機械翻訳自動評価指標の比較

今村 賢治 †§, 隅田 英一郎 †, 松本 裕治 §
† ATR 音声言語コミュニケーション研究所
§ 奈良先端科学技術大学院大学
{kenji.imamura,eiichiro.sumita}@atr.jp, matsu@is.aist-nara.ac.jp

1 はじめに

機械翻訳の分野においても、対訳コーパスの充実に伴い、コーパスベースの手法が盛んになってきている。これは、対訳コーパスから、翻訳規則あるいはモデルを自動獲得し、それを用いて翻訳を行うものである。

しかし、自動獲得された規則中には、自動獲得のエラーや、コーパスに内在する翻訳の多様性により、冗長な、あるいは問題のある規則が含まれることが多い。そのような問題のある規則は、誤訳や曖昧性の増大を招き、翻訳品質を低下させる。

フィードバッククリーニング (Imamura et al., 2003) は、冗長規則による翻訳品質低下の問題を、機械翻訳の自動評価を利用して解決を図った手法である。翻訳品質を低下させる規則を自動評価を用いて特定し、自動評価値を向上させるように規則の削除を行う。自動評価方法が主観評価と充分に相関があるならば、クリーニング後の翻訳品質は向上する。

Imamura et al. (2003) は自動評価方法として、BLEU (Papineni et al., 2002) を用いていたが、この他にも自動評価方法はいくつか提案されている。クリーニング後の翻訳品質は、自動評価方法 (これをクリーニング指標と呼ぶ) に依存して変化すると考えられる。本稿の目的は、これら自動評価方法をクリーニング効果という観点から比較し、フィードバッククリーニングに適したクリーニング指標を特定することである。

2 フィードバッククリーニング

フィードバッククリーニングでは、冗長規則の削除を、組み合わせ最適化問題として捉えている。すなわち、自動評価値を最適化用評価関数値と考え、これを向上させるように、 generate and test を繰り返すことにより実行される。組み合わせ最適化方法には、山登り法を用いている。フィードバッククリーニングの構成を図 1に、手順概要を以下に示す。

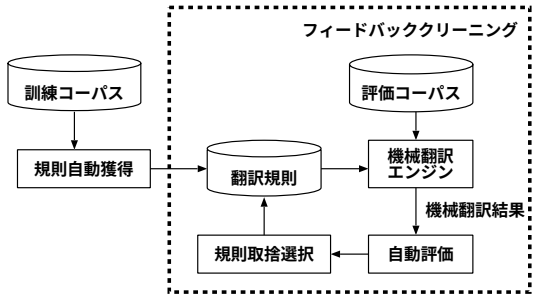

図 1: フィードバッククリーニングの構成

1. まず、訓練コーパスから翻訳規則を自動獲得する。これをベースルールセットと呼ぶ。
2. 現在のルールセットを用い、翻訳エンジンは評価コーパス (訓練コーパスとは異なる) 全体を翻訳し、自動評価する。
3. 次に、ルールセット中の規則１つ１つについて、その規則を削除し、評価コーパスを翻訳して、自動評価値を求める。
4. もし、規則を削除することにより、自動評価値が向上するならば、その規則を削除する。
5. 以上、ステップ 2〜 4を、自動評価値が変化しなくなるまで繰り返す。

なお、山登り法による組み合わせ最適化を適用した場合、評価コーパスサイズに比例して大量の文を翻訳しなければならない。たとえば、評価コーパスサイズが１万文であり、ルールセットが 10 万規則を含んでい

[0]本研究は通信・放送機構の研究委託「大規模コーパスベース音声対話翻訳技術の研究開発」により実施したものである。

たとすると、翻訳回数は 10 億回以上である。この問題を回避するために、フィードバッククリーニングでは、機械翻訳エンジンが使用した規則を抽出し、その規則が使われた文だけを再翻訳することにより、実行可能な時間でクリーニングを行っている。

