経験的知識を活用する変換主導型機械翻訳

古瀬 蔵[†,☆] 隅田英一郎[†,☆] 飯田 仁[†,☆]

本稿では，実際の言語現象から獲得した経験的知識を活用して翻訳処理を行う変換主導型機械翻訳(TDMT)を提案する．TDMTでは，必要に応じて形態素処理，解析，文脈処理，生成など他のモジュールと情報のやりとりを行いながら，変換モジュールがいろいろなレベルの言語表現の経験的知識を入力文に適用して翻訳結果を作り上げる．変換中心の翻訳メカニズムは，対訳決定，構造的曖昧性除去，入力文の正規化など翻訳処理のさまざまな部分で，用例に基づく枠組みに従って経験的知識を最大限に活用できる．したがって，TDMTは，1) 文法的に説明が難しい表現を処理できる，2) 意味距離計算により高速な処理が可能，3) 知識の記述や追加が容易など，用例に基づく枠組みが持つさまざまな利点を翻訳処理全体に反映させることができる．変換中心の翻訳メカニズムの下，経験的知識を活用することにより，入力の多様性に対応できる頑健性，翻訳結果を高速に出力する効率性などが要求される話し言葉翻訳システムを構築することができる．筆者らは日本語対話文を英語へ翻訳するプロトタイプシステムを作成し，TDMTの評価を行った．評価は，翻訳訓練やコーパス分析を通じて構築した経験的知識を使って「国際会議に関する問い合わせ会話」を対象とする対話文の翻訳実験により行った．その結果，TDMTによる高品質で効率的な話し言葉翻訳の実現可能性を確認することができた．

Transfer-Driven Machine Translation Utilizing Empirical Knowledge

Osamu Furuse,[†,☆] Eiichiro Sumita[†,☆] and Hitoshi Iida[†,☆]

This paper introduces a technique for Transfer-Driven Machine Translation (TDMT) using empirical knowledge stored from actual translations. Translation is performed by a transfer module which applies empirical knowledge to an input sentence. When necessary, other modules, such as lexical processing, analysis, generation, and contextual processing, help the transfer module to improve translation quality. This cooperative transfer-centered mechanism can utilize empirical knowledge in the example-based framework to solve a wide range of translation problems such as selection of target expression, structural disambiguation, and normalization of input sentences. For this reason, the whole translation process in TDMT gains the following advantages of the example-based framework; 1) ability to handle non-standard expressions and grammar, 2) high-speed processing using calculations of the semantic distance between linguistic expressions, 3) ease of describing and updating knowledge. By utilizing empirical knowledge with the cooperative transfer-centered mechanism, TDMT can be useful for spoken language translation, in which robustness and efficiency are especially crucial. A Japanese-English translation system for spoken dialogues concerning international conference registrations has been implemented. We have evaluated TDMT using empirical knowledge guided by translation training and corpus investigation. Experiments with the system have shown TDMT to be a promising technique for high-quality and efficient spoken language translation.

1. はじめに

外国語に習熟するためには，文法的な知識を学習するだけでなく，実際に使用されている言語表現にできるだけ多く接し，基本表現や頻出表現を大量に学習することが必要である．特に会話場面においては，経験的知識☆として蓄積した基本表現や頻出表現を活用することが多い．近年，大規模コーパスの利用が可能に

† ATR 自動翻訳電話研究所
ATR Interpreting Telephony Research Laboratories
☆ 現在 ATR 音声翻訳通信研究所
Presently with ATR Interpreting Telecommunications Research Laboratories

☆ 本稿では，実際の言語表現から獲得された知識を「経験的知識」と呼ぶ．

なり，機械翻訳においても，アナロジーに基づく手法[10]，統計的手法[1]など経験的知識に着目した手法が盛んに研究されている．

本稿では，経験的知識を活用して翻訳処理を行う変換主導型機械翻訳 (Transfer-Driven Machine Translation，以下，TDMT と記す)[3] を提案する．TDMT は変換モジュールを中心に各モジュールが翻訳のために協調した処理を行って翻訳結果を作り上げる．TDMT の変換中心の翻訳メカニズムは経験的知識を最大限に活用することができ，高品質で効率的な話し言葉翻訳システムを構築することができる．

TDMT は，用例に基づく (Example-Based) 枠組みに従って経験的知識を活用する．用例に基づく枠組みでは，経験的知識として蓄積された用例の中から，入力表現とベストマッチする用例を意味距離計算により求め，ベストマッチした用例の対訳情報を使って翻訳結果を作る．用例に基づく枠組みは，1)実際の言語表現から獲得した経験的知識を活用できるため文法的に説明が難しい表現を扱える，2)単純な意味距離計算により高速な処理が可能，3)知識の記述や追加が容易など，さまざまな利点がある．これまで，名詞句や単文など特定表現を翻訳する手法[14),16]，既存のトランスファ方式の機械翻訳システムに組み込まれた変換手法[15),18]，意味距離計算と DP マッチングを組み合わせた並列構造解析手法[7]などにおいて，用例に基づく枠組みが利用されている．しかし，これらの手法が経験的知識を活用しているのは翻訳処理の一部である．一方，TDMT は，文から語句までいろいろな言語単位の経験的知識を入力文に適用し，対訳決定，構造の曖昧性除去，入力文の正規化など翻訳処理のさまざまな部分において経験的知識を活用して翻訳結果を作り上げるので，用例に基づく枠組みの利点を翻訳処理全体に反映させることができる．

変換中心の翻訳メカニズムの下，経験的知識を活用することにより，入力の多様性に対応できる頑健性，翻訳結果を高速に出力する効率性などが要求される話し言葉翻訳システムを構築することができる．筆者らは，日本語対話文を英語へ翻訳するプロトタイプシステムを作成し，TDMT の評価を行った．評価は，翻訳訓練（経験的知識の追加により翻訳対象分野の例文を翻訳できるようにすること）やコーパス分析を通じて構築した経験的知識を使って，「国際会議に関する問い合わせ」を対象とする対話の翻訳実験により行った．実験の結果，TDMT による高品質で効率的な話し言葉翻訳の実現可能性を確認することができた．

以下では，日英方向の翻訳を例にとり，2章で用例に基づく対訳決定処理，3章で TDMT の翻訳処理の中心である変換処理，4章で変換と他の処理の協調，5章でプロトタイプシステムによる翻訳実験とその評価について述べる．

