参照訳・編集文を用いない機械翻訳出力の自動評価

相田　太一　山本 和英
長岡技術科学大学
{aida, yamamoto}@jnlp.org

1　はじめに

ニューラル機械翻訳 (NMT) は幅広く用いられており、2020 年の東京オリンピック開催に向けて、さらに注目が集まっている。近年の NMT の性能は目覚ましい進歩を遂げているが、NMT の出力した訳文は全て正しいというわけではなく、誤って翻訳された文も含まれている。特に翻訳者などは高い品質の訳文を要求するため、NMT によって出力された全ての訳文に対して品質評価をする必要がある。

NMT から出力された訳文を自動的に評価するための手法として、主に以下の 2 つが存在する。

- 正解となる参照訳を用いる手法
- 参照訳を用いない手法

訳文の正解となる文である参照訳を用いた訳文の自動評価手法として代表的なのが、BLEU [1] である。BLEU は参照訳と訳文の n-gram の一致数を元に品質を算出する。人手評価とも高い相関があり、現在も機械翻訳のタスクの評価尺度として用いられている。

参照訳を用いない訳文の自動評価手法における代表的なタスクとして、機械翻訳に関する WMT[1] というワークショップの中に Quality Estimation (QE) が存在する。このタスクでは、実際に NMT を用いる場合には手元には原文と NMT から出力された訳文のペアしかなく、正解となる参照訳を持ち合わせていないということから、参照訳を用いずに訳文を評価する。評価モデルの訓練時には、原文、NMT から出力された訳文、訳文を人手で編集した文を用いてモデルを訓練する。その後、原文と訳文だけを用いて、単語レベルの QE では挿入、置換・削除、誤訳を引き起こす原文かどうかのタグを、文レベルの QE では訳文から人手で編集した文への編集距離を予測することで訳文を評価する。このタスクでは、代表的なベースラインで数多くの特徴量を用いる QuEst++[2] や、原文と訳文から特徴を抽出する predictor と抽出した特徴を元に単語のタグや編集距離を推定する estimator を用いる手法 [3, 4] が提案され、それらを含めた複数の手法を組み合わせた手法 [5] が単語レベルの QE で最高性能、文レベルの QE で最高性能に迫る結果を達成した。

しかし、訳文を自動評価する上記の手法は、評価モデルの訓練時に訳文を人手で編集した文、または評価時に正解となる参照訳が必要となる。そこで、我々は上記のどちらも用いずに訳文の品質推定を行う。

[1]http://www.statmt.org/wmt19/

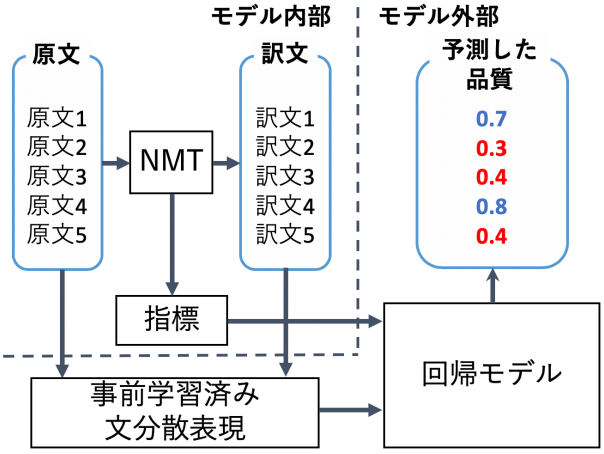

図 1: 提案手法

2　手法

概要図を以下の図 1 に示す。従来の訳文の自動評価手法では、訳文を人手で編集した文を用いて訓練させた評価モデルを用いるほか、訳文の正解となる参照訳を用いて品質を評価するなどしていた。そこで、我々は NMT の指標及び外部の事前訓練済み文分散表現を用いて、参照訳や人手で編集した訳文を用いずに機械翻訳の出力である訳文を自動評価する手法を提案する。