3 機械翻訳品質の自動評価

フィードバッククリーニングに使用可能な自動評価方式は、評価結果を単一スコアとして出力するものである。本稿では、そのような自動評価として、BLEU (Papineni et al., 2002), NIST (Doddington, 2002), Word Error Rate(WER) (Jurafsky and Martin, 2000), Position independent word Error Rate (PER) (Och, 2003) を対象とする。これらはいずれも、機械翻訳結果と、参照訳（同じ原文を人間が翻訳したもの）との類似度を測定することにより、翻訳品質を数値化する。そして、評価コーパス全体について、総和または平均を算出し、システム全体の評価値を出力する方式である。

3.1 BLEU スコア

BLEU は、機械翻訳結果と参照訳との類似度を、両者の n-gram 一致数を基に、以下の式で算出する。

$$BLEU = \mathrm{BP}_{BLEU} \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right) \qquad (1)$$

ただし、

$$p_n = \frac{\sum_i \text{翻訳文 } i \text{ と参照訳 } i \text{ で一致した } n\text{-gram 数}}{\sum_i \text{翻訳文 } i \text{ 中の全 } n\text{-gram 数}}$$
$$w_n = 1/N$$

p_n は、評価コーパス全体について、翻訳文と参照訳を比較し、n-gram(たとえば 2-gram) の一致率を算出しているものである。これを 1-gram から N-gram について幾何平均を求めることにより、スコアを算出する。N は、通常 4 が用いられる。1-gram は、単語訳の正しさを表す指標となっており、高次の n-gram は、翻訳の流暢さを表す指標であるので、BLEU スコアは両者を組み合わせた指標となっている。また、BP_{BLEU} は、翻訳文が参照訳より短い場合に与えるペナルティ (brevity penalty) で、翻訳文が参照訳より長い場合は 1 である。

このように、BLEU スコアは 0 〜 1 の実数で表現され、値が高いほど良好な翻訳文であると判断される。

3.2 NIST スコア[1]

NIST スコアは、BLEU と同様に機械翻訳結果と参照訳の類似度を、両者の n-gram の一致数を基に、以下の式で算出する。

$$NIST = \mathrm{BP}_{NIST} \cdot \frac{\sum_{n=1}^{N} \left(\sum_{\text{翻訳文 } i \text{ と参照訳 } i \text{ に共通する } w_1 \ldots w_n} Info_i(w_1 \ldots w_n) \right)}{\sum_i \text{翻訳文 } i \text{ 中の全 } n\text{-gram 数}} \qquad (2)$$

ただし、

$$Info(w_1 \ldots w_n) = \log_2 \frac{\text{評価コーパス中の } w_1 \ldots w_{n-1} \text{ 数}}{\text{評価コーパス中の } w_1 \ldots w_n \text{ 数}}$$

NIST スコアは、0 以上の実数で表現され、値が高いほど、良好な翻訳であると判断される。N は通常 5 が用いられる。なお、BP_{NIST} は、計算方法は異なるが、BP_{BLEU} と同様に、翻訳文の長さが参照訳より長い場合は 1 である。

BLEU との最大の相違点は、個々の n-gram に情報量に基づく重みがつけられている点である。一般には、機能語列より内容語列の方が情報量が高いため、内容語の翻訳が正しいときに高いスコアを出す傾向がある。

しかし、高次の n-gram になるに従い、コーパス中に存在する数は減少し、$Info$ も減少することが多い。なぜなら、4 単語連続が一致したときの 5 単語目のバラエティは、1 単語が一致したときの 2 単語目のバラエティに比べ、少ない傾向があるためである。つまり、NIST スコアは、語順の正しさより、単語訳の正しさを重視した自動評価スコアであると言える。

3.3 Word Error Rate (WER), Position independent word Error Rate (PER)

Word Error Rate は、音声認識の自動評価に標準的に用いられているスコアである。機械翻訳の自動評価に用いる場合、翻訳文と参照訳との DP-matching 結果

