機械翻訳の自動評価のための擬似参照訳作成

吉村綾馬　松村雪桜　山岸駿秀　小町守
首都大学東京
{yoshimura-ryoma, matsumura-yukio, yamagishi-hayahide}@ed.tmu.ac.jp
komachi@tmu.ac.jp

1　はじめに

現在使われている機械翻訳の自動評価手法の多くは、機械翻訳の出力と翻訳者が翻訳した文（参照訳）とを比較して単語の表層の一致率に基づいて評価している。そのため、意味的に正しい文を出力できていても参照訳と表層が異なると低く評価されてしまう問題がある [1]。

この問題を改善するために、多くの自動評価手法では複数の参照訳を用いて様々な表層を持った文を評価できるようになっている。実際に、複数の参照訳で評価した方が人手評価との相関が高くなることが知られている [2]。しかし、参照訳を人手で作成するには時間とコストがかかるため、通常のデータセットでは原言語文 1 文につき参照訳 1 文であることが多く、適切な評価を行えていない可能性がある。

そこで、本研究では擬似参照訳を自動で生成する手法を提案する。具体的には、評価する文に対応する原言語文を被評価システムとは別の翻訳システムで翻訳し、その出力を参照訳として用いる。実験の結果、元の参照訳のみを用いて評価したスコアと人手評価との相関よりも、提案手法で生成した参照訳を元の参照訳に加えた複数の参照訳を用いた評価スコアと人手評価との相関の方が高いことがわかった。

この論文の主な貢献は以下である。

- 被評価システムとは別の翻訳システムで参照訳に対応する原言語文を翻訳して得られた複数の文を元の参照訳に加えて評価したスコアと人手評価との相関が、元の参照訳のみで評価したスコアと人手評価との相関より上がることを示した。
- 被評価システムの学習に使ったデータとは別のドメインのデータを用いて参照訳を生成するシステムを学習しても、提案手法で評価したスコアと人手評価との相関が上がることを示した。

2　関連研究

Lepage ら [3] は参照訳の言い換えを自動で取得する手法を提案している。この研究では自動で生成した参照訳の質を文法性、意味の同等性、語彙および構文の変動量などの観点から分析することで、自動で取得した文を参照訳として使える可能性を示唆している。しかし、本研究とは異なり実際に自動で生成した参照訳を用いて機械翻訳の評価を行っていない。

Kauchak ら [1] は人手で作成した元の参照訳を被評価システムの出力に使用されている単語で置き換えて参照訳の言い換えを生成している。本研究とは異なり、元の参照訳を言い換えて得られた新たな参照訳のみで評価を行い、元の参照訳で評価した場合と比較している。

どちらの研究も参照訳を生成するために、元の参照訳を使用している。それに対して本研究では、原言語文のみを使用しており、元の参照訳の情報は使用していないが、自動評価スコアと人手評価スコアとの相関が高くなるような参照訳を生成できている。

3　参照訳の自動生成

図 1 に提案手法の概略図を示す。提案手法の手順は以下の通りである。

1. 参照訳を生成するための任意の機械翻訳モデルを準備する。
2. 評価データの原言語文を上述した機械翻訳モデルで翻訳する。
3. 翻訳して得られた複数の参照訳を元の参照訳に加える。

