対訳データの階層的なグループ化に基づく英日翻訳

5 J －2

中尾 嘉孝[*1] 平野 志奈[*1] 野原 ゆかり[*1] 西垣 万亀子[*1] 白井 諭[*2]

[*1]ＮＴＴアドバンステクノロジ（株） [*2]ＮＴＴコミュニケーション科学研究所

1 はじめに

現在、翻訳ソフトは内容把握や翻訳家の下訳作成などに利用されている。しかし、これらに英文記事のヘッドラインを翻訳させると意味の分かる訳を得ることができない。理由としては、ヘッドラインは、be 動詞の省略や to 不定詞による未来表現を多用するため[2][3]、ルール型翻訳方式に基づくシステムでは対処できないということが考えられる。

そこで本稿では、ヘッドラインの文は比較的短く、また使用される形式が限られている点に着目し、実例型翻訳方式の適用を検討した。対訳データを実例型翻訳方式のベースとして利用するシステムは数多く発表されているが（[1]等）、予め人手によりルール化しておく必要がある。これに対し本稿は、トランスレーションメモリ・テンプレート型翻訳・句レベル変換規則（複合語、副詞句など）を組み合わせた翻訳方式を提案する。ヘッドラインでは紙面のスペース上の制約から short word が使用されるが、これらの語は多義語であるためルール型翻訳方式では対応が容易ではない。これに対し本方式では文単位のテンプレートを利用することにより、訳し分けの制御ができると考えられる。

2 システムの構成

ルール型翻訳方式は汎用的であるが、翻訳結果がぎこちない場合が多い。用例翻訳では適合すれば良い訳が得られるが、汎用的ではない上に多数の用例との照合が必要である。従って本システムでは名詞句や慣用表現を変数化して、適用範囲を広げ、典型的な用例を予め整理してテンプレート化を行う。以下に本システムの方式提案と概略図を示す。

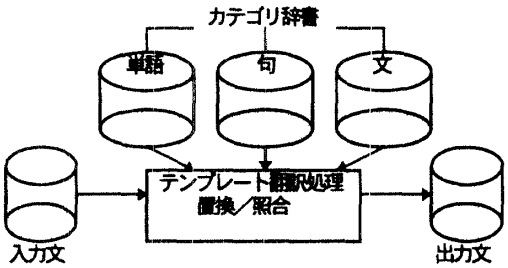

図１：テンプレート翻訳処理

本システムは、テンプレート作成処理、テンプレート翻訳処理、テンプレート学習処理から構成される。テンプレート翻訳処理は、カテゴリ辞書を用いて原言語の入力文を目的言語に翻訳するメイン処理である。

テンプレート作成処理では始めに対訳コーパスを入力として、コーパス中の対訳文に対し単語単位の変数化を行ない、文テンプレートを作成する。次にテンプレート縮約処理を行い、変数化された文テンプレートの集合から固定的な対訳変数の組合せを認定し、句単位の変数を自動生成し、文テンプレートに対し句単位の変数化を行うことで、文テンプレートを精錬する。

対訳コーパス中で同一文に対して複数の対訳文が存在する場合には、テンプレート作成処理によって同一文に対する複数の文テンプレートが生成される。この場合、学習処理を行うことで、入力コーパスをもとに単語、句の各レベルのカテゴリまたは文テンプレートの頻度情報が計算され、より出現頻度の高い表現を優先して訳出することが可能となる。

3 カテゴリ辞書の体系

カテゴリ辞書は単語・句・文の３つの階層で構成され、二言語間の対となる単語・句・文がそれぞれのレベルでカテゴリ識別子によってグループ化されて格納される。ここで言う文のカテゴリ辞書とは前章で述べた文テンプレートのことである。文テンプレートはそれ自身再帰的に一つのカテゴリとしてとらえることができるため、本稿では文テンプレートを単語・句と同列のカテゴリ辞書として定義している。図２にカテゴリ辞書の作成手順を示す。

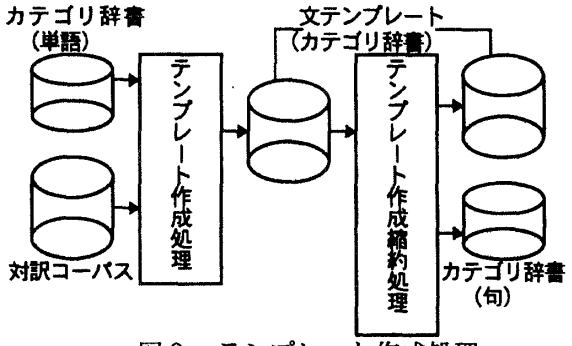

図２：テンプレート作成処理

An English-to-Japanese Translation System Using Categorized Bilingual Data

Yoshitaka Nakao[*1], Shina Hirano[*1], Yukari Nohara[*1], Makiko Nishigaki[*1] and Satoshi Shirai[*2]

[*1]NTT Advanced Technology Corporation and

[*2]NTT Communication Science Laboratories

3－1　カテゴリ辞書（単語）

以下に<direction>、<nation>等としてグループ化されたカテゴリの内容を示す。原言語と目的言語の単語はｎ対ｍの対応づけが可能である。

<direction>(eastern：東部,東の)(western：西部,西の)
(northwestern：北西部,北西の)...
<nation>(Africa：アフリカ)(Brazil：ブラジル)
(Colombia：コロンビア)(India：インド)...
<vehicle>(car：自動車,車)(bus：バス)(train：電車,列車)...
<blast>(bomb：爆弾)(blast：爆発)(explosive：爆発物)...

