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

佐　藤　理　史

名古屋大学大学院工学研究科電子情報システム専攻
ssato@nuee.nagoya-u.ac.jp

1. はじめに

ある一定量を超えるテキストには、見出しやタイトルが付けられるのが普通である。見出しやタイトルは、そのテキストの内容を端的に伝える役目を果たしており、特に新聞記事のような大量の情報群において、読者が読みたい記事を取捨選択する際の強力な支援ツールとして機能している。しかしその一方で、この機能を支える仕組みの解明、すなわち、「見出しは、非常に短いのにもかかわらず、なぜ、このように情報の取捨選択を効果的に支援するのか」や、見出しの優劣の判定基準、すなわち、「どのような見出しが優れた見出しなのか」といった問いに対する答は、ほとんど明らかになっていない。

このような背景より、本論文では、次の 3 つの疑問に答えることを主眼に、ウェブニュースの見出しを分析する。

- 見出しは、どんな「構造」をしているか。
- 見出しの「短さ」は、どのように達成されているか。
- どうすれば、「見出し風」のテキストを自動生成できるか。

分析対象には、ウェブページにおいて横幅制限のあるボックス内に表示され、ニュース記事本文を表示するページに直接リンクが張られている見出し（以後、これをボックス見出しと呼ぶ）を選ぶ。ボックス見出しは、長さのばらつきが小さいことから、長さに対するかなり強い制約の下で作成されていると推察される。

2. 分析対象

サンケイウェブ (http://www.sankei.co.jp) の NewsMinute のデータを、2005 年 6 月 12 日から 2006 年 12 月 4 日にかけて収集した（このウェブページは、すでに消滅している）。NewsMinute は、適宜更新されるため、1 時間に 1 回、その時点で掲載されているニュース記事をすべて収集し、それらから完全に同一の記事を削除してデータ化した。ニュース記事の総数は、40,043 件である。ニュース記事の例を表 1 に示す。各ニュースの最初の文が、ボックス見出しである。

ここでは、収集したボックス見出しのうち、2006 年 11 月 6 日から 11 月 10 日までの 5 日間、総計 225 件の見出しを詳細に分析した。

表 **2**　見出しの長さの分布

			アサヒドットコム	
文字数	サンケイ	グー	全体	第一要素
3	0	0	0	1
4	0	0	0	1
5	0	0	0	3
6	3	0	0	8
7	4	0	0	11
8	5	0	0	13
9	12	0	1	37
10	21	0	2	43
11	46	0	4	65
12	59	4	3	58
13	74	16	9	67
14	1	57	15	62
15	0	183	16	85
16	0	0	17	62
17	0	0	23	45
18	0	0	32	56
19	0	0	42	39
20	0	0	48	30
21	0	0	62	32
22	0	0	69	24
23	0	0	75	17
24	0	0	107	8
25	0	0	131	13
26	0	0	137	16
27	0	0	3	0

3. 見出しの長さ

まず、ボックス見出しの長さについて調査した。比較のため、同時期の次の 2 種類のニュース記事の見出しの長さも調査した。

- グー (http://www.goo.ne.jp) で RSS として提供されるニュース記事の見出し 260 件
- 朝日新聞のウェブ (http://asahi.com) で RSS として提供されるニュース記事の見出し 796 件

なお、グーの見出しはボックス見出しに相当する（ウェブではボックス見出しとして表示される）が、朝日新聞の見出しはボックス見出しではない。

見出しの長さは、全角換算の文字数（0.5 文字は切り上げ）として計測した。なお、アサヒドットコムは、収集した見出し全体、および、見出しの最初の要素（全角空白までの部分）、の両方を調査対象とした。

計測結果を表 2 に示す。この表から、ボックス見出しの長さが比較的良く揃っていることがわかる。グーは、ボックスの幅が 15 文字であり、ほとんどの見出しが 13 文字

表 1　サンケイウェブの NewsMinute のニュース記事の例

日時	ニュース記事
061106.07:16	青森、岩手両県で震度３。震源地は岩手県沖で、震源の深さは約５０キロ。Ｍ４．７と推定。津波の心配はないという。
061106.08:26	伊産ワインの新酒が解禁。業界は仏産のボジョレに負けじと販売に本腰。ローマのレストランで愛好家らが初物味わう。
061106.09:08	米ニューズとＳＮＳで提携。ソフトバンク。日本語版マイスペースを手掛ける合弁会社を月内に設立、サービス開始へ。

