Vol.2012-NL-206 No.5
Vol.2012-SLP-91 No.5
2012/5/10

ヘッドライン同定のための単語重要度の提案

宇高雅人[†1]　山村毅[†2]

本研究は，テキストに付けられたヘッドラインが適切であるかどうかを判断することの第一歩として，web 記事 (インターネット上の新聞記事) を対象にヘッドラインとその本文であるテキストを同定する手法を開発する．これを行うために，コーパスにおける単語の大域的頻度を用いた新たな単語重要度を導入した．新しい重要度の評価として，実際の web 記事を用いたヘッドライン同定実験を行い，従来の重要度計算手法である tf-idf 法と，提案手法，さらに人手での実験という 3 つの実験結果を比較した．ニュースサイトからランダムに選んだ 515 記事及び，似たまたは同じ内容を選んだ 303 記事に対して，提案手法は，それぞれ 78%，55%の正解率を示した．tf-idf 法では，それぞれ 76%，50%となり，従来の tf-idf 法よりも，ヘッドライン同定に有効であることが分かった．また人間と同等以上の性能を示した．

Proposing a word importance measure for head line identification

Masato Utaka [†1] and Tuyoshi Yamamura[†2]

This paper develops a method to identify a head line with a text for web articles (newspaper article on the Internet), as the first step of judging it whether a headline attached to the text is appropriate. We introduced new word importance measure using the global frequency of the word in the corpus. We performed the identification experiment using the real web articles. We compared the performance of our method with that of the conventional , tf-idf method and that of identification by human subjects. As a result the proposed method outperformed the tf-idf method; our method correctly identified 78% of 515 randomly selected web articles and 55% of another 303 articles that were similar to each other, whereas the tf-idf method correctly identified, 76% and 50%, respectively. It was also comparable or even superior to human subjects.

1. はじめに

本研究では，テキストに付けられたヘッドラインが適切であるかどうかを判断することの第一歩として，web 記事を対象にヘッドラインとそのテキストの同定手法を開発する．

ヘッドラインは，テキストを読む/読まないの判断をすることなどに利用できるため，これまでいくつかの研究がされている．

望月ら[1] は，指定された要約率の中で元テキストの情報をできる限り含め，作成されたものが自然で読みやすくなるような要約の作成手法を提案している．構文情報と語彙的結束性の情報を考慮し，同一単語の繰り返しによる冗長性を抑える手法，文意の維持のための必要な部分の補完をする手法，及び，内容に一貫性のある要約作成手法について述べている．これらの手法により作成された要約文は，一人の被験者が作成した重要箇所抽出と，よい類似をみせたと報告している．

大森ら[2] は，インターネット上で配信されている新聞記事を携帯端末向けに要約することを目的とし，文節ごとに tf-idf を算出し重要度の低い文節を順次削除することでヘッドラインの生成を試みている．作成されたヘッドラインは，人手で作成された携帯端末向け記事と名詞の一致率において 40%である．

池田ら[3] は新幹線の車内配信ニュースにおける特徴的な表現のうち，文体の体言止めや助詞止めなどの表現に着目し，ニュース記事に文末の整形を行って文を短縮する方法 (新幹線要約) を提案している．人手で作成された要約と，可読性と内容網羅性に関して 95%の正解率であると報告している．

廣嶋ら[4] は SVM を用いた統計的手法について研究している．これは重要語選択と文生成のモデルを用いて，可読性と内容網羅性を考えて単語をつなぎ，ヘッドラインを生成するものである．作成されたヘッドラインを，可読性と内容網羅性の観点から主観評価し，単純な tf-idf やトライグラムを用いた方法より幾分か優れていると報告している．

千田ら[5] は，専門知識のない人への対応に不慣れな技術開発担当者等の資料作成の手助けとして，新聞の見出しの付け方の分析結果に基づいた表題作成支援手法を開発している．新聞記事において用いられる 3 つのポイント (専門用語の意味に近い平易な用語の使用, 開

†1 愛知県立大学大学院情報科学研究科
Graduate College of Science and Technology,Aichi Prefectural University
†2 愛知県立大学情報科学部情報
School of Information Science and Technology,Aichi Prefectural University

Vol.2012-NL-206 No.5
Vol.2012-SLP-91 No.5
2012/5/10

発目的の主張，技術の長所の主張) による表題作成支援手法の開発と評価を行っている．提案手法を用いて作成した表題と提案手法を用いずに作成した表題を用いて，一般人 108 人へのアンケートを行い，関心度を 5 段階で調査した結果，91%以上の人が提案手法で作成した表題の方により関心があると述べたと報告している．

以上のように，従来の研究はいずれもヘッドラインを生成することに焦点を当てており，それらにおいては，生成したヘッドラインが正しいかどうかを，人手で作成したヘッドラインとの比較，又は，テキストとの主観的評価によって行っている．すなわち，ヘッドラインがテキストの内容を表しているかどうかを直接扱うものではない．