2. 用例に基づく対訳決定処理

用例に基づく枠組みは，入力語句と用例の間の意味距離計算という一貫した方法により原言語表現を最尤の目的言語表現へ変換する．意味距離計算は単純であり，高速な対訳決定処理を実現できる．本章では用例に基づく対訳決定処理について説明する．

2.1 用例を使った変換知識の記述

TDMT の変換知識は，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表現する経験的知識である．原言語表現と目的言語表現の対応関係は必ずしも一対一ではない．原言語表現は訳し分けの鍵となる部分を持ち，この部分を具体化する語句によって対訳が決定される．そこで，原言語表現の訳し分けの鍵となる部分の用例* と，対応する目的言語表現との関係を実際の言語表現から獲得し，TDMT の変換知識として次のように記述する**．

原言語表現＝＞目的言語表現$_1$ ($E_{11}, E_{12}, \cdots$),
:
目的言語表現$_n$ ($E_{n1}, E_{n2}, \cdots$).

この記述は，目的言語表現に n 通りの可能性があり，用例が E_{ij} であるとき，対訳は目的言語表現 i となることを示す．E_{ij} は語句の組である．TDMT の変換処理において，入力表現の語句とベストマッチする用例を意味距離計算により求め，ベストマッチした用例の目的言語表現を，対訳として選択する．入力表現の語句に最も意味的に近い用例が E_{ij} であれば，目的言語表現 i を原言語表現の対訳として選択する．このように用例は原言語表現の訳し分け条件となる．

2.2 意味距離計算

入力文の語句と用例の間の意味的な近さを求める方

* 「用例」という言葉は，「対訳用例（原言語表現とその対訳表現の対の具体例）」という意味で使われることがあるが，本稿では，「原言語表現の訳し分けの鍵となる部分についての具体例」という意味で使用する．また，本稿では，説明を簡潔にするために，用例の記述を省略したり，目的言語表現の記述を一つだけにすることがある．

** 本稿では，具体的に記述されているもの以外にも可能性があることを“：”により表す．

法として，TDMT は，隅田らの意味距離計算[16]を採用している．この方法では，まず，シソーラス（類語辞典[13]に準拠）の概念階層における意味概念間の位置関係によって，入力文の単語 i と用例の単語 e の間の意味距離 $d(i,e)$ を計算する．意味距離は，0 から 1 までを値域とし，0 に近いほど i と e は意味的に類似していることを示す．意味距離はシソーラスの与え方によって値が変わるが，プロトタイプシステムで現在使用しているシソーラスに基づいた意味距離の例を示す．

d（会議事務局，事務局）$=0.00$.

d（こちら，私）$=0.44$.

d（会議事務局，鈴木）$=1.00$.

原言語表現の訳し分けの鍵となる部分についての入力文の表現と用例を，それぞれ，I と E とする．I と E の間の意味距離は，I と E を構成する単語間の意味距離を基にして計算する．I および E が，次のように t 個の語句の組として構成されているとする．

$I=(i_1,\cdots,i_t)$.

$E=(e_1,\cdots,e_t)$.

I と E の間の意味距離を次のように計算する．

$$d(I,E)=d((i_1,\cdots,i_t),(e_1,\cdots,e_t))$$

$$=\sum_{k=1}^{t} d(i_k,e_k)\cdot w_k \qquad (1)$$

w_k は，翻訳における k 番目の要素の重みを示し，0 から 1 までを値域とする☆.

2.3 パタンレベルの変換知識を使った対訳決定処理

TDMT では言語表現の表現形式により変換知識を，ストリング，パタン，文法の三つのレベルに分類し，入力文の性質に応じて，これらのレベルの変換知識を使い分けて翻訳処理を行う．TDMT で用例に基づく枠組みを最も活用するのは，パタンレベルの変換知識を使った対訳決定処理である．以下，本節でパタンレベルの変換知識を使った対訳決定処理について，次節でストリングレベルと文法レベルの変換知識について説明する．

パタンレベルの変換知識は，文法属性を表現しない X のような記号（以下，「可変部」と呼ぶ）と表層語句とにより言語表現を表す．可変部は訳し分けの鍵となる部分であり，可変部を具体化する用例を訳し分け条件として記述する．例えば，次の変換知識は，「X は Y です」にさまざまな英語表現が対応することを示す．

X は Y です ＝＞

X' be Y'（(私，鈴木)，(ここ，事務局)，…），

X' may be paid by Y'（(費用，現金)，…），

X' will be done by Y'（(講演，様)，…），

：

X' は原言語表現 X に対応する目的言語表現を，(私，鈴木) は，X=「私」，Y=「鈴木」という用例を表す．「私は鈴木です」の場合，「X は Y です」に対する目的言語表現は "X' be Y'" である．可変部 X，Y を具体化する入力文の語句の組を I とする．例えば，入力文が「こちらは会議事務局です」の場合，I は（こちら，会議事務局）となる．

w_k の値を一律に 0.5 とすると，I と用例（私，鈴木）の間の意味距離は次のようになる．

d（(こちら，会議事務局)，(私，鈴木)）

$=d$（こちら，私）$\cdot w_1+d$(会議事務局，鈴木) $\cdot w_2$

$=0.44\times0.5+1.00\times0.5$

$=0.72$.

I と各用例との距離が以下のようになっており，I との意味距離が最小の用例は（ここ，事務局）であるとする．

d（(こちら，会議事務局)，(私，鈴木)）$=0.72$.

d（(こちら，会議事務局)，(ここ，事務局)）$=0.25$.

d（(こちら，会議事務局)，(費用，現金)）$=0.89$.

d（こちら，会議事務局)，(講演，様)）$=1.00$.