から 15 文字に「調節」されている。サンケイウェブは調節度が若干甘いが、大多数の見出しは 10 文字から 13 文字の範囲に入る。

これに対して、ボックス見出しではないアサヒドットコムの見出しは、長さにはかなりのばらつきがある。おそらく、見出しは 30 文字以下というゆるい制約で編集されているものだと推測される。この表の「アサヒ全体（見出し全体)」と「アサヒ第一要素（見出しの最初の要素)」の２つの列の違いから、アサヒドットコムの見出しの多くは、複数の要素から構成されていることがわかる。そして、第一要素のピークは、15 文字のところにある。

ボックス表示の幅（最大文字長）は、ウェブページの構成上の要請（できるだけ短いテキストで）と伝達すべき情報からの要請（その長さで見る気にさせるのに十分な情報を伝達できる）との兼ね合いで設計されたと考えられる。サンケイウェブ、グー以外でも、ヤフー (http://www.yahoo.co.jp) やエキサイト (http://www.excite.co.jp/) でも、ボックス見出しは、同様に 13 文字から 15 文字のものが多い。

『記者ハンドブック』[1] によれば、ニュース記事の見出しの字数は、「主見出し、脇見出しとも 12 字以内、三本目は 11 字以内」とある。紙媒体の記事では複数の見出しを許すので状況は多少異なるが、見出しの字数という点では、既存の紙媒体もウェブも大きな違いはないと考えてよい。

以上のことから、次のような帰結を得る。

(1) 日本語で、ニュース記事が伝える情報を短い見出しとして要約する場合、**10** 文字台前半 **(12** から **15** 文字**)** が目安となる。

(2) この分量のテキストで、読者が記事を取捨選択できるだけの情報を伝達することができる。

4. 見出しの構造分析

4.1 構造分析の方針

次に調べるべきことは、このような短い見出しの内部がどのような構造となっているかということであろう。

見出しの構造を分析するためには、何らかの基本方針が必要である。ここでは、次のような考え方を採用する。

(1) ニュースはコトを伝える。ゆえに、見出しもコトを伝える。

(2) 見出しの長さは短いので、見出しが伝えるコトは、ほとんどの場合、１つだろう。

(3) １つのコトは、単文で表現できる（要約できる)。ゆえに、見出しに対応する、(日本語の文として形の整った）単文を考えることができる。

(4) 文の骨格は、述語、補足語、修飾語、主題の 4 つの要素から構成される [2]。

(5) 見出しと単文の対応に基づき、見出しを上記の構成要素に分解する。こうして得られたものを、見出しの構造と考える。

表 3　見出しの構造解析例

見出し	＝	全国で銅の窃盗相次ぐ
文	＝	全国で銅の窃盗が相次いでいる
補足語	＝	全国で ＝ 全国で
補足語	＝	銅の窃盗が ＝ 銅の窃盗
述語	＝	相次いでいる ＝ 相次ぐ
見出し	＝	上越新幹線が一時見合わせ
文	＝	上越新幹線が運転を一時見合わせている
補足語	＝	上越新幹線が ＝ 上越新幹線が
補足語	＝	運転を ＝ ϕ
連用修飾	＝	一時 ＝ 一時
述語	＝	見合わせている ＝ 見合わせ
見出し	＝	エスカレータで１１人けが
文	＝	エスカレーターで１１人がけがをした
補足語	＝	エスカレータで ＝ エスカレータで
補足語	＝	１１人が ＝ １１人
述語	＝	けがをした ＝ けが
見出し	＝	千葉の薬局に２人組強盗
文	＝	千葉の薬局に２人組の強盗が押し入った
補足語	＝	千葉の薬局に ＝ 千葉の薬局に
補足語	＝	２人組の強盗が ＝ ２人組強盗
述語	＝	押し入った ＝ ϕ

表 4　見出しの大分類

分類	見出し数
１コト	132
アンカー＋１コト	60
２コト	23
アンカー＋２コト	2
１コト未満	2
その他	6

(5) 見出しと単文の対応に基づき、見出しを上記の構成要素に分解する。こうして得られたものを、見出しの構造と考える。

上記の考え方を要約すれば、「見出しをある特別な形式の文とみなし、文として構造分析する」ということである。見出しの構造分析例を表 3 に示す。なお、ここでは、連体修飾語は独立した要素とはせず、補足語内に含めた。

4.2 見出しの大分類

上記の考え方に沿って見出しを分析した結果、１つのコトを伝える見出し以外に、２つのコトを伝える見出しや、伝達内容が１つのコトに満たない見出しがあることがわかった。これらを考慮し、見出しを次のように分類した。