今回予測する品質は、機械翻訳の代表的な評価指標である BLEU とした。ただし、BLEU 計算において全ての n-gram カウントの初期値を 1 としてスムージングした。今回用いた指標を以下に示す。

文の尤度: 訳文 $\boldsymbol{y}(y_1, y_2, ..., y_l)$ を予測する際の対数尤度を文長 (今回は単語数) で正規化して用いる。

$$log\text{-}likelihood = \frac{\sum_{y_i \in \boldsymbol{y}} \log P(y_i)}{|\boldsymbol{y}|} \tag{1}$$

未知語: NMT から出力された訳文 $\boldsymbol{y}$ の中に未知語 $< unk >$ が含まれる割合を用いる。

$$unk = \frac{\sum_{y_i \in \boldsymbol{y}} Count(< unk >)}{|\boldsymbol{y}|} \tag{2}$$

原文、訳文の長さ: 原文 $\boldsymbol{x}$ および訳文 $\boldsymbol{y}$ の単語数を文の長さとして用いる。ここでは、文の長さをそのまま用いるのではなく、対数に変換して用いる。

$$len_{(src)} = \log(|\boldsymbol{x}|),\ len_{(out)} = \log(|\boldsymbol{y}|) \tag{3}$$

モデルの不確かさ: これまではモデルの入力・出力に着目してきたが、予測を行った NMT 自身も評価する必要があると考えている。これは予測の不確実性、モデルの不確かさなどと呼ばれており、モデルの出力を信頼するかどうかの基準となっている。例えば郵便番号をもとに分類を行うシステムにおいて、ある入力についてモデルが出力とともに高い不確かさを出した場合、その入力に対して人手で分類を行うことで、致命的なエラーを回避することができる。

通常、モデルの不確かさの予測には Bayesien 的な手法を用いる。ニューラルネットワークのモデルでも Bayesien Deep Neural Network という形でモデルの不確かさの予測が可能だが、最適化が難しいため、Gal ら [6] によって Monte Carlo Dropout (MC Dropout) が提案された。これは通常訓練時に行う dropout を生成時に行い、モデルの形を変えながら複数回同様の出力をさせ、その時の尤度を用いて最終的なモデルの不確かさを算出する手法である。モデルがある出力に対して自信があればモデルの構造が多少変わっても尤度に大きな変化は見られないが、自信がなければモデルの構造が多少変わるだけで尤度が大きく変わる。これにより、各出力に対するモデルの不確かさを算出することができる。

自然言語処理の分野において、MC Dropout は意味解析 [7] や逆翻訳で生成した擬似的なコーパスの重み付け [8] などに用いられ、性能向上に貢献している。そこで、今回我々はモデルの不確かさを訳文の品質推定に用いるため、MC Dropout を行った。モデルの不確かさの算出方法は期待値、分散、エントロピーなどが存在するが、今回は MC Dropout を逆翻訳に用いた先行研究において最も効果の高かった Combination of Expectation and Variance (CEV) を用いる [8]。ここで、入力する原文を $\boldsymbol{x}$ 、出力された訳文を $\boldsymbol{y}$ 、訳文生成における尤度を $P(\boldsymbol{y}|\boldsymbol{x})$ とする。また、MCDropout を用いてサンプリングする回数を K とし、今回の実験では $K = 20$ とした。MCDropout でサンプリングした K 個の尤度 $P_1(\boldsymbol{y}|\boldsymbol{x}), ..., P_K(\boldsymbol{y}|\boldsymbol{x})$ を用いて、期待値 $\mathbb{E}[P(\boldsymbol{y}|\boldsymbol{x})]$ 、分散 $\mathrm{Var}[P(\boldsymbol{y}|\boldsymbol{x})]$ はそれぞれ以下の式で近似できる。

$$\mathbb{E}[P(\boldsymbol{y}|\boldsymbol{x})] \approx \frac{1}{K}\sum_{k=1}^{K} P_k(\boldsymbol{y}|\boldsymbol{x}) \tag{4}$$

$$\mathrm{Var}[P(\boldsymbol{y}|\boldsymbol{x})] \approx \frac{1}{K}\sum_{k=1}^{K} P_k(\boldsymbol{y}|\boldsymbol{x})^2 - \mathbb{E}[P(\boldsymbol{y}|\boldsymbol{x})]^2 \tag{5}$$

