ヘッドライン同定における数値表現重要度補正の効果の検証

宇高 雅人　　指導教員：山村 毅

1 はじめに

近年，インターネットの幅広い普及により膨大な情報が蓄積されるようになり，結果として情報の氾濫を引き起こしている．このため効率的な情報検索や取捨選択に対する技術の開発に注目が集まっており，その一つにヘッドラインの利用がある．

ヘッドラインとは，新聞記事などでテキストの内容を簡潔に示したもののことである．一般的には，これを用いれば効率的な情報取得が可能になる．しかし，ヘッドラインがテキストの内容を正しく反映していない場合は，情報取得の効率が悪くなるだけでなく，場合によっては誤った情報を獲得してしまうこともある．

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を判断することの第一歩として，インターネットの新聞記事 (以下 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．

2 先行研究

宇高ら [1] は，単語の頻度から，その単語を含むテキストの予測される長さを求め，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って重要度を計算する単語重要度の計算手法を提案した．また，その計算手法を用いたヘッドライン同定手法も提案している．これは，同定したいテキストと複数のヘッドラインがあったときに，各ヘッドラインに対して，そこに含まれる単語でテキストに存在するものの重要度を計算，合計して単語あたりの平均を求め，それが最も高い値を示したヘッドラインをテキストの適切なヘッドラインだとするものである．

この手法を用いて，実際の web 記事 561 記事に対するヘッドライン同定実験を行ったところ，80.2%(450 記事) の精度であった．

3 先行研究の同定手法に対する考察

3.1 判定基準の緩和

本来のテキストとヘッドラインの組とは別のヘッドラインが同定手法によって選ばれ，さらにその間違って選ばれたヘッドラインがテキストの内容と釣り合っている (真の意味では正解ではないが適切とみなしても良い) 場合があった．以後，このような場合を容認可能と呼ぶことにする．図 1,2 に例を示す．図 1 は対象テキスト，図 2 はヘッドラインの候補で，1 が正解のヘッドライン，2 がシステムの選んだヘッドラインである．これらを見比べた時，2 つのヘッドラインのどちらが正解かを判断するのは人でも容易ではないので，先行研究は判定基準が厳しすぎたと考えられる．本研究の目的 (対象テキストの内容を表しているヘッドラインを同定すること) を考えた時に，容認可能なものを正解と判断するのはごく自然なことだという事ができる．

そこで，不正解と判定されたものの中で容認可能なものを正解とみなすことで，精度を再評価した．具体的には，不正解と判定された 111 個について，著者を含む被験者 3 人に容認可能かどうかの判定をしてもらった (意見が分かれたものについては，3 人による協議で判定した)．この結果 58 個が容認可能と認定

> 福島県が進めてきた収穫後のコメの放射性物質の検査が終わり、コメの作付けが行われたすべての市町村で国の基準を下回って出荷が認められました。福島県は、今後、コメの安全性を消費者にピーアールするなど風評被害対策を強化することにしています。(以下略)

図 1　対象テキスト

> 1. 福島 コメの風評被害対策強化へ
> 2. 福島のコメ 国の基準を下回る

図 2　ヘッドライン候補

され，全体としての正解数は 508(90.6%, 不正解 53) となった．

3.2 数字と接尾辞の重要度

容認可能なものを正解とみなすことで，真に対処すべき不正解の数は 53 個になる．これらを詳しく分析してみると，数字や接尾辞の重要度が，相対的に大きくなっていることがわかった．図 1 は「＜はぐれザル＞弘前中心部迷走中　市が注意呼びかけ／青森」と「独”人工衛星落下２０～２５日”」というヘッドラインに含まれる一部の単語の重要度を表したものである．この表より，テキストの内容を表す「注意」「ザル」よりも，より一般的な「市」や「２」の方が高い重要度を持っていることがわかる．

表 1　数字と接尾辞の重要度比較

単語	重要度	単語	重要度
注意	16.8	2	26.4
ザル	10.7	市	16.3

4 提案手法

3 で述べたように先行研究では，数字と接尾辞の単語重要度が他の単語より大きくなってしまうという問題があった．そこで，この問題を改善するための方法を提案する．

4.1 チャンキング

従来方法では，文章を形態素解析した結果得られた単語についてその重要度を計算しているが，数字や接尾辞は一つの単語としてバラバラに取り扱われるので，このままだとその重要度が大きくなりすぎてしまう．そこで，これらの重要度の上昇を抑えるためにチャンキング [2] を利用する．具体的には，以下の処理を行う．

1. 数字同士が連続して存在する場合にそれらを繋げる．
2.「名詞・接尾・助数詞」と数字が連続して存在する場合にそれらを繋げる．

4.2 数値表現の補正

チャンキングによって数字と単位を繋げることにより，連語としての出現頻度は減少するのでヘッドラインとテキストを同

定するときの目安となる特定性が大きく上がることが予測される．数値表現 (単位や日付) は記事を特定するための重要な特徴であるため，これによってヘッドライン同定の精度の向上が期待できる．しかし，実際には特定性は上がるが，数値表現の重要度はあまり大きくならないことがわかった．そこで，数値表現に対して重要度の補正を行うことでこの問題に対処する．具体的には，ヘッドラインに数値表現があった場合，その数値表現がテキストに存在するならプラスの補正 ($+\alpha$) を，存在しないならマイナスの補正 ($-\alpha$) を与える．