(1) 「１コト」見出し

１つのコトを伝える見出し。最も典型的であり、単文（無題文）に対応する。

(2) 「アンカー＋１コト」見出し

「アンカー」とは、そのニュースが「何についての」ニュースなのかを知らしめる役割を果たす表現をいう。書き手は、読み手が「アンカー」によって、それが指し示すコト・モノを想起することを仮定しており、「アンカー」の後に続く「コト」は、それに対する新情報を伝える。「アンカー」は、通常の文では、「は」でマークされることが多い（この場合は、有題文となる）が、他の格助詞をとることもある。

- 「東京円、１１７円台後半」
- 「米中間選挙、全米で開票進む」
- 「輸入牛肉、未申告部位が混入」

(3) 「２コト」見出し

関連する２つのことを伝える見出しで、複文に相当する。

- 「北海道で竜巻、８人死亡」＝北海道で竜巻が発生し、８人が死亡した
- 「郵便局強盗、５００万円奪う」＝郵便局に強盗が入り、５００万円を奪った

(4) 「アンカー＋２コト」見出し

「２コト」見出しの有題バージョン。

- 「中田はフル出場も得点なし」＝中田は、フル出場したが、得点はしなかった

(5) 「１コト未満」見出し

述語が完全に省略され、キーワードだけから構成される見出し。

- 「６３００万円の限定ベンツ」(が発売された)
- 「ＫＡＴ－ＴＵＮ盗写ＤＶＤ」(を頒布目的で所持していた女子大生が逮捕された)

(6) その他

発言の一部を引用、など。

調査対象の 225 件の見出しを分類した結果を表 4 に示す。この表より、「１コト」と「アンカー＋１コト」が大多数 (192/225=85%) を占めていることがわかる。

4.3 述語と述語相当語

「１コト」見出しの基本構造（骨格）は、次のいずれかである。

(1) いくつかの補足語 + 述語

(2) いくつかの補足語 + 述語相当語（名詞）

前者は述語に相当する部分が明示的に存在する見出し、後者は明示的に存在しない見出しである。実際の構造では、基本構造に修飾語や末尾表現が追加されることがある。

述語は、動詞（基本形か連用形)、サ変名詞、形容動詞語幹のいずれかの形をとる。このうち、動詞の基本形を除き、名詞相当であり、見出しは、みかけ上、名詞句となる。

明示的な述語を持たない見出しの最後の補足語に着目すると、それらに次のような類形が観察される。

(1) 動作性を含む名詞

- 「政府の教育再生会議が 初会合」＝政府の教育再生会議が初会合を開いた

この例に示すように、対応する文では述語にはならないが、その名詞自体が動作性を帯びている。

(2) 数値・期日

- 「貿易黒字、１５００億ドル に」
- 「青森、岩手両県で 震度３」
- 「２被告に 懲役１７－１５年」

(3) 気象・災害

- 「北海道奥尻島で 突風」

(4) 犯罪

- 「千葉の薬局に２人組 強盗」

(5) 人

- 「米クレイ賞に理研の 渡辺氏」

上記の名詞は、比較的内容性に乏しい述語を伴って文を構成するが、事態（コト）の中核は、その名詞が担うと考えるのが自然である。本論文では、これらの名詞は述語性を帯びているとみなし、述語相当語として扱う。

「１コト」および「アンカー＋１コト」見出し（但し、市場概況を伝える 27 個を除く）165 個に対して、述語または述語相当語を持つものの数を調べた結果を表 5 に示す。この表に示すように、ほとんどの見出しは、述語または述語相当語を持つ。すなわち、見出しは、見かけ上、名詞句に見えるが、文として分析するのが適切である。

表 5 述語および述語相当語を持つ見出し数

述語	118
述語相当	43
その他	4

表 6 末尾表現

末尾表現	意味
へ	未来
か	不確定
も	「〜もありえる」
の見通し	確実性の高い予想
の勢い	確実性の高い予想（選挙）
の恐れ	不確定な危険性

4.4 末尾表現

見出しは、述語の後ろに助詞などを伴うことがある。これを末尾表現と名付ける。これらの表現は、ある種のモダリティを表し、通常の文における助動詞・終助詞に相当する。表 6 に例を示す。

4.5 構成要素数

表 5 に示す 165 個の見出しを構成要素に分割し、その数と要素の長さを調べた。構成要素への分割では、述部(述語および述語相当語；末尾表現を含む)、アンカー部(直後の読点を含む）はそれぞれ 1 つの要素として扱い、残りの部分は、助詞または述語の連体修飾があれば分割するという方針で行なった。その結果を表 7 に示す。この表に示すように、見出しは 3 要素が最も多く、2 要素