（ここ，事務局）を用例として持つ "X' be Y'" が，最尤の目的言語表現として選択され，「こちらは会議事務局です」の翻訳結果として "this is the conference office." を得ることができる．

用例の追加は訳し分けの条件を詳細化し，変換処理の精度の向上につながる．TDMT の変換知識は，表

☆ 隅田ら[16]は，k 番目の要素の表現を固定したときの目的言語表現の分布に基づいて w_k の値を計算している．例えば，パタン「XのY」における X の重みは，「京都のY」のようにXの表現を固定した各表現に対する「X の Y」の目的言語表現のばらつきを基に計算される．全体的にばらつきが少なければ，「XのY」の目的言語表現を決定するときに与えるXの影響は大きいので，Xの重みに大きな値を与える．ただし，プロトタイプシステムでは，佐藤の MBT 1 b[14] と MBT 2[15] 同様，w_k の値を一律に $1/t$ としている．重みを一律にすると複数の目的言語表現が同距離になりやすいが，用例を十分に与えることによりこの問題をカバーできる．本プロトタイプシステムでは目的言語表現の決定について良好な結果を得ている．

層を反映した形なので記述しやすく，用例の追加による変換処理の高精度化が容易にできる．

2.4 ストリングレベルと文法レベルの変換知識

以下，ストリングレベルと文法レベルの変換知識について説明する．

・ストリングレベルの変換知識

語句や文の対応関係を表層語句のみで表す．

ありがとうございました ＝＞ Thank you.

こちら ＝＞ this☆.

・文法レベルの変換知識

品詞など文法カテゴリーにより言語表現を表し，各文法カテゴリーを具体化する語が用例を構成する．次の変換知識は普通名詞が3個並んだ場合の名詞句の翻訳に使われる．CN は普通名詞を表す．

$CN_1\ CN_2\ CN_3$ ＝＞

CN_3' of CN_1' ((研究会，開催，期間),…),

$CN_2'\ CN_3'$ for CN_1' ((発表，申込み，用紙),…),

：

「研究会開催期間」と「発表申込み用紙」は，「$CN_1\ CN_2\ CN_3$」を具体化し，それぞれ，“the time of the workshop” (CN_3' of CN_1')，“the application form for the presentation” ($CN_2'\ CN_3'$ for CN_1') という対応英語表現を持つ．

3. TDMT の変換処理

本章では，TDMT の翻訳処理全体のメカニズムについて述べた後，TDMT の翻訳処理の中心である変換モジュールにおける処理について説明する．

3.1 変換中心の協調的メカニズム

TDMT では，変換モジュールにおいて，文から語句までさまざまな言語的単位の変換知識を入力文に適用し，翻訳結果を作り出す．変換モジュールは，必要に応じて，形態素処理，解析，文脈処理など他のモジュールと情報のやり取りを行う．従来のトランスファ方式は解析モジュールが導く意味表現に基づいて変換以降の処理が規定されることが多い．例えば，慣用句や定型表現を翻訳するために，多層レベルの解析処理結果をあらかじめ求めた後，浅いレベルから変換知識を適用していく方式[4]，浅いレベルから解析と変換を試み，失敗すれば次のレベルの処理を行う方式[6] などが提案されているが，これらの方式は，処理の順序が固定されており，入力文の性質に応じた柔軟な処理を実現しているわけではない．一方，TDMT では，**図1** に示すように，変換モジュールを中心に各モジュールが翻訳のために協調した処理を行い翻訳結果を作り出す．この枠組みにより，入力文の性質に応じていろいろな翻訳戦略を使い分ける柔軟な処理，すなわち，単純な文に対しては変換知識の適用のみで反射的に翻訳結果を返し，複雑な文に対しては構文，意味，文脈などの知識を駆使して翻訳を行うことが可能になる．

3.2 変換モジュールにおける処理

変換モジュールは，用例に基づく枠組みに従って経験的知識を最大限に活用する．TDMT の変換モジュールにおける処理の流れは次のとおりである．

（a） 入力文に対し，変換知識の原言語側を組み合わせた原言語構造を作る．

（b） 原言語構造の部分構造ごとに意味距離計算に基づき最尤の部分構造へ変換し，目的言語構造を作る．

（c） （b）で得られた目的言語構造の中から，意味距離の総和に基づき最尤の構造を決定する．

（a）と（c）については，従来の機械翻訳方式では解析モジュールで行う処理であるが，TDMT では変換知識の情報を用いて変換モジュールで行う．以下，（a）

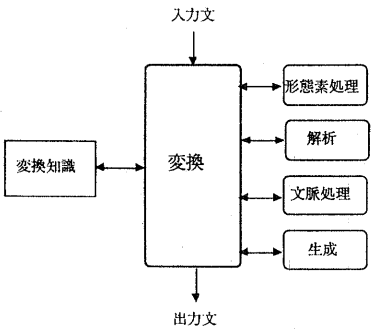


図 1 変換主導型機械翻訳の基本構成
Fig. 1 Configuration of Transfer-Driven Machine Translation.

☆ プロトタイプシステムでは，名詞，動詞などの内容語については，ストリングレベルの変換知識でデフォルトの対訳語句を与える．デフォルトの対訳が適当でない場合はパタンレベルの変換知識を使って細かく訳し分けする．例えば，「こちら ＝＞this」は，「こちら」に対するデフォルトの対訳語句が “this” であることを示しているが，「こちらに送る」を英語に翻訳すると “send us” であり，“this” は「こちら」に対する対訳語句として適当ではない．そこで，パタンレベルの変換知識「X に Y ＝＞…」に，用例（こちら，送る）では「こちら ＝＞us」と変換知識を修正することを記述する．この用例が有効な訳し分け条件となった場合，デフォルトでない “us” が「こちら」に対する対訳語句となる．

～(c)について順に説明し，最後に具体例を示す．

3.2.1 変換知識適用による原言語構造の作成

変換モジュールは，まず，変換知識の原言語部分を組み合わせて，入力文をカバーする原言語構造を作る．他の，用例に基づく機械翻訳システム[15),18)]は，依存構造などの解析結果に変換知識を適用するが，TDMT では入力文に直接，変換知識を適用する*．

以下に，TDMT における変換知識の適用メカニズムについて述べる．変換知識の大部分を占めるパタンレベルの変換知識に関しては，原言語の形態素から適用可能な変換知識へのインデックスを用意する．例えば，形態素「の」には変換知識「X の Y=>…」へのインデックスを与える．入力文を構成する各形態素からのインデックスにより適用可能な変換知識の候補を絞り込むことができる**．

次に，変換知識の原言語部分を組み合わせて，入力文の原言語構造を作り上げる．変換知識の組み合わせ方を制限するために，変換知識を，複文レベルのような大きなものから，単語レベルのような小さなものまで，言語的単位により分類する．**表1**に主な言語的単位を示す．

