変換主導機械翻訳の超並列化の検討

6P－3

大井 耕三　隅田 英一郎　古瀬 蔵　飯田 仁　北野 宏明†
ATR 音声翻訳通信研究所　†カーネギー・メロン大学

1　はじめに

変換主導機械翻訳 (Transfer-Driven Machine Translation: TDMT)[1] は、用例主導機械翻訳 (Example-Based Machine Translation:EBMT)[2] の用例検索の枠組に基づき、対話文の高精度な翻訳を実現している。一方、対話文の翻訳にはリアルタイム性が要求される。そこで本稿では、超並列 EBMT の実験結果 [3][4] と逐次型の TDMT の実験結果の分析を通して、超並列連想プロセッサ IXM2[5] を使った TDMT の超並列化の検討を行なう。

2　TDMT と超並列 EBMT

2.1　TDMT の概要

TDMT の基本構成を図 1に示す。

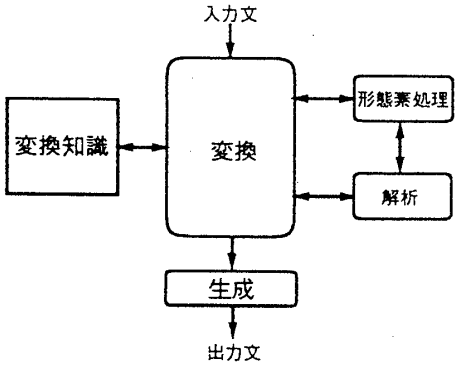

図 1: TDMT の基本構成

TDMT では、変換モジュールが入力文に最も類似した変換知識 (用例) を検索し、これを利用して翻訳を行なう。変換モジュールは、必要に応じて、形態素処理、解析、生成などのモジュールと情報のやり取りを行なう。変換知識 (用例) では、実際の翻訳に関する経験的な知識を、意味的にまとまった単位毎に、原言語表現 (SE) と目的言語表現 (TE) の対応として次のように記述している。

$$SE \Rightarrow TE_1 \ (E_{11}, E_{12}, \ldots), \quad \vdots \quad \vdots, \quad TE_n \ (E_{n1}, E_{n2}, \ldots)$$

対応関係は必ずしも一対一ではないので、各々の目的言語表現を選択するために原言語の単語リスト (E) を条件として記述している。この単語リストと対応する入力の単語リストとの間の意味距離計算によって最適な目的言語表現を選択する。

TDMT は、経験的知識を活用するので文法的に説明が難しい表現が扱え、変換中心の機構により効率的な処理が可能となっており、「国際会議に関する問い合わせ」のタスクでの翻訳実験で高精度な翻訳を実証している [1]。

2.2　超並列 EBMT

我々はすでに EBMT の用例検索を超並列計算機で高速化することを試み、良好な結果を得ている。EBMT の高速化を実現するために、EBMT の処理全体の中で最も時間がかかる用例検索を、超並列連想プロセッサ IXM2[5] を用いて実現した。IXM2 は連想メモリマシンで、64 台の連想プロセッサと 9 個の通信プロセッサからなる。各連想プロセッサは、インモス社のトランスピュータに連想メモリを装備したもので、連想メモリ上のデータに対する検索・書き込みなどが並列に行なえる。連想メモリに EBMT の用例を格納し、用例検索の並列処理により高速化が実現できる。

IXM2 の連想プロセッサ 1 台を使った実験では、用例検索は、1000 件の用例の場合、逐次計算機 SPARCstation2 の約 12 倍高速化できている [3][4]。

3　超並列 TDMT のデザイン

超並列 TDMT の構成の決定にあたり、国際会議及び旅行に関する問い合わせ対話の標準的な表現を集めたテスト文 746 文 [6] を用いて、逐次型 TDMT の処理時間の測定を行なった。用例検索は、TDMT においても全処理中で最も時間を要する部分であるため (図 3(a) 参照)、TDMT 全体の高速化ができる。表 1は、テスト文の翻

原言語表現 (SE)	割合 (%)	累積割合 (%)
(?X は ?Y)	28.88	28.88
(?X の ?Y)	19.43	48.31
(?X を ?Y)	14.97	63.28
(?X に ?Y)	11.89	75.17
(?X で ?Y)	9.08	84.25
(?X が ?Y)	6.57	90.82
(?X から ?Y)	1.24	92.06
(?X と ?Y)	0.79	92.85
(?X には ?Y)	0.68	93.53
(?X も ?Y)	0.66	94.19
⋮	⋮	⋮

表 1: 各原言語表現の用例検索時間の割合

Feasibility Study on Massively Parallel Implementation of Transfer-Driven Machine Translation
Kozo Oi, Eiichiro Sumita, Osamu Furuse, Hitoshi Iida and Hiroaki Kitano†
ATR Interpreting Telecommunications Research Laboratories
†Carnegie Mellon University