図３：カテゴリ辞書（単語）例

3－2　カテゴリ辞書（句）

句のカテゴリは１つまたは複数の連続した単語のカテゴリの集合として表現される。以下に<region>等としてグループ化されたカテゴリの内容を示す。

<region>(<direction><nation>：<nation><direction>)
(<nation>：<nation>)
<accident>(<vehicle><blast>：<vehicle><blast>)
(<blast>：<blast>)

図４：カテゴリ辞書（句）例

3－3　カテゴリ辞書（文）

図５に文のカテゴリの例を示す。<kill-1>が縮約前の文のカテゴリ、<kill-2>が縮約後の文のカテゴリである。例えば、"northwestern Colombia"が文レベルの辞書では<direction>、<nation>の二つのカテゴリ識別子に分割されているのに対し、縮約処理の結果、一つの<region>で抽象化することができ、①②は同型の文テンプレートの利用が可能になる。

① Bomb kills four in northwestern Colombia｜爆弾によりコロンビア北西部で４人死亡
<kill-1>(<blast> kills <num> in <direction> <nation>：<blast>により<nation><direction>で<num>人死亡)
<kill-2>(<accident> kills <num> in <region>：<region>で<accident>により<num>人死亡)
② Car bomb kills three in India｜インドで自動車爆弾により３人死亡
<kill-1>(<vehicle><blast> kills <num> in <nation>：<nation>で<vehicle><blast>により<num>人死亡)
<kill-2>(<accident> kills <num> in <region>：<region>で<accident>により<num>人死亡)

図５：カテゴリ辞書（文）例

4 システムの動作について

本システムは次のようにして訳文を生成する。入力文に対しカテゴリ辞書との照合を繰り返し行い、単語または単語列をカテゴリと置換していく。この時、まず単語のカテゴリ化を行い、次に句のカテゴリ化を可能な限り行い、文のカテゴリを得る。また頻度情報を参照することにより、文、句、単語の順に各カテゴリの対訳を選択し、埋め込んでいくことにより翻訳結果を得る。例えば、"Car bomb kills three in India"が入力されると図６のようになる。

入力文："Car bomb kills three in India"
単語のカテゴリ化：<vehicle> <blast> kills <num> in <nation>
句のカテゴリ化：<accident> kills <num> in <region>
文のカテゴリ化：<kill-2>
文の対訳選択：<region> で <accident> により <num> 人死亡
対訳の埋め込み１：インドで<accident> により３人死亡
対訳の埋め込み２：インドで自動車爆弾により３人死亡
(各カテゴリ辞書の内容は３章を参照)

図６：翻訳例とカテゴリ辞書の内容

5　単言語コーパスへの適用

この方式は、頻度情報により訳出の仕方を制御することができるため、単言語コーパスに基づいてシステムを改良できる可能性を持っている。例えば、図６において「スペインでバス事故により８人が死亡」やこれに類似する日本語表現が多数追加された場合を考える。この場合、対訳入力ではないため、カテゴリ辞書の項目を英日のペアで追加することはできない。しかし、カテゴリ辞書のほかに、「格助詞訳出不要」といった情報が登録されている機能語辞書があれば、次のように訳出スタイルを改良することができる。入力日本文に対し、カテゴリ辞書を用いてカテゴリ置換を行い、さらに機能語辞書を参照すれば<kill-2>の日本語と一致することがわかるので、この訳し方を別訳として追加する。すなわち、単一言語コーパスを入手すれば、訳出スタイルを改良できることになる。また、カテゴリ辞書は基本的に英日、日英のどちらの方向にも適用できるため、それぞれの単言語コーパスで訳出スタイルの改良を行うことにより、受け入れることができる表現のバリエーションを増やす効果も考えられる。

6　おわりに

本稿では、新聞記事のヘッドライン翻訳を対象に、トランスレーションメモリ・テンプレート翻訳・句レベル変換規則を組み合わせた英日翻訳システムとそれに用いるカテゴリ辞書の構成を提案した。カテゴリ辞書により類似表現の同一性が推定できるため、単言語コーパスにより訳出スタイルの改良が可能になる。人間の翻訳家は主として非対訳のコーパスから翻訳知識を得ると考えられるので、この方式はそれを目指すものであるといえる。

また本稿では英日翻訳を考えたが、カテゴリ辞書は逆方向にも適用可能であると考えられる。今後は日英翻訳への適用についても検討する予定である。

参考文献

[1] Kaji, H., Kida, Y., and Morimoto, Y., "Learning Translation Templates form Bilingual Text," Proc. of *14th COLING*, 1992
[2] 阿部、桝田、「英字新聞を読むための表現辞典」、語学春秋社、1992
[3] 藤井、「ニュース英語の翻訳プロセス－異文化間コミュニケーションとしての一考察」、早稲田大学出版部、1996