式 (4, 5) を用いて、 CEV の計算式を以下の式 (6) で表す。

$$CEV = \Big(1 - \frac{\mathrm{Var}[P(\boldsymbol{y}|\boldsymbol{x})]}{\mathbb{E}[P(\boldsymbol{y}|\boldsymbol{x})]}\Big)^2 \tag{6}$$

外部の事前学習済み文分散表現: 外部の事前学習済みの文分散表現として、多言語 Universal Sentence Encoder (USE) [9] を用いた。USE とは、Transformer のエンコーダーまたは Deep Average Network を用いて文脈から文を予測するように訓練された文埋め込み表現である。また、多言語 USE とは、マルチタスクデュアルエンコーダーモデルと呼ばれる手法により、Transformer のエンコーダーで構成される USE を 2 つ用いて複数の言語で同時に訓練することで、様々な言語に対応した文埋め込み表現である。

今回我々は事前学習済みの多言語 USE で原文と訳文の文分散表現 $\boldsymbol{h_{src}}$、$\boldsymbol{h_{out}}$ を獲得し、cos 類似度を算出した。

$$USE\_sim = \frac{\boldsymbol{h_{src}} \cdot \boldsymbol{h_{out}}}{|\boldsymbol{h_{src}}||\boldsymbol{h_{out}}|} \tag{7}$$

上記の手法で得られた指標を基に、回帰モデルを用いて BLEU を予測した。

3 実験

今回の実験にあたり、コーパスは ASPEC-JE [10] を用いた。ASPEC は科学技術論文を翻訳して作成されたコーパスで、ASPEC-JE は日英翻訳のための対訳コーパスである。ASPEC の統計量を表 1 に示す。

NMT のモデルは fairseq [11] の Transformer を用いた。回帰モデルは線形回帰と XGBoost を用い、比

表 1: 使用した ASPEC コーパスの規模（単位は文数）

言語対	日本語-英語
訓練	999,680
開発	1,790
テスト	1,812

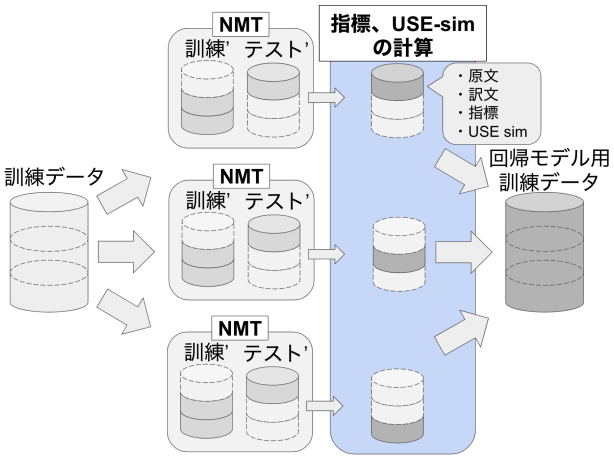

図 2: 回帰モデルの訓練データ獲得

較対象として我々の既存手法 [12] を採用した。回帰モデルの訓練データを集めるために、図 2 のような手法を用いた。NMT の訓練データを K 分割し、分割単位の 1 つをテストデータ、残り $K-1$ 個の分割データを訓練データとして、K 通りのデータセットを用いて K 個の NMT を訓練させる。各 NMT でテストデータを予測する際に、文の尤度、未知語といった指標を算出する。予測後、訳文と分割したテストデータの参照訳から BLEU を算出する。最終的に、指標が付与された K 個の分割されたテストデータを結合することで、今回用いる指標が付与された回帰モデルの訓練データを獲得することができる。今回は訓練データを 5 分割し、回帰モデルの訓練データを集めた。