TDMT では，構造をできるだけグローバルに捉えるよう，「失礼ですが」のような長単位の語句，「全然～ない」のような長距離依存構造に着目した変換知識を用意すると共に，以下のアルゴリズムに従って，大きな言語的単位から小さな言語的単位のものへと変換知識を入力文に適用し，原言語構造を作る***．このアルゴリズムは，[0]から開始し，[1]～[4]を繰り返す．

[0] 入力文がストリングレベルの変換知識（以下，ストリング）とマッチすれば，成功として終了，そうでなければ，パタンレベル変換知識（以下，パタン）を最も大きい言語的単位のものに初期設定し，入力文について[1]を実行する．

[1] パタンとのマッチングを試み，マッチングに成功すれば，パタンの可変部を満たす各表現について[2]を行う．マッチするパタンが全くなければ[3]を実行する．

[2] ストリングとのマッチングを試みる．成功すれば，この部分のマッチングは終了する．そうでなければ[1]を実行する．

[3] 適用を試みるパタンの言語的単位を一段小さくし[1]を実行する．これ以上言語的単位を小さくできないなら，[4]を実行する．

[4] 文法レベルの変換知識とのマッチングを行い，成功すれば，この部分のマッチングは終了する．そうでなければマッチング失敗として終了する．

すべてのマッチングが失敗せずに終了するような変換知識適用の経路が完成すれば，入力文の原言語構造を得ることができる．例えば，「会議に申し込みたいのですが」に対しては，「X たいのですが」，「X に Y」，「会議」，「申し込み」という変換知識の原言語側を組み合わせた原言語構造(1)を得る．

(1) (((会議) に (申し込み)) たいのですが).

アルゴリズムの[1]において複数のパタンが適用可能な場合があるが，横型探索によりすべての可能な原言語構造を求め，3.2.3 項で述べる方法によって最尤のものを選択する．変換知識適用の経路を作れなければ，原言語構造を得られず，翻訳処理は失敗する．

3.2.2 意味距離計算による目的言語表現への変換

意味距離計算に基づき原言語構造の各部分構造をそれぞれ最尤の目的言語表現に変換し，原言語構造を目的言語構造に写像する．例えば，原言語構造(1)の各部分構造に対しては，意味距離計算の結果を基に "I would like to X'", "Y' for X'", "conference", "apply" へと変換し，目的言語構造 (1′) を作る．

(1′) (I would like to ((apply) for (conference))).

用例との意味距離計算は，可変部の主部 (head) に相当する語句について行う．例えば，入力文「会議に

表 1 言語的単位による変換知識の分類
Table 1 Transfer knowledge classification based on linguistic unit.

言語的単位	具体例
複文	X たら Y=> if X' Y'
単文	X たいのですが => I would like to X'
格関係	X を Y=> Y' X'
名詞句	X の Y=> Y' of X'
単語	会議 => conference

* 丸山は，原言語の構文木に対する変換知識の高速マッチングアルゴリズム[8)]，大規模なパタン集合の高速な処理アルゴリズム[9)]を提案している．前者は渡辺の用例に基づく変換方式[18)]，後者は TDMT における変換知識の入力文への適用に有効であると論じている．

** インデックスによる絞り込みはパタンマッチング可能な変換知識を全て抽出する緩やかなものであり，これにより可能な原言語構造の導出を抑制することはない．「会場への道順」において形態素「へ」は「X へ Y=>…」と「X への Y=>…」を適用可能な変換知識として抽出するが，「X へ Y=>…」を使って原言語構造を作ることは失敗する．

*** このアルゴリズムの考え方は Mu の解析手法[11),12)]と類似している．

申し込みたいのですが」に変換知識「X たいのですが＝>…」を適用する場合，可変部 X にあたる「会議に申し込み」の主部「申し込み」について用例との意味距離計算を行う．

したがって，変換知識の原言語部分に主部に相当する可変部をあらかじめ指定しておくことになる*．「X に Y」の主部は Y であると変換知識「X に Y=>…」に指定しておけば，「会議に申し込み」において，Y に相当する語句「申し込み」が主部であることが決まる．

3.2.3 意味距離の総和による最尤構造の決定

入力文への変換知識の適用に複数の組み合わせ方が存在し，原言語構造の曖昧性が生じることがある．複数の原言語構造がそれぞれの目的言語構造を作り，複数の翻訳結果を生成する恐れがある．この場合，変換で生じる意味距離の総和が最小なものを最尤な構造として選択し構造の曖昧性を除去する．例えば，「京都ホテルの1万円の部屋」はパタン「X の Y」の適用に関して次の2通りの構造がある．

（2） ((京都ホテルの（1万円の部屋)).

（3） ((京都ホテルの1万円) の部屋).

「X の Y」は，次の変換知識が示すようにいろいろな目的言語表現を持つ．

X の Y => Y' of X' ((京都, ツアー), …),
Y' for X' ((ホテル, 登録), …),
X' Y' ((円, 部屋), …),
: .

図2と**図3**は(2)，(3)それぞれの構造の木表現である．(2)の構造では意味距離 0.00 の "Y' at X'" と意味距離 0.00 の "X' Y'" に変換し，意味距離の総和が 0.00 の英語構造(2′)を得る．(3)の構造では意味距離 0.00 の "X' Y'" と意味距離 0.28 の "Y' of X'" に変換し，意味距離の総和が 0.28 の英語構造(3′)を得る．

(2′) ((ten thousand yen room) at Kyoto hotel).

(3′) ((ten thousand yen of Kyoto hotel) room).

意味距離の総和が最小である(2′)が最尤の英語構造として選ばれ，生成処理を経て，"ten thousand yen room at Kyoto hotel" という「京都ホテルの1万円の部屋」の翻訳結果を得る．

このように，用例は翻訳の傾向を捉えるだけでなく，語の共起データとして構造の曖昧性除去に利用することもできる．

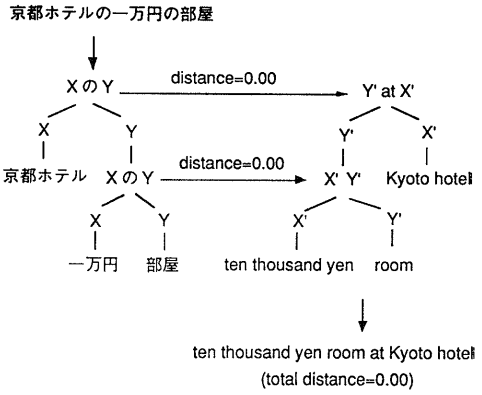


図2 「(京都ホテルの（1万円の部屋))」の変換
Fig. 2 Transfer of (Kyoto hotel no (ichi-man yen no heya)).

