# 機械翻訳システムの安定性評価

髙橋 寛治　　竹野 峻輔　　山本 和英
長岡技術科学大学
{takahashi, takeno, yamamoto}@jnlp.org

## 1 はじめに

良い機械翻訳システムを評価する軸として、本稿ではシステムの安定性 (Stability) の評価を提案する。

近年のニューラル機械翻訳 (NMT) はめざましい性能の向上を見せており、これまでの句に基づく統計的機械翻訳 (PBMT) と比べて、非常に流暢な出力をする。一方で、NMT では PBMT ではあまり見られなかった問題が散見される。例えば、流暢ではあるが、妥当性[1] を損なった翻訳を出力することがある。また、句読点の有無などの入力の変化に対する出力の変化が大きいという課題もある[2]。統計的機械翻訳に比べてルールベース機械翻訳は、規則的な安定した出力をすると分析されている [5]。

はたして、句読点の有無で出力が大きく変化する機械翻訳システムは使いやすいだろうか。意味が等価な範囲での入力の変化が出力に大きな変化を与える場合、事前に入力文の形式を定める必要が生じるなどシステムとして取り扱いにくい。そのために、機械翻訳システムの入力の揺らぎに対する出力の安定性を評価することで、扱いやすいシステムを定量的に評価することができる。この考えは、制御理論における安定性を参考にしたものである。

本稿では機械翻訳システムの出力の安定性の評価を試みる。本稿での提案および貢献を以下に示す。

- 機械翻訳システムに対する評価尺度として安定性という概念の提案
- 入力文の変化がどの程度出力に影響を与えるかを見ることによる安定性の評価手法の提案
- 安定性評価のための評価セットの構築

[1]妥当性とは、入力文の意味を出力文でどの程度保持しているかのことである。また、流暢性とは意味に関係なく文法的に正しい文を出力しているかどうかの程度である。

[2]ニューラル機械翻訳の課題として ”Large changes in translation from small changes in source sentence.” と指摘されている。https://twitter.com/marcfede/status/793107988273696769

## 2 機械翻訳システムの安定性

### 2.1 安定性とは

安定性は制御工学において用いられるシステムの評価観点の一つである。これは、あるシステムについて入力を変化させたときの、出力の変化度合いを測定する評価指標である。

この考えに習い、機械翻訳システムにおける安定性とは入力の揺らぎに対する出力の変化の度合いとする。ここでの揺らぎとは言い換えや表記ゆれや句読点の有無、同義表現などを介し意味的な変化を伴わないが表層の単語列や文字列が変化することが考えられる。安定性の高い機械翻訳システムとは、入力が意味の変わらない範囲で変化した際に出力の変化が小さいシステムのことである。

人が機械翻訳システムを利用する際に句読点の有無などで出力が大きく変化すると扱いづらいシステムであると考える。安定性の評価とは、システムの良さを数値化する一つの手段と言える。

### 2.2 安定性の評価

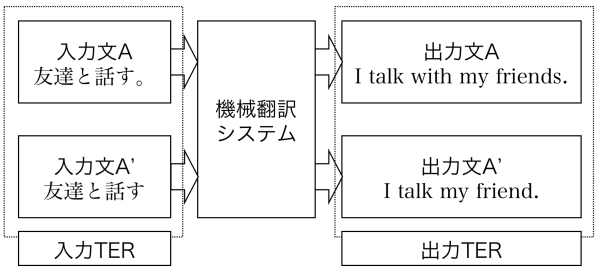


図 1: 安定性評価の基本的な考え

安定性評価の概略図を 1 に示す。考え方は非常に簡素である。入力文 A と微細な変更を加えた入力文 A ’

を機械翻訳システムに入力する。出力文 A と A ’ がシステムの出力となる。ここで、入力文 A と A ’ の変化および出力文 A と A ’ の変化を TER(翻訳編集率:Translation Edit Rate)[2] で数値化することにより、安定性を評価する。すなわち、入力の変化に対して、出力がどの程度変化するかを数値化するということである。

