日英機械翻訳における利用者登録語の意味属性の自動推定

6P－7　池原 悟* 白井 諭* 横尾昭男* Francis Bond* 小見佳恵**

*ＮＴＴ情報通信網研究所　**ＮＴＴアドバンステクノロジ

１．はじめに

機械翻訳システムの利用時には、通常、翻訳対象に合った利用者辞書が必要となるが、その作成作業の軽減が望まれる。特に、高品質翻訳を狙った日英機械翻訳システムＡＬＴ－Ｊ／Ｅでは、各単語に対して、約3,000種の分解精度を持つ単語意味属性の付与が必要であり(1)、一般の利用者が、このような精密な情報を付与するのは困難であった。

そこで本報告では、利用者が登録したい日本語単語（複合語を含む）に対して英語訳語を与えるだけで、システムがシステム辞書の知識を利用して、その単語の意味属性を自動的に推定して付与する方法を提案する。また、自動付与された意味属性を専門家の付与した意味属性と比較評価し、訳文品質に与える効果を示す。

２．単語意味属性付与の方法

ここでは、登録単語は名詞とし、日本語表記と英語訳語が与えられたとき、システム辞書（システム側で既に用意された辞書）の情報を使って、意味属性を推定する方法を示す。

２．１　主名詞の判定方法

登録された語が単語一語で構成されるときは、その語を主名詞とする。登録された語が複数の単語で構成されるとき(複合語)は、主名詞を以下の方法で決める。

（１）日本語表記から決める場合

見出し語の後方から文字単位に最長一致法で単語辞書を検索し、漢字２文字以上（カタカナ、ひらがな、英字の場合は字種の変わり目まで）の名詞単語が得られたとき、これを主名詞とする。

（２）英語訳語から決める場合

訳語全体が英語辞書にある場合は、それを主名詞とする。ない場合は、訳語中に in, on, with などの前置詞又は that, which などの関係詞（ストップワード）の有無（後方修飾の有無）を調べ、ある場合は、それを含む後方修飾部分を削除する。次に、残った英語全体に対して後方から単語単位の最長一致法で英語辞書を検索し、辞書内に一致する語（一般には語の組）があればそれを主名詞とする。

２．２　名詞種別の判定方法

ＡＬＴ－Ｊ／Ｅでは、名詞の意味属性には、一般名詞意味属性(約2,800種)と固有名詞意味属性(約200種)があり、一般名詞には一般名詞意味属性、固有名詞には一般名詞意味属性と固有名詞意味属性の両方を付与することが必要である。

そこで、登録された単語の英語側の主名詞に着目し、その先頭文字から１文字以上が大文字の場合は固有名詞とし、それ以外は一般名詞とする。但し、全ての文字が大文字の場合は、一般名詞とする。

２．３　意味属性の推定方法

（１）日本語表記から推定する方法

利用者登録語の日本語見出し語（複合語では、その主名詞を含む部分）がシステムの日英対照辞書の見出し語として既に存在する場合は、システム辞書の意味属性をそのまま利用者登録語の意味属性とする。表１で、利用者登録語の「治療」、「放射線治療」は、システム辞書（表２）に「治療」があるので、意味属性は《治療》となる。

表１　利用者辞書の例

日本語見出し（利用者登録）	英語訳（利用者付与）	単語意味属性（自動推定）
治療	cure	《治療》
放射線治療	radiotherapy	《治療》
手当て	treatment	《治療》
医療	medical treatment	《治療》
数値制御ロボット	numerical controlled robot	《産業機器》
照明付き机	desk with lighty unit	《家具》

表２　システム辞書（日英対照辞書）の例

日本語見出し語	英語訳	意味属性
治療	treatment	《治療》
制御ロボット	controlled robot	《産業機器》
机	desk	《家具》

（２）英語表記から推定する方法

利用者登録語の英語訳語（複合語では、主名詞を含む部分）がシステムの日英対照辞書の訳語に既に存在する場合は、それに対応する日本語のシステム辞書の意味属性を、そのまま利用者登録語の意味属性とする。表１で、利用者登録語の「手当て」、「医療」は、その訳語（または主名詞訳語）「treatment」がシステム辞書（表２）にあるので、意味属性は《治療》となる。

２．４　意味属性の付与

（１）一般名詞と固有名詞の扱い

上記の意味属性の推定では、登録語が一般名詞の場合は、システム辞書の持つ一般名詞意味属性を抽出し、固有名詞の場合は、システム辞書の持つ固有名詞意味属性と一般名詞意味属性の双方を抽出する。

（２）複数意味属性の扱い

前述の方法では、システム辞書には一般に複数の意味属性が付与されていること、日本語表記だけでなく英語表記からも意味属性が抽出されることのため、一般に一語に対して複数の意味属性が抽出されることになる。意

Automatic Acquisition of Semantic Attributes for User Dictionaries, Satoru Ikehara, Satoshi Shirai, Akio Yokoo, Francis Bond and Yoshie Omi, NTT Network Information Systems Laboratories, 1-2356 Take, Yokosuka-shi, 238-03 Japan

味属性としては、これらの得られた意味属性すべてを（ただし重複する属性は重複を除いて）登録する。

3．意味属性推定精度の評価

3．1　実験の条件

新聞記事100文（平均42文字／文）を対象に、利用者辞書登録語の意味属性自動推定品質と、その訳文品質に対する効果を調べるため、翻訳実験を行った。原文100文中、システム辞書にない語（未知語）を含む文は53文、未知語の総数は77語（一般名詞26語、固有名詞51語）であった。以下では、専門家(ｱﾅﾘｽﾄ)が付与した場合を正解とし、自動付与方式の精度と効果を評価する。

