対象言語のフリーテキストを用いた複合名詞の訳語選択

5B−4

加藤　直人
NHK放送技術研究所

1．はじめに

英語ニュース文には，形容詞＋名詞や名詞連続等の複合名詞が多数出現する．このような複合名詞すべてをあらかじめ機械翻訳用辞書に登録しておくことは困難である．したがって翻訳の際に，複合名詞を構成するそれぞれの語の訳の中から適切な訳語を選択し，複合名詞全体として適切な訳を作らなければならない．

最近，訳語選択に例文や知識ベースを使った手法が提案されている[佐藤91][野美山91]．しかし，これらの手法では，細かい意味情報を付与したデータを人手によって構築しなければならない．したがって，大量のデータを作成するのは困難であり，実験レベルに留まっている．一方，原文のままであるフリーテキストの大規模なデータとして，新聞や事典等が電子化されてきている．最近のハードウエア技術の発展にともない，これらの中から高速にデータの検索ができるようになってきた．

本稿では対象言語のフリーテキストを使うことにより，複合名詞の適切な訳語選択をする手法について述べる．本手法は，データを作るという，人手を使った面倒な作業を必要としないので，実現が比較的容易である．

以下では，英日翻訳する場合を考える．

2．複合名詞翻訳の問題点

複合名詞 "advanced cruise missile technology" を例にとって訳語選択の問題を考える．小学館の『最新英語情報辞典』[情報辞典89]には「先進巡航ミサイル技術」と訳してある．

さて，複合名詞を構成する語が次のようにそれぞれ訳語を持っているとする．

"advanced"
【形容詞】「先進の」「進歩した」「高等の」「ふけた」「普通より高い」
"cruise"
【名詞】「船旅」「巡航」「遊覧」
"missile"
【名詞】「ミサイル」「誘導弾」「飛び道具」
"technology"
【名詞】「テクノロジー」「技術」「科学技術」「応用科学」

この複合名詞は文法的には形容詞＋名詞ということになる．またこの例のように，それぞれの語は複数の訳語を持つ．それぞれの第一訳語を単純に使えば，この複合名詞は「先進の船旅ミサイルテクノロジー」と誤訳されてしまう．すなわち，次の問題がある．

1）複合名詞を構成する名詞の訳語選択をしなければならない．

ある語が複数の訳語を持つ場合，この訳語を頻度順に並べたり，前回使用したものを第一訳語にすることによって，通常，訳語選択をしている．しかし，それぞれが単独で使われた場合のみを想定して訳語の順番を決めているので，複合名詞を構成するときにはこの第一訳語が必ずしも適切ではない．また，分野を限って訳語選択することも考えられるが，例えば"technology"の訳語「テクノロジー」，「技術」・・のように訳語が類語であると，分野を限っても区別できない．

2）日本語では，例えば形容詞「先進の」が名詞句を修飾する場合，形容詞語尾「の」を取り除かなければならない．

この問題に関して自動的に対処しているものはなく，通常，後処理の中で人手によって削除されている．したがって，このような語が多いと人間にかかる負担も大きくなる．

3．フリーテキストを使った複合名詞の訳語選択

本手法は，大まかに2段階に分かれている．始めに，複合名詞を構成する語の訳語の組み合わせすべてによって，訳語候補を作る．次にこれらとフリーテキスト中の語とを照合し，同じものがあればそれを訳語とする．具体的なフローチャートを図1に示す．

図1を簡単に説明する．始めに構文解析を行ない複合名詞を抽出する．複合名詞が形容詞を含んでいるならば，「の」（例，「先進の」），「な」（例，「単一な」），「的な」（例，「世界的な」）等の形容詞

Machine translation of compound nouns by matching with texts in target language
Naoto KATOH
NHK Science and Technical Research Laboratories