[1] NIST(National Institute of Standards and Technology) は、米国の公共機関名であるが、そこで開発され、NIST 主催の機械翻訳ワークショップで採用された自動評価法であるため、通称 NIST スコアと呼ばれる。

表 1: コーパスサイズ

		英語	日本語
訓練コーパス	対訳文数	149,882	
	形態素数	868,087	984,197
評価コーパス	対訳文数	10,145	
	形態素数	59,533	67,554
テストコーパス	対訳文数	10,150	
	形態素数	59,232	67,193

を基に、以下の式で算出する。

$$WER = \frac{\sum_i (\text{挿入語数}\, i + \text{削除語数}\, i + \text{置換語数}\, i)}{\sum_i \text{参照訳}\, i \,\text{の語数}} \quad (3)$$

つまり、挿入コスト、削除コスト、置換コストがすべて 1 のときの編集距離を正規化したものである。編集距離に基づく機械翻訳の自動評価法には、安田 et al. (2002)、Akiba et al. (2003) などがあり、主観評価との相関が確認されている。

機械翻訳の自動評価では、たとえ正しく翻訳されていても、翻訳文と参照訳の語順が著しく異なる場合があるため、語順を無視してエラー率を算出する方法も用いられている。これが Position independent word Error Rate で、以下の式で算出する。

$$PER = 1 - \frac{\sum_i \text{翻訳文}\, i \text{・参照訳}\, i \,\text{間の一致語数}}{\sum_i \text{参照訳}\, i \,\text{の語数}} \quad (4)$$

WER, PER とも、 0 〜 1 の実数で、値が低いほど、よい翻訳であると判断される。

4 実験

本稿では、英日翻訳を対象に評価を行う。

4.1 実験条件

対訳コーパス 本実験では、 BTEC (Basic Travel Expression Corpus)(Kikui et al., 2003) を用いる。これは、旅行会話に頻出する表現を集めた基本表現集である。 BTEC コーパスのうち、約 17 万文を、表 1のように、訓練、評価、テストコーパスに分割した。

翻訳システム 本実験で用いた翻訳システムは、 HPAT (Imamura, 2002) である。これは、構文トランスファ方式の翻訳システムであり、トランスファの規則 (変換規則) を、対訳コーパスから階層的句アライメント (今村, 2002) を用いて自動獲得する。獲得された規則は、基本的には同期文脈自由文法規則である。

表 1の訓練コーパスから獲得された規則数 (ベースルールセット) は、 105,588 規則であった。

評価方法 評価方法は、クリーニング指標による自動評価 (BLEU, NIST, WER, PER) と、以下の 2 種類の主観評価を使用した。なお、自動評価ではテストコーパスすべてを用い、主観評価ではテストコーパスのうち 510 文を、評価者 1 名が評価した。また、自動評価に用いた参照訳は、いずれも 1 原文あたり 1 つである。

1. 一対比較
   一対比較は、ベースルールセットによる翻訳結果を基準とし、クリーニング後のルールセットによる翻訳結果が向上したか否か、あるいは同品質であるか、評価者が 1 文毎に比較する形式で行い、向上率を算出した。

2. 4 段階評価

   各翻訳結果を評価者が A(完全訳) 〜 D(不可訳) の 4 段階で評価した。

   A: 完全訳。訳文としてまったく問題なし。
   B: 部分訳。文法的に間違っていたり、訳文では情報が欠けていたりして不自然さを伴うが、原文で伝えたい主要な情報が容易に復元できる。
   C: 理解可能訳。訳文はかなり情報が欠けているが、文脈情報や訳文の断片的情報から原文で伝えたい情報を復元できる。
   D: 不可訳。原文が伝えたい情報が復元できない。

4.2 実験結果

実験結果を表 2に示す。なお、表中の括弧付き数字は、各評価方法でのクリーニング指標の順位を表す。

自動評価による品質測定結果 自動評価では、 BLEU を除き、クリーニング指標自身で測定したときのスコアがいずれも最高値を示した。しかし、各指標の順位を総和して判断した場合、 WER, NIST, BLEU, PER の順で翻訳品質が向上した。

