単語属性変換で作成した疑似負例データを用いた自動機械翻訳評価

高橋 洸丞 [1]　石橋 陽一 [1]　須藤 克仁 [1,2]　中村 哲 [1]
[1] 奈良先端科学技術大学院大学　[2] 科学技術振興機構さきがけ
{takahashi.kosuke.th0, ishibashi.yoichi.ir3, sudoh, s-nakamura}@is.naist.jp

概要

本研究では、誤りを含む文の評価に対して評価性能の向上を狙って、WMT の metrics shared task コーパスに含まれる単語属性を反対のものへ変換することで疑似負例データを作成し、疑似負例データで評価モデルの追加学習を行った。

その結果、提案手法は 20 年度の metrics コーパスにてコーパス全体での人手評価とのピアソンの相関係数が向上した。また事例分析では、提案手法が評価スコアの決定時に原言語文の内容との合致率をより重要視している傾向が見られた。

1　はじめに

近年の自動評価モデルは多くが Bidirectional Encoder Representations from Transformers (BERT) [1] という大規模な言語モデルを利用したものである。BERTscore[2]、BLEURT[3]、C-SPEC[4]、COMET[5] などは BERT をエンコーダとして使用しているモデルで、20 年度の WMT 評価用タスクの metrics shared task にて人手評価との高い相関を記録した。しかし、[4] で述べられているように BERT をベースとした評価手法は、品質が悪く人手評価のスコアが低い翻訳文に対して相関が低くなる。そこで [6] は低品質な翻訳文の評価を向上させるために、独自の人手評価の基準を設けて新たな人手評価が付けられたコーパスを作成した。

本研究は、低品質な翻訳文に対する評価性能を向上させるため、[6] とは異なり、疑似的な誤りを含む翻訳文コーパスを作成する。そしてどのような疑似的な誤りを作るのかを決めるために、低品質な翻訳文とその評価誤りの関係について、17 年度の WMT metrics shared task において、事前に BLEURT と C-SPEC を用いた評価実験をした。その結果、人手評価とのピアソンの相関係数を著しく下げる事例の多くは表 1 のように、名詞の翻訳誤りと述語構造の誤りに由来するものであった。そこで、本研究では名詞の評価誤りを改善するために、疑似的に誤った名詞また動詞を含むコーパスを作成する。疑似負例データの作成は [7] の手法に従い、単語の属性変換モデルを反対の意味・属性を持つ名詞そして動詞に適用することで、本来の翻訳文に含まれる単語とは逆の意味の単語に置き換わるように設定する。提案する評価モデル (C-SPECpn : Cross-lingual Sentence Pair Embedding Concatenation fine-tuned on pseudo-negatives) は、学習済みの C-SPEC を作成した疑似負例データで追加学習したものである。

2　関連研究

機械翻訳に向けた自動評価システムは多くが WMT の metrics shared task のコーパスを用いて、訓練や評価性能の評価を行う。このワークショップが他の評価タスクと異なる特徴的な点は、大量の人手評価が機械翻訳文につけられているところであり、提出された評価システムの良し悪しが、人手評価との相関の高さによって決められる。人手評価には 2 種類の手法が取り入れられており、DA(Direct Assesment)[8] が WMT15-20 の翻訳結果に、MQM(Multidimensional Quality Metrics)[9] が WMT20 の翻訳結果にアノテーションされている。DA は 0-100 の整数値で参照訳文に対して翻訳文の出来を評価する手法で、高性能な翻訳システムの評価には信頼性が欠けるとされている [10]。一方で MQM は翻訳誤りの種類や程度を誤翻訳されている箇所にアノテーションし、その誤りの種類や程度によって各翻訳文の評価スコアを決定する。

近年、WMT の評価タスクで人手評価と高い相関を示した BLEURT[3]、C-SPEC[4]、COMET[5] などは、DA や文単位の MQM などの人手評価を学習に際して必要とする。これらの評価モデルは訓練済

表 1　WMT17 の metrics shared task における BLEURT と C-SPEC の評価結果の例。

参照訳文	They want to compete with delivery services.
システム訳文	They want to compete with airlines.
人手評価 (DA)	-0.861
BLEURT[3]	-0.344
C-SPEC[4]	0.595
参照訳文	I'm glad we got some of those dirty licks caught on tape.
システム訳文	I'm glad we lick some of the dirty things they made.
人手評価 (DA)	-1.548
BLEURT[3]	-0.088
C-SPEC[4]	-0.405

みの BERT モデルをエンコーダとして用いて、参照訳文やシステム訳文は文単位の分散表現に符号化し、線形層を通してスコアを出力する。この内、C-SPEC と COMET は参照訳文に加えて原言語文もシステム訳文の評価に使用することで評価性能を向上させるように設計されている。BLEURT も同様に 21 年度 WMT で提出されたモデルでは原言語文を扱うように変更されている。