参照訳生成のための機械翻訳システムはニューラル機械翻訳 (NMT) や統計的機械翻訳 (SMT) などのシステムの種類や用意するシステムの数に制限はない。シ

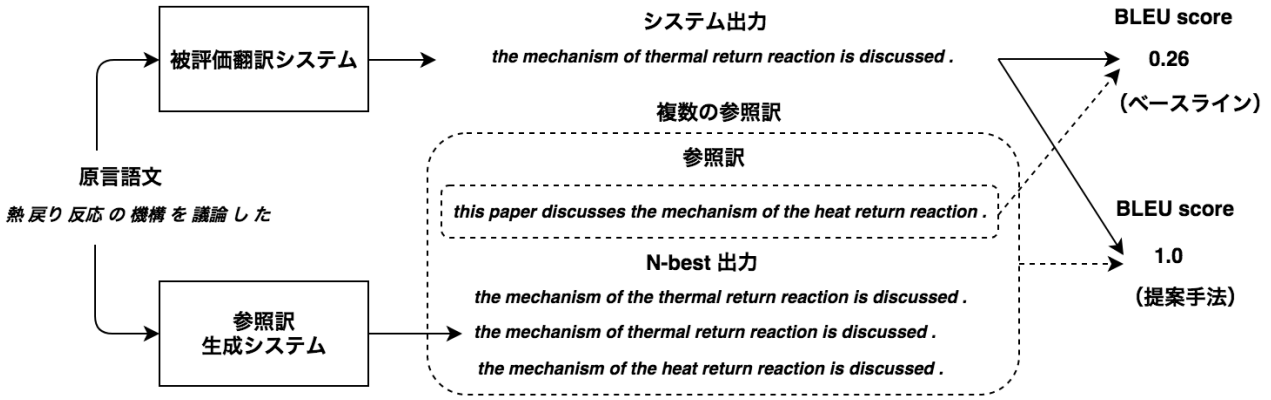

図 1: 提案手法の概略図

ステム数が 1 の場合、複数の参照訳を得るためにビームサーチを用いて N-best を出力する。

4 実験

4.1 参照訳生成システム

参照訳を生成するための機械翻訳システムは Luong ら [4] の global dot attention 型の NMT を基に実装したシステムを使用した。[1] モデルのハイパーパラメータは隠れ層の次元数を 512、埋め込み層の次元数を 512、語彙サイズを 30,000 として実験を行った。最適化手法は AdaGrad を用いた。

学習データと評価データのドメインの違いによる影響を調べるために、参照訳生成のための翻訳システムの学習データには Asian Scientific Paper Excerpt Corpus (ASPEC) [5] と Japan Patent Office corpora (JPC) [6] の 2 種類を用いた。日本語の単語分割には形態素解析 MeCab[2]（バージョン 0.996, IPADIC）を用い、英語の単語分割には Moses の tokenizer.perl[3] を用いた。なお、原言語および目的言語の学習用データから 1 文あたり 60 単語を超える文対を削除した。

さらに、得られた NMT の性能を評価するために元の参照訳での BLEU スコアを測定した。各コーパスの文数とそれぞれの BLEU スコアを表 1 に示す。

システム数は 1 つなので、出力の上位 N 件をビームサーチによって得る。本研究では最大の N を 5 に設定して、出力結果 5-best までを元の参照訳に加えて複数の参照訳を得る。

また、参照訳を生成するシステムの多様性による影響を調べるために、Workshop on Asian Translation (WAT) 2015 の ASPEC 日英コーパスで学習した NMT 以外の 3 システムの出力それぞれ 1 つずつの計 3 つを参照訳として元の参照訳に加えて評価する実験も行った。3 システムのモデルはそれぞれ、SMT, 用例ベース機械翻訳 (EBMT), ルールベース機械翻訳 (RBMT)+SMT である。[4]

表 1: 各コーパスの文数と学習した NMT の BLEU スコア

corpus	train	dev	test	JA-EN	EN-JA
ASPEC	977,367	1,790	1,812	18.52	27.52
JPC	996,712	2,000	2,000	10.15	11.28

4.2 評価データ

評価には WAT2017 における ASPEC の日英機械翻訳の評価データを用いた。このデータは 200 文の対訳文と各システムの出力およびそれに対する 2 人分の人手評価（1〜5）がある。人手評価は 2 人分の評価スコアの平均値を正規化して使用した。日英翻訳では 3 システム分、英日翻訳では 5 システム分のデータを用いた。表 2 に、正規化された人手評価スコアのヒストグラムを示す。

4.3 評価方法

本実験では、文レベルかつ複数の参照訳で評価を行える、SentenceBLEU [7]（これ以降 BLEU とする）, RIBES [8], METEOR [9] を用いて評価を行いそれぞれの自動評価手法でのスコアを文ごとに測定した。RIBES は翻訳で単語の順番が変わるような場合に有効な手法である。今回は日英及び英日翻訳の評価のた

[1] https://github.com/yukio326/nmt-chainer
[2] http://taku910.github.io/mecab/
[3] https://github.com/moses-smt/
[4] http://lotus.kuee.kyoto-u.ac.jp/WAT/evaluation/ DataID はそれぞれ 655, 829, 529 のデータを使用

