2C−7

共起関係情報の帰納的学習
― 英日機械翻訳への応用 ―

藤田澄男
(株) 日本コンピュータ研究所

1 はじめに

英語・日本語間のように構文的にも語彙的にも隔たった言語間の機械翻訳では、共起関係による制約情報を利用して、解析・変換・生成を行う必要がある。そのために、予め辞書中に大量かつ的確な共起情報を記述しておかなければならない。

しかしながら、共起情報は、対象とする文章の分野などに依存するので、予め用意していた情報が、かえって翻訳に悪影響する場合もある。そのために、共起情報を、テキストから学習する方式がいくつか提案されている[1]。しかし、これまでの方式では、基本的にテキストに出現した単語と単語の共起を学習するものだった。

本発表では、英日機械翻訳システムで、動詞の訳語を決定するための共起情報を、対訳データから帰納的に学習する方式を提案する。

2 システム構成

図１に本方式の処理の流れを示す。本方式では文翻訳後に、ユーザが後編集として訳語選択を行い、その結果の訳文を確定訳とする。このため、システムの出力した訳文を負の事例、確定訳文を正の事例とし学習することができる。こうして得られた対訳データと翻訳時の解析情報から共起情報を抽出することができる。これに対して、帰納的一般化を行い、一般化された学習情報を学習辞書に登録する。

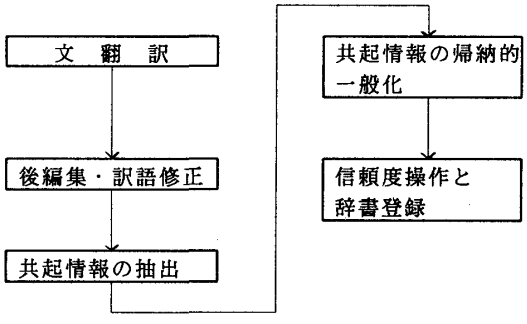

図１　処理の流れ

3 共起情報の表現と抽出

共起情報は、素性構造として、英文側及び和文側双方の共起を、対応をとって表示する。図１と図２に共起情報の例を示す。ｅｃｏｃとｊｃｏｃがそれぞれ動詞の格要素に対する英文共起と和文共起であり、動詞パターンによって決まる要求される格(subj,objなど)についての格制約のリストである。

```
[gv   = [word=offer
         vptn=VP6A]
 jgv  = [word=申出る
         infl=下一]
 ecoc= [subj= [word=he
               semc=動作主]
        obj = [word=resignation
               semc=動詞的]]
 jcoc= [subj= [word=彼
               part=は]
        obj = [word=辞職
               part=を]]]
```

図２　"He offers his resignation."に対する共起情報

```
[gv   = [word=offer
         vptn=VP6A]
 jgv  = [word=申出る
         infl=下一]
 ecoc= [subj= [word=he
               semc=動作主]
        obj = [word=help
               semc=動詞的]]
 jcoc= [subj= [word=彼
               part=は]
        obj = [word=援助
               part=を]]]
```

図３　"He offers his help."に対する共起情報

なお動詞パターンは、Hornbyによる分類を使った。

Inductive Learning of Information on Co-occurrence Relation
- Application to English-Japanese Machine Translation -
Sumio FUJITA
Computer Institute of Japan, Ltd.

このような英和文内の単語間の共起は、図４のような翻訳時の中間解析木と対訳文、辞書データを参照することによって得られる。一方、ｓｅｍｃには、格要素に現れた単語の訳語に対応する意味素性を設定する。

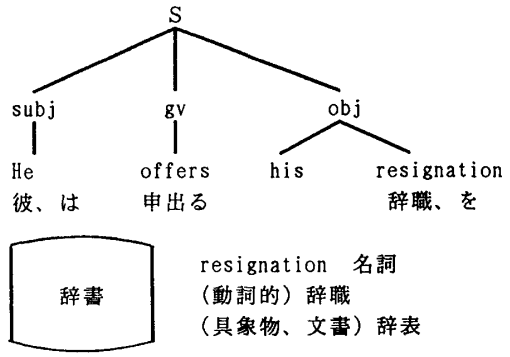

図４　共起情報抽出に必要な翻訳時情報

後編集による訳語の修正は、これらの格要素に現れた単語の訳を、辞書中の他の訳語の中から選択するか、または、意味素性と共に新たに入力することにする。これによって、英和の共起の対応を認識することができる。

