重要度を考慮した，語句の削除によるニュース記事ヘッドラインの生成手法の提案

松川　維吹樹　　指導教員：山村　毅

1　はじめに

現代社会には，多くの情報が存在し，自分にとって必要ではない情報も少なからず存在する．ニュース記事のヘッドラインは，記事本文の内容を簡潔に表す文とされ，その記事を読む/読まないの判断をする際，極めて有用である．このヘッドラインは記事本文の作成者によって作成されることが多く，労力を必要とする．また，作成する人によって品質が変わる問題もある．そこで，ヘッドラインを自動生成することができれば労力や品質の問題が改善できると考え，前迫 [1] はニュース記事に対する調査を行い，不要箇所の削除と変換による要約手法によって，ヘッドラインを自動生成する手法を提案している．また，松川 [2] はシステムとしてそれを実装した．

本研究では，既存のシステムの問題点から，単語ごとの重要度を考慮した語句の削除によるヘッドラインの生成を行う手法を提案するとともに，重要度を考慮することでヘッドラインの自動生成にどのような効果があるかについての考察を行う．

2　先行研究

前迫 [1] は多くの新聞記事とヘッドラインを比較し，不要な部分の削除や語句の変換による要約によって，自動的にヘッドラインを推定する手法を考案した．前迫が提案した手法は以下の 5 つの処理を順番に適用していくというものである．

引用文処理　記事中に多く出現する文の形式に，次のようなものがある．

- A が B 日 “…” ということを示した．
- A が B 日 “…” と発表した．
- A が B 日 “…” ことを述べた．

これらの文から “…” を抽出する．

文末変換処理　「サ変名詞＋する」をサ変名詞のみに変換する．

時格要素削除　時を表す語を削除する．

連体修飾句・節削除　名詞を修飾するものを削除する

助詞削除　単に助詞のみを削除する．

3　先行研究の問題点

先行研究の手法で作成したヘッドラインを分析すると，必要以上に削除を行ってしまい重要な単語が削除されていたり，また削除するべきところを削除せず残していたりと，処理の適用の制御ができていなかった．そこで，重要な単語が削除されることを防ぐため，単語ごとの重要度を考慮した語句の削除を行うことによりヘッドラインを生成する手法を提案する．例えば，「伊調馨選手の国民栄誉賞の授与式を，首相官邸で行うと発表した」という文があるとする．先行研究の手法だと，「授与式首相官邸で行う」というヘッドラインになる．しかし，正解では「国民栄誉賞の」という部分が残っており，重要な部分を削除している．そこで表 1 のような重要度リストを参照し，適当な閾値 (例の場合は 6.0) で区切り，重要ではない部分を削除することで「国民栄誉賞の授与式を，首相官邸で」というヘッドラインになる．

表 1　単語ごとの重要度の例

単語	重要度
伊調馨選手	5.2
国民栄誉賞	10.9
授与式	8.5
首相官邸	6.3
行う	0.9
発表	1.5
し	0.2

4　提案手法

先行研究と同じ前処理，依存構造解析を行う．この際，形態素解析で得られた結果に対して，単語ごとに重要度を付与するとともに，依存構造解析で得られた結果に対して，文節ごとに単語の重要度の高いものを付与する．そして，閾値を超えないような重要ではない文節を削除していくことでヘッドラインを生成する．ただし，最終文節にくる主動詞は重要であると考え必ず残し，重要であるとされた文節から最終文節に至るまでの文節は全て削除しないようにする．図 1 に本研究のシステム構成図を示す．

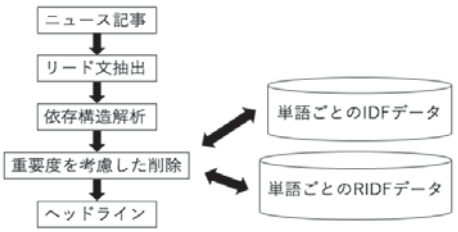

図 1　本研究のシステム構成図

5　重要度

5.1　IDF

IDF とは，文書頻度 (DF) の逆数をとったもので，文書頻度に基づく重み付けとしてよく知られている．出現頻度が高いもの，すなわち多数の文書に共通して出現する語句には低い重みを与え，出現頻度が少ない，すなわち特定の文書に偏って出現する語句は，その記事を特徴付けるため高い重みが与えられる．文書の総数を N，索引語 t の文書頻度を DF(t) とした時 IDF(t) は以下の式 (1) で定義される．

$$IDF(t) = \log \frac{N}{DF(t)} \tag{1}$$

5.2　RIDF

索引語に 1 番求められることは，その索引語が文書の内容を表しているものでなければならないことである．しかし，IDF による重み付けは，我々の直感に合わない部分もあり，必ずしも

