多段階交叉位置決定手法を用いて生成される翻訳例の評価 *

5 J − 7

工藤 晃一　荒木 健治　桃内 佳雄　栃内 香次†

北海学園大学 ††　北海道大学 ‡

1 はじめに

我々は、現在、学習型の機械翻訳システムの性能の向上を目指すために、遺伝的アルゴリズムを用いた実例からの帰納的学習による機械翻訳手法（GA−ILMT）の開発、研究を行なっている[1]。GA−ILMTの有効性は既に確認されている。しかし、学習部で生成される新翻訳例の個数と精度が十分ではないため、翻訳の精度と品質に問題が残されている。そこで、我々は、この問題を解決するために、新翻訳例の生成の精度を向上させる多段階交叉位置決定手法[2]を開発し、従来よりもこの精度を向上させることに成功した。本稿では、本手法で生成される新翻訳例の評価結果及び、本手法の今後の展開について述べる。

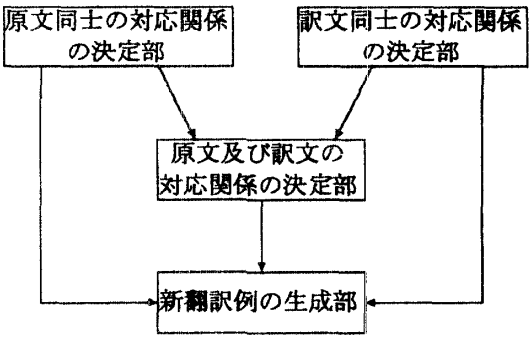

図 1: システム構成

2 概要

本手法は、対象となる２つの文の単語の対応関係を確実性の高い順に用いて、遺伝的アルゴリズムの適用における交叉位置を決定する手法である。本手法で使用される情報は、確実性の高い対応関係を決定するために、字面、読み又は英単語の原形、訳語、上位概念、品詞の順に多段階で用いられる。また、上位の段階で決定した対応関係は、下位の段階では処理対象としない。

*Evaluation for Generation of New Translation Examples using Multi-Stage Decision Method

†Koichi Kudo, Kenji Araki, Yoshio Momouchi, Koji Tochinai

††Hokkai-Gakuen University

‡Hokkaido University

3 処理過程

本手法に基づいたシステムは、図１に示されるように構成される。入力文は、原文が英文であり、訳文は日本語文である。各処理部について説明する。

3.1 原文同士の対応関係の決定部と訳文同士の対応関係の決定部

原文同士及び訳文同士の単語の対応関係の決定部では、次のように多段階で対応関係を決定する。

(1) 出現位置が同じで字面が一致する単語の対応関係を決定。

(2) 出現位置が異なり字面が一致する単語の対応関係を決定。

(3)
- 原文では、原形が同じ単語の対応関係を決定。(Brill Tagger)
- 訳文では、読みあるいは表記が一致する単語の対応関係を決定。(帰納的学習を用いた形態素解析)

(4) 同一の単語の訳語として存在する単語の対応関係を決定。(英和・和英電算辞書)

(5) 上位概念が一致する単語の対応関係を決定。(Word Net, 分類語彙表)

(6) 決定済みの対応関係に挟まれている一語の対応関係を決定。

(7) 品詞の一致する単語の対応関係を決定。(Brill Tagger)

以上の７段階で対応関係を決定する。番号が若いほど確実性の高い対応関係である。括弧内は対応関係の決定に使用した辞書とツールを示す。上位の段階で決定された対応関係は下位の段階では、処理対象としない。

