知識を導入したGA-ILMTの有効性の確認

工藤晃一[†]・荒木健治[†]・桃内佳雄[†]・栃内香次[††]

[†]北海学園大学工学部　[††]北海道大学工学部

1 はじめに

近年、様々な機械翻訳システムが考案され、また、実用化されているが、依然として翻訳の精度及び品質に問題が残されている。そこで、我々は、学習型の機械翻訳手法の実用化のために、遺伝的アルゴリズムを用いた実例からの帰納的機械翻訳手法 (GA-ILMT) を開発し、性能評価実験によりこの機械翻訳手法の有効性を確認している [1]。しかし、GA-ILMT の有効性は確認されてはいるが、依然として翻訳精度に問題があり、実用には至っていない。これは、与えられた翻訳例から遺伝的アルゴリズムの過程の一つである交叉によって生成される新翻訳例の精度が低いためである。遺伝的アルゴリズムの適用により、冗長性の高い翻訳例を生成してしまうことが原因である。そのため、生成される誤った新翻訳例を減少させ、同時に生成される正しい新翻訳例を増加させて、生成される新翻訳例の精度を向上させることが必要である。

そこで、我々は交叉位置の決定に知識を導入し、生成される新翻訳例の精度を向上させる多段階交叉位置決定手法を開発した。本手法は、様々な知識を用いて単語の対応関係を決定し、その結果から交叉位置を決定して新翻訳例を生成する手法である。遺伝的アルゴリズムに対する知識の導入により、その多様性に制限を加えることになるが、新翻訳例の精度を向上させることが可能であり、その有効性は確認されている [2]。本稿では、多段階交叉位置決定手法 GA-ILMT システムを使用し、以前より入力文を増加させた評価実験と考察結果に基づき、知識を導入した GA-ILMT の実用性の向上と有効性を確認する。

2 GA-ILMT の処理過程

2.1 処理過程

GA-ILMT の基本的な処理過程を図 1 に示す。() 内は、各処理部の遺伝的アルゴリズムの適用過程を示している。また、英文とその日本語訳文を組とした翻訳例を染色体に対応させ、またその染色体を翻訳例を構成している遺伝子に対応させる。

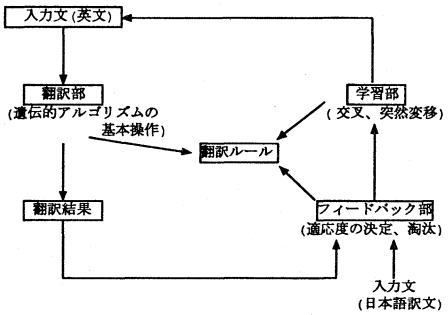

図 1: GA-ILMT の処理過程

本システムは、英日機械翻訳システムである。まず、原文として、英文を入力すると、翻訳部において、獲得されている翻訳ルールに対する遺伝的アルゴリズムの基本操作を用いて翻訳結果を生成する。次いで、フィードバック部において、ユーザーが与えた正しい日本語訳文を用い、翻訳部で用いられた翻訳ルールに対する適応度の決定と淘汰を行う。そして、学習部において、与えらた翻訳例 (英文とその日本語訳文) に対して、交叉と突然変異を行い多様な新翻訳例を生成し、また、与えられた翻訳例及び生成された新翻訳例を用いて翻訳ルールを抽出する。今回、生成される新翻訳例の精度を向上させるために知識を導入するのは、学習部の新翻訳例の生成の部分である。

2.2 GA-ILMT の新翻訳例の生成

従来の GA-ILMT における新翻訳例の生成について述べる。図 2 にその例を示す。従来の手法 (多段階交差位置決定手法を適用しない GA-ILMT) による新翻訳例の生成では、それまでに入力された翻訳例から英文

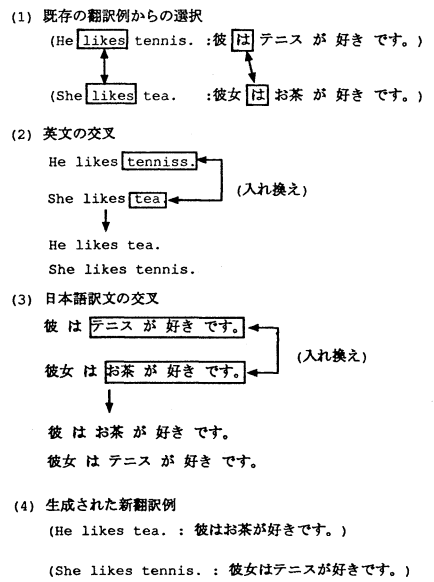

図 2: 従来の新翻訳例の生成例

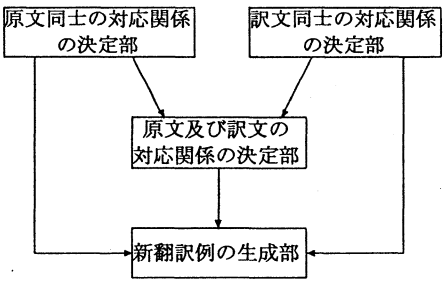

図 3: システム構成

と日本語訳文のそれぞれにおいて、字面が完全に一致する文字列を共通部分に持つ 2 つの翻訳例を選択し、共通部分を交叉位置として一点交叉を行う。図 2 に示す例では、英文において「likes」、日本語訳文においては、「は」と「が好きです」が共通部分として抽出される。したがって、英文においては「likes」、日本語訳文においては「は」が交叉位置として抽出される。また、「が好きです」を交叉位置とすると同じ日本語訳文が生成されるので翻訳ルールとしての登録は行わない。

