バイリンガル･コーパスを用いた対話文翻訳のための局所文脈解析

6P－6

側嶋康博
ATR音声翻訳通信研究所

1. はじめに

従来の機械翻訳では文脈処理が行われないため、文脈に依存した表現は訳し分けは困難であった。これまで、対話文翻訳のための文脈処理として、ヒューリスティックに特定領域のプランを作成する手法[1]、発話行為タイプ (IFT) の推移を n-gram など統計的に処理する手法[2] などが提案されているが、「文」を単位とした入力言語のモノリンガル解析に中心が置かれている。しかし、2. で述べるように、現実の翻訳では、文対応の割合は高くない。

本研究は、用例主導翻訳[3]の考えを文脈処理、特に訳語の選択に応用しようとするもので、IFT を付与したバイリンガル･コーパスを、対話文翻訳のための文脈知識として利用している。隣接するバイリンガル用例知識と翻訳対象とを比較して得点化し、最高得点の訳語を選択するこの局所文脈解析により、高頻出の「はい」「いいえ」「そうです」「が」「けれども」など、対話の進行に重要な表現の訳語を高精度で選択することができた。本稿では、この局所文脈解析の概要と、この処理を用いた翻訳実験の結果について報告する。

2. 対話文の特徴とキーボード会話

本研究で対象としたのは二者が行う目的指向型会話であり、具体的には、国際会議に関する問合せを日英各言語をそれぞれ母国語とするオペレータが翻訳者を介してキーボードで対話を行った模擬会話実験結果 (177会話) である。キーボード会話では音声対話に頻出する言いよどみや言い直し、整合性が欠如したものが実験者自身の意志で除去され、かつ会話特有の表現は残されている。しかし、文と文の対応は７割程度にとどまり、様々な対応関係が存在する。ただし、ここで文に制限せず、独立した節の対応も含めると、約93%が両言語で対応していることがわかった。これらの対応をもとに、文に制限されない一般化した対話モデルを考案した。

Bilingual Corpus-based Local Context Analysis for Dialogue Translation
Yasuhiro Sobashima
ATR Interpreting Telecommunications Research Laboratories

3. 対話の基本モデルと翻訳単位

対訳データは、発話ごとに図2 の各行のように節または文単位で対応付けを行い、さらに、信号、接続、内容の各要素に分割した。

```
J: {信号} <接続> (内容) <接続> (内容) <接続>
E: {signal} <joint> (message) <j.> (m.) <j.>
```

図1　発話の構成

信号は発話先頭の呼びかけや応答の語で、接続は句読点を含む接続(助) 詞類に対応する。また、内容は接続要素にはさまれた意味伝達部分である。

```
J:{はい、}<それでは>(用紙をお送りします)<ので、>
E: {Okay,} <then> (I'll send it out soon)<.>
J: (ご住所をお願いします)<。>
E: (Could you give me your address)<?>
```

図2　発話の要素を用いた対応記述

１つの内容が対象言語の１内容に対応しないもの (7%) もあるが、簡便のため、２言語で対応する複数個を１つの内容と扱い、信号と接続を含めたこのセットを翻訳単位 (TU) と設定した。すなわち翻訳単位は、信号、(前置) 接続、内容、(後置) 接続の各要素で構成される。177 のキーボード模擬会話 (日本語 4517文、英語 4738文、TU数 5419) すべてについて、この対応記述を行い、各内容に20 分類の発話行為タイプ (IFT) を付与した。

情報提供			要求		
事象	A1	過去	情報		
	A2	現在		X1	選択疑問
	A3	未来		X2	提供依頼
心情	B1	感謝歓迎	判断		
	B2	陳謝危惧		Y1	YN疑問
	B4	希望要望		Y2	確認
	B4	その他	行為		
	B5	別れ挨拶		Z1	依頼
判断	C1	理解受諾		Z2	禁止
	C2	正誤肯否		Z3	許可
	C3	決意選択			
	C4	義務必要			
	C5	可能許可			

図3　内容の分類 (IFT)

4. 局所文脈解析と得点計算

表現Aの訳$A'_1, A'_2, ..., A'_n$から最適解を決定するのに各々の得点$a_1, a_2, ..., a_n$を計算した後、最高得点のものを採用する。ここで、表現XとYの近さと翻訳候補の得点を次のように定義する。

近さ (X, Y) = u (c v + 1)