3　提案手法 : C-SPECpn

提案手法の評価モデルは C-SPEC と同じモデル構造で、BERT 系モデルとして XLMRoBERTa[11] を使用する。対となる原言語文と参照訳文そしてシステム訳文が、システム訳文-原言語文、システム訳文-参照訳文の 2 つのペアとしてそれぞれ XLMRoBERTa に入力され、文対の分散表現 (ベクトル) に落とし込む。次に得られた 2 つのベクトルを結合し、多層パーセプトロンにより回帰の形で最終的な評価スコアを出力する。訓練時は、WMT15-20 で人手評価として採用されている DA を標準化したものを教師データとし、平均二乗誤差 (MSE : Mean Squared Loss) によりモデルのパラメータをアップデートする。

具体的な評価モデルの訓練は次のような 4 ステップに分けて実行した。また、各評価ステップの始めでは、多層パーセプトロンのパラメータをランダムに初期化しており、次の訓練ステップに引き継がれるのは XLMRoBERTa のパラメータのみである。

訓練ステップ 1　WMT15-16 の人手評価である DA コーパスで訓練を行う。この訓練ステップは、評価モデルの性能を安定させることを目的としており、WMT15-16 は機械翻訳の性能があまり高くないので DA に含まれるノイズの影響が小さく、学習時の MSE ロスが低くなりやすい。

訓練ステップ 2　WMT15-17、WMT18-20、WMT15-20 の 3 つのコーパスからいずれかの DA が付与されたデータで、再度評価モデルの追加訓練を行う。3 つのコーパスに条件を分けた理由は、2018 年度以降の DA データはノイズを多く含むと、[3] にて述べられており、ノイズを含む可能性のある 2018 以前と以降で評価モデルの比較をするためである。

訓練ステップ 3　3 つ目の訓練ステップでは、作成した負例コーパスで回帰ではなく分類問題としてモデルの追加訓練をする。

訓練ステップ 4　WMT20 年度の MQM と呼ばれる人手評価が付与された文単位のデータに対して追加訓練を行う。

3.1　単語の属性変換による負例コーパスの作成

鏡映変換に基づく埋め込み空間上の単語属性変換 [12] を用いてデータ拡張を行った。この手法は、2 つの MLP のパラメータにより決定される、単語の埋め込み空間中の鏡面により、単語の属性に関わる前知識なしで、特定の属性を持つ単語を反対の意味を持つ単語へ変換するものである。例えば、*queen* という単語の性別属性を反転させると *king* という単語に変換することが可能である。また、ターゲットの属性を持たない単語の場合、その単語は変化せず、*apple* に対して性別属性で単語属性変換を適用しても、*apple* が出力される。

そして、本研究では負例データの作成時に、性別と対義関係の 2 つの属性で、名詞や動詞と辞書に登録された単語を逆の属性を持つ単語に変換する。単語の属性変換は、2 つ目の訓練ステップで使用されたコーパス内の英語の参照訳文において、全ての文

中に含まれる単語について適用する。例を挙げると、ある参照訳文 *"It is our duty to remain at his sides", he said, to applause.* に対して対義関係の属性変換を実行すると、*"It is our duty to change at his sides", he said, to whisper.* という文が得られ、*remain* → *change* 、*applause* → *whisper* と変換されている。そして、参照訳文中に一つもターゲットの属性を持つ単語が存在しない場合、その参照訳文は一切変化が加えられないため、負例コーパスから排除した。

3.2 負例コーパスでの追加訓練

作成した負例データには人手評価のスコアがついていないため、それまでの訓練ステップ 1、2 と同様な訓練を行うことができない。そこで本研究では、3 種類のシステム訳文の条件を設定し、分類問題として評価モデルの追加訓練を行う。それぞれの条件における評価モデルの真の入力は以下の通りであり、訓練時の評価モデルは、入力がこれらの内どれに当てはまるのかを学習していく。

1. システム訳文-原言語文、システム訳文-参照訳文 (変化なし)
2. 参照訳文-原言語文、参照訳文-参照訳文 (参照訳文に置き換え)
3. 負例文-原言語文、負例文-参照訳文 (負例文に置き換え)

分類カテゴリ 2 では、18 年度以降は高性能な翻訳システムによる翻訳文が多く、システム訳文が正解に近い場合でも、評価モデルが些細な違いを識別できるように、正解訳文同士の入力ペアを導入した。また、分類問題の設定が簡単で分類の正解率が高く、訓練ステップ 2 のコーパスにおける過学習を防ぐために、各負例文は一度のみ学習に用いた。