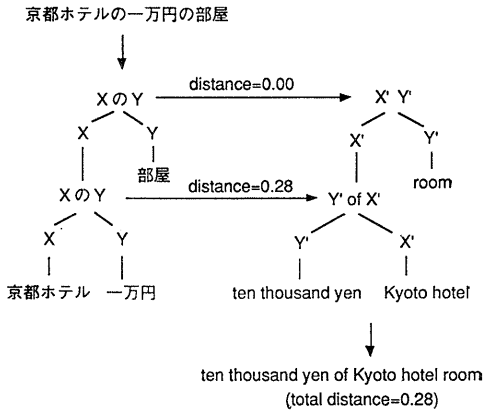


図3 「((京都ホテルの1万円) の部屋)」の変換
Fig. 3 Transfer of ((Kyoto hotel no ichi-man yen) no heya).

3.2.4 変換処理例

文(4)に対する TDMT の変換処理の流れを**図4**に示す．

（4） 「京都ホテルの1万円の部屋をお願いします.」

変換モジュールは，文(4)の各形態素からのインデックスにより変換知識を絞り込んだ後，言語的単位の最も大きい変換知識を文全体に適用しようとする．すなわち，次のような単文単位の変換知識を適用する*．

* ただし，日英翻訳のプロトタイプシステムでは，日本語の性質から表現の最後の語句を主部として捉え，用例との意味距離計算を行うので，主部の指定は省略している．

* 国際会議問い合わせに関する対話コーパスにおいて，「Xをお願いします」の対訳の多くは "X' please" でカバーされる．しかし，Xの内容によっては，"speak" や "reserve" など具体的な動詞を使った分野特有の典型的表現に訳す必要がある．

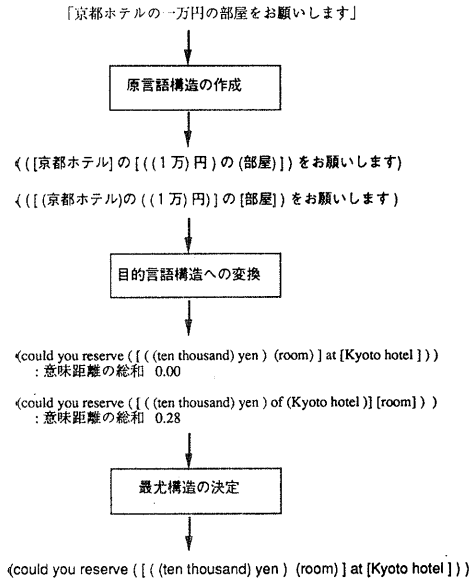


図 4 変換処理の例
Fig. 4 Example of the transfer process.

X をお願いします ＝＞
X' please ((人数), …),
may I speak to X' ((先生), …),
could you reserve X' ((部屋), …),
:

次に，この変換知識の可変部 X を具体化する表現「京都ホテルの1万円の部屋」に変換知識を適用する．この表現には名詞句単位の変換知識「X の Y＝＞…」を適用できるが，3.2.3 項で述べたように2通りの構造的曖昧さが生じる．

「X の Y」の可変部 X, Y に対し，以下のような単語単位の変換知識を適用することにより，入力文全体が変換知識の原言語部で構成された二つの原言語構造を得る．

京都ホテル＝＞Kyoto hotel.
X 円＝＞X' yen.
1万＝＞ten thousand.
部屋＝＞room.

二つの原言語構造を意味距離計算によってそれぞれ目的言語構造に変換する．そして，意味距離の総和が最小となる構造を最尤の目的言語構造として選択する．この構造から生成処理により英文 (4′) が出力される．

(4′) "Could you reserve a ten thousand yen room at Kyoto hotel?"

4. 変換と他の処理の協調

変換モジュールは他のモジュールと協調しながら3章で述べた処理を進め，適切な翻訳結果を効率的に得ようとする．本章では変換モジュールと協調する形態素処理，解析，文脈処理，生成の各モジュールについて述べる．

4.1 形態素処理

形態素処理モジュールは原言語辞書を参照して，形態素分割，および品詞活用などの情報検知を行う．各形態素は意味距離計算で使用するシソーラスコードを持つ．また，「お目にかかる」，「楽しみにする」など意味的にまとまった複数の形態素を一つの語と捉える形態素結合も行う．

4.2 解　　析

構造の複雑な文などに対しては変換知識の適用だけでは翻訳できないことがある．そのような場合，入力文を分析する解析が必要となる．解析モジュールは入力文に対して解析知識の適用を行い，その情報を変換モジュールに提供する．変換モジュールは送られてきた情報を基に変換知識の適用を行う．すなわち，変換と解析は協調して翻訳結果を作り上げる[3]．

解析知識は，変換知識と同様，実際の言語表現から獲得し，次のように用例を使って記述する．

原言語表現＝＞修正原言語表現$_1$ ($E_{11}, E_{12}, \cdots$),
:
修正原言語表現$_n$ ($E_{n1}, E_{n2}, \cdots$).

変換知識が原言語表現を目的言語表現に写像するのに対し，解析知識は修正原言語表現に写像する．解析知識には現在，1) 省略された機能語を補完する知識，2) 準体助詞，接続助詞，文末表現などのゆれを標準的な原言語表現に正規化する知識，3) 連体修飾表現，wh 構文，呼応など特定の言語現象を顕在化させる知識などがある☆．これらの知識はいずれも，変換知識を適用できるよう入力文を修正する．

解析知識は翻訳処理の頑健化にもつながる．例えば，日本語の話し言葉において助詞は頻繁に省略される．「私鈴木です」という文は，次の解析知識により「私は鈴木です」という助詞「は」が補完された文に正規化される．

☆ 解析知識は，入力文に対し過度の修正を行う可能性がある．そこで，プロトタイプシステムでは危険対の出現がありうる場合，原言語表現と同一の修正原言語表現を用例とともに用意し，解析知識の原言語表現部とマッチしても必ずしも修正が起きないようにしている．

代名詞 固有名詞=>代名詞 は 固有名詞.

