E-020

見出しにおける文の成分に関する調査
Investigation of relationship between sentence elements in an article and the headline

浅倉 優介† 丹羽 弘充† 山本 けい子‡ 田村 哲嗣‡ 速水 悟‡
Yusuke Asakura Hiromitsu Niwa Keiko Yamamoto Satoshi Tamura Satoru Hayamizu

1. 研究背景・目的

近年，インターネットの普及に伴い，Web 上に存在する情報が急激に増加しており，必要な情報を必要な量だけ入手することが可能である．しかし，その内容を 1 つ 1 つ確認するには，膨大な労力と時間が掛かってしまう．

この問題に対する解決策として，記事の見出しに注目する．見出しは，その記事の重要な情報を端的に表現するためのものであるため，内容の確認を容易にする一方，情報が更新され急増していく中，人手で見出しを作成することは容易ではないと考えられる．

見出しに関する先行研究として，見出しの文字数に着目した研究[1]や，タイトルに対応する文を抽出し，その文から不用なものの削除や変換を行い，見出しの自動生成を行っている研究[2]がある．

本研究では，文の成分に着目し，インターネット上のニュース記事の見出しを自動生成するために，ニュース記事本文において何文目の文の成分が，見出しに使用されている単語を多く含んでいるのか調査を行った．

2. 予備実験

ニュース記事は 5W1H(Who,What,When,Where,Why,How)という規則を基に作成されている．これは記事の 1 文目に重要な情報を記載するという考えに基づくものである．しかし，2 文目以降にも重要な情報が記載されている場合がある．ニュース記事の見出しに使われている単語が，記事本文の何文目に記載されているか調査を行ったところ，以下の式によって近似することができることが分かった．予備実験で用いたデータは毎日新聞コーパスの 2002 年～2005 年までの 4 年間分の記事である．

$$\mathrm{Score}(x) = \begin{cases} 5.15 & x = 1 \\ 2.78/x^{0.28} & \text{otherwise} \end{cases} \qquad (1)$$

Score(x)は，見出し中に含まれる単語が x 文目に出現する個数である．式(1)は，見出し中に使用されている単語が，記事本文の 1 文目に多く出現し，以後減少する傾向を表す．この事から，見出しの自動生成には，記事本文の 1 文目を使用することが良いと考えられる．

†岐阜大学大学院工学研究科
Graduate School of Engineering,Gifu University
‡岐阜大学工学部
Faculty of Engineering,Gifu University

3. 文の成分

本研究では，文の成分に着目して調査を行った．ここで，文の成分 5 つを以下に示す．

(1)主語:文の主体となる文節
(2)述語:主語に対する説明する文節
(3)修飾語:他の文節の内容を詳細に説明する文節
(4)接続語:文節と文節をつなぎ合わせる文節
(5)独立語:どの文節とも関係がない，独立した文節

目的語は，(3)の修飾語に含まれている．本研究では，上の文の成分のうち，主語，目的語，述語に着目して調査を行った．理由として，その 3 つを組み合わせて見出しを生成することにより，内容を明確に表現できると考えたためである．

4. 実験

4.1 評価対象

本研究における評価対象は，以下の 3 つであり，機械的に抽出を行った．

(1) 主語
　記事本文中の主語として，”A は”，”A が”の A の部分(名詞)が，見出しに含まれているか否かを調査する．
(2) 目的語
　記事本文中の目的語として，”B に”，”B を”の B の部分(名詞)が，見出しに含まれているか否かを調査する．
(3) 述語
　記事本文中の述語として，”C する”(サ変動詞)における C の部分(名詞)と一般動詞が，見出しに含まれているか否かを調査する．

ここで，(1)，(2)の名詞部分の評価方法の詳細について，以下に例を示す．

記事の本文中に主語として”原油先物相場は”，見出しが”NY 原油，大幅続落”とする．図 1 より，名詞部分を複合語としてではなく，形態素解析結果を対象とする．これは，見出しは文字数が少ないため，長い複合語は使用されている頻度が少ないと考えられるためである．