辞書中に予め登録されている共起情報も同様な形式とする。英文共起の格制約には、選択制約として共起しうる名詞(名詞句内の主辞)の意味素性を記述しておく。但し特定の単語間の共起については、その単語を記述しておく。これを利用して動詞の多義を解消して訳語を決定することができる。

翻訳時には、まず原言語内の主動詞とその動詞パターンが決定される。主動詞ｇｖの辞書項目中にはその動詞パターンについて、英文共起ｅｃｏｃ$_1$・・・ｅｃｏｃ$_n$と和文共起ｊｃｏｃ$_1$・・・ｊｃｏｃ$_n$がある。そのうちの制約が満たされるｅｃｏｃ$_i$に対応する和文共起ｊｃｏｃ$_i$の情報が適用される。

４　帰納的一般化

図２・図３のような単語間の共起情報に対して一般化を行い、主動詞とその選択制約との共起情報を得ることができる。　抽出された共起情報について、次のような一般化の処理を行う。

1)　辞書中から同一動詞、同一動詞パターンかつ同一動詞訳語を持つ共起情報のうちから一つを選ぶ。
2)　ｅｃｏｃ・ｊｃｏｃ内の各格制約についてマッチングしない部分を変数にする。
3)　意味素性が、マッチングしていなければ、別の共起情報について、1)以降を繰り返す。

図５に、図２と図３の共起情報を一般化した結果を示す。”offer”の目的語に現れた二つの単語の意味素性がマッチングしたので、その単語と訳語を変数にして、一般化している。意味素性どうしのマッチングでは、意味素性を階層的に記述して、二つの異なる素性を、その共通の上位範疇に一般化する。一般化できる共起情報が、辞書中に存在しない場合は、抽出した単語間の共起情報をそのまま学習する。

```
[gv  = [word=offer
        vptn=VP6A]
 jgv = [word=申出る
        infl=下一]
 ecoc= [subj= [word=he
               semc=動作主]
        obj = [word= __
               semc=動詞的]]
 jcoc= [subj= [word=彼
               part=は]
        obj = [word=__
               part=を]]]
```

図５　一般化された共起情報

５　信頼度の操作

負事例の共起情報が、すでに辞書中に登録されていた場合には、それの優先度を下げる。一方、正事例の共起情報が、登録済みの場合は、それの優先度を上げる。この操作によって、間違った情報を学習しても、それは長期的には、修正される。

学習されて、一般化された共起情報が、負事例になった場合は、その正事例の共起情報をより優先度を上げて登録する。

e.g.

1)	take flu	感冒(＋病気) にかかる
2)	take cancer	癌(＋病気) にかかる
3)	take cold	風邪(＋病気) をひく

1)と2)で、” take N(+病気):Nにかかる” という一般化された共起情報を得る。ところが、3)ではそれが負事例になって、"take"の訳語が、"ひく"に修正される。このような場合、最後の事例を特殊事例とみなして、"take cold:風邪をひく"という情報を、そのまま登録する。すでに一般化された学習情報を特殊化することはしない。

６　おわりに

英日機械翻訳システムで、共起情報を帰納的に学習する方式を提案した。本システムは、現在プロトタイプを作成して評価をおこなっている。

本方式で、動詞の訳が、その目的語や補語によって変わる場合などに効果的な学習ができる。一方、英文と和文の間に単語の対応がない固定表現のようなものは、学習の対象外である。

今後の課題として1)一般化の方式の最適化、2)名詞句内の共起の学習への応用、3)確定訳の教示方法の改善、等が挙げられる。

参考文献

[1] 中島ほか：テキストからの共起関係自動抽出の試み、情報処理学会第３８回全国大会、pp.325-326 (1989)

[2] A S Hornby:Oxford Advanced Learner's Dictionary of Current English,KAITAKUSHA(1974)