4.3 ヘッドライン同定手法

ここで，ヘッドライン同定手法について簡単に説明する．まずヘッドラインとテキストをそれぞれ単語集合に分解する．次に，前処理を行なって，名詞，動詞，未知語のみを取り出し，次いで，チャンキングを行なって，「名詞・接尾・助数詞」以外の接尾辞の削除と数値表現の生成を行う．単語の重要度を計算した後，数値表現に補正をかけ，ヘッドライン適合度を計算し判定をする．

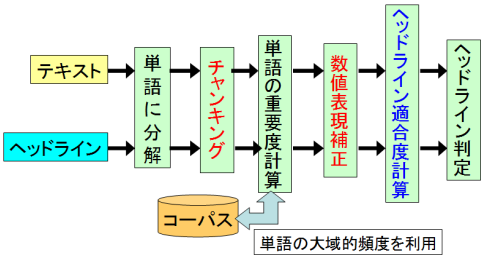

図 3　ヘッドライン同定手法の概念図

5 評価実験

5.1 評価方法

4.3 で述べた方法を計算機上に実装し，先行研究と同じ web 記事 (561 記事) に対してヘッドライン同定実験を行う (単語重要度計算の元になる単語の大域的頻度もこれらの記事集合から計算する)．4.2 で述べた数値表現の補正値 (α) を 1〜60 まで変えて正解数を求める．

5.2 実験結果

図 4 に補正値を変えた結果を示す．図より，最も正解数が多かったのは，補正値が 12 から 19 のときで，523 記事 (93.2%) であった．これは，先行研究の結果 (90.6%) より 2.6 ポイントよく，数値表現の補正がうまく働いていることがわかる (有意水準 5% の適合度検定で有意差あり)．

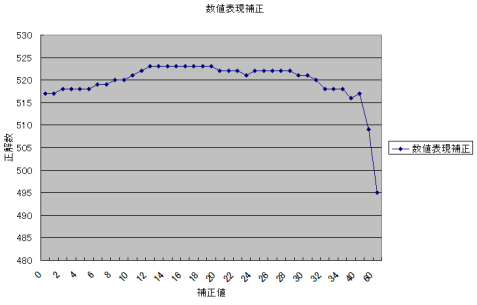

図 4　数値表現補正の実験結果

5.3 固有表現全体の補正

数値表現への補正が有効であったことを受けて，数値表現は固有表現の一部であるので，固有表現全体に補正を加えることで，さらなる精度向上が見込めるのではないかと考えた．そこで，今度は数値表現だけでなく，固有表現全体にも補正をかけて同様の実験を行い，数値表現の補正の場合と比較した．結果を表 2 に示す．表からわかるように，精度は向上しなかった．固有表現全体を補正しても数値表現に補正する場合と同様の効果が得られるが，補正のような付加的な処理は不必要ならば，しないほうが良いので，数値表現にのみ補正を加えるほうが良い．

表 2　数値表現補正と先行研究と固有表現補正の比較

実験の種類	補正値	容認可能含む正解数 (%)
先行研究	0	508(90.6)
数値表現補正	15	523(93.2)
固有表現補正	11	522(93.0)

5.4 考察

ヘッドライン同定の失敗例から不正解の原因を調査した．まず，あげられるのは同義語の問題である．これは，ヘッドラインの単語がテキストで同義語に言い換えられることで起こる問題である．具体的には，ヘッドラインの単語「W杯」がテキストでは「ワールドカップ」となることにより，重要度が計算されないというものである．これに対して，例えば共起頻度を用いるようなことなどが考えられる．

次いで，省略表記の問題がある．これは，単位を省略して表記するために起こる問題である．具体的には，ヘッドラインの単語「2，3号機」があったとき (テキストでは「2号機」「3号機」となっている)，提案手法のチャンキングでは，「2」と「3号機」に分かれて出力してしまう．この出力から重要度計算と数値表現補正を行うと，「3号機」には計算も補正も適用されるが，「2」には両方とも適用されない．このような問題に対処するには，数字と数字の間に「,」が存在し，数字の後に単位が続くような場合において，「,」の前の数字にも単位が出力されるような工夫が必要である．

6 まとめ

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．

評価実験の結果，先行研究の手法より，数値表現に補正を加えることで高い精度を得ることができた．固有表現全体に補正をかけたが，数値表現の場合以上の効果は得られなかった．

今後の課題としては，同義語や数値表現の中の漢数字への対処，省略表記への対応を行う必要がある．

参考文献

[1] 宇高 雅人, 山村 毅: ”ヘッドライン同定のための単語重要度の提案 ”, 情報処理学会研究報告 Vol.2012-NL-206，No.5，pp.1-5，2012.

[2] 工藤 拓，松本 裕治:Support Vector Machine を用いた Chunk 同定，自然言語処理研究会 2000-NL-140，pp.9-16(2000).