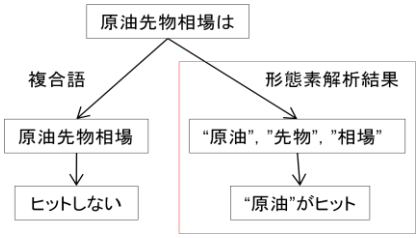

図 1. 評価対象の例

4.2 評価方法

本調査はニュース記事本文において何文目の文の成分が，見出しにどの程度含まれているかというものだが，詳細として，ある記事において，1 文目に出現して 2 文目にも出現した場合，その記事は，すでに 1 文目に出現した記事とはせず，単純に，1 文目に出現している文書数はいくつか，2 文目に出現している文書数はいくつか，という方法で評価した．今回は全記事において 1～5 文目までとし，以下の式を用いて評価を行った．

$$p(x) = \frac{a_x}{z_x} \times 100 \ (1 \leq x \leq 5) \qquad (2)$$

z_x:x 文目まで存在する文書数
a_x:x 文目から抽出された候補が見出しに存在している文書数
p(x):x 文目から抽出された候補が見出しに存在している割合

4.3 実験データ

本実験のデータの概要を表 1，表 2，表 3 に示す．使用したデータは，asahi.com，毎日 jp，nikkansports.com の 2008 年 5 月～2009 年 6 月に収集したニュース記事であり，数字はジャンルにおける記事数を示している．

表 1 asahi.com の使用したデータ

ビジネス	文化	国際	社会
2467	474	1950	6143
政治	スポーツ	全ジャンル	
1697	2328	15059	

表 2 毎日 jp の使用したデータ

経済	事件	ライフ	政治
2982	4796	3920	2287
スポーツ	海外	全ジャンル	
661	2665	17311	

表 3 nikkansports.com の使用したデータ

芸能	社会	スポーツ	全ジャンル
2574	17990	4086	24650

4.4 実験結果

実験結果を図 2 に示す．縦軸は x 文目から抽出された候補が見出しに存在している割合，横軸は全てのサイトにおけるそれぞれの文の成分と全ての文の成分である．ただし，図 2 は 4.2 に記した評価方法より，1 文目～5 文目の結果を全て足すと 100 を超える．

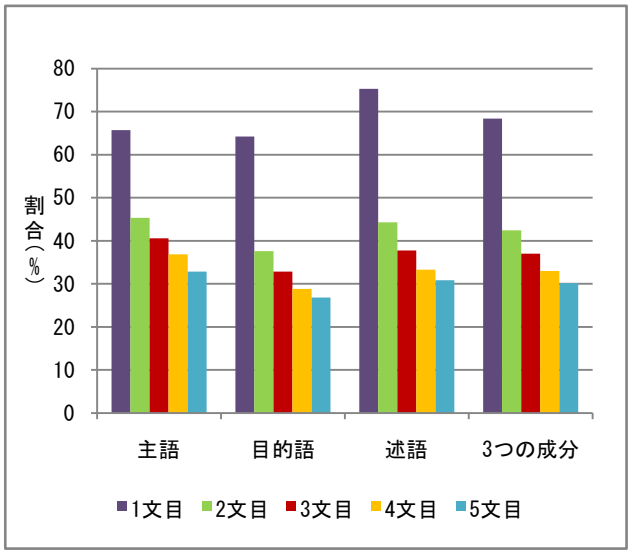

図 2. 記事本文の文の成分が見出しに含まれている割合

5. 考察

図 2 の結果より，記事の第 1 文目に記載されている文の成分が見出しに使用されている割合が高いことが分かる．これは，2 に記したように，記事の 1 文目に重要な情報が記載されているためだと考えられる．

また，見出しと記事本文を比較した結果，記事本文に記載されている名詞が，見出しでは略称や言い換え表現によって記載されている記事がいくつかあることが分かった．これは，記事の内容を極力少ないテキスト量で表すためであると考えられる．

また，図 2 の結果では文の成分による違いは少ないことが分かった．

6. まとめ

本研究では，見出しの自動生成を行うために，見出しに使用されている文の成分が，記事本文のどの位置(何文目)に存在しているのかを調査した．結果より，記事の 1 文目に記載されている文の成分が見出しに使用されている割合が高いということが分かった．これらの調査をもとに，今後は見出しの自動生成に関する研究へと発展させていく．

参考文献

[1] 佐藤 理史，”13 文字で何が伝えられるか:ウェブニュースボックス見出しの分析”，言語処理学会 第 14 回年次大会 C3-7,pp508-511(2008)
[2] 前迫 綾，竹川 美希，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，O-497(2008)
[3] 吉田 文彦，”日本語の報道記事を対象とする事象データ抽出システム”，東海大学紀要文学部，(2002)
[4] 形態素解析エンジン MeCab，工藤 拓 http://mecab.sourceforge.jp/