TER は、機械翻訳の出力を参照訳に近づくよう編集した際の編集距離を評価するものである。単語の削除、挿入といった編集操作だけでなく単語のシフト操作も考慮するため、単純な編集距離よりも適切な編集操作を評価できる。そのため、安定性を評価する尺度として適していると考えた。

また、本稿では上記安定性を評価するとともに、BLEU による評価も行う。これにより安定性と翻訳の良さを評価できると考える。

# 3 評価データの作成

本章では機械翻訳システムの安定性評価のためのデータの作成について説明する。安定性は、入力の変化に対する出力の変化を評価するので、入力文に変更を加える。基本方針は、入力文の意味が変わらない範囲での編集とする。

## 3.1 対象

評価データの作成には、京都フリー翻訳タスクで用いられる京都関連文書対訳コーパス (1) を用いる [3]。このコーパス Wikipedia 中の京都に関する記事から対訳コーパスが抽出加工されたものである。このうちの日本語側の入力文「kyoto-test.ja」に対して加工を行う。

上記コーパスからコーパスから無作為に選んだ 100 文に対して、評価データの作成を行う。

## 3.2 入力文の加工

下記 5 つの加工を行う。それぞれの変更は独立に加える。なお、変更が加えられないものに対しては何もしない。

**文中の「、」の削除**

入力文中の読点「、」を削除する。

**文末の「。」の削除**

文末の句点「。」を削除する。

**並び替え**

意味が変化しない範囲での入力文内での並び替えを行う。「〜と〜」といった並列句も並び替えの対象となる。

例 1) 「公園で」と「友達と」の入れ替え

私が公園で友達と遊ぶ。
→ 私が友達と公園で遊ぶ。

**表記の変更**

入力文中の単語の表記を変更する。なお表記の変更を行っても、入力文の意味は変化しない。

例 2) 「引っ越し」の表記ゆれ

彼が引っ越しする
→ 彼が引越する。

**同義語による換言**

入力文中の単語を同義語により換言する。

例 3) 「現在」と「今」の言い換え

現在も機能している。
→ 今も機能している。

## 3.3 作成したデータ

作成したデータ数について表 1 に記載する。また、データは (`http://www.jnlp.org/SNOW/E12`) で公開する。

表 1: 作成したデータ数

| 加工方法 | 文数 |
|---|---|
| 無加工 (ベースライン) | 100 |
| 「。」の削除 | 85 |
| 「、」の削除 | 100 |
| 並び替え | 62 |
| 表記ゆれ | 30 |
| 同義語による換言 | 64 |

[3] Creative Commons Attribution-Share-Alike Licence 3.0 に基づいた再配布可能な貴重な対訳データであるため、加工した評価データが公開できるため用いた。

# 4 実験

## 4.1 実験設定

3 章で作成したデータを入力文として評価実験を行う。NMT には注意機構付き Sequence-to-Sequence モデルの実装 (3) を利用する。PBMT には Moses(2) を用いる。機械翻訳モデルの学習には単語分割および小文字化を適用した KFTT コーパスを用いる。

## 4.2 TER による評価

TER による安定性評価の結果を図 2 に示す。出力 TER とは、ベースラインの出力と加工が加えられた入力に対する出力間の TER である。TER が高いと安定性が低いことを示す。NMT と PBMT を比較すると、NMT の TER が高いため PBMT と比べて安定性が低いと言える。