ここで、

u = 1　XとYのIFTが一致
= 0　XとYのIFTが不一致
v = 1　XとYの形態パターンが一致
= 0　XとYの形態パターンが不一致
c = 定数 (>0)

得点 $(A'_i) = \Sigma_j$ 近さ (X, Y_j) f

ここで、

X: テキスト(生成言語側)の直前表現
Y_j: 用例中のA'_iの直前表現
f: Y_jとA'_iが隣接した回数(頻度)

図4　得点計算

この局所文脈解析の特徴は、各訳の得点が、生成される言語側で決定されることである。例えば「そちらは会議事務局ですか。」(Is this the conference office?)の次に「はい、そうです。」という入力があった場合、その候補訳(yes, it is./yes, we are./okay, I'll do so./..)それぞれについて、得点計算を行う。そこで"is this ..."の後に「そうです」の訳となった"yes, it is.."が文脈用例にあれば、高得点を得て、選択されることになる。もし、同じ形態パターンがない場合でも、IFTが等しかったものが選択されるため、システムは頑強さを保つ。

用例主導翻訳では、用例に同じものがあれば距離0(無限大の得点に相当)を与えるため訳し分けが困難だが、本手法では有限の値を頻度に乗じて加算し、隣接条件で訳し分けが可能となっている。

5. 翻訳実験結果

IFTと形態を用いた局所文脈解析の有効性を確認するため、形態パターンのテンプレートを作成して内容の対応付けを行い、解析・変換・生成用のデータとしてオープン(実験対象は未学習データのもの)とクローズ(学習済み)の実験を行った(図5)。

実験では、1文字ずつ入力を行い、形態辞書を検索しながらテンプレートの作成を試み、翻訳単位ができると候補訳すべてについて隣接の得点計算を行って最高得点の表現を文字列で出力する。オープン、クローズいずれの実験も、4会話から177会話まで用例学習データを増やしていきながら正答訳出率を調査した。図5の横軸は、用例のTU数で、対数目盛りになっている(オープンはTU数181~5325、クローズはTU数196~5406)。

クローズの場合は当然高い正答率(94～100%)を示したが、オープンでも、学習量の増加にしたがって、正答率が増加している。グラフ中の絶対値は入力TU数に対する正解率であり、相対値は出力したTU数に対するものである。

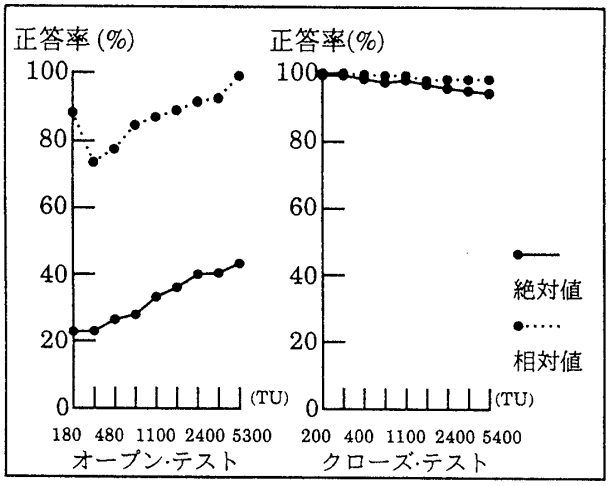

図5　実験結果

6. まとめと課題

文脈に依存した表現の選択を、バイリンガル·コーパスを使用して、高精度で行うことができた。また、文単位に依存しない対話モデルを用いたため、同時通訳的な翻訳出力が得られた。だだし、1階層のテンプレートを使用しているため、この手法だけで一般の翻訳を行うには膨大な用例が必要となる。しかし、このような特定トピックの対話であれば2000～3000TU程度の学習で、出力中正答率9割以上の翻訳が可能であることもわかった。

この局所文脈解析の手法は従来システムの前処理として使用できるが、代替表現を多く登録して、適用率を高めることも検討したい。

謝辞

本研究を進めるに当たり、飯田室長、隅田、古瀬両氏に、数々の有益なご助言をいただきました。感謝致します。

参考文献

[1] 山岡, 飯田:「階層型プラン認識モデルを利用した次発話予測手法」(信学論 D-II Vol. J76-D-II No. 6)

[2] 永田:「統計的な対話モデルの試みとその音声認識への応用」(人工知能学会研究会資料 SIG-SLUD-9202-11)

[3] 古瀬, 飯田:「変換と解析の協調的処理による翻訳手法 - 変換主導翻訳」(情報処理学会 研究報告 92-NL-87)