上記の手法で収集した訓練データで回帰モデルを訓練させ、モデルで予測した BLEU は ASPEC の参照訳を用いて実際に算出した BLEU の値と Pearson の相関係数を計算し評価を行った。

4 結果・考察

実験結果を表 2 に示す。得られた結果から、XGBoost で予測した時が最も BLEU との相関が高いことがわかる。また、指標別で結果を比較すると、文の尤度や MCDropout によるモデルの不確かさ CEV を用いた際に相関が高く、未知語や外部の事前学習済み文分散表現である USE を用いた場合は相関が低くなっている。

表 2: 結果

手法	指標	相関係数
そのまま用いる [12]	対数尤度	0.323
	未知語	−0.092
	原文の長さ	−0.144
	訳文の長さ	0.155
	CEV	−0.225
	USE-sim	0.100
線形回帰	対数尤度	0.323
	未知語	0.091
	原文の長さ	0.144
	訳文の長さ	0.155
	CEV	0.225
	USE-sim	0.100
	All	0.379
XGBoost	対数尤度	0.333
	未知語	0.116
	原文の長さ	0.160
	訳文の長さ	0.167
	CEV	0.381
	USE-sim	0.138
	All	**0.393**

表 3: それぞれの指標が BLEU 予測に与える影響

手法	指標	相関係数
線形回帰	All	0.379
	− 対数尤度	**0.261**
	− 未知語	0.374
	− 原文の長さ	0.380
	− 訳文の長さ	0.374
	− CEV	0.369
	− USE-sim	0.376
XGBoost	All	0.393
	− 対数尤度	0.383
	− 未知語	0.390
	− 原文の長さ	0.386
	− 訳文の長さ	0.390
	− CEV	**0.381**
	− USE-sim	0.391

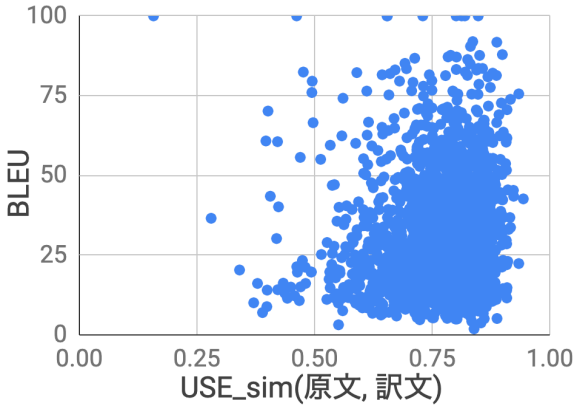

図 3: USE-sim と BLEU の関係

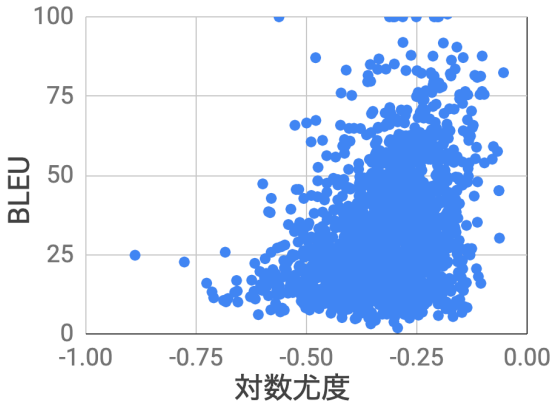

図 4: log-likelihood と BLEU の関係

上記の結果をもとに、各指標を 1 つ欠いた状態で同様に BLEU の推定を行い、それぞれの指標が予測に与える影響を調査した。その結果を表 3 に示す。得られた結果より、文の尤度またはモデルの不確かさ CEV を欠いた時に予測結果と BLEU との相関が低くなっている。したがって、この 2 つの指標が BLEU を予測する際に重要であることがわかる。