索引語の重み付けとして適当であるとは言えない．そこで，これを訂正するために，ポアソン分布からの推定値と実際の値との差を用いて単語の重要性を評価する尺度としたものが RIDF(残差 IDF) である．単語 t の RIDF は，文書における実際の IDF 値とポアソン分布から推定される IDF 値の差となり，以下の式 (2) で定義される．

$$\begin{aligned} RIDF(t) &= (\text{実際の } IDF) - (\text{ポアソン分布で推定された } IDF) \\ &= \log \frac{N}{DF(t)} - \log \frac{N}{N(1 - P(0;\lambda))} \\ &= \log \frac{N}{DF(t)} + \log \left(1 - \mathrm{e}^{-\frac{F(t)}{N}}\right) \end{aligned} \tag{2}$$

6 実験

6.1 実験方法

毎日新聞の 2005 年の記事 200 記事を用いて実際にヘッドラインを生成した．これを，配信記事に付けられているヘッドラインと比較し，再現率，適合率，F 値を求めた．また，重要度として，5 章で述べた IDF を用いた場合と，RIDF を用いた場合を比較した．

6.2 IDF を用いた結果

IDF を考慮した実験の結果を表 2 に示す．

表 2　IDF を利用した結果

閾値	再現率 (%)	適合率 (%)	F 値 (%)
1.26	85.5	28.0	40.9
1.36	85.2	28.4	41.2
1.48	83.9	28.5	41.1
1.66	82.3	28.8	41.3
1.96	74.1	28.4	39.4
2.01	73.1	28.5	39.2
2.06	72.3	28.5	39.1
2.11	70.7	28.0	38.6

表 2 より，閾値を上げていくことで再現率は下がっていくことがわかる．また，適合率に関しては，閾値によって大きな差は見られない．F 値をみると閾値 1.66 のときが最も高く，41.3 %であった．

6.3 RIDF を用いた結果

RIDF を考慮した実験の結果を表 3 に示す．

表 3 より，閾値を上げていくことで再現率は下がっていくことがわかる．また，適合率に関しては，閾値によって大きな差は見られない．F 値をみると閾値 2.55 のときが最も高く，41.3 %であった．

6.4 考察

IDF の最も良い時，RIDF の最も良い時，先行研究の結果を表 4 に示す．

再現率で比較すると，重要度を考慮することで先行研究より高くなる結果が得られた．また，適合率で比較すると，重要度を考慮した手法は先行研究の手法を超えることができなかった．しかし，総合的な評価である F 値では先行研究より 3 %良い結果が出ている．

表 3　RIDF を利用した結果

閾値	再現率 (%)	適合率 (%)	F 値 (%)
1	88.8	26.8	40.0
1.5	88.5	26.8	40.0
2	87.6	26.8	39.9
2.5	85.4	28.3	41.1
2.55	85.4	28.4	41.3
3	83.7	28.6	41.2
3.5	80.5	28.7	40.8
4	74.4	28.4	40.8

表 4　実験結果 (本研究 + 先行研究)

	再現率 (%)	適合率 (%)	F 値 (%)
IDF	82.3	28.8	41.3
RIDF	85.4	28.4	41.3
先行研究	63.1	30.1	38.2

重要度を考慮した結果，先行研究より F 値がよくなったことから，重要度を考慮することはヘッドライン生成において有用であることがわかる．また，本研究の手法の再現率が高かった理由としては，重要語から最終節に至る文節を全て残したためであると考えられる．しかし，適合率が高くないことから，余分な単語を残してしまっていることが伺える．また，記事 1 つ 1 つに着目すると，正解に数字表現が含まれている場合が 57 記事あるのに対して，本研究の手法では 21 記事でしか数字表現を残せていない．これは数字の出現頻度が高く，重要度が低いというデータになっているためだと考えられる．このため，数字に関しては別の処理を用意するべきだと考えられる．

7 まとめ

既存のヘッドライン生成手法に対して，重要度を考慮するヘッドライン生成手法を提案した．重要度として IDF と RIDF を用いヘッドラインを生成した結果，最も高い F 値が 41.3 %という結果を得ることができ，先行研究より良い結果を得ることができた．また，結果と考察から，重要度を考慮することはヘッドライン生成において有用であるが，数字表現に弱いという結論が得られた．今後の課題としては，正解とするヘッドラインが人手で生成されていて個人差があることから，正解とするヘッドラインを一定にすることが考えられる．また，本研究の手法は再現率が高く，これはいらない単語を多く含んでいるとも考えられるので，要約率などを設定し，設定した単語数まで削除することで改善されるかもしれない．

参考文献

[1] 前迫綾, 竹川美樹, 山村毅: ”ニュース記事のタイトルの自動生成”, 電気関係学会東海支部連合大会講演論文集, O-497, 2008

[2] 松川維吹樹, 山村毅: ”語句の削除，変換を用いた，ニュース記事ヘッドラインの自動生成”, 第 15 回情報学ワークショップ (WiNF2017), PC-24, 2017