E-50

用例ベース翻訳D^3のための文分割

Sentence Splitting for Example-based machine translation, D^3

土居 誉生† 隅田 英一郎†
Takao Doi Eiichiro Sumita

1．はじめに

我々は用例ベース翻訳の一つとして単語列編集距離を使った翻訳方式（DP-match Driven transDucer、以下 D^3 と呼ぶ）を提案した[1]。本方式では対訳コーパスを直接利用することにより、人手による翻訳ルール、パターンの記述を不要とする。翻訳システム開発のコスト削減が可能となり、タスク移植性、多言語適用性の向上が期待できる。

旅行会話タスクにおける日英翻訳へ当方式の適用実験を行ない高い翻訳品質を得た。翻訳一対比較法[2]によるとTOEIC 750 点の日本人と同等の翻訳品質のレベルである。一方、入力文が長くなったときの翻訳品質の劣化が大きいという短所が明確になった。今回この短所への対応として入力文を分割して翻訳する方式を提案し有効性を検証する実験を行なった。

2. D^3の概要

D^3 は、対訳コーパス、対訳辞書、類語辞書を用いて翻訳を行なう。入力文と類似する用例を基に次の手順により訳文を生成する。

(1) 用例検索

入力文と距離が最小の原文を持つ用例を対訳コーパスから検索する。与えられた閾値未満の距離の用例が存在しなければ検索および翻訳処理は失敗（FAIL）となる。文同士の距離は次の式で定義している。

$$\text{距離} = (2 \times \text{置換単語間距離} + \text{削除挿入単語数}) \,/\, (\text{入力文長} + \text{用例原文長})$$

ここで単語間距離は類語辞書に基づき 0 と 1 の間の値をとる。文長は文の単語数を意味する。

(2) 翻訳パターン生成

用例原文の入力文と異なる箇所を変項とし、変項と用例訳文中の単語、変項と入力文中の単語の対応をとり翻訳パターンを生成する。対応を決める際に対訳辞書、類語辞書を参照する。

(3) 翻訳パターン選択

同じ距離のパターンが複数ある場合は一つ選択する。選択基準には変項の対応率、用例数、変項中の単語の出現頻度を用いる。

(4) 訳文生成

用例訳文中の変項対応箇所を対応する入力文中の単語の訳語で置き換える。この際対訳辞書を利用する。

3. D^3の課題 －文長と翻訳品質－

旅行会話に関する基本文集[4]をテスト文と学習文に分けて使用し実験を行った。学習文つまり用例に使用した文数は 152,172、テスト文は 1,524 文である。テスト文の平均文長は 6.5 語である。対訳辞書には旅行会話用に作成した辞書[3]、類語辞書には角川類語辞書[5]を用いる。翻訳結果に対して日本語のできる英語ネイティブにより 4 段階の評価を与える[3]。評価レベルを次に示す。

A: 問題なし
B: 主要な情報が容易に復元できる
C: 主要な情報がなんとか復元できる
D: 主要な情報が復元できない

また用例検索に失敗し翻訳結果の得られない場合は FAIL と表記する。用例検索の距離の閾値は 1/3 とする。

図 1 に実験結果を示す。横軸は文長、縦軸は評価レベル別の文数である。文が長くなると翻訳品質が悪くなることが分かる。文長が 10 以上になると特に劣化が大きい。

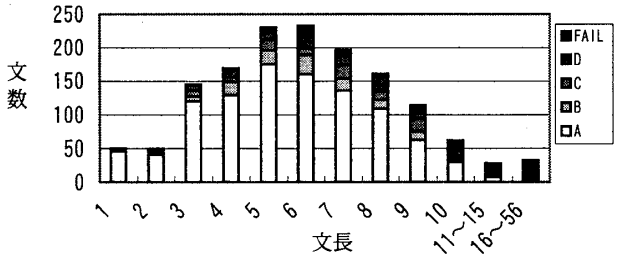

図 1．文長と翻訳品質

4. 分割翻訳の枠組み

D^3 において翻訳結果が得られないのは、入力文との距離が閾値未満の類似用例が見つからない場合である。特に入力文が長くなれば類似用例の存在しない可能性が大きくなる。距離と翻訳品質の間には明確な相関関係が存在する[1]ため、むやみに閾値を大きくするのは適当ではない。一方、入力文から部分を切り出せば類似用例が見つかり翻訳が成功する可能性がある。翻訳結果の得られない入力文への対策として、入力文を分割して翻訳する分割翻訳を提案する。

4.1 分割翻訳

入力文を複数の部分に分割する、各々の部分の翻訳結果を連結して入力文全体の翻訳結果とする。分割の判定には次の基準を使う。

① 翻訳結果の得られない部分の単語数が小さい方が良い
② ①が同じならば、分割数が小さい方が良い
③ ②が同じならば、部分-類似用例間距離の合計が小さい方が良い

このように分割方法は、分割部分の翻訳可能性に注目し、単語の連接情報や文法的知識は仮定していない。この基準による最良の分割翻訳を如何に見つけ出すかは探索問題となるが、高速化など実装手法についてここでは議論しない。

† ATR 音声言語コミュニケーション研究所
ATR Spoken Language Translation Research Laboratories

4.2 分割翻訳の例

分割翻訳の例を示す。この例では入力文は2分割され、各部分の翻訳結果を連結した結果が出力される。

［入力文］

"はい合計百九十五ドルになりますカードでお支払いですね"

