翻訳システム用の辞書ツール

柏岡秀紀
ATR音声翻訳通信研究所

鷹尾 和享
東洋情報システム

1. はじめに

機械翻訳の研究において、どのような翻訳方式を採用するか、どのように処理を進めるかは重要な課題である。それと共に、実際に、システムを運用、あるいは、評価、活用する際に、辞書がどの程度、整備されているか、また、維持・管理の容易さや語彙の拡充性は、重要な課題である。

ATR 音声翻訳通信研究所で作成してきた翻訳システムTDMT(Transfer Driven Machine Translation)では、用例を利用した翻訳方式を採用し、辞書として、形態素辞書、対訳辞書、意味コード辞書の 3 つを利用している。これらの辞書は、それぞれ異なった情報を有しており、各知識に詳しい別々の作業者により、作成、管理されてきた。そのため、3 つの辞書を組み合わせた時に、一貫性に欠けることがあり、その場合、エントリーに過不足が生じる。また、すべての項目に対し、複数の作業者で一貫性をチェックし、修正を加えるには、コストがかかる。もし、一人の作業者だけで、すべてチェックするには、その作業者がすべての辞書に精通していなければならない。たとえ精通していても、何らかの計算機による支援が無ければ、困難な作業となる。

そこで、本稿では、これら3 つの辞書を同時に維持/管理するためのツールを紹介する。TDMT では、多言語の翻訳システムを構築しているが、本稿で紹介するツールは、英日翻訳システム用に作られている。本ツールでは、作業者がすべての辞書に精通していなくても十分に処理できるように、品詞の推定、意味コードの推定をシステムが行ない、既存の辞書の検索もできるようにしている。

2. 辞書の概要

TDMT は、用例翻訳手法を取り入れた翻訳システムである。入力文を形態素解析した後に、変換知識を利用して、原言語の構造解析と目的言語の構造生成をおこない、目的言語の構造から翻訳文を生成している。各処理を進めるにあたり、以下の 3 種類の辞書を利用している。

1. 形態素辞書
   形態素解析用の辞書
2. 変換辞書(対訳辞書)
   デフォルトの対訳を記す辞書
3. 意味コード辞書
   変換処理で用例との意味距離を計算するために利用される辞書、意味コードとして、角川新類語辞典[1]のコードを利用している。

以下それぞれの辞書について説明する。

2.1. 形態素辞書

入力文字列を、形態素解析するための辞書で、5 つの要素からなる。

（出現形 正規形 品詞 活用 数）

ここで扱う形態素解析は、TDMTという翻訳システムを前提とした解析のため、通常の形態

素として扱われる単位より長いものをエントリーとして認めている。特に、決まりきった文句などは、長い単位で登録されているものも多い。

実際の辞書の記述例を以下に示す。

("limited express" "limited express" CN NIL NIL)
("makes sense" "make sense" V NIL 3S)

2.2. 対訳辞書

入力単語のデフォルトの訳を記述している辞書で、原言語、目的言語とも品詞を伴ったエントリーとして扱っている。そのため、以下の 4 つの要素で一つのエントリーを表現する。

(英語品詞 正規形の英語形態素
日本語品詞 正規系の日本語形態素)

形態素辞書と同様、翻訳に適した単位のエントリーを単語としている。現在、TDMT の制約として、一つのエントリーに対して、一つの訳語しか対訳辞書には記述できない。

2.3. 意味コード辞書

原言語である英語の構造解析と同時に目的言語である日本語の構造生成を行なうための、変換知識の選定基準として用例と入力の意味距離を計算するために、入力形態素の意味コードが必要となる。そのために、形態素辞書、対訳辞書と同様、英語形態素とその品詞に対して、意味コードリストを定義するために、3 つの要素で記述する。

(英語形態素 品詞 意味コードリスト)

意味コードは、リストとして複数のコードが記述できる。これは、同じ形態素、品詞であっても異なる意味を有するものが多く、状況がわからなければ、意味を一意に決めることができない。現状のTDMTでは、複数の意味で距離を計算し、その平均値で変換知識を選択している。

3. 辞書メンテナンス作業

辞書のメンテナンス作業は、既存の辞書エントリーの修正作業、新規追加作業に大きく分けて考えることができる。

3.1. 修正作業

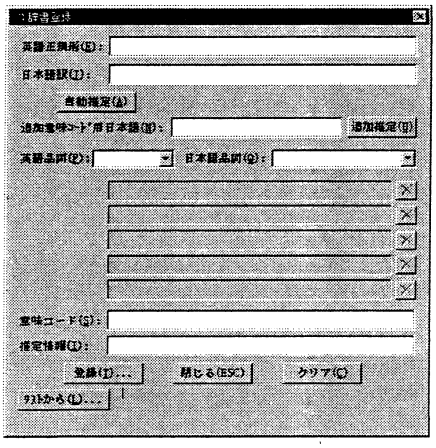

図 1 メンテナンス初期画面

既存の辞書エントリーの修正作業は、以下の手順により行なわれる。

1. エントリーとなる英語形態素を入力

図に示される初期画面で、英語正規系のフィールドに調べたい英語形態素を入力し、自動推定のボタンを押す。既存の辞書から該当する部分が検索され情報が提示される。

表示されている情報の必要に応じた修正

各フィールドに表示されている情報は、すべて書き換えることが可能である。また、意味コードについては、推定処理により得られるコードや類語辞典から得られるコードを簡単な操作で取り入れることができるようにしている。