解析知識の適用結果「私は鈴木です」に対し，変換知識「X は Y です=>…」の適用が可能になり，"I am Suzuki." という英文を得ることができるようになる．

4.3 文脈処理

その文だけでは，翻訳結果を一意に決定できない，あるいは，翻訳するための情報が不足していることがある．そのような場合，文脈処理モジュールにより，翻訳に必要な文脈情報を抽出する必要がある．変換モジュールと文脈処理モジュールを協調させるメカニズムとして，以下のことが考えられる．

(a) 変換知識における目的言語の訳し分け条件を文脈情報で記述し，文脈処理モジュールがその情報を変換モジュールに伝える．

(b) 解析モジュール同様，文脈処理モジュールが文脈情報に基づいて入力文を修正し，変換モジュールにその情報を伝える．

現在，プロトタイプシステムの文脈処理モジュールでは，簡単な(a)の機能が実現されている．例えば，「はい」は次のように前文情報を訳し分け条件とし，文脈処理モジュールによる前文参照の結果に基づいて「はい」を最尤の目的言語表現に変換する．

はい=>yes (前文が Yes-No 疑問文),
sure (前文が命令文),
hello (前文が「もしもし」),
:

4.4 生成

生成モジュールは目的言語構造を文字列化し目的言語文を生成する．適切な文を生成するためには目的言語の文法に従って形態素合成，語順の適正化，冠詞生成* などを行う必要がある．そのため，変換知識の目的言語表現部に生成情報を記述し，生成モジュールでその情報を利用する．例えば，英語における格の語順は日本語の場合に比べて制約が強い．日英翻訳のプロトタイプシステムでは，正しい語順の英語文を出力するために，変換知識の英語側部分に格情報を付与している．

5. プロトタイプシステムによる翻訳実験

TDMT の有効性を確認するため，プロトタイプシステムを，日本語語彙 1,500 語の規模で，Genera 8.1 リスプマシン上に実現した．以下，プロトタイプシステムが持つ経験的知識，訓練データ文* を使った翻訳処理時間の実験，非訓練データ文**を使った翻訳正解率の実験について述べる．

5.1 プロトタイプシステムの経験的知識

筆者らは，「国際会議に関する問い合わせ」に関する対話コーパス[2]の約 17,000 文の日本語文と対訳英語文の頻度調査や分析を行って，変換知識や解析知識を構築した．さらに，これらの経験的知識を補強するために，上記コーパスとは別の訓練データ 825 文を使った翻訳訓練により，経験的知識の追加を行った．現時点においてプロトタイプシステムが持つ経験的知識の概要を**表 2** に示す．

5.2 訓練データ文を使った処理時間の実験

訓練データ 825 文の約 98% に相当する 808 文について，翻訳訓練により正しい翻訳結果が得ることができた．翻訳誤りの原因の多くは，翻訳するための情報を文脈処理により抽出する必要があったことである．

図 5 に，訓練データ 825 文の翻訳処理時間の分布を示す．時間計測は CPU 性能 10 MIPS のリスプマシン XL 1200 で行った．翻訳処理時間は最大 9.4 秒，平均 1.9 秒であり，変換中心の翻訳メカニズムと意味距離計算によって効率的な処理が実現されている．

入力文の単語数の増加は，翻訳処理時間の増加に必ずしも結びついていない．この主な理由は原言語構造

表 2 プロトタイプシステムの経験的知識
Table 2 Empirical knowledge of the prototype system.

知識	種類数	具体例
変換知識		
・ストリングレベル	1783	失礼します => Good-bye.
・パタンレベル	490	X を Y=>Y' X'.
・文法レベル	45	普通名詞$_1$ 普通名詞$_2$ => 普通名詞$_1$′ 普通名詞$_2$′.
解析知識	338	代名詞 固有名詞 =>代名詞 は 固有名詞.

* プロトタイプシステムの冠詞生成メカニズムは単純であり，5 章で示す翻訳正解率は冠詞の問題を無視した数字である．

* 「国際会議に関する問い合わせ」についての10対話 225 文と，基本単語を使った例文 600 文の合計 825 文より成る．1 文当り平均語数は 9.2 語である．プロトタイプシステムにとって既知のデータである．

** 50 対話 1,056 文より成る．1 文当り平均語数は 7.8 語である．会議参加申し込み，参加費問い合わせ，参加キャンセル，会場への交通案内，ホテル案内の五つのトピックについて訓練データ文の対話を 5 人の被験者に示し，さまざまな言い回しでほぼ同じ内容の対話を作成させた．作成した文がシステムの扱う 1,500 語の範囲になるよう，使用単語について若干の補正を行った．プロトタイプシステムにとって未知のデータである．

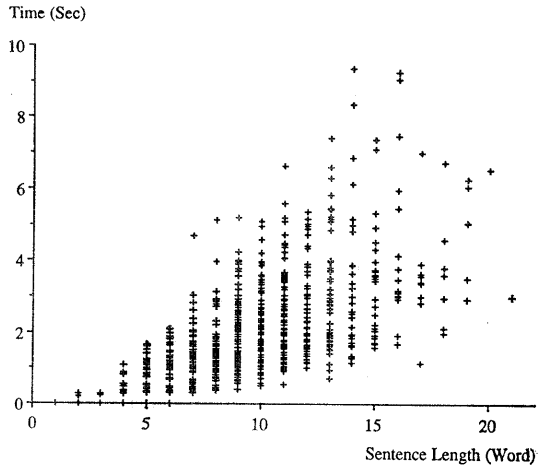
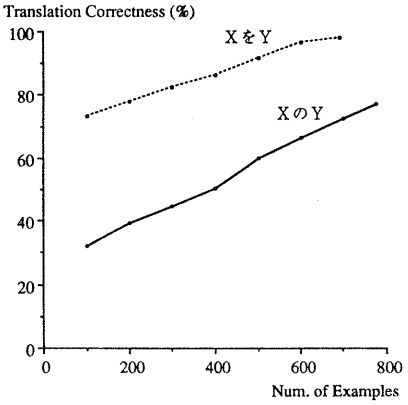
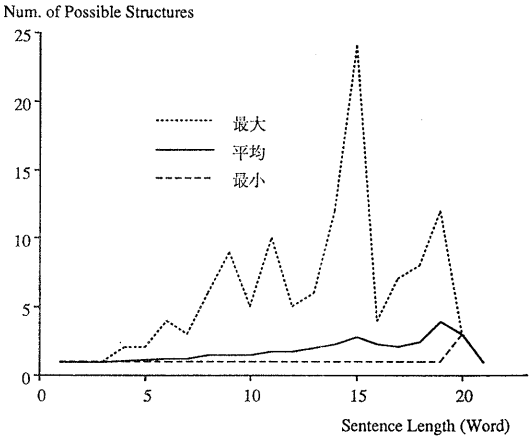