［分割結果］

"はい合計百九十五ドルになります／カードでお支払いですね"

［翻訳結果］

"it's one hundred ninety five dollars in total, will you be paying with your credit card"

5. 分割翻訳実験

次に平均文長の長いテストデータを用いて分割翻訳実験を行なった。旅行会話に関するバイリンガル模擬会話の言語データベース[4]を実験対象として選んだ。テスト文数は330であり平均文長は11.4語である。テスト文を入力とし分割を行わない従来の翻訳、今回提案した分割翻訳をそれぞれ実行し結果を分析した。

表1．翻訳結果の品質

評価	分割無	分割有
A（A）	93（28.2%）	97（29.4%）
B（AB）	40（40.3%）	56（46.4%）
C（ABC）	32（50.0%）	66（66.4%）
D	38	111
FAIL	127	0

表1では評価レベル毎の文数、A、AB、ABCの割合を分割翻訳無／有の場合で示している。分割翻訳無しに比べ有りでは翻訳成功率（ABCの割合）が16.4%向上する。ABの割合、Aの割合もそれぞれ6.1 %、1.2%向上する。分割無しで翻訳出力の得られなかった127文で見ると翻訳成功率は42.5%である。

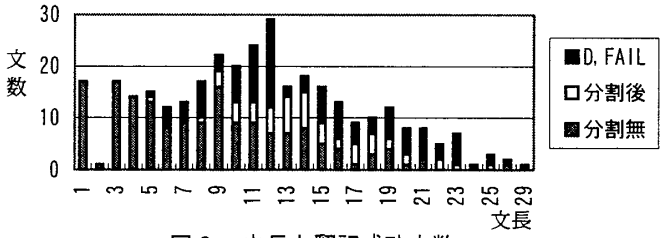

図2．文長と翻訳成功文数

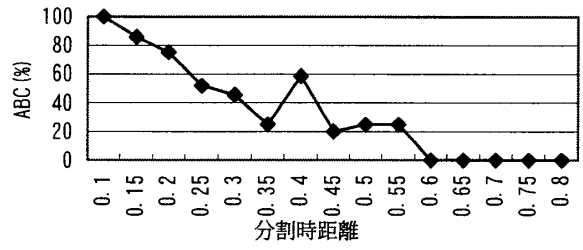

図3．分割時距離と翻訳成功率

図2は文長別の翻訳成功文数を示す。分割無しの翻訳で評価レベルABCとなった文数、分割無しではFAILだが分割翻訳によりABCとなった文数、分割翻訳の結果でもDまたはFAILとなった文数をグラフ中で区別している。分割翻訳によりABCとなった文が分割翻訳の効果である。特に文長9以上で効果が見られる。

図3は分割時距離と翻訳成功率との関係を示す。ここで分割時距離を次の式で定義する。入力文と用例との距離を分割翻訳の場合に一般化している。

分割時距離 ＝ Σ（部分長 × 部分と類似用例との距離）／入力文長

類似用例の得られない部分では類似用例との距離は1とする。分割時距離と翻訳品質との相関は明瞭であり距離が大きくなると品質は悪くなる。

6. まとめ

用例ベース翻訳において入力文の分割翻訳方式を提案し、その有効性と課題を確認した。長い文に弱いというD^3の短所の改善が見られ、全体として翻訳成功率が向上した。一方で翻訳結果の信頼性の観点からは、悪い訳（評価D）を出すよりFAILの方が良いとの考え方もある。その点に関しては、分割翻訳用に一般化した距離と品質に相関が確認できたので、求められる信頼性に応じて分割翻訳の閾値を設けるなどの方策が考えられる。

実験結果の翻訳誤りの原因には分割翻訳の問題と分割翻訳に限らないD^3の問題とがある。前者には、分割の特徴と翻訳品質の関連データの収集と分析、区切箇所を判断する別の指標☆の導入などが考えられる。後者には今後様々な改良を行なう予定である。

謝辞 本研究は通信・放送機構の研究委託により実施したものである。

参考文献

[1] Sumita, E. 2001 Example-based machine translation using DP-matching between word sequences, Proc. of DDMT workshop of 39th ACL

[2] 菅谷史昭ら,2001.音声翻訳システムと人間との比較による音声翻訳能力評価手法の提案と比較実験, 電子情報通信学会論文誌 D-II Vol. J84-D-II No.11

[3] Sumita, E. et al. 1999 Solutions to Problems Inherent in Spoken-language Translation: The ATR-MATRIX Approach, Proc. of MT Summit VII

[4] Takezawa, T. et al. 2002. Toward a Broad-coverage Bilingual Corpus for Speech Translation of Travel Conversations in the Real World, Proc. of LREC-2002

[5] 大野晋, 浜西正人, 1984, 類語新辞典, 角川書店

[6] Berger, A.L. et al. 1996. A Maximum Entropy Approach to Natural Language Processing, Association for Computatial Linguistics

[7] 竹沢寿幸ら, 1999. 発話単位の分割または接合による言語処理単位への変換手法, 自然言語処理 Vol.6 No. 2

[8] 中嶋秀治ら, 2001. 音声認識過程での発話分割のための統計的言語モデル, 情報処理学会論文誌 Vol.42 No.11

[9] 金淵培ら,1994. 日英機械翻訳のための日本語長文自動短文分割と主語の補完, 情報処理学会論文誌 Vol.35 No.6

☆ [6][7][8][9]など文分割に関する研究では語の連接の特徴から区切箇所の指標を求めることが多い。