4Q－7

対訳コーパスから細粒度翻訳知識の自動獲得手法

任　福継[1]，簡　幼良[2]，范　莉馨[2]，栃内　香次[3]
[1]広島市立大学情報科学部，[2]大連理工大学計算機学部，[3]北海道大学工学部

1．はじめに

我々はこの数年来、実用的機械翻訳システムの研究・開発を行ってきたが[4,5]、最大の問題点は翻訳知識の獲得であった。機械翻訳システムの翻訳の質は、これらの知識の量と質に大きく依存する。これらの知識の総量は膨大なものであり、機械翻訳システムの開発にかかわる作業のうちのほとんどは、これらの知識の作成に費やされることになる。

対訳コーパスには両言語の情報や両言語間の翻訳に有用な知識が含まれ、これらの情報や知識が機械翻訳システムの翻訳品質の向上に有効に作用すると考えられる。コーパスからの知識の自動獲得法に関し、既に多くの研究があり[1-3]、大規模な対訳コーパスから各種各様な翻訳知識の自動獲得が期待されているが、現実には、有用な翻訳知識を完全に自動でコーパスから取り出すのは容易でない。

本稿では、人間とコンピュータ各々の利点を活かす観点から、対訳コーパスを別の視点から利用する方法を提案する。これは、対訳コーパスから完全に自動的に翻訳知識を獲得することができない現実から、人間の介入により対訳コーパスを有効的に利用するアプローチである。具体的には、現段階では困難であり、かつ機械翻訳システムに強く要求される文節の対応と制約条件を人間に任せ、その他をコンピュータによって自動的に獲得するという方法である。

我々は提案する方法に基づき、日中対訳コーパスからの翻訳知識獲得システムを構築し、それにより獲得した知識を我々により開発した日中機械翻訳システムJC_TRANSに組み込んで翻訳実験を行った。実験結果から本論文で提案する手法の有効性を確認することができた。

2．対訳コーパス情報付けツール

対訳コーパスの文に構文情報や意味情報を付けるとき、要求される情報が多すぎたり、細かすぎると、一般のユーザが使えなくなるし、たとえ付けることができても、ユーザの認識判断の差異により、付けられた情報が利用できなくなる。一方、要求する情報が少なすぎなら、情報の不足で知識の獲得が困難になり、期待な翻訳結果が得られない。このような理由で、我々は機械翻訳における知識に渉及する情報を分析し、対訳コーパスに半自動的情報付けツールＴＣＴＬを開発した。図１にＴＣＴＬの一画面を示す。

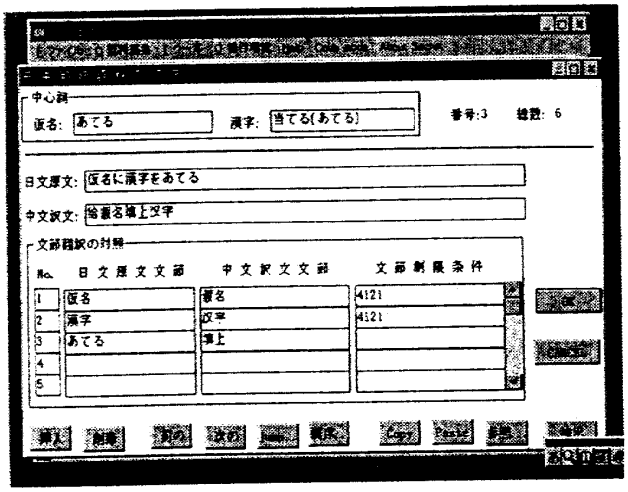

図１　コーパスに情報付けツール

3．翻訳知識の獲得手法

3.1　知識の種類(GetType)

日中機械翻訳の視点から、（１）名詞性複合語の知識、（２）名詞の知識、（３）用言の知識という３つの種類の知識を対訳コーパスから獲得するのが望ましい。

3.2　翻訳知識獲得アルゴリズム

次に用言の知識獲得を例として、情報付け対訳コーパスから翻訳規則（翻訳知識とも呼ぶ）を生成するアルゴリズムを説明する。

ある中心詞govに対し、その情報付け対訳コーパスを下記の＜１＞と仮定する。

原言語文字列　NS = b1 b2 ... bi ... bn

目的言語文字列 CS = c1 c2 ... cj ... cm　　<1>

ここに、biは原言語文の文節であり、この例はnつの文節から構成される原言語文、mつの単語あるいはフレーズから構成される目的言語文からなる。目的文の構成要素cjとは原言語文中のある文節の独立語訳語、あるいは独立語訳語以外の目的言語の文字列である。

なお、目的言語と原言語の訳語対応ペア集合Pは<2>式のように記述される。

P = ｜bi, cj, Li ｜ bi JNS,cj JCS,Li JLimit｜　<2>

ある制約条件Liの下で、文節biの訳語はcjである。ここでの制約条件Liは属性の集合であり、テスト可能属性集合中の１つあるいは１以上のテスト属性になる。また、Limitは＜３＞のような要素からなる。

