

機械翻訳言い換えシステムにおける学習機能の拡張

鈴木 良生 田添 丈博

鈴鹿工業高等専門学校 専攻科 電子機械工学専攻

椎野 努

愛知工業大学 情報科学部 情報科学科

1. 背景・目的

現在、英日機械翻訳システムは、短い英文の機械翻訳にはそれなりの精度がある。しかし、長文や複雑な文では直訳に近い硬い表現や、日本語として意味を取りづらい翻訳結果となってしまうことがある。

本研究は、不自然な日本語となった機械翻訳文に対して「言い換え」を行い、より自然な日本語訳を出力するシステムの提案を行う。今回は、学習機能の拡張を行うことにより学習成功数の向上を目指す。

2. 機械翻訳言い換えシステム

2.1. システム構成

システムの構成を図1に示す。本システムは、ユーザから入力された英文を既存の英日機械翻訳システムへの入力とし、出力される機械翻訳文に対して「言い換え」を行う。これにより自然な訳文を生成する。

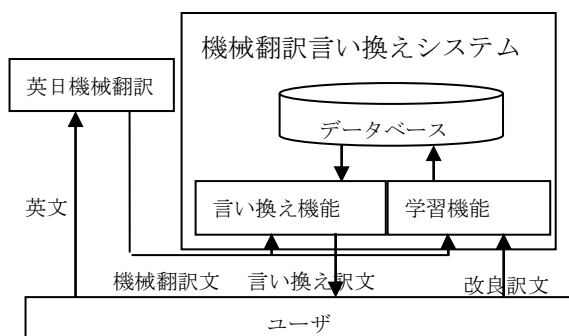


図1. システム構成

言い換えを行うためのデータは、機械翻訳文に対してユーザがより自然であると言える改良訳文を入力する。これを学習機能により学習することで、言い換えデータを生成する。この言い換えデータを用いて「言い換え」を行う。

2.2. 学習機能

言い換えデータの学習は機械翻訳文と改良訳文の2つに対し、構文解析を行った結果より文節単位で係り受け構造のマッチングを行う。

マッチング方法は、それぞれの訳文に完全一致した文節を「一致文節」として、一致文節を検索する。このとき、ある文節の一致文節が複数通り考えられる場合、どの文節同士が対応しているのかを機械的に判断することが難しい。よってこのような場合、マッチングを行わないこととする。一致文節が1対1であるならば、一致文節の係り受け関係にある文節を検索する。文節の完全一致による学習例を示す。図2に例文を示し、構文解析を行った結果を図3、登録される言い換えデータを図4に示す。

[英文] I saw an old doctor.
 [機械翻訳文] 私は古い医者に会った。
 [改良訳文] 私は年老いた医者に会った。

図2. 例文

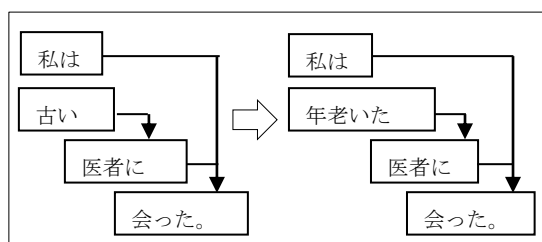


図 3.構文解析結果

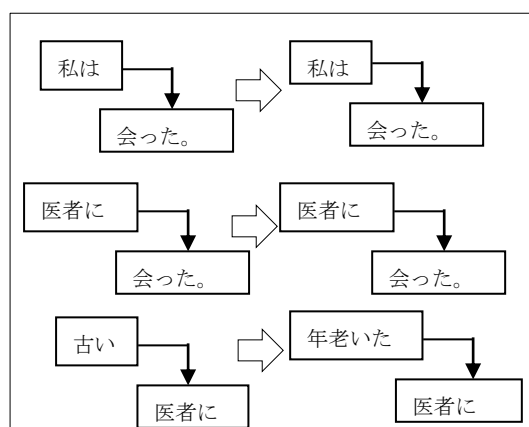


図 4.学習結果

3.学習機能の拡張

3.1 曖昧一致文節の導入

従来のマッチング方法は完全一致する文節を「一致文節」として考える方法であった。この考え方の欠点として、助詞などが多少異なるだけで一致できず、学習できなくなってしまうという問題点があった。今回、自立語のみのマッチングを行う。この時、多少文節が異なっても一致文節と考えることから、これを「曖昧一致文節」と呼ぶ。これにより学習成功数の向上を目指す。