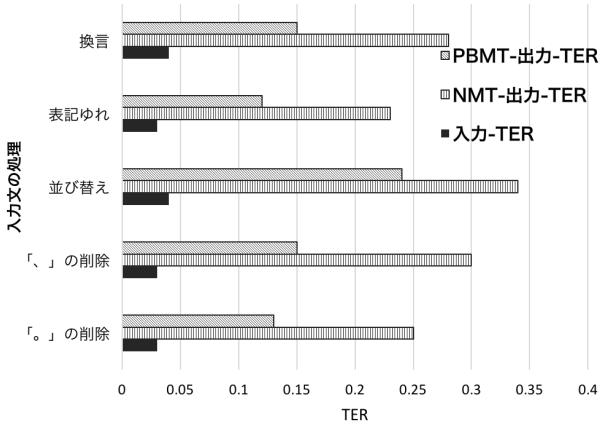


図 2: TER の変化

## 4.3 翻訳性能も含めた定量的な評価

4.2 節では、NMT が PBMT に比べて安定性が低いことが示された。一般に NMT は PBMT よりも高い BLEU 値を出す傾向にある。翻訳性能と安定性の関係はどうであろうか。

翻訳性能を含めた評価を表 2 に示す。評価は BLEU, RIBES および NIST で行った。NMT のほうが PBMT に比べて、BLEU および RIBES スコアが高い。NMT は出力が変化しやすいが翻訳精度が高いと言える。

## 4.4 翻訳性能も含めた定性的な評価

定性的な評価を行うために一例を取り上げる。文末の句点を削除した際の影響を示す。

**入力文原文** 1922 年（大正 11 年）4 月 1 日 - 宇治川仮信号所を宇治川仮信号場に変更。

**参照訳** april 1 , 1922 : the ujigawa temporary signal box ( 宇治川仮信号所 ) was changed to the ujigawa temporary signal station ( 宇治川仮信号場 )

**ベースライン (NMT)** april 1 , 1922 : the uji-gawa temporary signal station was changed to the ujigawa temporary signal station .

**句点の削除 (NMT)** april 1 , 1922 : uji temporary signal station was changed to 川 temporary signal station .

**ベースライン (PBMT)** in april 1 : the uji river , the uji-gawa river , a temporary signal station was changed to a temporary signal station .

**句点の削除 (PBMT)** in april 1 : the uji river , the uji-gawa river , a temporary signal station was changed to a temporary signal station

NMT と PBMT を比べると、NMT は句点の削除により出力の変化が PBMT より多いことが分かる。図 2 にもあるように、句点の削除に対して NMT のほうが PBMT より TER の変化が大きい。これらより、TER を用いた安定性の評価は適切に行えていると言える。また定量的に NMT は PBMT に比べて入力の揺らぎに出力が影響されやすい安定性の低いシステムと言える。

# 5 関連研究

機械翻訳の自動評価に関する研究は従来から盛んに取り組まれている。自動評価尺度は参照訳とシステム出力文の一致度合いを何らかの手法で数値化するものである。代表的な評価尺度である BLEU[1] は、n-gram 一致の適合率に基づいて評価する。単語の誤り率に基づく WER や編集距離を利用した TER[2] がある。これらは機械翻訳システム出力文の品質について参照訳との比較から評価したものである。

別の観点からの評価として、単語アラインメントを用いて流暢さを評価したもの [6] やタスクに適した機

表 2: 評価値

| 加工方法 | 入力 TER | NMT | | | | PBMT | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | 出力 TER | BLEU | RIBES | NIST | 出力 TER | BLEU | RIBES | NIST |
| ベースライン | 0.0 | 0.0 | 0.20 | 0.71 | 4.78 | 0.0 | 0.17 | 0.66 | 4.83 |
| 「。」の削除 | 0.03 | 0.25 | 0.20 | 0.72 | 4.72 | 0.13 | 0.16 | 0.59 | 4.73 |
| 「、」の削除 | 0.03 | 0.30 | 0.20 | 0.70 | 4.67 | 0.15 | 0.16 | 0.66 | 4.77 |
| 並び替え | 0.04 | 0.34 | 0.18 | 0.64 | 4.52 | 0.24 | 0.14 | 0.61 | 4.57 |
| 表記ゆれ | 0.03 | 0.23 | 0.17 | 0.65 | 4.05 | 0.12 | 0.15 | 0.60 | 4.3 |
| 同義語による換言 | 0.04 | 0.28 | 0.17 | 0.66 | 4.37 | 0.15 | 0.13 | 0.64 | 4.48 |