表 7　構成要素数とその長さ

		要素長				
構成要素数	記事数	3	2	1	述部	平均
2	41			7.39	3.71	5.55
3	98		5.14	3.80	2.85	3.93
4	26	3.73	3.46	2.58	2.54	3.08

と 4 要素と続く。当然のことながら、平均要素長は、構成要素が多くなるほど短くなる。

3 要素ということは、「何ガ 何ヲ ドウシタ」を盛り込めるということである（例：「アイオワ州知事が/出馬/表明」）。構成要素長が短かければ、さらに要素を盛り込むことができる（例：「ヒラリー氏/早々に/再選/決める」）。一方、構成要素長が長ければ、どれかの要素を省略することになる（例：「ネットいじめ動画を/放置」）。もちろん、その場合は、構成要素により多くの情報が盛り込まれることなる。長い構成要素は、多くの場合、複合名詞の形式をとる。

表 7 が示すもう一つの興味深い事実は、構成要素の平均要素長は、要素数にかかわらず、先頭要素が最も長く、徐々に短くなっていき、末尾の述部が最も短くなるという点である。平均要素長と伝達情報量に相関があると考えれば、「できるだけ前の方に、情報量が多い要素を持ってくる」という法則が働いているとみなすことができる。

5. 見出しの短さの秘密

5.1 要約、プロトコル化、省略

すでに述べたように、見出しはある特別な形式の文とみなすのがよい。すなわち、見出しは、伝達すべき情報の単文要約である。複雑な内容の情報も、その中核的内容は単文に要約できるという事実が、短い見出しを作ることを可能としている。

比較的よく現れる情報タイプに対しては、見出しの定型化（プロトコル化）が進む。これにより、非常に効率の良い情報伝達が可能となる。

- ＜市場＞、＜状況（金額）＞
  ＝「東京円、１１７円台後半」
- ＜地域＞、震度＜数字＞
  ＝「青森、岩手両県で震度３」

限られた文字数では、単文の全ての要素を詰め込むことができないことが多い。そのような場合は、重要な要素を優先して残し、それ以外は思い切って省略する。見出しは、それ単体では、読者を見る気にさせれば十分であり、それだけで完結した情報を伝達する必要はない。それを逆手にとって、情報不足により「おや？」と思わせ、見る気にさせるという方法もある。実際、表 1 の最後の記事では、「提携」の動作主格である「ソフトバンク」は見出しには存在しない。このような見出しには、「どこが？」と思わせる効果がある。

5.2 縮　　約

それぞれの構成要素においては、その要素自身を短くする工夫がなされる。典型的には、つぎのようなものがある。

- 述語
  - テンス・アスペクト等の省略、名詞化
    「全壊家屋の下から猫 見つかる」（見つかった）
    「タンカーが浅瀬に 乗り上げ」（乗り上げた）
  - サ変名詞の使用
    「ラムズフェルド国防長官辞任」（した）
    「さくらやを連結子会社化」（した）
- 連用修飾語
  - 漢語副詞の複合述語化
    「日本人イラストを 初 採用」
    「ロナウジーニョが 連続 受賞」
- 補足語
  - 最後の補足語の助詞の省略（複合名詞化）
    「０８年に ハイブリッド車 生産」
  - 省略形の使用
    「米選抜 が７２年ぶりの全勝」
- 連体修飾語
  - 「の」の省略、複合名詞化
    「日本アジア航空機客室 で煙」

6. おわりに

本研究で得られた知見は、次のようにまとめることができる。

- 見出しの長さは 10 文字台前半である。
- この長さで、「コト」の中核を伝え、関心のある人を読む気にさせるのに十分な情報を伝達できる。
- 見出しは、短くするために特殊な形式を取っているが、通常の文に似た構造を持つ。
- 見出しの短さは、単文要約、プロトコル化、省略、縮約の合わせ技で達成されている。

論文の冒頭で設定した疑問のうちで残されているものは、どのようにすれば「見出し風」のテキストを自動生成できるかという疑問である。見出しは特殊な形式な文であり、その構造は通常の文とかなりよく対応する。このことから、通常の文を「見出し風」の特殊な形式に変換するというアプローチが考えられる。その特殊形式は、プロトコル化、省略、縮約の結果として現れているので、これらの操作を形式化すれば、文から見出し風のテキストを生成することが可能となる。

謝辞　本研究は、科学研究費補助金 萌芽研究「オンラインニュース見出しの言語構造と情報構造の解明」(課題番号 18650032）の支援を受けた。

参　考　文　献

1) 共同通信社. 記者ハンドブック 第 10 版. 共同通信社, 2005.
2) 益岡隆志, 田窪行則. 基礎日本語文法–改訂版–. くろしお出版, 1992.