4 WMT20 年度 MQM コーパスでの実験

評価実験は WMT20 の文単位 MQM コーパスの内、全体の 10%で行われた。ただし、予め訓練に使用された DA コーパスと重複しないようにした。実験結果は、評価モデルのスコアと人手評価とのピアソンの相関係数をまとめたものである。

4.1 実験結果

評価モデルごとの実験結果を表 2 に示す。C-SPEC と比べて、C-SPECpn の WMT15-17、WMT18-20 で訓練されたモデルの方がピアソンの相関係数が高くなった。そして、3 つの訓練コーパスでは WMT18-20 で訓練されたモデルが最も高いピアソンの相関係数を記録した。一方で、WMT15-20 で訓練された C-SPECpn は C-SPEC に及ばない結果となった。これは WMT15-20 がコーパスサイズが大きく、負例データも同様に大きくなるので、モデルが過学習をしてしまったのではないかと推察する。

4.2 人手評価の値域ごとのピアソンの相関係数

本研究の目的であった低品質なシステム訳文への提案手法の有効性を調べるために、5 段階に人手評価のスコアを分け、それぞれの値域でピアソンの相関係数を算出した。その結果を図 1 に示す。提案モデルの C-SPECpn は、[-25, -20)、[-10, 0) の区間において C-SPEC よりも高い相関係数を記録した。このことから、作成した負例データによる追加訓練は、深刻な誤りを含むシステム訳文や、軽度な誤りを含むシステム訳文に対して、評価モデルの頑健性を高めることが示唆される。また、21 年度の WMT metrics shared task[13] では、システム評価や文単位の評価においてトップを争う結果を記録した。

5 WMT21 年度 challenge set での事例分析

これまでの実験結果に加えて、WMT21 年度の評価用 challenge set タスクにおける C-SPEC と C-SPECpn の比較を事例分析により行った。challenge set とは、特定の翻訳誤りに対して評価モデルの性能を比較するためのタスクで、文意の極性変換、誤りを含む参照訳、また独英間の翻訳で発生する翻訳誤りを集めたものである [13]。

評価事例の一部を表 3 に示す。C-SPECpn は分類問題を追加で訓練しているので、C-SPEC とは出力の値域が異なり、人手評価が公開されていないので判断が難しいが、C-SPECpn は C-SPEC よりも原言語文との類似度が高そうなときにスコアが高くなる傾向があった。例文 1 は参照訳が誤っている例で、参照訳文に *the flooding* という名詞が出現しているが、原言語文では水という表現に留まっている。このとき、システム訳文に対して、C-SPECpn は C-SPEC よりも高いスコアをつけている。また例文 2 は独英間の翻訳で起こる翻訳誤りの例で、原言語文の *konntet* が翻訳されずにそのままシステム訳文に出現しているが、C-SPECpn は C-SPEC よりも

表 2　WMT20 MQM コーパスでのピアソンの相関係数。C-SPEC は負例データなし、C-SPECpn は負例データありで学習したモデル。en-de は英語からドイツ語、zh-en は中国語から英語へ翻訳されたデータ。avg は en-de、zh-en のピアソンの相関係数の平均、 all は 2 つの言語対をまとめてコーパス全体でのピアソンの相関係数を示している。

評価モデル	en-de	zh-en	avg	all
C-SPEC trained on WMT15-17	0.609	0.773	0.691	0.787
C-SPEC trained on WMT18-20	0.612	0.805	0.708	0.813
C-SPEC trained on WMT15-20	0.603	0.798	0.700	0.808
C-SPECpn trained on WMT15-17	**0.626**	0.809	0.717	0.817
C-SPECpn trained on WMT18-20	0.619	**0.824**	**0.721**	**0.829**
C-SPECpn trained on WMT15-20	0.309	0.715	0.512	0.724

