多段階交叉位置決定手法を用いた新翻訳例の生成

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、コンピューターネットワークの急速な発達により、一般の人たちにも日本語以外の言語にふれる機会が大幅に増えてきた。そのため、翻訳を大量に、かつ迅速に行うために計算機を使って翻訳を行う機械翻訳システムが考案され、実用化されている。しかし、現存する機械翻訳システムは、翻訳の精度及び品質に問題があり実用性も十分とは言えない。

そこで、我々は学習型の機械翻訳システムの精度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）の研究を続けている[1]。ＧＡ－ＩＬＭＴは、翻訳例から翻訳ルールを抽出して学習し、その翻訳ルールに遺伝的アルゴリズムを適用して翻訳を行なう手法である。また、この手法では、与えられた翻訳例に対して遺伝的アルゴリズムを適用し、新翻訳例の生成を自動的に行なう。新翻訳例の生成過程では、字面情報による共通部分を持つ２つの翻訳例が選択され、共通部分を交叉位置として決定し、次いで、これらの２つの翻訳例に対して一点交叉が行われて新翻訳例が生成される。それゆえ、与えられた翻訳例より多くの翻訳例を学習させることが可能となる。

しかし、ＧＡ－ＩＬＭＴの有効性は既に確認されているが、依然として精度、品質に問題が残されている。これは、学習部で生成される例文が依然不足し、その精度が十分ではないためである。そこで、我々は生成される新翻訳例の精度の向上を目指すために、”多段階交叉位置決定手法”を提案する。本手法を新翻訳例の生成で適用し、ＧＡ－ＩＬＭＴにおける問題解決を試みた。本稿では、本手法の概要とその有効性を確認するために行った実験結果について述べる。

2. 処理過程

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、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、シソーラス、品詞の順に各段階で用いられる。また、上位の段階で決定した対応関係は、処理対象からはずれる。従って、上位で決定される対応関係ほど、確実性が高い。本手法に基づいたシステムでは、２つの翻訳例の原文と訳文における全ての確実性の高い対応関係を決定するために、２つの翻訳例の原文同士及び訳文同士の単語の対応を決定し、次いで、これらの情報を用いて、２つの翻訳例における原文と訳文の単語の対応関係を決定する。そして、対応関係の結果を用いて一点交叉を行い、２つの新翻訳例を生成する。従って、本システムは、図１に示されるように構成される。処理の対象となる翻訳例は、原文が英文、訳文が日本文である。本システムで使用する前に入力文は形態素解析を行っておく。原文は、tagger[2]を、訳文には、帰納的学習による形態素解析手法[3]を用いる。

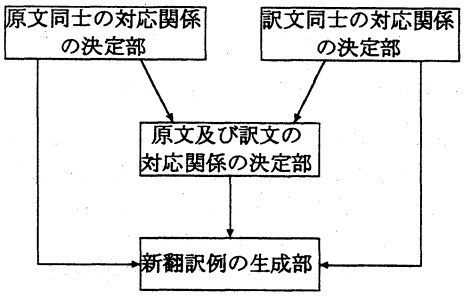

図 1: システム構成

2.1 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、読みあるいは表記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。つぎに、各段階で決定した対応関係によって決まる対象になった文同士の類似度の得点について述べる。この得点は、各段階における対応関係が存在した個数で決まり、次の式で計算される。

$$alpha = \frac{100.0}{\text{２つの翻訳例の原文又は訳文の単語の合計}}$$

$$\text{得点} = alpha \times (2.0 \times N1 + 1.8 \times N2 + 1.6 \times N3 + 1.2 \times N4 + 0.8. \times N5 + 0.4 \times N6 + 0.2 \times N7 - 0.5 \times N8)$$

N1:（１）の段階の対応関係数　*N2*:（２）の段階の対応関係数
N3:（３）の段階の対応関係数　*N4*:（４）の段階の対応関係数
N5:（５）の段階の対応関係数　*N6*:（６）の段階の対応関係数
N6:（７）の段階の対応関係数　*N8*:対応関係がなかった単語数

2.2 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[4]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。

2.3 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。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必ずある。そして、原文と訳文における対応関係が存在する。

図２では、原文同士、訳文同士の対応は、（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"が交叉位置として一点交叉が行われる。そして、以下のような翻訳例が生成される。

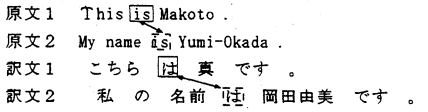

図 2: 翻訳例の対応関係の例

例)

