

機械翻訳用辞書への易しい単語登録

3G-1

- 名詞とサ変名詞の意味分類自動判定方式 -

嶺野和夫

富士通(株)

1. はじめに

機械翻訳用辞書に単語を登録する際に必要な情報として単語の意味分類がある。ここでは、名詞の意味分類を自動的に決定する実用的手法について紹介し、その有効性を検証する。

2. 辞書登録の現状分析

機械翻訳システムのチューニング作業の中で、最もコストを要する作業が言語辞書への未登録語の登録である。

現在の日英翻訳システムATLAS II (MSP版)は基本辞書と私用辞書とを持ち、辞書エディタを使用して詳細な辞書情報を各言語辞書毎に登録することができる。辞書登録の現状として、実際に当社の海外向け各種マニュアル・ドキュメント類の翻訳業務に使用しているシステムの私用辞書を分析した結果を下記に示す。

【日本語私用辞書の全登録単語24,839語】

名詞 (93.2%) 23,143語	サ変名詞 (5.3%) 1,336語
--------------------------	--------------------------

その他 360語 (1.5%)

この結果から名詞とサ変名詞の登録を簡易化すれば、辞書登録作業の98.5%をカバーできることが判る。

3. 名詞の簡易登録

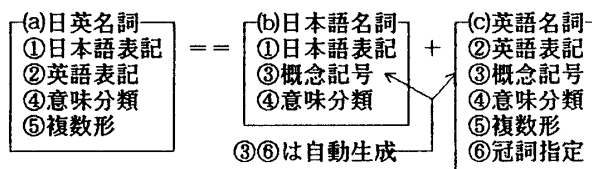
ATLAS IIの名詞登録には下記の(b)と(c)の項目を指定する必要がある。この登録を簡易化するためには(1)~(4)の解決が必要であり、一部はG版ATLASで実用化された。

(1)日英同時登録：概念記号を一意に自動生成することで、日英両辞書に同時登録する。

(2)複数形の生成：複数形変化規則を適用して自動生成。

(3)冠詞の生成：冠詞の標準規則を適用して自動生成。

(4)意味分類生成：4.で提案する方法で生成する。



4. 意味分類の自動判定方式

日本語には、文字そのものに特定の意味を有する性質が

あり、特に漢字についてはその性質が強いと考えられる。

漢字は意味の有る部首等からなりたっている。単独でも意味を有する漢字が、複数個結合された単語の場合には更に意味が限定され、単語同士が結合された単語も、結合される前の意味を継承しつつ更に意味が限定されると考える。この時、意味の継承される文字は単語の後ろに位置することが多い。(例：本体装置)このことに着目して後ろ部分の意味分類を新用語に継承させるのが今回の自動判定方式である。次に本方式のアルゴリズムを解説する。

(1)新単語の表記で辞書を検索する。⇒(2)辞書中に名詞またはサ変名詞として登録されていない場合は先頭文字を削除しながら(1)(2)を繰り返す。ただし、漢字以外の1文字が残った場合には処理を中止する。⇒(3)名詞またはサ変名詞として登録されていた場合は辞書中の意味分類を新単語の意味分類に継承させる。この時、辞書中に複数の同一表記単語が存在する場合は頻度値の最大の単語を使用する。

5. 本方式の評価

2.で述べた実業務で登録された私用辞書中の名詞とサ変名詞(24479語)について、本方式を適用して評価した。今回の評価は私用辞書中の個々の単語について、基本辞書中の単語によって意味分類の判定が可能であるかを検定している。評価結果は次の通りであった。

基本辞書の検索結果	◎自動判定成功	×自動判定失敗
完全に一致：1329語	1292語：5.3%	37語：0.2%
存在する：20963語	19685語：80.4%	1278語：5.2%
存在しない：2187語	0語：0%	2187語：8.9%
合計：24479語	20977語：85.7%	3502語：14.3%

6. おわりに

以上の様に、全登録単語の98.5%を占める名詞とサ変名詞について、本方式により85.7%の確率で意味分類を自動判定できることが判った。また、辞書整備が進むにつれて検索対象の単語が増えることから成功率の向上が期待できる。今後は他の方式との組み合わせによる判定精度の向上や、他の品詞への適用等を検討し、総合的にワープロ並みの、易しい機械翻訳用辞書登録を実現していきたい。

Easy word registering for dictionaries in machine translation.
(Automatic decision method for meaning classification of
general nouns and Sahen nouns.)

Kazuo MINENO
FUJITSU, Ltd.