図 5 訓練データ 825 文の処理時間
Fig. 5 Translation time of 825 trained sentences.



それぞれの用例数について，用例の無作為抽出，翻訳正解率の調査を各 10 回行なった．図の翻訳正解率は 10 回の翻訳結集の平均値である．

図 7 用例数と翻訳正解率の関係
Fig. 7 Relation between number of examples and translation correctness.



可能な原言語構造の数が最大（24通り）の文
「新幹線で京都まで来るのはそれほど大変ではありません」

図 6 文の単語数と原言語構造の曖昧性の数の関係
Fig. 6 Relation between sentence length and structural ambiguity.

の曖昧性の数が抑制されていることである．原言語構造の曖昧性の数が抑制されると，目的言語構造の曖昧性の数も抑制される．**図 6** に，入力文の単語数と原言語構造の曖昧性の数の関係を示す．変換知識の適用が表層中心であり，3.2.1 項で述べたように言語的単位により変換知識の組み合わせ方を制限していることから，入力文の単語数の増加に対する原言語構造の曖昧性の数はゆるやかな増加に留まっている．全体の約 6% に相当する 51 文は 5 秒以上の処理時間がかかっている．これらの文では，複文構造における副詞句の係り先の問題などにより原言語構造の曖昧性の数が大きく，翻訳処理における意味距離計算の回数が膨大になっている．

5.3 非訓練データ文を使った翻訳正解率の実験

コーパス分析と翻訳訓練により構築した経験的知識を持つプロトタイプシステムが，訓練していない未知のデータに対し有効かどうかを評価するために，非訓練データ 1,056 文を使って翻訳正解率の実験を行った．**表 3** に非訓練データ文に対する翻訳正解率，**表 4** に翻訳誤りの典型例を示す．

非訓練データ文に対して 60〜82% の翻訳正解率を得た．翻訳誤りの多くは用例を追加することにより解消される．プロトタイプシステムは，現在，「X の Y」，「X を Y」に関する変換知識について，それぞれ 774, 689 種類の用例を持つ．その中から無作為抽出により各変換知識の用例の数を 100 種類単位で増加させ，非訓練データの中の該当する表現について翻訳正解率を調べた．結果を**図 7** に示す．いずれの表現についても，用例数の増加に対して翻訳正解率も単調に増加している．換言すれば，翻訳訓練により経験的知識を追加すれば，翻訳正解率が向上する．「X を Y」については，約 700 用例で 100% 近い翻訳正解率が得られているが，「X の Y」のように対訳の可能性が多様な表現については，経験的知識がまだ不足してお

表 3 非訓練データ 1,056 文の翻訳正解率
Table 3 Translation correctness of 1,056 untrained sentences.

トピック	特徴的な表現	文数	正解数	翻訳正解率
会議参加申し込み	住所，名前などの問い合わせ	172	137	79.7%
参加費問い合わせ	料金，支払方法に関する陳述，依頼	185	114	61.6%
参加キャンセル	代理出席の提案，断り	195	118	60.5%
会場への交通案内	値段，道筋，手段などの問い合わせ	157	128	81.5%
ホテル案内	ホテル情報の陳述，予約の依頼	347	248	71.5%
総計		1,056	745	70.5%

表 4 翻訳誤りの典型例
Table 4 Typical incorrect translations.

・目的言語表現の選択誤り
京都国際ホテルに予約する → reserve with Kyoto international hotel
登録料の六万円は振り込む → sixty thousand yen transfers
・原言語構造の選択誤り
参加料は今申し込むといくらですか → if I apply for the attendance fee now, how much?
・翻訳のための情報の不足（文脈処理が必要）
どのくらいですか → how long is it?（当該文脈では "how much is it?" の方が適切だった）

り，さらに用例を追加する必要がある．

6. 考 察

本章では，5章で示した 1,500 語規模のプロトタイプシステムの実験結果をふまえて，より高品質，高速，分野適応性の高い，大規模な話し言葉翻訳システムを構築するための課題について考察する．

6.1 協調的メカニズムによる高品質な翻訳処理

実験結果から，変換中心の翻訳メカニズムと用例に基づく枠組みによって，用例を十分に用意できれば高品質な翻訳が可能であることがわかった．しかし，用例に基づく処理では解決できない問題も話し言葉翻訳にはある．対話では翻訳するための情報が一文内では得られないことがある．このような場合は，文脈処理が必要になる．また，自然な対話文を翻訳結果として作り出すためには，生成処理を充実させなければならない．TDMT の翻訳の質をさらに改善するためには，文脈や生成などの処理と変換処理の間の協調的メカニズムを実現する必要がある．

6.2 超並列による高速な翻訳処理

現在の知識量においては，意味距離計算は高速な翻訳処理を実現している．しかし，システムの性能を向上するためには経験的知識の追加が必要であり，それに伴って意味距離計算の回数が増大し，処理時間が膨大になってしまうという懸念がある．意味距離計算については超並列による高速化の効果が報告されており[17]，経験的知識が大量になっても高速な TDMT の翻訳処理を実現することができる．

6.3 自動獲得による経験的知識の拡張

経験的知識を活用することにより，「国際会議に関する問い合わせ」という特定の分野における対話文の翻訳に対応できる見通しが得られた．しかし，他の分野を翻訳対象とする場合，使われる表現や対訳の傾向が変わるので，経験的知識をある程度修正しなければならない可能性がある．したがって，いろいろな分野の翻訳に対応するため，経験的知識をコーパスから自動獲得する技術[5]が必要になってくるであろう．