本研究では，ヘッドラインがテキストの内容を表しているかを判定する手法を開発する．具体的には，複数のヘッドラインの中からテキストに対応するヘッドラインを同定する方法を開発する．

2. 単語重要度とヘッドライン同定手法

2.1 単語からの予想テキスト長を用いた重要度計算

単語重要度計算において，局所的重み，大域的重みにはいずれも文書の長さに影響を受けてしまうという欠点がある．その欠点を小さくする処理として文書正規化係数があるが，その一つであるコサイン正規化も逆の問題 (短い文書が優先される) を起こしてしまう．このコサイン正規化の問題を解決するためにピボット正規化というものもある．これはテスト・コレクション[⋆1]中の検索質問文に対し，長さ l の文書が適合している確率 (適合確率) と長さ l の文書が検索される確率 (検索確率) を調べ，両者の食い違いを解消するように索引語の重みに修正をかけるという考え方である．しかし，この方法は前述したコサイン正規化に比べ，計算が一気に複雑化することが問題となっている．

そこで本研究では，単語から，その単語を含むテキスト (対象テキスト) の相対的な長さを求め，それが実際のテキストの長さとどのような関係にあるかにより重要度を決定する手法を提案する．

いま，語彙数 N のコーパスにおいて，ある単語 w の出現回数 (大域的頻度) が $F(w)$ であったとする．あるテキスト d における単語 w の出現回数を $tf(w,d)$ とするとき，

$$\tilde{l}(w,d) = tf(w,d) \cdot \frac{N}{F(w)} \tag{1}$$

は，テキスト d における単語 w の相対頻度が，コーパス全体におけるそれと同じであると仮定した場合の，テキスト d の予想される長さを表す．この $\tilde{l}$ が，実際のテキスト d の長さ $l(d)$ と大きく離れているのならば，単語 w は，そのテキスト d で，偏って使用されている，すなわち，重要視されていることになる．そこで，それらの比，

$$f(w,d) = \frac{\tilde{l}(w,d)}{l(d)} \tag{2}$$

を，単語 w のテキスト d における重要度とする．ただし，この値をそのまま使うと，大域的頻度 $F(w)$ が小さい単語について，その重要度が大きくなりすぎる傾向があるので，実際には，$\tilde{l}$ として，式 (3.1) の $\frac{N}{F(w)}$ の対数をとった次式を用いる．

$$\tilde{l}(w,d) = tf(w,d) \cdot \log \frac{N}{F(w)} \tag{3}$$

この計算手法は，従来問題であったテキストの長さに影響を受けることがなく，さらに，本来のテキストの長さに対しての影響を無くすための処理であるピボット正規化に比べ式が簡便である等の利点をもつ．

2.2 ヘッドライン同定手法

2.2.1 概　　要

図 1 に提案するヘッドライン同定手法の概要を示す．まず，ヘッドライン同定をしたい web 記事のテキストと，候補となる (複数の) ヘッドラインを用意し，これらに前処理を行って単語集合に変換する．次に，web 記事のコーパスであらかじめ求めておいた，単語の大域的頻度を利用して，テキストの単語集合と (一つの) ヘッドラインの単語集合との適合度を計算する．最も大きな適合度となったヘッドラインを，テキストのヘッドラインとして判定する．

以下各処理について詳しく述べる．

2.2.2 前　処　理

形態素解析を行って単語 (原形) に分解する．得られた単語集合に対してストップワード処理を行って，不要な単語 (機能語) を取り除く．本研究では，形態素解析に茶筌を用いている．

2.2.3 適合度計算

前処理を施した結果求まったテキストの単語集合と，ヘッドラインの単語集合との適合度を計算する．具体的には，ヘッドラインの単語を $w_i(i = 0, \cdots n)$ とするとき，(2) 式で，

⋆1 テスト・コレクションには，検索質問文の集合と各検索質問文に適合する文書集合が与えられている．そのため，適合確率が求められる．

Vol.2012-NL-206 No.5
Vol.2012-SLP-91 No.5
2012/5/10

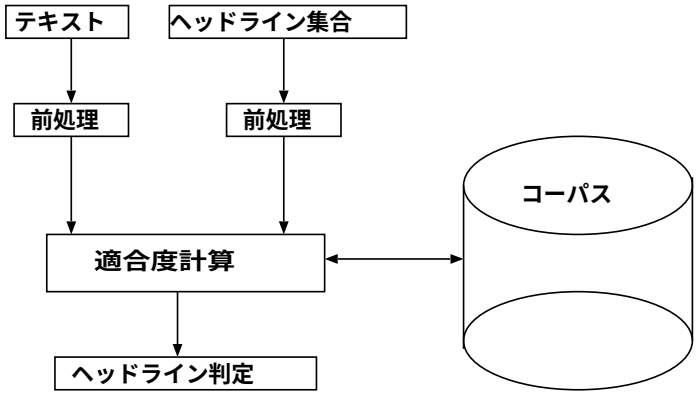

図 1 提案手法の概要

w_i のテキスト d での重要度 $f(w_i, d)$ を計算し，これを用いて，以下のように，ヘッドラインの単語集合のテキストでの重要度の平均値を，ヘッドラインとテキストの適合度とする．