表 2: 自動評価と人手評価とのケンドールの順位相関係数

	ASPEC					JPC				
	JA-EN			EN-JA		JA-EN			EN-JA	
	BLEU	RIBES	METEOR	BLEU	RIBES	BLEU	RIBES	METEOR	BLEU	RIBES
single (baseline)	0.304	0.253	0.349	0.371	0.319	0.304	0.254	0.349	0.371	0.345
+ N-best	0.331	0.326	0.356	0.378	0.348	0.309	0.271	0.360	0.392	0.381

表 3: 自動評価と人手評価とのケンドールの順位相関係数。Single + 3 は WAT2015 の NMT 以外の日英 3 システムの出力を元の参照訳に追加した場合。(* は統計的有意差があることを示す。($p < 0.05$))

	BLEU	RIBES	METEOR
single (baseline)	0.304	0.254	0.349
+ 3-systems	*0.370	*0.385	*0.419

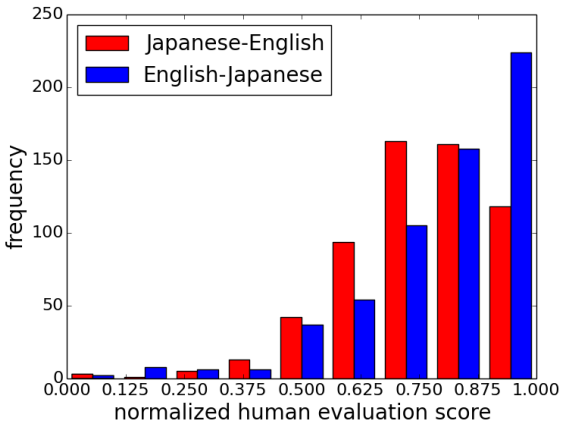

図 2: 正規化された人手評価スコアのヒストグラム

め、RIBES を用いた。METEOR は単語の表層の違いを考慮して評価する手法である。単語単位ではなく文単位でも改善が認められるかどうかを調べるために METEOR を用いた。METEOR は日本語の評価には対応していないため、日英での実験の評価のみ行っている。

各自動評価手法で得られた文ごとのスコアと各文の人手評価スコアの相関を測定した。相関係数はケンドールの順位相関係数を用いた。N-best において、どの N を使うかを決定するために、評価データを test データと dev データに分けた。日英翻訳システムでは 3 分割交差検定、英日翻訳システムでは 5 分割交差検定を行って最終的な評価スコアを計算した。さらに、ベースラインでの評価と WAT2015 の NMT 以外の 3 システムを用いた評価に有意差があることを示すために、Wilcoxon の順位和検定を行った。

5 結果

表 2 に ASPEC と JPC でそれぞれ学習した翻訳システムで実験を行って得られた評価スコアと人手評価スコアとの相関を示す。全ての言語と評価手法で提案手法がベースラインよりも相関が高くなっている。

表 3 に ASPEC の日英で学習した NMT 以外の 3 システムで実験を行って得られた評価スコアと人手評価スコアとの相関を示す。全ての評価手法で提案手法がベースラインよりも相関が高くなっている。さらに、一つのシステムの N-bset を用いるよりも、複数のシステムの出力を合わせて用いた方が相関が高くなっていることがわかる。

6 考察

ベースラインで評価した BLEU スコアと人手評価スコアとの差よりも、提案手法で評価した BLEU スコアと人手評価スコアとの差の方が小さくなった事例（成功例）と大きくなった事例（失敗例）を図 4 に示す。成功例ではシステム出力と参照訳の意味は同じであり、人手評価も 1.0 と高くなっている。しかし、システム出力と参照訳では態も使用語彙も異なっているため元の参照訳では BLEU スコアが低くなってしまっている。一方で、生成した 3-best のなかにシステム出力と同じ文が存在するため、参照訳に 3-best を加えた複数の参照訳での評価では BLEU スコアが 1.0 となり、人手評価との誤差が小さくなっている。

失敗例ではシステムの出力の人手評価は 0.625 であったが、3-best のなかにシステム出力と表層が似たような文が入っているので 3-best を加えた複数の参照訳で評価した BLEU スコアは 1.0 となり、不当に高く評価してしまっている。図 2 より、今回用いたデータセットでは人手評価スコアが高い事例数が多いことから、表 4 の成功例のように人手評価スコアに近く BLEU で評価できる事例数が多くなると考えられる。したがって、ベースラインの評価では低い順位だった事例が適切な高い順位になったことで相関が高くなったのではないかと考えられる。実際に表 4 の成功例のように、