原文 1' This is Yumi-Okada . 訳文 1' こちら は 岡田 由美 です 。

原文 2' My name is Makoto . 訳文 2' 私 の 名前 は 真 です 。

条件を満たす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。

3. 評価実験

3.1 実験方法

本システムに入力する翻訳例は、中学１年用教科書ガイド・ワンワールド[5]に掲載されている英文とその日本語訳文の９３４組と中学１年生用教科書ガイド・ニューホライズン[6]に掲載されている英文とその日本訳語文の８００組を使用する。入力は、２つの教科書の同じＬｅｓｓｏｎごとに行う。例えば、ワンワールドのＬｅｓｓｏｎ１とニューホライズンのＬｅｓｓｏｎ １を入力する。 そして、２つの翻訳例の対応関係の決定と新翻訳例の生成は、各Ｌｅｓｓｏｎ ごとにワンワールドの翻訳例とニューホライズンの翻訳例との全ての組み合わせについて実験を行う。 今回は、原文対応の得点と訳文対応の得点がそれぞれ２５点以上の翻訳例の対のみ翻訳例の生成を行う。 点数制限をつける理由は、ある程度対応関係存在する翻訳例の対を選択して翻訳例の生成を行うためである。

3.2 評価方法

本システムで生成された翻訳例の精度と正誤別の生成個数について調べる。同様に、ＧＡ－ＩＬＭＴの手法で生成された翻訳例についても行う。また、精度は、次の式で計算される。

$$\text{精度\%} = \frac{\text{正しい新翻訳例数}}{\text{総新翻訳例数}}$$

3.3 実験結果

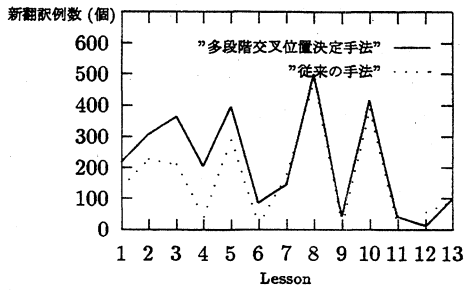

図 3: 正しい新翻訳例数の推移

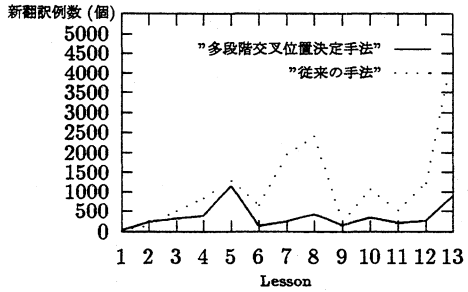

図 4: 誤った新翻訳例数の推移

実験結果は、表１、図３、４、５に示す。従来の手法で生成された新翻訳例の正誤別の合計は、正しい新翻訳例の合計が、2146 個、誤った翻訳例の合計が 5332 個である。この結果に対して多段階交差位置決定手法では、正しい新翻訳例合計が、689 個増加し 2835 個となり、その増加率は 32.1、誤った翻訳例の合計が 617 個減少し、4715 個となった。そして、従来の手法で生成した新翻訳例の全体の精度は、12.3%。これに対して、多段階交叉位置決定手法で生成した新翻訳例の精度は、25.2 ポイント向上して 37.5%となった。

4. 考察

4.1 新翻訳例の精度と生成個数について

表１及び図３、４、５より、従来の手法を使用した翻訳例の生成と比較すると、本手法を使用して生成された正しい新翻訳例数は増加し、誤った翻訳例数は減少して、全体の精度が増加している。これは、新翻訳例の生成時に対応関係の決定における類似度の得点にある程度の制限をつけているため、類似性がある程度高い翻訳例を交叉して新翻訳例を生成するため誤った翻訳例が減

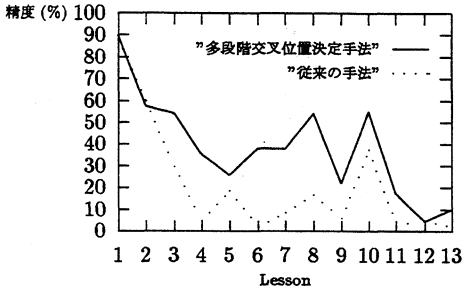

図 5: 新翻訳例の精度の推移

表 1: 新翻訳例の正誤別の総計と精度の表

使用した手法名	正 (個)	誤 (個)	精度%
従来の手法	2146	5332	12.3
多段階交叉位置決定手法	2835	4715	37.5

少し、正しい新翻訳例が増加して、精度が上昇したと考られる。また、翻訳例の訳文と原文において交叉位置となるそれぞれの単語の対応関係が決定されているため、従来よりも確かな対応関係がある交叉位置で交叉が行われることも精度の増加の要因の一つである。そして、従来の字面のみの対応関係による交叉位置の決定だけでなく、様々な対応関係によって交叉位置を決定することにより、従来の手法では生成できなかった正しい新翻訳例が増加したと考えられる。

