衛星放送・英日機械翻訳システムの辞書整備

6B−3

畑田のぶ子*　江原暉将*　山本ゆき**
(*NHK放送技術研究所　**NHK情報ネットワーク)

１．はじめに

ＮＨＫでは衛星放送に英日機械翻訳システムを試用している。ここでは翻訳者が専門辞書に不足な用語を登録し、翻訳作業をおこなつている。約１年５ケ月間に翻訳者が登録した３１２９語について分析した。現在この分析をもとに辞書整備をおこなっている。分析内容と辞書整備に際し、考慮すべき課題について述べる。

２．衛星放送翻訳システムと辞書整備の概要

衛星放送で機械翻訳の対象となっているニュースは、文化的なトピックスが多く、一日あたり約５０文、放送時間は５分である。

現システムでは翻訳者が名詞、動詞、形容詞、副詞を辞書登録でき、登録された語はすでに組み込まれている語より優先的に翻訳に使用される。

辞書整備の概要を図１に示す。今回は翻訳者による用語登録を中心に報告する。

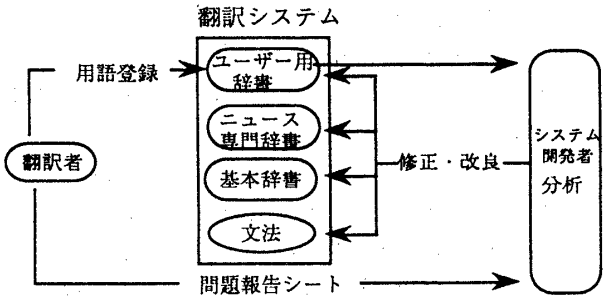

図１．辞書整備の概要

３．登録語の分析

１）登録語の品詞別状況

品詞別状況を図２に示す。

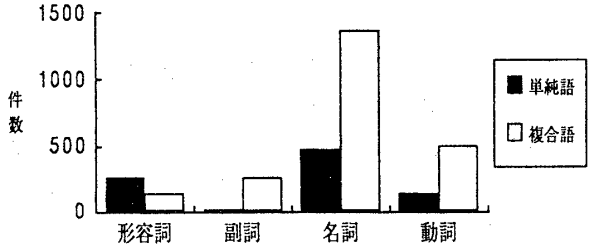

図２．登録語の品詞別状況

Dictionary Refinement of English-Japanese MT System for Satellite Broadcasting

N. Hatada (1), T. Ehara (1), Y.Yamamoto (2)
(1) NHK Science and Technical Research Laboratories.
(2) NHK Joho Network.

翻訳者の登録語数は全体で３１２９語、その内５８％が名詞、２０％が動詞、１３％が形容詞、９％が副詞であり、名詞が非常に多い。登録語の内７０％は複合語である。

２）登録語（見出し語）のカバー率

見出し語で現辞書でカバーされないていなかった語（＝未知語）の割合は７６％である。複合語にその割合が多く、９３％ある。

品詞別割合でみると、未知語であったものの割合は副詞に多く、８４％である。これは種々雑多な用語を副詞として登録しているためである。名詞、動詞は７７％、形容詞は６０％である。

３）登録語の品詞・訳語のカバー率

見出し語がすでに現辞書にあるにもかかわらず、新たに登録されたものは、現辞書の品詞の不足か訳語の不足が主な原因である。

品詞不足のものは形容詞に特に多く、約９％ある。動詞１.４％、副詞０.７％、名詞０.５％である。訳語不足のものは、形容詞１８％、動詞、副詞、名詞１２％である。

見出し語がすでに辞書に登録されていて、同一の品詞や訳語があるにもかかわらず、翻訳者が再登録しものについては下記の理由が考えられる。

・訳語の順位やデフォルトの訳語の設定が翻訳者の希望するものと異なっていた。
・構文解析により満足する訳語が選択されなかった。
・辞書データの不備により、意味マーカー等の設定が不足または不適当であった。

４）登録語の品詞別特徴

<u>形容詞</u>

見出し語が現辞書にあるものの割合が一番多く現辞書の品詞・訳語の不足が原因と思われる。複合語の登録の割合は非常に少なく、他の品詞と比べると割合が逆転している。

・ハイフンを含む語が９６件あり、名詞に次いで多い。(ex. adult-to-adult (a) :成人間の,family-run (a) :家族経営の)