2. 登録作業

登録時に、既存の辞書と再び照合を行ない、登録事項の一覧を提示する。表示は、既存の情報と変化があるか、修正された情報か、新しい情報かが区別できるように色を変えて表示される。その時点で、再び、修正することが可能である。

3.2. 新規追加作業

新規追加作業では、登録したい英語形態素の正規形と日本語形態素を入力し、推定処理を行ない、必要な情報を可能な限り自動的に埋め込んだ形で、ユーザに提示する。作業手順としては、修正作業とほぼ同様である。異なる点は、最初の入力に日本語形態素が必要となる。

英語形態素としては、一つのエントリーとして登録するので、当然、複合語や翻訳の都合で取り込んだ長い単語列に対しても、一形態素として扱い、一品詞を割り当てる。しかし、対訳として入力される日本語は、一形態素とする必要はない。そこで、複数の形態素となる場合は、形態素の間に空白を入れて入力することで対応した。

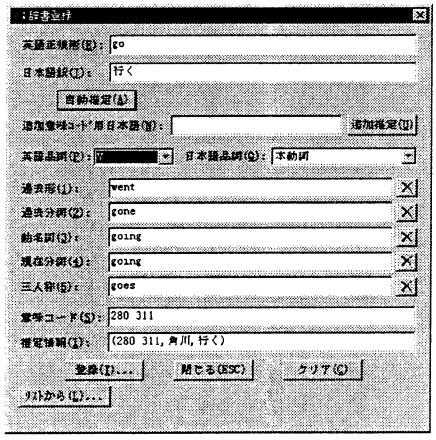

図 2 “go” の入力画面

また、形態素辞書では、英語形態素の品詞により、語形の変化が異なる。語形が変化した語を登録するかどうかは、辞書の構成、規模などから、議論すべきことであるが、現状、TDMT の辞書では、表現形が異なるものはすべて登録するという方針をとっている。たとえば、“go”という活用する形態素を登録する場合、図2に示すように過去形“went”や 過去分詞“gone”を形態素辞書に登録できるように、過去形や分詞の表現形を登録するフィールドが現れる。

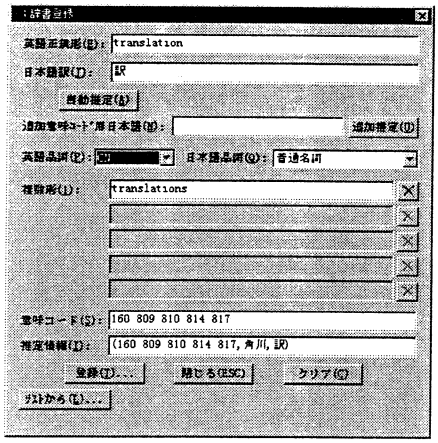

図 3 “translate” の入力画面

また、登録する語が、名詞の場合にも、図 3 に示されるように、複数形を登録するフィールドが現れ、適宜、修正できるようになっている。

この機能は、既存のエントリーを修正する場合にも有効で、登録で活用の漏れがあった場合、容易に修正できるようになる。

4. ツールの機能

ここでは、エントリーの登録、修正の際に処理される品詞の推定処理や、意味コードの推定について説明する。

4.1. 品詞推定

入力された形態素から品詞を推定するために、英語、日本語で、それぞれアドホックなルールを利用している。利用しているアドホックなルールは、これまでの作業者に、インタビューする形で作成した。

活用形の仕方や単複の語形変化についても、同様に、アドホックなルールを作成し、自動的にとりあえず、フィールドが埋まるように推定している。

4.2. 意味コード推定

意味を推定するために、英語の形態素列と日本語の形態素列を利用する。アドホックルールにより、形態素列の中から主要と思われる単語を選び、英語、日本語、両者の既存の意味コード辞書にエントリーとして存在するかどうかを調べる。存在していれば、そこに与えられている意味コードを利用する。既存の意味コード辞書に現れていない場合、主要語の語形が、準備しているテンプレートに一致すれば、それに従い、意味コードを付与する。準備しているテンプレートの例を以下に示す。

英語:"… network"

日本語: "…剤"

また、本ツールには、まだ組み込んでいないが、異なる体系の辞書を利用して意味コードを推定することが可能である[2,3]。

5. おわりに

翻訳システム TDMT で利用している形態素辞書、対訳辞書、意味コード辞書の三種類の辞書について概要を示し、これらの辞書をメンテナンスするためのツールを紹介した。従来は、メンテナンスのためには、個々の辞書に精通している作業者が必要であり、別々にメンテナンスされていたため、エントリーの一貫性を保つのが困難であった。本ツールで実現した推定機能や検索機能により、すべての辞書を一人の作業者がメンテナンスする時の負荷を軽減した。

また、現在、開発されている多くの機械翻訳システムの辞書は、多くが独自のシステムに依存した形式で記述されている。幸いなことに、共通に利用するためのフォーマット[4]が、提案されている。将来的には、この UPF のフォーマットにも対応できるようにしていきたいと思う。

参文考献

[1]大野晋,浜西正人:"角川類語新辞典" CD-ROM 版,角川出版(1989).

[2]鷹尾,柏岡,白井:"異なる辞書を利用した意味コードの自動付与",情報処理学会第 59 回(平成 11 年後期)全国大会講演論文集,情報処理学会(1999).

[3]NTT コミュニケーション科学研究所監修:"日本語語彙大系"全 5 巻,岩波書店(1997).

[4]亀井他:"UPF:機械翻訳ユーザ辞書の共通フォーマット",言語処理学会第 4 回年次大会発表論文集,言語処理学会(1998)