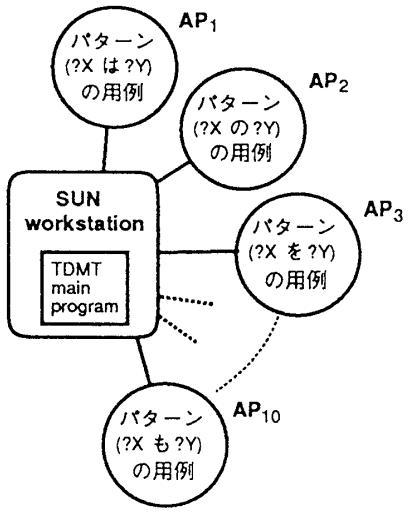

図 2: 超並列 TDMT の一構成案

訳において、変換知識中の各原言語表現の用例検索に要した時間の割合を示したもので、上位 10 個の原言語表現で全体の用例検索時間の約 94% を占めている。従って、この上位 10 個の原言語表現の用例検索を IXM2 上で実現すれば、用例検索時間の大幅な短縮が可能となる。

図 2は、超並列 TDMT の一構成案で、表 1の上位 10 個の原言語表現に対応する用例を IXM2 の連想プロセッサ ($AP_1, AP_2, AP_3, \cdots, AP_{10}$) に格納し、それらを SUN workstation に接続した形態である。

4 逐次型 TDMT と超並列 TDMT の性能分析

対話文の準リアルタイム翻訳を実現するには、1 文当りの翻訳時間は 2,3 秒以内にする必要があると考える。逐次型 TDMT のテスト文 746 文の実験結果では、約半数が 2 秒以内、他のほとんどは 2 ～ 10 秒の範囲で、25 文が 10 秒以上となっている (表 2参照)。

翻訳時間 (秒)	文数
0 ～ 2	379
2 ～ 4	237
4 ～ 6	64
6 ～ 8	29
8 ～ 10	12
10 ～	25

表 2: 逐次型 TDMT の翻訳時間の分布

図 3(a) は、表 2の各分布における翻訳時間の平均時間を、用例検索時間と他の処理時間に分けて示したものである。処理時間が長い文ほど用例検索が占める時間の割合が増加している (50% → 80%)。このように用例検索は全処理中で最も時間のかかる処理となっている。

この逐次型 TDMT の処理時間の分析と IXM2 を使うと用例検索の時間が 1/12 になること (2.2節) を考慮して超並列 TDMT の処理時間を計算すると、テスト文の約 97% が 2.5 秒以内の翻訳時間となる (図 3(b) 参照)。従って、超並列 TDMT は対話文の翻訳に十分なスピードを実現できると考えられる。

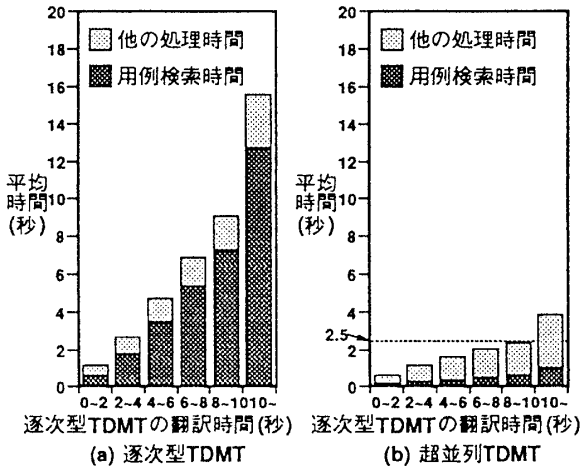

図 3: 翻訳時間 (逐次型 TDMT vs. 超並列 TDMT)

5 まとめ

対話文のリアルタイム翻訳を目指して、超並列 EBMT の実験結果と逐次型の TDMT の実験結果の分析を通して、超並列連想プロセッサ IXM2 を使った TDMT の超並列化の検討を行なった。

超並列 TDMT では、国際会議及び旅行に関する問い合わせ対話の標準的な表現を集めたテスト文 746 文の約 97% が 1 文当り 2.5 秒以内で翻訳でき、音声翻訳に要求されるリアルタイム応答実現の見込みが得られた。

現在 IXM2 を使った TDMT を実装中であり、次稿にその評価を報告する予定である。

参考文献

[1] 古瀬, 飯田 : "変換と解析の協調的処理による翻訳手法 — 変換主導型翻訳手法 —", 情報処理学会研究報告, 92-NL-87 (1992.1)

[2] Sumita, E. and Iida, H. : "Example-Based Transfer of Japanese Adnominal Particles into English", IEICE TRANS. INF. & SYST., Vol.E75-D, No.4 (1992.7)

[3] 大井, 隅田, 飯田, 樋口, 北野 : "用例主導型機械翻訳の超並列連想プロセッサ IXM2 による高速化", 情報処理学会第 46 回全国大会, 5B-5 (1993.3)

[4] Sumita, E., Oi, K., Furuse, O., Iida, H., Higuchi, T., Takahashi, N. and Kitano, H.: "Example-Based Machine Translation on a Massively Parallel Processor", Proc. of IJCAI-93, (1993).

[5] Higuchi, T., et al.: "IXM2: A Parallel Associative Processor for Knowledge Processing", Proc. of AAAI-91 (1991)

[6] 浦谷, 他 : "話し言葉の日英翻訳システムの評価法", 情報処理学会第 46 回全国大会, 6B-4 (1993.3)