英文記事ヘッドラインの特徴について

4B－1

白井 諭*1 大山芳史*1 中尾嘉孝*2 西垣万亀子*2 上田洋美*2 小見佳恵*2

*1NTTコミュニケーション科学研究所　*2NTTアドバンステクノロジ(株)

1 はじめに

新聞記事において，日本文記事の見出しと英文記事のヘッドラインとでは書式に大きな違いがある。日本文見出しはキーワードを含む断片的な表現が多用されるのに対し，英文ヘッドラインは文に準じた形式で書かれている。このため，見出しとヘッドラインの間の翻訳には単純に機械翻訳を適用することはできないと思われる。

日本文見出しや英文ヘッドラインは記事の内容を簡潔に表現したものであるから，これらの間の翻訳ではなく，記事の本文に基づいて見出しやヘッドラインを生成することが考えられる。記事の本文においては，第１文に本質的な事柄を盛り込む傾向がある。従って，見出しやヘッドラインと第１文との相関性が強いことが予想される。

筆者らが開発中の日英翻訳システム[池原 87]は，事実報告的な記事のリード（要約または第１段落）に対し，未知語を登録すれば 6～7 割の訳文合格率(英訳文だけで意味がわかる文の割合)が得られる。特に，第１文の訳文合格率は８割に達するが，これは 5W1H 的な要素の省略がなく文意が明確であることが要因であると思われる。以上から，日本文記事を英訳し，その結果に基づいて英文ヘッドラインを生成する方法が考えられる。

本稿では，英文ヘッドラインを英文記事本文に基づいて生成する方法を検討する。具体的には，英文ヘッドラインの形式の分類，ヘッドラインと英文記事本文および第１文における単語の出現傾向，ヘッドラインと本文および第１文の相関性の分析を行なう。それらの結果に基づいて，英文ヘッドラインを生成するための基本方針を整理する。

Characteristics of English Newspaper Article Headlines

Satoshi SHIRAI*1, Yoshifumi OOYAMA*1, Yoshitaka NAKAO*2, Makiko NISHIGAKI*2, Hiromi UEDA*2 and Yoshie OMI*2

*1NTT Communication Science Laboratories and *2NTT Advanced Technology Corporation

2 ヘッドラインの形式

本稿で用いた英文記事は，日本経済新聞社のデータベース Japan News and Retrieval に日経四紙の主要記事の英訳として登録されている 1995 年 8 月の 2,875 記事である。

英文ヘッドラインの表現形式は表 1 のように分類される。このうち出現数の多い述語形式が現在形のものは出来事の報告であり，to 不定詞は計画の内容である。これは，[高橋 96]の方法により対応する日本文記事を取り出すと，No.1 の例では「自由米市場創設を提言」と「同友会が関税化で」が，No.2 では「自工振、車情報盛り込んだＣＤ－ＲＯＭ発売」が見出しであることから確認される。なお，No.9 は「東証株式１部大引け」のことであり，具体的な内容には触れていないので，別に１項目を立てた。

表1　英文ヘッドラインの分類

No	述部形式	出現数(割合)	英文ヘッドラインの例 注
1	現在形	1,174 (40.8%)	Business leaders recommend loosening controls on rice
2	to不定詞	1,040 (36.2%)	Automobile industry group to offer car guidebook CD-ROM
3	過去分詞	94 (3.3%)	U.S. credit rating firms worried about Japan's banks
4	現在分詞	78 (2.8%)	City banks greatly expanding extension of housing loans
5	助動詞	34 (1.2%)	BOJ may have helped Cosmo with credit line
6	過去形	28 (1.0%)	Tokyo gov't told Cosmo to cut interest rates before failure
7	(形容詞)	21 (0.7%)	Eurobond market active after deregulation
8	(名詞句)	21 (0.7%)	Plain paper-based modeling system now on the market
9	(表題的)	385 (13.4%)	Tokyo Stocks 1st Sec Cls
合　計		2,875 (100%)	注 "～: MITI" や "Earnings: ～" のような付加部分は取り除いた

3 単語の出現傾向

ヘッドライン，記事第1文，記事本文（記事第1文を含む）の別に，単語ごとの出現度数を集計した結果を表2に示す。ここでは，簡単のため，固有名詞等の大文字化に伴う違いは無視して集計した。