表 4: 提案手法で人手評価スコアに近く BLEU で評価できた例と不当に高く評価している事例

成功例	
原言語文	熱 戻り 反応 の 機構 を 議論 し た
システム出力	the mechanism of thermal return reaction is discussed .
参照訳	this paper discusses the mechanism of the heat return reaction .
3-best	the mechanism of the thermal return reaction is discussed .
	the mechanism of thermal return reaction is discussed .
	the mechanism of the heat return reaction is discussed .
人手評価スコア: 1.0 ベースラインでの BLEU スコア: 0.26 提案手法での BLEU スコア: 1.0	

失敗例	
原言語文	ステロイド 内服 , ステロイド 軟膏 と 亜鉛 華 軟膏 の 外用 を 施行 し た 。
システム出力	steroid internal use , steroid salve and external use of zinc ointment were performed .
参照訳	oral administration of steroid , and external use of steroid and zinc ointment were performed .
3-best	steroid internal use , steroid salve and external use of zinc ointment were enforced .
	steroid internal use , steroid salve and external use of the zinc ointment salve were enforced .
	steroid internal use , steroid salve and external use of the zinc ointment were enforced .
人手評価スコア: 0.625 ベースラインでの BLEU スコア: 0.457 提案手法での BLEU スコア: 1.0	

人手評価スコアに近く BLEU で評価できている事例数を数えると 600 事例中 581 事例あった。

直感的には参照訳を生成するシステムが被評価システムよりも高い性能であるべきだが、今回の実験で学習した翻訳システムは被評価システムよりも弱い。今回の実験では被評価システムよりも弱い翻訳システムを用いても人手評価との相関が上がるということを示している。さらに、被評価システムの学習に使用した ASPEC ではなく、JPC を用いて参照訳生成システムを学習した場合の実験結果はベースラインよりも相関が高くなっていることから、提案手法は参照訳生成システムの学習データは、被評価システムの学習データと異なる場合でも使用することができる。

7 おわりに

本研究では、機械翻訳の自動評価をより適切に行うための擬似参照訳自動生成手法を提案した。提案手法では、評価する出力文に対応する原言語文を別の翻訳システムで翻訳した出力を参照訳として用いる。実験結果は、両方向の翻訳と複数の自動評価尺度で提案手法がベースラインでの人手評価との相関よりも高くなった。

今後は、どの N-best までを使うかを、生成した参照訳の多様性を考慮して決める方法を検討したい。また、今回の提案手法では原言語文の情報しか使っておらず、参照訳の情報を使っていないため、原言語文に加えて参照訳の情報も使用して参照訳を自動生成する手法も検討したい。

参考文献

[1] David Kauchak and Regina Barzilay. Paraphrasing for automatic evaluation. In *NAACL*, 2006.

[2] Andrew M Finch, Yasuhiro Akiba, and Eiichiro Sumita. How does automatic machine translation evaluation correlate with human scoring as the number of reference translations increases? In *LREC*, 2004.

[3] Yves Lepage and Etienne Denoual. Automatic generation of paraphrases to be used as translation references in objective evaluation measures of machine translation. In *IWP*, 2005.

[4] Thang Luong, Hieu Pham, and Christopher D. Manning. Effective approaches to attention-based neural machine translation. In *EMNLP*, 2015.

[5] Toshiaki Nakazawa, Manabu Yaguchi, Kiyotaka Uchimoto, Masao Utiyama, Eiichiro Sumita, Sadao Kurohashi, and Hitoshi Isahara. ASPEC: Asian scientific paper excerpt corpus. In *LREC*, 2016.

[6] Satoshi Kinoshita, Tadaaki Oshio, Tomoharu Mitsuhashi, and Terumasa Ehara. Translation using JAPIO patent corpora: JAPIO at WAT2016. In *WAT*, 2016.

[7] Chin-Yew Lin and Franz Josef Och. Orange: a method for evaluating automatic evaluation metrics for machine translation. In *COLING*, 2004.

[8] Hideki Isozaki, Tsutomu Hirao, Kevin Duh, Katsuhito Sudoh, and Hajime Tsukada. Automatic evaluation of translation quality for distant language pairs. In *EMNLP*, 2010.

[9] Michael Denkowski and Alon Lavie. Meteor 1.3: Automatic metric for reliable optimization and evaluation of machine translation systems. In *EMNLP*, 2011.