3.2 原文・訳文の対応関係の決定部

原文の単語と訳文の単語の対応関係の決定には、英和辞書[3]を使用する。この辞書から訳文の単語を検索し、対応する原文の単語を探す。

3.3 新翻訳例生成部

新翻訳例の生成は、対応関係が決定された単語を交叉位置として翻訳例に対して一点交叉を適用して行う。

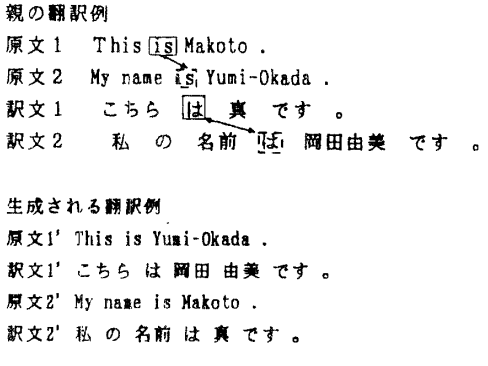

図 2: 翻訳例の対応関係の例

図２では、原文同士、訳文同士の対応は、3.1節の (2) の段階で決まる。また、原文と訳文の対応関係の決定において"is"と"は"の対応関係が決定する。原文において対応関係にある単語と訳文において対応関係がある単語が、各翻訳例において原文と訳文における対応関係が一致する場合、これらの単語を交叉位置として一点交叉を行う。例では、原文１と原文２では、"is"、訳文１と訳文２では、"は"を交叉位置として一点交叉が行われる。そして、図２のような新翻訳例が生成される。対応関係が存在する単語の対は、全て交叉位置とし、新翻訳例を生成する。

4　新翻訳例の評価

新翻訳例は、中学校１年の教科書 [4] [5] の翻訳例を使用して実験した実験結果を用いる。その評価結果は、表１に示す。表１における、生成される新翻訳例の評価方法は、次のようになっている。（原文が英文であり、訳文は日本語文である。）

正: 訳文が、原文の正しい訳文である。
誤 (1): 原文が文法的に誤り、訳文は正しい。
誤 (2): 訳文が文法的に誤り、原文は正しい。
誤 (3): 原文と訳文は文法的に正しいが、訳文が原文の訳文として誤っている。
誤 (4): 原文、訳文が共に文法的に誤っている。

5　考察

本手法が、従来の手法よりも新翻訳例の精度及び生成個数において有効であることは、既に確認されている [2]。しかし、誤った翻訳例が多く生成される。誤った翻訳例を評価するために、誤った翻訳例を４つに分類する。表１より、誤 (2) の誤りがもっとも多いことがわかる。英文では、文法的な語順、単語の単数複数形、動詞の活用形の規則が厳しい。したがって、新翻訳例が交叉して生成されるとき、新翻訳例の英文の語順の文法的な誤り、また、単語の活用の正しくない部分が交換などが原因となり新翻訳例の英文の誤りが多くなった考えられる。これに対して日本語文は、英文よりも語順や単語の活用に対して規則が緩いため、新翻訳例の訳文の誤りが少ないと考えられる。また、誤 (3) の発生の原因は、交叉時に交換される単語が日本語文と英文で一致しないためであると考えられる。全体の割合で見ると誤っている翻訳例の数が多い。正しい翻訳例は、全体の 36.4%とまだ低い。GA-ILMT では、正しい翻訳例の割合が多いほど翻訳性能が向上するので、さらに誤った翻訳例を減少させ、正しい翻訳例を増加させる必要があると考えられる。

表 1: 本手法の新翻訳例の評価

	正	誤 (1)	誤 (2)	誤 (3)	誤 (4)
個数 (個)	2703	215	2356	1748	402
全体の割合 (%)	36.4	2.9	31.8	23.7	5.4

6　おわりに

今後は、今回の誤った翻訳例の評価を生かし、正しい翻訳例の生成の精度を向上させる改良を加え、本手法の性能向上を目指していく。また、本手法をＧＡ－ＩＬＭＴに組み込んで実験を行い、実際の翻訳における本手法の有効性について確認する予定である。

参考文献

[1] 越前谷博, 荒木健治, 桃内佳雄, 栃内香次: 遺伝的アルゴリズムを用いた実例からの帰納的学習による機械翻訳手法, 情報処理学会論文誌, Vol.2, No.8, pp.1565-1579, (1996).

[2] 工藤晃一, 荒木健治, 桃内佳雄, 栃内香次 : 多段階交叉位置決定手法を用いた新翻訳例の生成, 言語処理学会第３回年次大会発表論文集, pp.589-592, (1997).

[3] 久保正治 : 英和・和英電策辞典 gene, 技術評論社、東京, (1995).

[4] 教科書ガイド教育出版ワンワールド, 日本教材, 東京, (1991).

[5] 教科書ガイド東京書籍版ニューホライズン, あすとろ出版, 東京 (1991).