文節のマッチングを行う際に構文解析した文節を、助詞などを除き、動詞などは原形へと変換した自立語のみへ変換する。これを用いてマッチングを行う。

自立語のみの文節への変換を、例として図5のように示す。「関連した」は、関連(名詞)+し(動詞)+た(助動詞)と分解できる。こ

れを自立語は原形に、付属語は省略とすると、関連(名詞)+する(動詞の原形)とできる。

3.2. n 対 m のマッチング

従来は2対2の文節による学習であったが、図6のように複数対複数の学習へと拡張する。

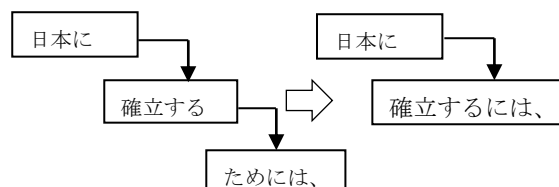


図 5. n 対 m のマッチング

n 対 m のマッチング方法では、構文解析によって得た係り受け構造より、複数の係り受け先、元を取得する。レーベンシュタイン距離を用いてこれを比較し、距離数を得る。この距離数が最小である文節の関係を学習する。このとき、レーベンシュタイン距離とは、情報理論において2つの文字列がどの程度異なっているかを示す数値のことを指す。

図6を例に示す。この時、

確立するためには、 と 確立するには、

を比較すると2文字異なることが分かる。よって、この時のレーベンシュタイン距離は2となる。また、

確立する と 確立するには、

を比較すると3文字異なることが分かる。よって、この時のレーベンシュタイン距離は3となる。よってレーベンシュタイン距離が最短である

確立するためには、 と 確立するには、

を選択する。

4.実験

4.1 実験方法

日英新聞記事対応付けデータ[2]からサンプル100文をデータとして用いて実験を行う。100文のデータによる実験においては、正解数を620と仮定する。この正解数は、10文で実験を行った際に正解数が62であったことから620と仮定するものである。

実験1 完全一致 2対2

実験2 曖昧一致 2対2

実験3 完全一致 n対m

実験4 曖昧一致 n対m

4.2 実験結果

実験結果を表1に示す。

表1. 実験結果

	学習数	学習成功数	再現率	適合率	F値
実験1	163	113	0.693	0.182	0.288
実験2	275	161	0.585	0.260	0.360
実験3	159	120	0.755	0.194	0.309
実験4	269	173	0.643	0.279	0.389

4.3 考察

今回の実験結果より、自立語のみのマッチングを行うことでF値を0.288から0.360へ向上させることができた。また、n対mの学習方法に関しても、F値を0.309、曖昧一致では0.389へと向上させることができた。

図6のような例を挙げると、従来の方法では、

- 適切な手段を->適切な措置を
- 雇用の増加を->雇用の増加の
- 促進すること->促進するために

を学習することができる。これに曖昧一致を導入することにより、

増加を-作成するように->

増加の-創出を

を学習することができるようになり、更に学習数を増やすことができた。しかし、学習数は増えたが、このように学習成功とはならないこともある。これに対し、n対mの学習を導入することにより、

増加を-作成するように->

増加の-創出を



増加を-作成するように->

増加の-創出を-図る

と正しい文節を学習することができた。

[翻訳文]4. 私達によってはこれを実行する適切な手段を同意した雇用の相当な増加を作成するように設計されている支持できる拡張を促進することに全体的な成長の作戦が取って、取る。

[改良訳文]4 我々は、雇用の大幅な増加の創出を図る持続可能な拡大を促進するために合意された、この世界的な成長戦略を実施すべく適切な措置をとりつつあり、また今後もとる決意である。

図6. 例文2

n対mの学習において、図7の例文を考えたとき、

- ① 雇用創出の-ための-機会を->
雇用創出の-機会を

という学習と

- ② 雇用創出の-ための->
雇用創出の-機会を

というレーベンシュタイン距離がどちらも3となる学習を考えた場合、今回のアルゴ

リズムでは、文節数が少ないものを学習することとしたため、②を学習した。しかし、図7の例文の場合、①を学習することが好ましい。よって、レーベンシュタイン距離の評価方法を改善する必要があると考えられる。

(<http://mastarpj.nict.go.jp/~mutiyama/jea/sample/p11-sample.txt>)

[翻訳文]私達は環境政策によって提供される雇用創出のための機会を強調する。
[改良訳文]我々は、環境政策により提供される雇用創出の機会を強調する。

図7. 例文3

5.まとめ

今回の実験から機械翻訳言い換えシステムへの曖昧一致文節の導入は有効であると考えられる。

今後の課題として、正解数に対し、学習成功数が少ないことが明らかである。よって、学習成功数を増やしつつF値を向上させる必要がある。考察にあるように、レーベンシュタイン距離が同数であるが文節数から学習を失敗してしまった例を改善するために、レーベンシュタイン距離の評価方法を改善する必要があると言える。更に学習失敗例を分析し、システムを改善する必要がある。また、今回実験データとして100文を用いたが、今後実験データ数を増加させることにより、言い換えデータ数の充実を目指すことが挙げられる。

参考文献

- [1]宮地洋太:英日機械翻訳における自然な訳文への言い換えシステム,C3-5,言語処理学会第16回年次大会(NLP2010)
- [2]日英新聞記事対応付けデータ,1対1対応の日英文