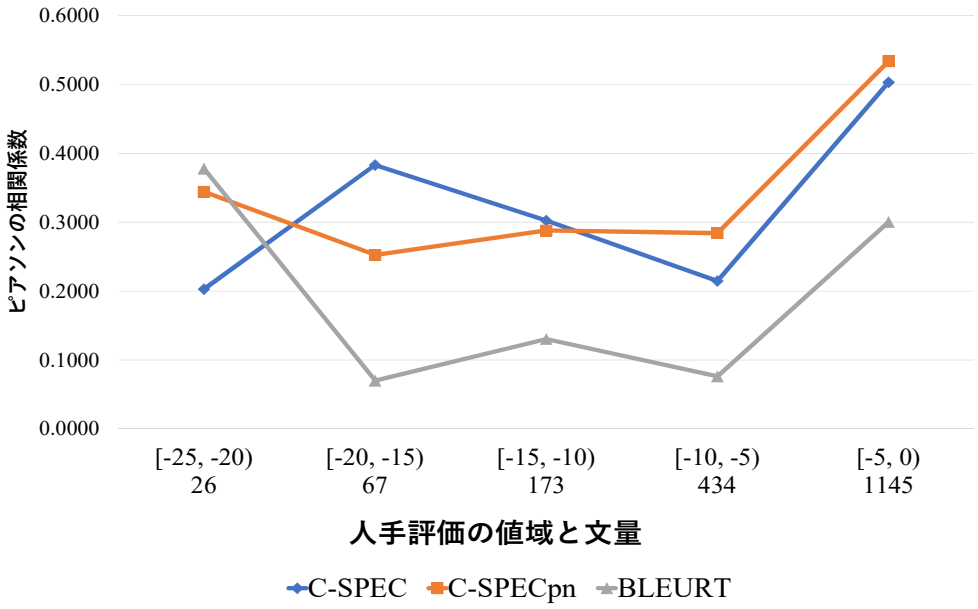

図 1　人手評価の値域におけるピアソンの相関係数。値域の下にはその値域に含まれる文の数が記されている。C-SPEC、C-SPECpn は共に WMT18-20 で訓練されたモデル、BLEURT は WMT15-19 で訓練された配布モデルを使用した。

高いスコアを出力している。

表 3　WMT21 challengeset での評価結果の例。

原言語文 1	希望水早点退
参照訳文 1	We hope that the flooding ends soon.
システム訳文 1	I hope the water will return early.
C-SPEC	-1.5684
C-SPECpn	-0.6392
原言語文 2	Ihr konntet denken.
参照訳文 2	You were able to think.
システム訳文 2	your konntet thinking.
C-SPEC	-8.2422
C-SPECpn	-1.4688

6　おわりに

本研究では、単語の属性変換を用いて疑似負例データを作成し、疑似負例データで評価モデルを追加訓練することで、評価性能が一部向上することを示した。また事例分析では、提案モデルである C-SPECpn は評価に際してより原言語文に重きを置いた評価となることがわかった。

謝辞

本研究は JST さきがけ (JPMJPR1856) の支援を受けたものである。また全ての実験は JST さきがけの支援の元で理研の miniRAIDEN を用いて行われた。

参考文献

[1] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pp. 4171–4186, Minneapolis, Minnesota, June 2019. As-

sociation for Computational Linguistics.

[2] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. In International Conference on Learning Representations, 2020.

[3] Thibault Sellam, Dipanjan Das, and Ankur Parikh. BLEURT: Learning robust metrics for text generation. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 7881–7892, Online, July 2020. Association for Computational Linguistics.

[4] Kosuke Takahashi, Katsuhito Sudoh, and Satoshi Nakamura. Automatic machine translation evaluation using source language inputs and cross-lingual language model. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 3553–3558, Online, July 2020. Association for Computational Linguistics.

[5] Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. COMET: A neural framework for MT evaluation. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 2685–2702, Online, November 2020. Association for Computational Linguistics.

[6] Sudoh Katsuhito, Takahashi Kosuke, and Nakamura Satoshi. Is this translation error critical?: Classification-based human and automatic machine translation evaluation focusing on critical errors. In Proceedings of the Workshop on Human Evaluation of NLP Systems (HumEval), pp. 46–55, Online, April 2021. Association for Computational Linguistics.

[7] 石橋陽一, 須藤克仁, 中村哲. 単語属性変換による自然言語推論データの拡張. 言語処理学会第 26 回年次大会発表論文集, 2021.

[8] Yvette Graham, Timothy Baldwin, and Nitika Mathur. Accurate Evaluation of Segment-level Machine Translation Metrics. In Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 1183–1191, Denver, Colorado, May–June 2015. Association for Computational Linguistics.

[9] Arle Lommel, Hans. Uszkoreit, and Aljoscha Burchardt. Multidimensional Quality Metrics (MQM) : A Framework for Declaring and Describing Translation Quality Metrics. Tradumàtica, No. 12, pp. 455–463, 2014.

[10] Markus Freitag, George Foster, David Grangier, Viresh Ratnakar, Qijun Tan, and Wolfgang Macherey. Experts, errors, and context: A large-scale study of human evaluation for machine translation, 2021.

[11] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 8440–8451, Online, July 2020. Association for Computational Linguistics.

[12] Yoichi Ishibashi, Katsuhito Sudoh, Koichiro Yoshino, and Satoshi Nakamura. Reflection-based word attribute transfer. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop, pp. 51–58, Online, July 2020. Association for Computational Linguistics.

[13] Markus Freitag, Ricardo Rei, Nitika Mathur, Chi-kiu Lo, Craig Stewart, George Foster, Alon Lavie, and Ondřej Bojar. Results of the WMT21 metrics shared task: Evaluating metrics with expert-based human evaluations on TED and news domain. In Proceedings of the Sixth Conference on Machine Translation, pp. 733–774, Online, November 2021. Association for Computational Linguistics.