$$F = \frac{1}{n} \sum_{i=1}^{n} f(w_i, d) \tag{4}$$

3. 評 価 実 験

3.1 概　　要

前章で提案した手法を用いて，実際の web 記事に対してヘッドライン同定実験を行った．その際，本研究で導入した単語重要度の有効性を調べるため，重要度を従来の tf-idf 法で計算した場合の結果と比較した，

実験に用いたデータは「NHK NEWS WEB[6]」からランダムに集めたニュース記事本文とそのヘッドラインの組 515 個，及び,「yahoo!JAPAN ニュース[7]」から同じ内容もしくは似ている内容を取り扱った複数の会社のニュース記事本文とそのヘッドラインの組 303 個の合計 818 個である．303 組のデータには一つの内容につき似た記事が 2 から 4 個ある．

3.2 実 験 結 果

818 記事に対してヘッドライン同定を行った結果，573 記事 (70.0%) について正しく同定することができた．

表 1 全 818 記事に対する結果

	提案手法	tf-idf 法
全体数	818	818
正解数	573	546
正解率 (%)	70.00	66.70

表 2 ランダムに選んだ 515 記事に対する結果

	提案手法	tf-idf 法
全体数	515	515
正解数	405	392
正解率 (%)	78.60	76.10

表 3 同じまたは似た内容の 303 記事に対しての結果

	提案手法	tf-idf 法
全体数	303	303
正解数	168	154
正解率 (%)	55.40	50.80

3.2.1 tf-idf 法との比較

ランダムに選んだ記事と似たもしくは同じ内容の記事を合わせた 818 記事に対して，ヘッドライン同定を行った結果を表 1 に示す. 表からわかるように，正解率は提案手法，tf-idf 法ともに，70%弱と比較的高い．又，提案手法の方がわずかながら性能が良い．

次に，ランダムに選んだ 515 記事に対してのみの結果を抽出したものを表 2 に示す．全体の場合と比べると，正解率はやや上がり，正解率は提案手法，tf-idf 法ともに，75%以上と高い．又，提案手法の方がわずかながら性能が良い．

一方，同じまたは似た内容の 303 記事に対してのみの結果を抽出したものを表 3 に示す. 当然ながら，先の場合と異なり，正解率は 50%と低い．しかし，こちらの場合でも，提案手法は tf-idf 法よりも良い性能を示している．

3.2.2 人手で行った場合との比較

次に人手でヘッドライン同定を行ってもらった結果との比較を示す.

まず，人手でヘッドライン同定を行ってもらった結果を表 4，表 5，表 6 に示す．表 4 は，818 記事からランダムに選んだ 64 記事に対しての結果，表 5 はランダムに選んだ 515 記事から選んだ 32 記事に対しての結果，表 6 は，似た 303 記事から選んだ 32 記事に対しての結果である．それぞれの実験の被験者は愛知県立大学の学生である．人数は，表中に示す通りである．

次に，提案手法および，tf-idf 法によるそれぞれ記事に対する結果と比較したグラフを順に，図 2，図 3，図 4 に示す．図 2 から分かるように，全体の結果では，提案手法は，人間

Vol.2012-NL-206 No.5
Vol.2012-SLP-91 No.5
2012/5/10

表 4 818 記事に対する人手の実験結果

実験人数	平均	正解率 (%)
32	22.4	70.00

表 5 515 記事に対する人手の実験結果

実験人数	平均	正解率 (%)
11	29.8	93.20

表 6 303 記事に対する人手の実験結果

実験人数	平均	正解率 (%)
21	15.0	48.21

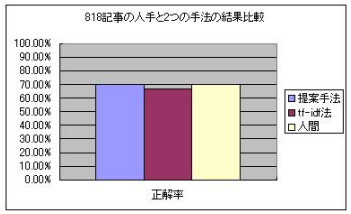

図 2 818 記事に対する各手法と人手の正解率の比較

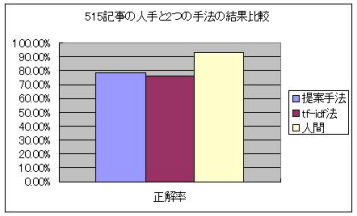

図 3 515 記事に対する各手法と人手の正解率の比較

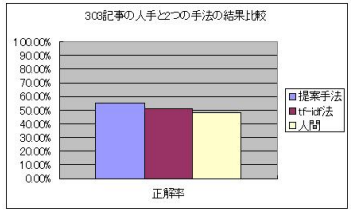

図 4 303 記事に対する各手法と人手の正解率の比較

と同等程度の性能を示している．図 3 のランダムに選んだ記事 515 記事に対する結果では，人間の正解率は 9 割を超えており，提案手法よりも極めて良いことがわかる．しかし，図 4 の似た内容の 303 記事に対する結果を見てみると，提案手法は，人間よりも良い性能であることがわかる．