[Limit : *MULTIPLE* BUNRUI, FSK, JBK, SEM, PINCI, POSI] <3>

ここに、BUNRUIは文節の種類、FSKは付属語単語、JBKは独立語単語、SEMは意味分類属性、PINCIは独立語品詞属性、POSIは文節の順序号である。

本手法では、形態素解析の結果と上述した対訳文に付けた情報により、自動的に翻訳規則を生成する。

翻訳規則の形式化定義は<4>のようになる。

R = TS(テスト):OS(動作) <4>

まず文の各構成要素の特徴を分析し、テスト条件TSを作成する。そして、原言語文と目的言語文の構造の対照により、文の構文構造を確立する。ここでの構文構造は概して中心語の各フレーム、翻訳モデル（翻訳フレーム）、訳語の選択などの動作OSを指す。最後に、TSとOSを合成し１つの翻訳規則を作成する。

3.3 知識獲得の例

日本語原文： 論文を直す。

中国語訳文： 修改论文。

形態素解析結果：論文.を／直す.／

この例について、下記のように情報付けを経て、<5>の翻訳知識を獲得する。

(b1) 論文 论文 [文書類]

(b2) 直す 修改

文節b1特徴：PINCI =n BUNRUI = 体言

JBK = 論文 FSK = 'を'

POSI = 0 SEM = c [文書類]

c は限定条件を表示する。

X.(FSC[0]isを&YYFL[0]is2112):CreatCase(X,S,OBJ) &ChangYY(S,修改) <5>

規則の意味は以下のように解釈する。

直す： PRED: VERB OBJ

VERB: 修改

OBJ: SEM = c [文書類]

FSK = 'を'

4. システム実現と考察

本文で提案する対訳コーパスからの翻訳知識獲得技術は我々の開発した日中機械翻訳システムJC_TRANSに応用され、WINDOWS95マシンで実現された。日本語版WINDOWS95で中国語を入力、中国語WINDOWSで日本語を入力するため、我々はそれぞれ日本語WINDOWS用中国語処理器と中国語WINDOWS用日本語処理器を開発した。翻訳システムでは、現在52795語が基本辞書に構築されており、他の5つの専門用語辞書を開発した。システムの翻訳規則ベースには約8000規則が組み込んでいる。

我々は専用のインターフェイスを開発した。ユーザはこのインターフェイスメニューでファイル編集、コーパス修正、翻訳作業などが容易に行える。

我々は日漢辞典と人手で獲得した日中対訳コーパスから、本文で述べたTCT-TOOLを用い、選択した対訳例文に情報を付け、さらに、本文で提案した知識獲得システムを利用し、翻訳規則を自動的に抽出した。そして、このようにして得た翻訳規則をJC_TRANSの規則ベースに組み込んで、機械翻訳実験を行い、本文で提案した手法の有効性を評価した。因みに、現システムでは、選択された2450キーワードで延べ6017ペアのコーパスから獲得した翻訳規則が含まれている。我々は日本語教科書（標準日本語、上、下冊）をテスト対象として翻訳実験を行い、正翻訳率95.6%という結果を得た。図2は実際の翻訳画面の一部を示す。

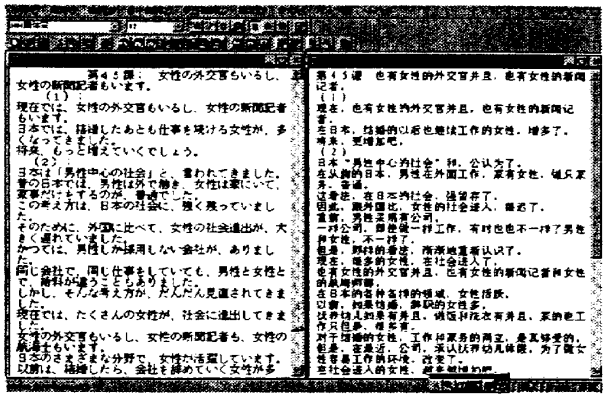

図２ システムの実行画面の一部

5. むすび

本稿では、人間の部分的介入により対訳コーパスから翻訳知識を獲得する手法を述べった。また、この手法に基づき実用化システムを構築した。これにより得られた翻訳知識を日中機械翻訳システムに実装して、翻訳実験を行った。「標準日本語」にある全部1582文をテスト対象とし、正翻訳率95.6%という良い結果が得られた。今後、翻訳規則の合併および規則の適用優先順位の判定手法を研究する予定である。

謝辞 本研究の一部は広島市特定研究費(95A414)及び文部省科研費により行われた。長い間有益な御討論、御助言、さらに実験の部分実施を頂く北海道大学工学部電子機器講座、大連理工大学自然言語研究室および広島市立大学自然言語処理学講座の各位に感謝致します。

参考文献

1) Kitamura,M. and Matsumoto,Y.: A Machine Translation System based on Translation Rules Acquired from Parallel Corpora, Proc. RANLP,99.27-44(1995).

2) Kaji,H., Kida, Y. and Morimoto, Y.: Learing Translation Templates from Bilingual Text, Proc. COLING-92, pp.672-678(1992).

3) 北村、松本：対訳コーパスを利用した翻訳規則の自動獲得、情処学論、Vol.37, N0.6, pp.1030-1040(1996).

4) 任，范，宮永，栃内：家族モデルを用いた文の分解に基どく日中機械翻訳システム，情処学論，Vol.32,No.10, pp.1149-1258 (1991).

5) 任，宮永，栃内：日中常用文型機械翻訳システム，信学誌D-2,Vol.J74, No.8, pp.1060-1069, (1991).