4.2 生成される翻訳例について

従来の手法では、字面が一致する単語の並びを交叉位置としてきた。その例を図６に示す。上記の例では、原文では"is my"、訳文では"の"が字面で一致し交叉位置となり、交叉位置より右側にある単語の列を入れ替えて翻訳例が生成される。一方、従来では行うことのできない交叉位置の決定が、本手法では可能である。

図７の例では、原文において"am"と"is"が原形が一致して対応関係が決定し、訳文では、それぞれ"は"が、

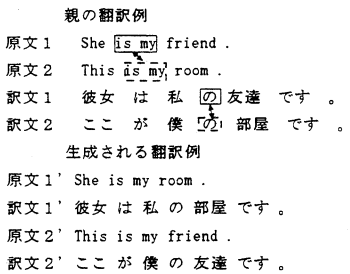

図 6: GA-ILMT における翻訳例の生成

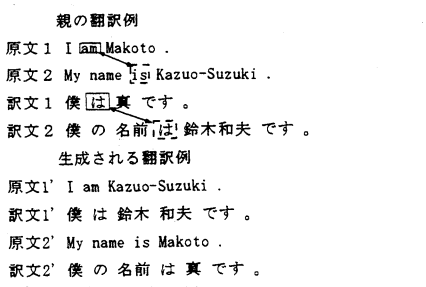

図 7: 多段階交叉位置決定手法を用いた翻訳例の生成

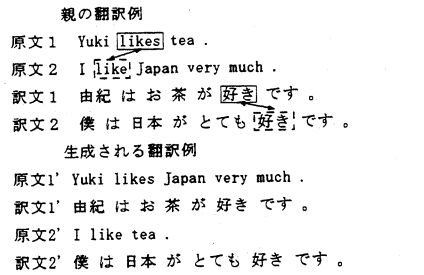

図 8: 誤った新翻訳例の生成例

位置が異なるが字面が一致するとして対応関係が決定する。そして、"am"と"is"は、"は"と対応関係が決定し、これらの単語を交叉位置として交叉して新翻訳例を生成する。従来の手法では原文に一致する字面の単語がないので翻訳例の生成を行うことができない。このように対応関係を利用して、従来の手法では生成できない翻訳例も生成可能となることが確認できる。

4.3 間違った翻訳例の生成について

図８では、原文において原文１の"likes"、原文２の"like"が、原形の一致によって対応関係が決定し、また、訳文において、訳文１、２の"好き"が同じ出現位置で字面で一致して対応関係が決定されている。そして、原文の"likes"、"like"が、原文と訳文の対応関係の決定で訳文の"好き"との対応関係が決定される。そして、これらの単語を交叉位置として交叉が行われ、誤った翻訳例が生成される。この原因は、原文と訳文の単語の文法的な語順が異なっているため、各翻訳例の原文と訳文における交換部分の対応関係に誤りが生じるからである。図８では、訳文では、　"好き"が交叉位置となり、訳文１、訳文２ともに"です。"が交換され、一方、原文では、原文１では"likes"、原文２では"like"が、交叉位置となり原文１の"Japan very much ."、原文２の"tea ."が交換される。しかし、原文、訳文の交換部分が明らかに対応していない。この問題は、従来の手法でも問題になっていたが、本手法においてもこの問題となってしまった。しかし、このようなことを防ぐには、交叉にかなり厳しい条件を加えなければならない。交叉に条件を付けるには遺伝的アルゴリズムに乗っ取ったシステムを作る上でこれは、好ましくない。このような誤った翻訳例は、淘汰によって消していくしかないと考えられる。

5. おわりに

実験結果より本手法を用いたシステムは、従来の手法よりも新翻訳例の生成の精度とその生成個数において有効であることが確認された。今回は、親の翻訳例の対の類似度を示す得点をある程度制限して実験を行った。しかし、この点数を増減させることにより、精度と生成個数が増減することが考えられる。よって、正しい新翻訳例の精度及び生成個数を増加させるための最適な値がある可能性が高い。この値を調査することが今後の課題の一つだと考えられる。また、対応関係に使用する知識や対応関係の決定に改良を加える余地があると考えられる。このように、本手法は、翻訳例の生成にある程度の有効性を示したが、精度をさらに増加させる必要があり、今後は、本システムのさらなる性能向上を目標とし、学習型機械翻訳の翻訳率向上を目指す。

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

[3] 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).

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

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

[6] 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).