3 多段階交叉位置決定手法

GA-ILMT の学習部に適用される本手法は、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決定し、その対応関係から交叉位置を決定する手法である。対応関係の決定には、形態素情報や英和辞書、シソーラス辞書等を使用して行なう。また、本手法では、英文と日本語訳文の両方で品詞情報を使用するため、入力文は、形態素解析を行っておく必要がある。原文は、Bril Tagger[4] を、訳文には、我々が開発した帰納的学習による形態素解析手法 [5] を用いる。多段階交叉位置決定手法の処理過程を図 3 に示す。

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、以下の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。
- 訳文では、読みが一致する単語の対応関係を決定。

(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。

(5) 上位概念が一致する単語の対応関係を決定。

(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。

以上の 7 段階で対応関係を決定する。また、上位の段階で決定された対応関係は、以下の段階では処理対象としない。

3.2 原文と訳文同士の対応関係の決定

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 [6] 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 新翻訳例の生成部

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。交叉位置となる単語は、原文における対応関係または、訳文における対応関係が存在し、かつ、原文と訳文における対応関係が存在する単語である。図 4 に、多段階交叉位置決定手法による新翻訳例の生成の例を示す。最初に原文同士の対応関係の決定部では、原文 1 の"sister"と原文 2 の"brother"が上位概念の一致により対応関係が決定する。次に訳文同士の対応関係の決定部は、訳

表 1: 翻訳結果の総計

翻訳手法	正翻訳（文）	誤翻訳（文）	精度%
GA-ILMT	134	365	26.9
本手法を適用した GA-ILMT	162	337	32.5

(全翻訳文数 499 個)

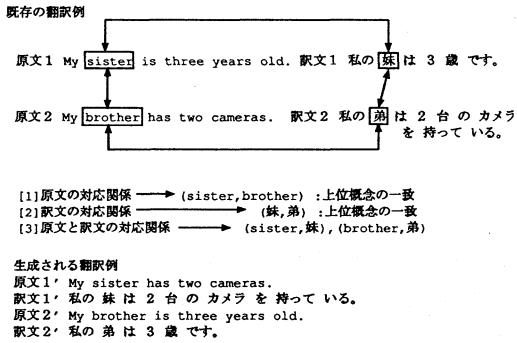

図 4: 多段階交叉位置決定手法による新翻訳例の生成

文 1 の"妹"と訳文 2 の"弟"もまた、上位概念の一致により対応関係が決定する。そして、原文と訳文同士の対応関係の決定部は、原文 1 の"sister"と訳文 1 の"妹"の対応関係を決定し、また、原文 2 の"brother"と訳文 2 の"弟"の対応関係を決定する。そして、これらの単語の位置が交叉位置となり一点交叉が行われ新翻訳例が生成される。

4 評価実験

4.1 実験方法

実験には、中学校 1 年生用教科書ガイド・ワンワールド [7] に掲載されている英文と訳文の 802 組を使用した。実験では最初の 303 文を学習させ、残りの 499 文を学習させながら翻訳を行なった。翻訳結果は、正翻訳と誤翻訳の 2 つ評価方法で分類する。正翻訳は、使用している利用者が正しい翻訳であると判断できる翻訳結果である。誤翻訳は、利用者が正しい翻訳であると判断できない翻訳結果である。また、翻訳不能の場合も誤翻訳に含む。正翻訳率は以下の式 (1) で決定される。

$$\text{正翻訳率 (\%)} = \frac{\text{正翻訳数}}{\text{全翻訳数}} \times 100 \quad (1)$$

4.2 実験結果

実験結果より、多段階交叉位置決定手法を適用した GA-ILMT の正翻訳率は、32.5%となった。多段階交叉位置決定手法を適用しなかった従来の GA-ILMT の正翻訳率は、26.9%である。したがって、本手法を組み込むことにより正翻訳率は、5.6 ポイント増加した。また、辞書中の総翻訳ルール数が 69848 個から 54024 個に減少した。