表 2: クリーニング指標別翻訳品質
(括弧内の数字は、各評価方法での順位を表す)

			クリーニング指標			
		ベース	BLEU	NIST	WER	PER
削除規則数		0	6,220	5,070	3,832	3,678
自動評価	BLEU	0.232	(3) 0.244	(2) 0.248	(1) 0.252	(4) 0.228
	NIST	4.88	(3) 4.98	(1) 5.12	(2) 5.11	(4) 4.76
	WER	0.713	(3) 0.695	(2) 0.684	(1) 0.673	(4) 0.730
	PER	0.512	(2) 0.488	(4) 0.496	(3) 0.489	(1) 0.469
主観評価	一対比較	-	(1) +5.88%	(3) +4.71%	(2) +5.69%	(4) +3.73%
	A	50.8%	(1) 54.3%	(3) 53.5%	(1) 54.3%	(4) 50.6%
	A+B	66.1%	(3) 68.2%	(2) 70.0%	(1) 70.4%	(4) 67.3%
	A+B+C	77.6%	(3) 80.8%	(2) 81.0%	(1) 81.2%	(4) 80.4%

主観評価による品質測定結果 主観評価では、A ランクでは、BLEU, WER の品質向上が大きく、A+B+C ランクでは、WER, NIST の品質向上が最も大きかった。A ランクは、訳語の正しさ、文法的正確性の両方を必要とするため、語順まで含めて翻訳文と参照訳が一致したときに高いスコアを示す、BLEU, WER が好成績を収めたと考えられる。一方、C ランクは、文法的に崩れていても訳語が正しいければ、理解可能と判断されるので、訳語の正確性を重視した自動評価法である NIST が好成績を収めたと考えられる。

5 まとめ

本稿では、フィードバッククリーニングへの効果という観点から、機械翻訳品質自動評価方法の比較を行った。その結果、いずれの指標も翻訳品質を向上させたが、完全訳を重視した場合、BLEU, WER の効果が高く、理解可能訳を重視した場合、WER, NIST の効果が高かった。この結果から、語順の制約、単語訳の正確性の両方を測定可能な新たな自動評価方法が必要であると判断される。

参考文献

Yasuhiro Akiba, Eiichiro Sumita, Hiromi Nakaiwa, Seiichi Yamamoto, and Hroshi G. Okuno. 2003. Experimental comparison of MT evaluation methods: RED vs. BLEU. In *Proceedings of MT Summit IX*, pages 1–8.

George Doddington. 2002. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In *Proceedings of the HLT Conference*, San Diego, California.

Kenji Imamura, Eiichiro Sumita, and Yuji Matsumoto. 2003. Feedback cleaning of machine translation rules using automatic evaluation. In *Proceedings of ACL-2003*, pages 447–454.

Kenji Imamura. 2002. Application of translation knowledge acquired by hierarchical phrase alignment for pattern-based MT. In *Proceedings of TMI-2002*, pages 74–84.

今村賢治. 2002. 構文解析と融合した階層的句アライメント. 自然言語処理, 9(5):23–42, 10 月.

Daniel Jurafsky and James H. Martin. 2000. *Speech and Language Processing.* Prentice-Hall, Inc.

Genichiro Kikui, Eiichiro Sumita, Toshiyuki Takezawa, and Seiichi Yamamoto. 2003. Creating corpora for speech-to-speech translation. In *Proceedings of EuroSpeech 2003*, pages 381–384.

Franz Josef Och. 2003. Minimum error rate training in statistical machine translation. In *Proceedings of ACL-2003*, pages 160–167.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of ACL-2002*, pages 311–318.

安田圭志, 菅谷史昭, 竹澤寿幸, 山本誠一, 柳田益造. 2002. 対訳コーパスを用いた翻訳品質自動評価法. 情報処理学会論文誌, 43(7):2108–2117.