3.3 考　　察

表 7，表 8 に 515 記事，303 記事での不正解がどのような記事で生じたかの内訳を示す．これらの表で「似ている」は同じ，もしくは似た内容の別の記事のヘッドラインを選んでしまった回数,「似ていない」は，内容に類似性のない別の記事のヘッドラインを選んでしまった回数,「数字」は「似ていない」の場合の中で主たる原因が数字によるものである回数である．

これらの表から，提案手法も tf-idf 法も，不正解のほとんどが「似た」記事に対するものであったことがわかる．又，わずかではあるが，提案手法の方が，似た記事に対する誤りが少ないことがわかる．しかし，その反面，提案手法は「似ていない」記事に対して，tf-idf 法よりも多くの誤りを犯している．これは，本手法では，数字のような固有表現を区別せずに重要度計算しているため，tf-idf 法に比べると過大に評価されてしまうからである．表中，提案手法では数字による不正解があるが，tf-idf 法ではない．

表 7 515 組の不正解の内訳

515*515	提案手法	tf-idf 法
似ている	93	102
似ていない	27	21
(数字)	(10)	(0)
計	120	123

表 8 303 組の不正解の内訳

303*303	提案手法	tf-idf 法
似ている	95	119
似ていない	40	30
(数字)	(13)	(0)
計	135	149

これは，例えば，個々の固有表現の出現頻度をその種類の出現頻度に置き換えることで対処できるものと考えられる．

4. おわりに

本研究では，web 記事を対象に，記事のテキストと，そのヘッドラインの適合度を計算する手法を提案した．そこでは，コーパスにおける単語の大域的頻度を用いた，新たな単語重要度を導入した．実際の web 記事を用いたヘッドライン同定の評価実験では，提案手法は，人間と同等以上の性能を示し，又，提案した単語重要度は，従来の tf-idf 法よりも，ヘッドライン同定に有効であることが分かった．

テキストに含まれる数字の扱いや平均値を求めるためのヘッドラインの長さで割る処理などが原因で，正しく同定できない場合もあったが，これについては形態素解析後の処理に数字と単位をまとめる処理やヘッドラインの単語のテキストに含まれる割合による処理などで改善できると考えられる．

ランダムで選んだ web 記事 515 組による実験では 78%，似たもしくは同じ内容の web 記事による実験では 55%とどちらも tf-idf 法の 76%，50%を上回る結果が得られた．今後，これらの正解率をより向上させるために，(1) 似たもしくは同じ内容の web 記事による実験において，似た内容のヘッドラインを選んでしまうことの対策や (2) テキストに使われている単語の類似度をヘッドラインに用いている場合への対処，(3) 提案手法の完全実装，が課題として挙げられる．

Vol.2012-NL-206 No.5
Vol.2012-SLP-91 No.5
2012/5/10

参 考 文 献

1) 望月　源，奥村　学：読みやすさの向上と冗長性の排除を考慮した重要箇所抽出型要約，情報処理学会研究報告. NL193-3.
2) 大森岳志，増田英孝，中川裕志：Wen 新聞記事の要約とその携帯端末向け記事に寄る評価，情報処理学会研究報告. NL153-1.
3) 池田諭史，大橋一輝，山本和英：「新幹線要約」のための文末整形，情報処理学会研究報告. 2004-NL-163(22).
4) 廣嶋伸章，長谷川隆明，奥　雅博：Web ページのヘッドライン生成のための統計的要約，自然言語処理， Vol.12, No.6, pp.113–128 (2005).
5) 千田恭子，篠原靖志，奥村　学：技術成果を効果的に伝える表題作成支援手法:開発と評価，情報処理学会論文誌， Vol.46, No.11, pp.2728–2743 (2005).
6) NHK: NHK オンラインニュース，NHK（オンライン），
入手先⟨http://www3.nhk.or.jp/news/⟩（参照 2011-10）.
7) YAHOO!JAPAN: YAHOO!JAPAN ニュース，YAHOO!JAPAN（オンライン），
入手先⟨http://headlines.yahoo.co.jp/hl⟩（参照 2011-10）.