記事第1文と記事本文全体では順位の変動はあるものの単語の出現傾向は似ているのに対し，ヘッドラインではかなり異なっている。ヘッドラインでは，theは80位（31回）aは84位（30回）とあまり使われず，FY95 (= fiscal 1995, 18位, 95回) やgov't (= government, 39位, 54回) のような簡略化が見られる。ヘッドラインの述部形式が現在形とto不定詞の記事を別に集計してみたが，若干の順位変動が見られるだけで，単語の出現傾向はよく似ていた。

表2　単語の出現度数

順位	ヘッドライン		記事第1文		記事本文全体	
	度数	単語	度数	単語	度数	単語
1	1,456	to	3,582	the	25,644	the
2	553	in	2,412	to	13,375	to
3	269	Tokyo	2,340	of	11,261	of
4	250	for	1,811	in	9,260	in
5	237	of	1,724	a	7,997	and
6	180	up	1,323	and	7,145	a
7	167	on	845	for	4,539	yen
8	144	stocks	765	Co.	4,436	for
9	139	yen	756	said	3,810	will
10	118	profits	687	yen	3,136	is
11	115	Cls	682	will	3,122	on
12	115	Mng-cls	673	on	2,292	by
13	111	new	609	its	2,901	at
14	107	Japan	534	by	2,538	from
15	105	sales	511	has	2,429	billion
16	101	market	503	Corp.	2,293	its
17	95	bank	503	from	2,251	year
18	95	FY95	483	year	2,248	with
19	92	firms	409	with	2,232	said
20	92	Sec	394	at	2,185	as
	延べ20,679語		延べ74,899語		延べ394,173語	
	平均単語長5.40		平均単語長5.29		平均単語長5.11	

4 本文との相関性

次に，ヘッドラインと記事中の各文との相関を共通単語数で測定した。ただし，前置詞と冠詞は共通語数には加えず，単数と複数，現在形と過去形の違いなどは共通語数に加えた。結果を表3に示す。

表から，全般的にはヘッドラインと第1文の相関性が特に高いことがわかる。この場合も，ヘッドラインの述部形式の違いによる相関性の差異はあまり認められなかった。

表3　ヘッドラインと記事中の各文の相関

文の位置	文数	ﾍｯﾄﾞﾗｲﾝと各文の共通単語数 (A)	照合対象の単語総数			
			ヘッドライン (B)	(A/B%)	記事中の各文 (C)	(A/C%)
1	2,875	10,183	18,725	54.4	60,273	16.9
2	2,860	3,708	18,633	19.9	48,946	7.6
3	2,820	3,572	18,375	19.4	45,992	7.8
4	2,683	3,236	17,458	18.5	42,343	7.6
5	2,315	2,538	14,906	17.0	35,620	7.1
6	1,766	1,751	11,334	15.4	26,448	6.6
7	1,214	1,229	7,803	15.8	18,158	6.8
8	776	744	5,074	14.7	11,645	6.4
他	1,524	1,372	10,130	13.5	24,010	5.7
計	18,833	28,333	122,438	23.1	313,435	9.0

5 考察

前節までの検討で，ヘッドラインと記事第1文は相関性が高いことが確認された。本稿のような単純な方法では，例えば図1で斜体で示した4単語以外に，意味的に対応する“up”と“grew”，“nearly 16%”と“15.9%”，“year on year”と“year earlier”が集計されないという問題がある。これを考慮するとさらに相関性は高くなると予想される。逆に，第1文からヘッドラインを作る際には，主要な要素を取り出すだけでなく，単純化のための置き換えが必要であることを示唆していると云える。

NEWS: *July truck sales* up nearly 16% *year* on year

Sales of medium and heavy-duty *trucks* with a load capacity of four tons or more grew 15.9% from a *year* earlier to 13,182 units in *July*, industry sources said Tuesday.　（以下略）

図1　ヘッドラインと第1文の相関分析の例

6 おわりに

本稿では，共通に使用される単語の調査により，英文ヘッドラインと英文記事第1文の相関性が強いことを示した。今後は，共通単語がヘッドライン語数の2割を下回った500記事を中心に意味的な対応について整理するとともに，第1文からのヘッドラインの生成方式の検討を進める予定である。

謝辞　本検討にご協力下さった井上浩子氏を始めとするNTTアドバンステクノロジの各位に感謝する。

参考文献

[池原 87] 池原, 宮崎, 白井, 林: 言語における話者の認識と多段翻訳方式, 情報処理学会論文誌 Vol.28 No.12, pp.1269-1279

[高橋 96] 高橋, 白井, 藤波, 池原, 上田, 松島: DBから抽出した日英新聞記事の自動対応付け, 言語処理学会第2回年次大会 B3-3, pp.201-204