菌性髄膜炎，他に分類されないもの

G00.0 インフルエンザ菌性髄膜炎

G00.1 肺炎球菌性髄膜炎

G00.2 レンサ球菌性髄膜炎

⋮

図 5: 病名マスターの階層構造 (ICD10)

今後研究を進めていくにあたっては，話題語の抽出精度の向上が目標となる．そのためには，以下の課題について解決する必要がある．

- 類似度の尺度算出方法の再検討
- 病名マスターの階層構造 (図 5) など，対象領域の事前知識の有効利用
- 文中からの重要語抽出技術の応用
- コーパス収集方法の改善

本研究を土台としてさらに話題抽出の精度を向上させ，音声認識や手書き文字認識，パッセージ検索等のシステムへの応用に役立てていきたい．

参考文献

[Salton 83] Salton, G. and McGill, M. J.: Introduction to Modern Information Retrieval, *McGraw-Hill* (1983)

[財団 06] 財団法人医療情報システム開発センター (MEDIS-DC)：ICD10 対応電子カルテ用標準病名マスター version 2.43 (2006), http://www.medis.or.jp/

[松尾 02] 松尾 豊, 石塚 満：語の共起の統計情報に基づく文書からのキーワード抽出アルゴリズム, 人工知能学会論文誌, Vol. 17, No. 3 (2002)

[松本 00] 松本 裕治, 北内 啓, 山下 達雄, 平野 善隆, 松田 寛, 高岡 一馬, 浅原 正幸：日本語形態素解析システム『茶筌』(2000)

[長沼 05] 長沼 潔, 速水 悟：医療分野における Web 文書からの話題抽出方法, 第 19 回人工知能学会全国大会 (2005)

[東京 91] 東京大学教養学部統計学教室（編）：統計学入門, 東京大学出版会 (1991)