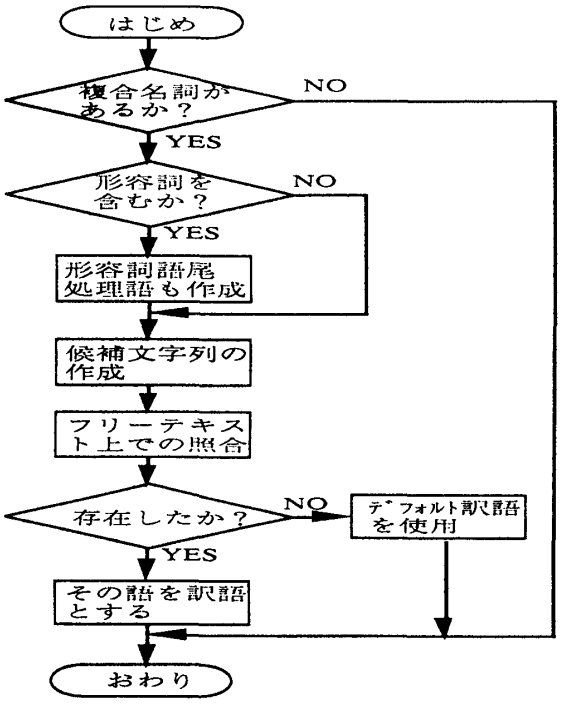

図１　複合名詞の訳語選択のフローチャート

語尾を取った翻訳結果も同時に作る．先の例では「先進の」に対して「先進」を，「高等の」に対して「高等」を作り，"advanced"の訳語候補とする．次にそれぞれが持つ訳語を使って，複合名詞の訳として得られる候補をすべて作る（図２）．フリーテキスト全文中にこれらの語が含まれているかどうかを検索する．検索の結果，検索に成功した語があればそれを複合名詞の訳語とする．失敗したならば，通常のデフォルトの訳語を使う．

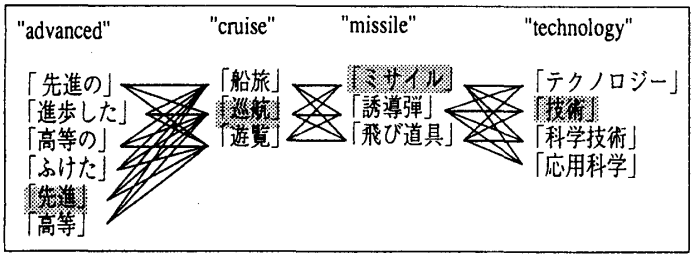

図２　複合名詞の訳語候補

4．実験

本方法に基づいて簡単な実験を行なった．入力はＡＰ電に出現した２語連続語（約16万語）の中で

・出現頻度が５回以上９回以下
・形容詞＋名詞
・形容詞の訳語の中に形容詞語尾「の」を含む
（ただし，"his", "all", "each"等は除いた．）

という条件を満たすものを用いた（サンプリング調査ではこの条件を満たす語は約2,400語であると推測される）．フリーテキストはＮＨＫ放送データベースの'92年８月分（約４Ｍバイト）を使った．また，フリーテキスト中の検索はFAST法[浦谷 89]によった．

その結果，28個の訳語が得られた．この中で正しく得られたのは約86％であった．ただし，例えば"Japanese people" は文脈によっては「日本の人々」の方がよい場合もあるが，今回は「日本人」で正解とした．結果の一部を図３に示す．

economic relations 経済関係
Japanese people 日本人
Israeli government イスラエル政府
French government フランス政府
special representative 特別代表
environmental issues 環境問題
overseas markets 海外市場
Japanese banks 日本銀行
British government イギリス政府
economic assistance 経済援助
American government アメリカ政府
southern provinces 南地方
Vietnamese government ベトナム政府

図３　実験によって得られた複合名詞

"Japanese banks"を「日本銀行」（「日本の銀行」が正しい）とするのは誤訳である．

5．おわりに

対象言語のフリーテキストを用いた複合名詞の訳語選択手法について述べ，実験でその有効性を示した．本手法は，連続する名詞をどこで切るかの決定や，複合名詞辞書を自動作成することにも応用できる．

今後は大規模な実験をし，この手法の有効性を確認したい．評価の際には，本来は複合名詞の訳として抽出されるべきものが抽出されなかったものの割合を求める必要もある．さらに，類語辞典を使って訳語候補を増やすことにより，機械翻訳用辞書にない場合でも訳語が生成できるようにすることも考えられる．また，テキストを圧縮すれば検索の高速化もはかれると思われる．

【参考文献】

[佐藤 91] 佐藤「MBT1: 実例に基づく訳語選択」人工知能学会誌，Vol.6, No.4, pp. 592-600(1991)

[野美山 91] 野美山「目的言語の知識を用いた訳語選択とその学習性」情報処理学会研究報告，NL-86-8(1991)

[情報辞典 89] 堀内他「最新英語情報辞典　第２版」小学館(1989)

[浦谷89] 浦谷「高速な複数文字列検索アルゴリズム：FAST」情報処理学会誌，Vol.30, No.9, pp. 1119-1125 (1989)