・...ed, ...ing の分詞形 の見出し語が７２件あり、基本辞書にその形での見出し語のないものは５９件である。
・固有名詞関連の見出し語は５８件ある。
・前置詞ではじまるものが２４件ある。
（これは前置詞句を叙述形容詞として登録しているためである ex. in good health (a) :良い健康状態, in question (a):問題になっている）

副詞

現辞書にない語の割合が一番が多く、８４％を占め、登録語の９５％が複合語である。
・時を表すもの、場所を表すものが多く、それぞれ１３３件、１９件ある
これらの中には一部分が変化するパターンが多く、月、曜日、季節、人称代名詞、場所などが変化する。(ex. late September (adv) :九月末に, late Sunday (adv) :日曜日遅く, later this month (adv) :今月末に)
・前置詞で始まるものは１４７件である。

名詞

登録語全体の５８％を占め、それに伴って未知語の数も多い。
・固有名詞、およびそれに準ずるものも多く、３０４件ある。
・定冠詞で始まるものも多く１６２件ある。
・ハイフンを含む語は品詞の中では一番多く、１０７件ある。

動詞

名詞に次いで登録語数が多く、登録語の２０％を占める。
・複合語４９０件は基本辞書に対してはすべて未知語である。ニュース専門辞書でカバーされているものが２９件ある。
・前置詞または副詞小詞で終わるものが１６６件あり、現辞書に動詞＋前置詞または副詞小詞の共起として登録されているものが６８件ある。(ex. bring back (vt):持ち帰る, close down (vt):閉鎖する, point out (vt):指摘する)
・上記の他に目的語、その他を伴っていわゆる慣用句として登録しているものが２９２件あり、基本動詞と関連するものも多い。
(ex. get a haircut (vi):散髪する, get sick (vi):病気になる, get cold feet (vi):おじけづく)
・基本動詞を含むものが１２１件ある。

4．辞書整備の課題

1）辞書の拡充

翻訳者の登録語はニュース関連の未知語を多く含む。これらの用語は順次システム辞書や専門辞書に組み込むことによって、辞書の拡充が可能である。

また、翻訳者が登録した語で翻訳にうまく反映されない語、反映されるが他に悪影響を及ぼすものについては原因を調査し、辞書の登録方法や翻訳文法の改良で対応する必要がある。

2）用語について

調査した用語の中には、（a）..ed, ...ing の分詞形 をもつ見出し語、（b）動詞で前置詞または副詞小詞で終わるもの、（c）基本動詞と関連するもの、（d）人称代名詞、指示代名詞、月、曜日、などの一部分が変化するパターンをもつものがある。

（a）については動詞の基本形と屈折形の両方が見出し語になつている場合があり、その時は優先順位の高いユーザー用辞書に屈折形を登録するのは問題がある。このような語は基本辞書に移し、ウェイトの調整をする必要がある。

現在は、基本辞書中にある動詞で...ing 形を見出し語にもつ用語をサンプリングし、人手で動詞、形容詞、名詞間でのウェイト調整の実験をおこなつている。この結果を用いて自動的にウェイトを調整する手法が望まれる。

（b）のうち句動詞については動詞と副詞小詞の共起として登録し、目的語が副詞小詞の前にきても良いようにすべきである。

（c）については基本動詞の訳し分けに関連する部分と慣用表現に関する部分がある。基本動詞の訳し分けとして広くとらえるべきである。

（d）一部分が変化するものについては、パターンを利用し、(i) ローカルな文法処理を組み込む方法(ii) 辞書に登録すべき用語を自動生成する方法について検討中である。

5．おわりに

衛星放送・機械翻訳の辞書整備は今回の述べたものと、図１の下側で示した問題シートの分析を用いたものがある。後者については別途報告したい。今後も４.の課題を基に辞書整備、文法の改良を行なっていく予定である。

参考文献

相沢、他：放送ニュースへの機械翻訳システムの適用、信学会、ＮＣＬ９１−２０，１９９１
H.Tanaka：A Method of Ttanslating English Delexical Structurers into Japanes, COLING 92 , 1992