7. おわりに

本稿では，経験的知識を活用して翻訳処理を行う TDMT を提案した．TDMT は変換モジュールを中心に各モジュールが翻訳のために協調した処理を行って翻訳結果を作り上げる．TDMT の変換中心の翻訳メカニズムは，経験的知識を最大限に活用することができ，用例に基づく枠組みの利点を翻訳処理全体に反映させることができる．TDMT は，変換中心の翻訳メカニズムの下，経験的知識を活用することにより，頑健性や効率性などが要求される話し言葉翻訳に有効である．

TDMT の有効性を評価するために，日本語対話文を英語へ翻訳するプロトタイプシステムを 1,500 語規模で作成し，「国際会議に関する問い合わせ」を対象とする翻訳実験を行った．翻訳実験の結果，TDMT によって高品質で効率的な話し言葉翻訳システムを構

築できる見通しが得られた.

謝辞 本研究の機会を与えてくださった ATR 自動翻訳電話研究所の樽松明社長（現在，電気通信大学教授）に感謝いたします.

参考文献

1) Brown, P.F., Cocke, J., Della Pietra, S.A., Della Pietra, V.J., Jelinek, F., Lafferty, J.D., Mercer, R.L. and Roosin, P.S.: A Statistical Approach to Machine Translation, *Computational Linguistics*, Vol. 16, No. 2, pp. 79-85 (1990).
2) 江原暉将，小倉健太郎，篠崎直子，森元 逞，樽松 明：電話またはキーボードを介した対話に基づく言語データベース ADD の構築，情報処理学会論文誌，Vol. 33, No. 4, pp. 448-456 (1992).
3) Furuse, O. and Iida, H.: Cooperation between Transfer and Analysis in Example-Based Framework, *Proc. of COLING-92*, pp. 645-651 (1992).
4) 池原 悟，宮崎正弘，白井 諭，林 良彦：言語における話者の認識と多段翻訳方式，情報処理学会論文誌，Vol. 28, No. 12, pp. 1269-1279 (1987).
5) Kaji, H., Kida, Y. and Morimoto, Y.: Learning Translation Templates from Bilingual Text, *Proc. of COLING-92*, pp. 672-678 (1992).
6) Kitano, H.: Φ DMDIALOG: A Speech-to-Speech Dialogue Translation System, *Machine Translation*, Vol. 5, No. 4, pp. 301-338 (1990).
7) 黒橋禎夫，長尾 眞：長い日本語文における並列構造の推定，情報処理学会論文誌，Vol. 33, No. 8, pp. 1022-1031 (1992).
8) Maruyama, H. and Watanabe, H.: Tree Cover Search Algorithm for Example-Based Translation, *Proc. of Fourth International Conference on Theoretical and Methodological Issues in Machine Translation*, pp. 173-184 (1992).
9) 丸山 宏：大規模文法の文脈自由解析，信学技報，NLC 92-9 (1992).
10) Nagao, M.: A Framework of a Mechanical Translation between Japanese and English by Analogy Principle, *Artificial and Human Intelligence*, Elithorn, A. and Banerji, R. (eds.). North-Holland, pp. 173-180 (1984).
11) Nagao, M., Tsujii, J. and Nakamura, J.: Machine Translation from Japanese into English, *Proc. of the IEEE*, Vol. 74, No. 7, pp. 993-1012 (1986).
12) Nagao, M.: Are the Grammars So Far Developed Appropriate to Recognize the Real Structure of a Sentence, *Proc. of Fourth International Conference on Theoretical and Methodological Issues in Machine Translation*, pp. 127-138 (1992).
13) 大野 晋，浜西正人：類語新辞典，角川書店 (1984).
14) 佐藤理史：MBT1：実例に基づく訳語選択，人工知能学会誌，Vol. 6, No. 4, pp. 592-600 (1991).
15) 佐藤理史：MBT 2：実例に基づく翻訳における複数翻訳例の組合せ利用，人工知能学会誌，Vol. 6, No. 6, pp. 861-871 (1991).
16) Sumita, E. and Iida, H.: Example-Based Transfer of Japanese Adnominal Particles into English, *IEICE Trans. Inf. & Syst.*, Vol. E 75-D, No. 4, pp. 585-594 (1992).
17) Sumita, E., Oi, K., Furuse, O., Iida, H., Higuchi, T., Takahashi, N. and Kitano, H., Example-Based Machine Translation on Massively Parallel Processors, *Proc. of IJCAI-93*, pp. 1283-1288 (1993).
18) Watanabe, H.: Similarity-Driven Transfer System, *Proc. of COLING-92*, pp. 770-776 (1992).

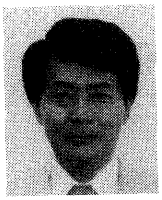
(平成5年4月2日受付)
(平成5年12月9日採録)

古瀬 蔵（正会員）

1982年九州大学工学部情報工学科卒業．1984年同大学院総合理工学研究科修士課程修了．同年日本電信電話公社武蔵野電気通信研究所入所．1990年 ATR 自動翻訳電話研究所に出向．引き続き，1993年より ATR 音声翻訳通信研究所に出向．自然言語処理，特に機械翻訳の研究に従事．電子情報通信学会会員．

隅田英一郎（正会員）

1980年電気通信大学計算機科学科卒業．1982年同大学院修士課程修了．同年日本アイ・ビー・エム(株)東京基礎研究所入所．現在(株)エイ・ティ・アール音声翻訳通信研究所主任研究員．主として自然言語処理（機械翻訳，情報検索，CAI）および超並列人工知能の研究に従事．電子情報通信学会会員．



飯田　仁（正会員）

1972年早稲田大学理工学部数学科卒業．1974年同大学院修士課程修了．同年日本電信電話公社武蔵野電気通信研究所入所．1986年ATR自動翻訳電話研究所に出向．引き続き，1993年よりATR音声翻訳通信研究所に出向．自然言語処理，特に対話理解，機械翻訳，音声言語統合処理の研究に従事．電子情報通信学会，人工知能学会，日本認知科学会，ACL各会員．