経験的な言語知識を利用する対話翻訳機構
－ 日英・日韓の対話翻訳システム －

古瀬 蔵　　赤峯 享　　河井 淳　　金 徳奉　　飯田 仁
ＡＴＲ音声翻訳通信研究所

1　はじめに

話し言葉の翻訳では、実時間でのコミュニケーションのための効率的な処理、文法から逸脱した表現などの多様ないい回しを扱える頑健な処理を実現することが不可欠である。我々は、言語データベースから学習した経験的な変換知識を入力文に適用することにより、翻訳結果を効率的に出力する変換主導型翻訳（Transfer-Driven Machine Translation, 以下、ＴＤＭＴと呼ぶ）[1]について研究している。ＴＤＭＴは、頑健な処理を実現するために用例に基づく手法[4]を活用している。用例に基づく手法は、意味距離計算により、入力と経験的な言語知識のベストマッチングを行ない、入力表現に対する最適な対訳表現を求めることができる。

筆者らは、経験的な言語知識の利用という統一的な処理機構による多言語間の対話翻訳の実現を目指し、言語構造の遠い日英間、近い日韓間についてＴＤＭＴによる双方向翻訳の研究を行なっている。本稿では、日英方向、日韓方向の対話翻訳システムについて報告する。以下、ＴＤＭＴの翻訳メカニズム、言語データベースからの言語知識の学習、対話翻訳システムの概要について述べる。

2　ＴＤＭＴの翻訳メカニズム

ＴＤＭＴでは、変換知識を用いた変換処理が翻訳処理の中心である。以下、変換知識、変換処理の概要について説明する。

2.1　変換知識

変換知識は原言語表現と目的言語表現の対応関係を意味的にまとまった単位で記述する。変換知識の言語表現はパタンによって表す。パタンは、任意の言語表現と照合可能な変項部分や表層語句などにより構成される。

変換知識は、原言語表現パタンごとに翻訳用例を収集、編集することにより作られる。例えば、日本語表現「Ｘの Ｙ」について、「ホテルの住所」→ *the address of the hotel* や「英語のパンフレット」→ *the pamphlet in English* などの翻訳用例を収集し、次のような日英の変換知識が作られる。

X の Y =>
　Y' *of* X' ((ホテル, 住所),(新幹線, 切符)...),
　Y' *in* X' ((英語, パンフレット), ...),
　Y' *for* X' ((明日, 天候), ...),
　　：

X' は X の対訳を示す。この変換知識は、「X の Y」が、Y' *of* X'、Y' *in* X' などさまざまな目的言語表現の可能性を持つことを表しており、それぞれの目的言語表現が選ばれる場合の変項部分 X、Y の具体的な語を (ホテル, 住所) のように併記する。

ＴＤＭＴは用例に基づく手法を活用している。用例に基づく手法では、変換知識の中から入力に最も意味的に類似する翻訳用例を意味距離計算により求め、その翻訳用例を模倣することにより翻訳結果を得る [4]。例えば、翻訳の入力が「日本語のパンフレット」とする。「X の Y」に関する変換知識の中で「英語のパンフレット」が意味的に最も近ければ、Y' *in* X' を使って *the pamphlet in Japanese* という翻訳結果を得る。

翻訳用例を大量に学習することが、変換知識における訳し分けの条件を詳細に記述し、目的言語表現の決定の精度を高めることになる。

2.2　変換処理

ＴＤＭＴでは、変換が翻訳処理の中心であり、形態素処理、生成、文脈処理などのモジュールが、変換モジュールと協調して、適切な翻訳結果を効率的に出力しようとする。

変換では、図 1に示すように、変換知識を用いて、Constituent Boundary Parsing [3] による原言語構造の導出、意味距離計算の結果に基づく目的言語構造への写像を行なう。

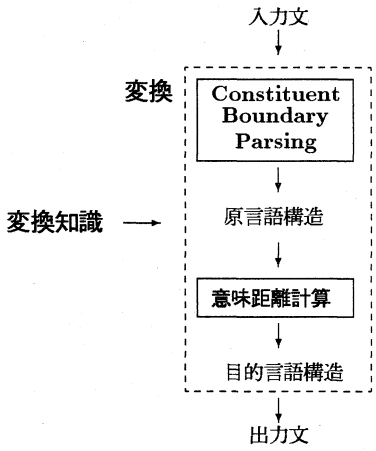

図 1: TDMTの変換処理

「京都に来てください」の日英翻訳を例にとる。変換知識の原言語表現である「X てください」や「X に Y」を組み合わせて原言語構造を作る。さらに、意味距離計算の結果をもとにして、原言語構造の各パタンが、最尤の目的言語表現である *Please X'* や *Y' to X'* に写像され、目的言語構造を得る。この目的言語構造から *Please come to Kyoto.* という英文を生成する。

