1S−4 英語ニュース文における ハイフンを含む語の局所解析

加藤　直人*　浦谷　則好*　中瀬　純夫**
*NHK放送技術研究所　**(株)カテナ・リソース研究所

1．はじめに

ニュース文では少ないスペースの中に多くの情報を含めなければならないため、名詞を修飾する句（形容詞句や同格の名詞句）が多く、これらはハイフンを含む語（ハイフン語）となりやすい。ハイフン語の中でも接頭語、複合語、様々なタイプの造語によるハイフン語が多く、辞書に登録しきれない。

本稿では、英語ニュース文に出現したハイフン語について述べ、その局所的な解析手法について考察したので報告する。

2．英語ニュース文のハイフン語

平成元年3月～6月の90日分の外電中（総語数約350万語、異なり語数約7万5千語）[1] に出現した頻度2回以上のハイフン語は総数で約 31300語であった。このうち、異なり語数は約3400語である。表1にハイフン語における異なり語数に対する頻度別語数の割合を示す。表2に出現頻度が多いハイフン語の上位10語を示す。ここで頻度を数える際に、例えば "45-year-old" と "57-year-old" などのような、数字部分のみが異なる語は同じ語として扱うため数字は "¥" で置き換えて頻度を数えた。また、屈折形は異なり語とした。

表1　頻度別割合

頻度(回)	割合(%)
100以上	1
10～100	15
5～9	18
4	10
3	18
2	38

表2　ハイフン語上位10語

順位	頻度(回)	ハイフン語
1	1899	¥-¥
2	1499	¥-year-old
3	475	cease-fire
4	435	pro-democracy
5	404	¥-member
6	393	¥-year
7	366	short-range
8	242	¥-dollar
9	231	state-run
10	214	¥-million-dollar

表1から出現頻度が少ない語ほど語数が多いことがわかる。頻度が少ないハイフン語には固有名詞を含む造語が多かった。

表2をみると、上位には数字を含むハイフン語が多いことがわかる。ハイフン語総数の約28%は数字を含むハイフン語であった。

3．ハイフン語の解析

3．1　辞書だけによるハイフン語解析

ハイフン語の中でも通常使われる語は辞書に登録されている場合もあり、特殊な解析を要しない語もある。そこで、ハイフン語もはじめの解析過程では辞書引きする。ただし、ハイフン語は複合語である場合もあるので次のような処理を行ないながら辞書引きする。

解析Ⅰ　ハイフンを含むままで辞書引き

ハイフンがついたままでハイフン語を辞書引きする。

我々の英日辞書は、一般語辞書（見出し語数約4万語）、ニュース専門語辞書（同約6万語）の2つであり、この中に約6千語のハイフン語があった。

解析Ⅱ　ハイフンを空白に置き換えて辞書引き

"human-right" は、ハイフンを空白に置き換えて "human right" で辞書引きすれば登録されている。

解析Ⅲ　ハイフンを取り去って辞書引き

例えば、"post-war" は、ハイフンを取り去って "postwar" で辞書引きすれば登録されている。

Local analysis for hyphenated words in English news
Naoto KATOH*, Noriyosi URATANI*, Sumio NAKASE**
*NHK Science and Technical Research Laboratories
** Catena-Resource Laboratories Inc.

3．2　形態素処理によるハイフン語解析

ここの解析では、ハイフン語を構成する単語の表層上の特徴や品詞情報を使って処理する。

解析Ⅳ　ＬＯＣＴ処理

我々の機械翻訳システムでは

ＬＯＣＴ（LOcal Context Translation）

と呼ぶ処理[2]の中でＣＦＧルールとＬＯＣＴ辞書を使い、局所的な情報のみでまとめあげられる定型的表現（固有名詞[3]、数量・時制表現等）を複合語と認定し、訳語を与えている。

一方、ハイフン語の中には

"¥-year-old"、"¥-milion-dollar"

のような数量表現があり、既にＬＯＣＴで処理している。

また、英語ニュース文のハイフン語には

"New York-based"、"Ethiopia-Sudan border"

のように、地名や国名等の固有名詞を含む定型的表現（「地名－過去分詞」、「国名－国名　名詞」）がある。そこで今回、固有名詞を含むハイフン語も解析できるようにＬＯＣＴ処理を拡張した。

ＬＯＣＴでは、アルファベットの大文字で始まる文字列で未知語である語を固有名詞として認定する機能がある。したがって、"Damascus" が未知語であっても "Damascus-based" は「Damascusに本拠を置く」と翻訳することができる。

解析Ⅴ　その他の場合のハイフン語処理

Ⅳまでの過程で処理できなかったハイフン語の中には、

1)　分詞を含む語

例　"government-funded"、"making-policy"

2)　接頭語（re-、co-、pre-等）を含む動詞

例　"re-elect"、"co-sponsor"

3)　接頭語（anti-、ex-等）を含む名詞

例　"anti-apartheid"、"ex-president"

等の特徴的なものがある。そこでこの特徴に注目して次のようなハイフン語解析を行う。

1)「名詞＋過去分詞」、「副詞＋過去分詞」、「名詞＋現在分詞」、「副詞＋現在分詞」を処理。

例えば「名詞（Ｎ）＋過去分詞（Ｖ）」の場合、「ＮによってＶされた」と翻訳する。

例：　"government-funded"

「政府によって資金を供給された」

この中で、「ｘ＋過去分詞」の "ｘ" が名詞と副詞の両方の品詞を持つ場合は、品詞についているペナルティの小さい方を選ぶようにした。

2)「接頭語（Ｐ）＋動詞（Ｖ）」は「ＰＶする」。

例：　"re-elect"　「再び選ぶ」

3)接頭語付きの名詞

例：　"anti-apartheid"　「反アパルトヘイト」

"ex-president"　「前の大統領」

Ｖでも処理されなかった語が未知語となり、訳は英単語のまま出力される。

4．結果

2．で述べたハイフン語3416語に対して、処理Ⅰ～Ⅴを行ない、各処理の結果を表3に示す。ただし、ハイフン語処理を行なった結果が品詞、訳語として適切であるかどうかは文中に出現した具体例によって判断されなければならないが、その評価はしていない。単にハイフン語のみを見ただけで判断した。

表3　処理結果

処理	処理率(%)
Ⅰ	18
Ⅱ	2
Ⅲ	3
Ⅳ	26
Ⅴ	14
合計	63

処理できなかったハイフン語には "half-half"のように一般的処理ができないものや、"white-controlled"（「白によって制御された」）のように訳語選択が失敗したもの（正解は「白人によって」）があった。

5．おわりに

以上、英語ニュース文に出現するハイフン語についてその処理方法を検討し、処理実験結果について述べた。

今後は文中での翻訳結果を見て、本処理の有効性について確かめたい。また現在、処理Ｖの規則は非常に少ない。処理の数を増やせば、処理率はさらに向上するものと思われる。さらに動詞を含む語の場合には意味マーカーに使うことによって適切な訳語が選択できるようにしたい。

【参考文献】

[1]浦谷他：「英語ニュースデータベースの構築」、情報処理学会第41回全国大会(1990)

[2]相沢他：「衛星放送ワールドニュースの英日機械翻訳」情報処理学会第40回全国大会(1990)

[3]加藤他：「英日機械翻訳における固有名詞処理」、情報処理学会第40回全国大会(1990)