3．2　名詞種別自動判定の精度

利用者辞書登録語77語中、アナリストが49語を固有名詞と判定したのに対して、本方式では、アナリストが一般名詞としたもの１語(「中部圏」)を含む46語を固有名詞と判定した。固有名詞を一般名詞と判定したのは、「PC9800」、「VOS3/ES1」などの４語であった。77語中、判定を誤ったものは５語であるから、本方式の名詞種別判定の精度は93.5%となる。

表３　単語別にみた自動付与とアナリスト付与との比較

意味属性の付与	ｱﾅﾘｽﾄが付与した属性との比較	一般名詞意味属性	固有名詞意味属性
自動付与された	全属性が一致	３８語(49.4%)	４２語(54.5%)
	余分に付与	２１語(27.3%)	０語(0.0%)
	一部付与不足	４語(5.2%)	０語(0.0%)
	全てが不一致	１０語(6.5%)	３語(3.9%)
	ｱﾅﾘｽﾄ付与なし	０語(0.0%)	１語(1.3%)
自動付与されなかった	ｱﾅﾘｽﾄ付与なし	１語(1.3%)	２７語(35.1%)
	ｱﾅﾘｽﾄは付与	３語(3.9%)	４語(5.2%)
合計		７７語（100%）	７７語（100%）

[注]名詞種別を正しく判定しても、システム辞書との関係で、意味属性が付与できなかった語もある。

3．3　意味属性自動付与の精度

単語別にみたときの自動付与とアナリスト付与の結果を表３、付与された意味属性全体の数とその内訳を表４に示す。アナリストの付与した意味属性が正解であると考えたときの適合率と再現率は、表４から表５の通り求められる。これらより以下のことが分かる。

①単語別にみて、自動付与方式とアナリストが完全に一致した単語は、一般名詞意味属性では（49.4+1.3＝）50.7%、固有名詞意味属性では、固有名詞49語中の42語で85.7%であり、特に、固有名詞意味属性の自動推定精度が高い。

②利用者登録語が原文中に含まれる割合は比較的少ないから、余分な意味属性が付与されていても、必要な意味属性が含まれていれば、翻訳品質の低下は少ないと推定される。これより、一般名詞意味属性の正解率も実効的には(49.4+27.3=)77%と考えられる。

③一般名詞意味属性では、抽出のもれた意味属性よりも、余分に抽出された意味属性が、かなり多くなっているため、適合率に比べて再現率はかなり高い。これは機械翻訳にとって有利と考えられる。

表４　属性数から見た自動付与とｱﾅﾘｽﾄ付与の比較

属性の種類＼属性付与の方法	自動付与した属性数	ｱﾅﾘｽﾄが付与した属性数	自動付与とｱﾅﾘｽﾄ付与が一致した属性数
一般名詞意味属性	１９４件	１２７件	７４件(+21)*
固有名詞意味属性	４６件	４８件	４２件
合計	２４０件	１７５件	１１６件(+21)*

＊（）内の数は、不一致ではあるが、属性の上下関係を考慮した場合、すなわち自動付与された属性がｱﾅﾘｽﾄ付与の属性の上位または下位にあるものの数を示す。

表５　自動付与した意味属性の適合率と再現率

意味属性種別	適合率	再現率
一般名詞意味属性	38.1%(49.0%)	58.3%(74.8%)
固有名詞意味属性	91.3%	87.5%
全体	48.3%(57.1%)	66.3%(78.3%)

（）内の数字は、属性の上下関係を考慮した場合を示す。

4．訳文品質の向上効果

新聞記事100文中、利用者辞書への登録語を含む53文に対して、以下の３つの場合の翻訳結果を求めた。

①アナリストが意味属性を付与した利用者辞書を使用
②意味属性を付与しない利用者辞書を使用
③意味属性を自動付与した利用者辞書を使用

実験結果によれば、①の結果に対し、②では53文中13文の訳文品質が低下した。本試験文では、原文(約2,100単語)中の利用者辞書登録語（未知語）は77語(3.7%)と、比較的少なく、大半が固有名詞であるにも拘らず、その意味属性の影響は比較的大きいことが分かる。

次に、③では、訳文品質の低下は３文にとどまった。これより、アナリストの付与した利用者辞書の意味属性は13文に対して訳文品質を向上させたのに対して、自動付与では10文の訳文品質を向上させたことがわかる。

自動付与方式で訳文品質を向上できなかった３文を見ると、その原因は、名詞種別の判定誤りが１件、正解の意味属性の上位または下位の属性を選択したものが、それぞれ１件であった。本方式では、名詞の種別も自動判定しているが、誤りの例から見て、名詞種別と意味属性の単純な分類(上位2～3段程度)を利用者に依頼することができれば、訳文品質を低下させるような意味属性付与の誤りは、ほぼ防ぐことができると推定される。

5．おわりに

利用者辞書登録語の単語意味属性を自動付与する方式を提案した。また、新聞記事翻訳への適用実験で、本方式によれば、専門家が付与した場合に比べて、大差ない訳文品質が得られることを確認した。今後は、技術マニアルや論文など、未知語が多く、より大きな効果の予想される文書を対象に、実験を進める予定である。

［参考文献］

(1)池原，宮崎，横尾：「日英機械翻訳のための意味解析用の言語知識とその分解能」，情処論，Vol.34，No.8（1993）