械翻訳尺度の検討したもの [3] がある。機械翻訳では安定性や一貫性といった評価は検討されていないが、情報検索を対象に単語分割一貫性の定量的な評価が提案されている [4]。単語分割の一貫性を評価することにより、検索漏れの問題が定量的に評価できるようになる。

我々は従来の自動評価尺度を組み合わせて利用することにより、安定性評価という観点を提案した。

## 6 まとめ

本稿では機械翻訳の評価の新たな観点として、機械翻訳システムの安定性の評価を提案した。機械翻訳をシステムとして見た際に、安定性を評価に加えることで使いやすいシステムかどうかを評価することを考えた。機械翻訳システムの安定性は、入力の言い換えによる出力の変化を見ることで確認できることを示した。また、評価用セットとして入力文の言い換えセットを構築した。評価セットは（`http://www.jnlp.org/SNOW/E12`）にて公開する。

## 使用したツールと言語資源

(1) The Kyoto Free Translation Task, Graham Neubig, `http://www.phontron.com/kftt`

(2) Open source statistical machine translation system Moses, `http://www.statmt.org/moses`

(3) Sequence-to-Sequence Learning with Attentional Neural Networks, Yoon Kim, `https://github.com/harvardnlp/seq2seq-attn`

## 参考文献


[1] Kishore Papineni, Salim Roukos, Todd Ward, and Wj Zhu. BLEU: a method for automatic evaluation of machine translation. *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pp. 311–318, July 2002.

[2] Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. A study of translation edit rate with targeted human annotation. *Proceedings of Association for Machine Translation in the Americas*, 2006.

[3] 杉山享志朗, 水上雅博, Graham Neubig, 吉野幸一郎, Sakriani Sakti, 戸田智基, 中村哲. 言語横断質問応答に適した機械翻訳評価尺度の検討. 情報処理学会 第 223 回自然言語処理研究会, 広島, Sep 2015.

[4] 高橋文彦, 颯々野学. 情報検索のための単語分割一貫性の定量的評価. 言語処理学会第 22 回年次大会, pp. 949–952, Mar 2016.

[5] 赤部晃一, Graham Neubig, 工藤拓, John Richardson, 中澤敏明, 星野翔. Project next における機械翻訳の誤り分析. 言語処理学会第 21 回年次大会ワークショップ『自然言語処理におけるエラー分析』発表論文集, Mar 2015.

[6] 吉見毅彦, 小谷克則, 九津見毅, 佐田いち子, 井佐原均. 単語アライメントを用いた英日機械翻訳文の流暢さの自動評価. 自然言語処理, Vol. 17, No. 1, pp. 7–28, Jan 2010.


## 付録

評価データの一例を掲載する。

**原文** その 第 二 次 長州 征伐 最中 の 7 月 2 0 日 、 将軍 ・ 家茂 が 大坂 城 で 薨去 する 。

**「。」の削除** その 第 二 次 長州 征伐 最中 の 7 月 2 0 日 、 将軍 ・ 家茂 が 大坂 城 で 薨去 する

**「、」の削除** その 第 二 次 長州 征伐 最中 の 7 月 2 0 日 将軍 ・ 家茂 が 大坂 城 で 薨去 する 。

**並び替え** その 第 二 次 長州 征伐 最中 の 7 月 2 0 日 、 大坂 城 で 将軍 ・ 家茂 が 薨去 する 。

**表記ゆれ** その 第 二 次 長州 征伐 最中 の 7 月 2 0 日 、 将軍 ・ 家茂 が 大阪 城 で 薨去 する 。

**同義語による換言** その 第 二 次 長州 征伐 最中 の 7 月 2 0 日 、 将軍 ・ 家茂 が 大坂 城 で 死ぬ 。