TDMTの変換処理は、原言語と目的言語のどの組合せについても共通である。

3 言語データベース

翻訳対象と同じトピックの言語データベースは、翻訳に有効な言語知識を与える。TDMTプロトタイプシステムの翻訳対象は、音声翻訳の使用が有効である場面を想定した旅行会話である。ATRでは、旅行会話についてバイリンガルの模擬会話収集を行なっており [2]、このデータをもとにシステムの言語知識の学習を行なっている。

また、基本表現を網羅するために、模擬会話に加えて、旅行会話基本表現集からの言語知識の学習も予定している。

以下、バイリンガル言語データベースの概要、言語データベースからの言語知識の学習について述べる。

3.1 言語データベースの概要

言語データベースの模擬会話は、相手の言葉が理解できない外国人旅行者と旅行業者がそれぞれの母国語でしゃべり通訳を介して意志の疎通を行なうという設定により作られ、録音される。通訳は音声翻訳システムとしての役割を演じている。日本人話者と米国人(韓国人)話者の日英(日韓)間バイリンガル会話を言語データベースとして収集しており、現在の語彙サイズは、日英間で約60万語、日韓間で約5万語規模である。いろいろないい回しを収集するために、多くの話者が会話収集に参加している。音声翻訳システムで処理することを想定して、割り込み、方言、乱暴な言葉使い、連続して10秒を超える発話などは禁止している。これらの制約はあるものの、収録された会話の文にはさまざまな話し言葉特有の現象を含んでいる [5]。

録音により収集されたバイリンガル会話は、テキストファイルとして書き起こされる。バイリンガルテキストデータがまず作られ、バイリンガルテキストデータから日本語部分、英語(韓国語)部分をそれぞれ抜き出したモノリンガルテキストデータが作られる。モノリンガルのテキストデータから形態素情報が付与されたTaggedデータが各言語ごとに作られる。

3.2 言語知識の学習

テキストデータとTaggedデータを持つ言語データベースを使って図2のように言語知識の学習を行ない、翻訳対象の文の翻訳に対応できるようにする。

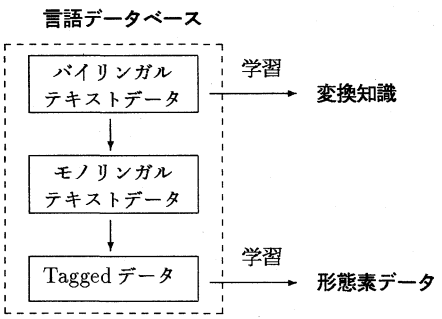

図 2: 言語知識の学習

これらの学習により、翻訳精度を高め、広範囲

に翻訳できるようにする。形態素データと変換知識の学習について、以下に示す。

3.2.1 形態素データの学習

Taggedデータから翻訳対象の語彙となる辞書や形態素解析の際の連接尤度に用いるn-gram情報など形態素データを学習する。形態素データは形態素処理や生成で使用する。テキストデータからTaggedデータの作成は人手で、Taggedデータから形態素データの学習は自動的に行なっている。

3.2.2 変換知識の学習

システムが翻訳ができるようにバイリンガルテキストデータを用いて変換知識の学習を行なう(この学習を以下、翻訳訓練と呼ぶ)。図 3は、ＴＤＭＴプロトタイプシステム上で「Xですか」、「XはY」などに関する変換知識について学習を行ない、「部屋は狭くないですか」という文が日韓翻訳できるようになったことを示している。

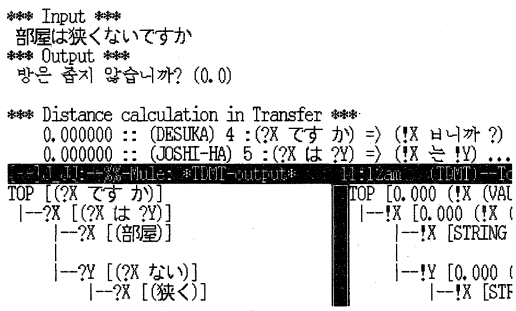

図 3: 日韓方向の翻訳訓練

バイリンガルデータからの変換知識の学習は、現在、人手で行なっている。

4 日英翻訳と日韓翻訳

旅行会話を翻訳対象とするＴＤＭＴのプロトタイプシステムは、日英間、日韓間など多言語の対話翻訳を実現できるよう構築されている。多言語翻訳に対応するため、形態素処理と生成について各言語固有のモジュールが用意されているが、システムの中心部分である変換モジュールは共通である。現在、プロトタイプシステムは、Common LISPで記述され、UNIXのMuleエディタ上で動作している。以下、モジュールの整備や言語知識の学習が先行している日英方向、日韓方向の翻訳について述べる。