5 考察

5.1 本手法の有効性について

表 1 より、GA-ILMT に多段階交叉位置決定手法を適用することにより正翻訳例の精度は、5.6 ポイント向上した。本手法を適用した結果、誤翻訳から正翻訳へ移行したものが 31 文、正翻訳から誤翻訳に移行したものが 7 文である。結果として、正翻訳が 28 文増加したことに相当する。これは交叉位置の決定に知識を導入することにより、従来よりも精度の高い新翻訳例が生成され、これに伴い新翻訳例から抽出される翻訳ルールの精度が向上するため、正翻訳数が増加したものと考えられる。図 5 は、多段階交叉位置決定手法の適用により翻訳可能となった例である。この例では、従来では交叉位置とすることができなかった位置で交叉が行なわれ、正しい新翻訳例が 2 個生成される。交叉後、翻訳ルールの抽出が行なわれ、単語の翻訳ルール (she:彼女) と (he:彼) と変数つき翻訳ルール (Is @0 your classmate? :@0 はあなたのクラスメイトですか?) が獲得される。この変数付き翻訳ルールが"Is Kayo your classmate?"の翻訳に使用され正翻訳が導出される。従来の手法では、この入力文の翻訳までにこの変数つき翻訳ルールは獲得することができないため、この入力文の翻訳は不可能であった。このように新翻訳例の精度向上により、翻訳ルールの精度も向上し良質な翻訳ルールを多く獲得することができるのである。以上より本手法の適用の有効性を確認した。

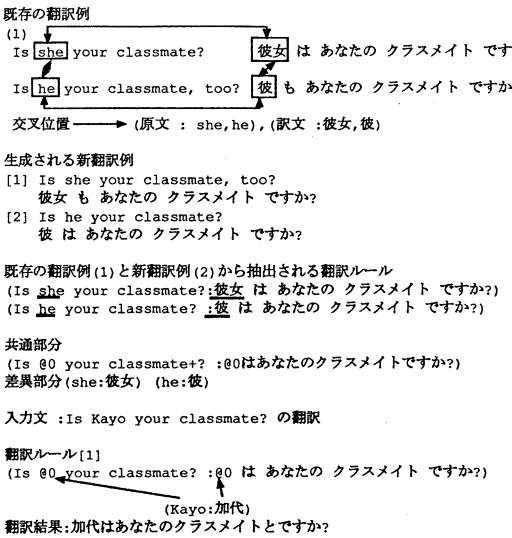

図 5: 本手法により翻訳可能となった翻訳の例

5.2 GA-ILMT への知識の導入

本手法の組み込みにより、GA-ILMT の正翻訳率は向上した。しかし、本手法を組み込むことにより 7 文が翻訳不可能または、誤翻訳となった。この原因は、生成される新翻訳例に存在すると考えられる。従来の GA-ILMT で生成可能で本手法を適用した GA-ILMT では生成不可能な正しい新翻訳例も存在するためである。これは、従来の原文と訳文における共通な字面による共通部分の全ての組み合わせを交叉位置としたランダムな新翻訳例の生成が本手法ではできなくなった影響である。したがって、従来の方法で生成可能であった新翻訳例が生成不可能となることに伴い、その新翻訳例から抽出されていた翻訳ルールも抽出不可能となり、この影響により 10 文が翻訳できなくなったと考えられる。これは、知識の導入による遺伝的アルゴリズムの多様性の制限によって翻訳ルールの多様性が失われた結果であると考えられる。しかし、多様性の制限により辞書中の翻訳ルール数が 69,848 個から 54,024 個に約 16,000 個減少させることから、翻訳率を向上させつつも膨大な辞書中の翻訳ルール数を減少させる効果がある。

6 おわりに

本稿では、知識を GA-ILMT に導入することの有効性を性能評価実験の結果より述べた。また、本手法の適用は遺伝的アルゴリズムの多様性に制限を課すことになり、その新翻訳例の多様性が失われることも確認された。しかし、多様性の損失の多くは誤った新翻訳例の多様性の損失であり、知識の導入による今まで生成不可能であった正しい新翻訳例の生成によって正しい新翻訳例の多様性の損失を補い、かつ、正しい新翻訳例の多様性を向上させることが可能であると考えられる。したがって、知識の導入により今まで問題であった新翻訳例の生成精度の低さによる正翻訳率の低下が解消される利点があり、GA-ILMT を実用的な機械翻訳システムにするためには、知識の導入が不可欠であることが確認された。

現段階では、依然として正翻訳率が不十分なので、多段階交叉位置決定手法の性能をさらに向上させ、実用的な学習型機械翻訳システムの実現に向けて研究を進める予定である。

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] K,Kudo.,K,Araki.,Y,Mmouchi and K.Tochinai: Multi-Stage Decision Method for Production of New Translation Examples, Proceedings of the IASTED International Conference Arti fi-cial Intelligence and Soft Computing, pp.125-128 (1997).

[3] 荒木健治, 栃内香次:多段階共通パターン抽出法を用いた翻訳例からの帰納的学習による翻訳, 情報処理北海道シンポジウム'91,pp.47-49(1991).

[4] Eric Brill,A CORPUS-BASED APPROACH TO LANGUAGE LEARNING,(1993).

[5] 荒木健治, 栃内香次：帰納的学習による語の獲得および確実性を用いたごの認識, 電子情報通信学会論文誌,Vol.J75-D-II,No.7,pp.1213-1221,(1992).

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

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