4.1 翻訳訓練

現在、システムは日英間で約3000語、日韓間で約1500語の辞書、日英方向で約900文、日韓方向で約320文の翻訳訓練による変換知識を蓄積している。翻訳訓練を行なった会話のトピックは、ホテルの予約、ホテルの紹介、ホテルでのサービス、乗物の切符購入、道案内、交通手段問い合わせ、観光ツアーの案内など旅行会話全般に渡っている。

翻訳訓練では、文脈処理の必要性や英語の冠詞生成などの問題のために、完全な出力結果を得られない場合もあったが、変換知識の更新により、訓練データに対しおおむね理解可能な翻訳結果をシステムは出力している。

図 4と図 5は、翻訳訓練の結果、システムが生成した日英翻訳と日韓翻訳の出力結果を示す。日本語入力文は言語データベースの中の実際の文である。

4.2 対訳の種類

日英翻訳も日韓翻訳も言語知識の学習方法、翻訳メカニズムは同じであり、ＴＤＭＴの統一した枠組での多言語翻訳が実現できる見通しを得た。一方、日英間と日韓間の言語的な距離が、学習した変換知識の内容に反映されていることが分かった。

表 1は、日本語の格(係)助詞表現が変換知識を学習した言語データにおいて、対訳の種類がいくつあったかを、日英翻訳、日韓翻訳それぞれについて調べた結果である。用例数は異なりである。

表 1: 日英翻訳と日韓翻訳における助詞の訳し分け

日本語パタン	日英方向		日韓方向	
	用例	対訳の種類	用例	対訳の種類
X は Y	194	25	66	1
X が Y	109	9	40	1
X の Y	216	32	88	2
X を Y	117	12	41	1
X に Y	106	18	55	5
X で Y	94	18	33	5

source : "ラスベガスには夜の九時までには着きたいんです"
target : "i would like to arrive at las vegas by nine o'clock at night" (0.0)

source : "九時十五分のバスに乗られましたら夜の八時までには着きますが"
target : "if you get on the bus leaving at nine fifteen , you will arrive by eight o'clock at night" (0.0)

source : "八時ですか"
target : "at eight o'clock" (0.0)

source : "じゃあだいたい十一時間掛かるんですね"
target : "well, does it take about eleven hours , right" (0.0)

source : "いいえ時差がありますから十二時間掛かります"
target : "no, there is the time difference and it will take twelve hours" (0.0)

図 4: 翻訳訓練による日英翻訳の出力

source : "こちらでホテルの予約はお願いできますか"
target : "여기에서 호텔 예약은 부탁할 수 있습니까?" (0.0)

source : "ご紹介はできるんですがご予約の方はお客様ご本人で直接していただきます"
target : "소개는 가능한데 예약은 손님 본인에서 직접 해 주시겠습니다." (0.0)

source : "じゃあとりあえず市内のホテルを紹介してください"
target : "그러면 우선 시내 호텔을 소개해 주십시오." (0.0)

source : "宿泊費があまり高くなくて交通の便利なところがいいんですけれども"
target : "숙박비가 너무 비싸지 않고 교통의 편리한 곳이 좋은데요." (0.0)

図 5: 翻訳訓練による日韓翻訳の出力

言語構造の近い日韓間は、対訳の種類が少なく、Open Data に対して、日英翻訳より翻訳精度が上回ることが予想される。

5 おわりに

言語構造の遠い日英間、近い日韓間の話し言葉翻訳を、統一した処理機構で行なう対話翻訳機構について述べた。経験的な言語知識の学習を続け、旅行会話のドメインが広範囲に達成できるよう、効率的で頑健な対話翻訳システムの実現を目指している。辞書のサイズは 10000 語規模を目標にしている。

今後は、言語データの中の翻訳訓練を行なっていない会話 (Open Data) に対して翻訳実験を行ないながら、訓練データや翻訳メカニズムなどの評価、改良を行なっていく予定である。

また、対話翻訳のための文脈処理、ＴＤＭＴプロトタイプシステムと音声処理との統合による音声翻訳システムの実現などが今後の課題である。

参考文献

[1] 古瀬、隅田、飯田: "経験的知識を活用する変換主導型機械翻訳" 情報処理学会論文誌、Vol.35, No.3, (1994).

[2] Furuse, Sobashima, Takezawa and Uratani : "Bilingual Corpus for Speech Translation", *Proc. of AAAI'94 Workshop 'Integration of Natural Language and Speech Processing*, (1994).

[3] Furuse and Iida: "Constituent Boundary Parsing for Example-Based Machine Translation", *Proc. of Coling '94*, (1994).

[4] Sumita and Iida: "Example-based transfer of Japanese adnominal particles into English" IEICE Transactions on Information and Systems, E75-D, No.4, (1992).

[5] 竹沢、田代、森元: "音声言語データベースを用いた自然発話の言語現象の調査" 人工知能学会研究会資料 SIG-SLUD-9403-3, (1995).