今回の実験で USE を用いて算出した原文と訳文の類似度 USE-sim が BLEU の推定において効果が低いことがわかった。この原因を調べるために、BLEU 推定において効果の低い USE-sim と、効果の高い対数尤度の分布を比較する。分布図をそれぞれ図 3、4 に示す。それぞれの分布を比較すると、USE-sim が図 3 のように全体的に原文と訳文の類似度が高いことを意味する 1.0 の近くに分布していることがわかる。このため、USE-sim は BLEU の推定において効果が低かったのであると考える。

5 おわりに

本研究では、NMT の指標を用いて評価時に参照訳を見ずに、また訓練時に訳文を人手で編集した文を用いずに訳文の品質を推定する手法を提案した。今回は訳文の品質として BLEU を用い、線形回帰と XGBoost で予測を行った。実験の結果より、XGBoost を用いた場合が最も BLEU との相関が高く、指標別に見ると文の尤度とモデルの不確かさの 2 つが BLEU 予測において重要であることがわかった。

謝辞 本研究は、平成 29–31 年学術研究助成基金助成金 挑戦的研究 (萌芽) 課題番号 17K18481 の助成を受けています。

参考文献

[1] Kishore Papineni, Salim Roukos, Todd Ward and Wei-Jing Zhu. BLEU: A Method for Automatic Evaluation of Machine Translation. In *Proc.* of ACL, pp. 311–318, 2002.

[2] Lucia Specia, Gustavo Paetzold and Carolina Scarton. Multi-level Translation Quality Prediction with QuEst++. In *Proc.* of ACL-IJCNLP, pp.115–120, 2015.

[3] Julia Ive, Frédéric Blain and Lucia Specia. deepQuest: A Framework for Neural-based Quality Estimation. In *Proc.* of COLING, pp. 3146–3157, 2018.

[4] Jiayi Wang, Kai Fan, Bo Li, Fengming Zhou, Boxing Chen, Yangbin Shi and Luo Si. Alibaba Submission for WMT18 Qualiti Estimation Task. In *Proc.* of WMT, pp. 809–815, 2018.

[5] Fabio Kepler, Jonay Trénous, Marcos Treviso, Miguel Vera and André F. T. Martins. OpenKiwi: An Open Source Framework for Quality Estimation. In *Proc.* of WMT, pp. 117–122, 2019.

[6] Yarin Gal and Zoubin Ghahramani. Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning. In *Proc.* of ICML, pp. 1050–1059, 2016.

[7] Li Dong, Chris Quirk and Mirella Lapata. Confidence Modeling for Neural Semantic Parsing. In *Proc.* of ACL, pp. 743–753, 2018.

[8] Shuo Wang, Yang Liu, Chao Wang, Huanbo Luan and Maosong Sun. Improving Back-Translation with Uncertainty-based Confidence Estimation. In *Proc.* of EMNLP-IJCNLP, pp. 791–802, 2019.

[9] Muthu Chidambaram, Yinfei Yang, Daniel Cer, Steve Yuan, Yunhsuan Sung, Brian Strope and Ray Kurzweil. Learning Cross-Lingual Sentence Representations via a Multi-task Dual-Encoder Model. In *Proc.* of RepL4NLP, pp. 250–259, 2019.

[10] Toshiaki Nakazawa, Manabu Yaguchi, Kiyotaka Uchimoto, Masao Utiyama, Eiichiro Sumita, Sadao Kurohashi and Hitoshi Isahara. ASPEC: Asian Scientific Paper Excerpt Corpus. In *Proc.* of LREC, pp. 2204–2208, 2016.

[11] Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier and Michael Auli. fairseq: A Fast, Extensible Toolkit for Sequence Modeling. In *Proc.* of NAACL-HLT, pp. 48–53, 2019.

[12] Taichi Aida and Kazuhide Yamamoto. Confidence Modeling for Neural Machine Translation. In *Proc.* of IALP, pp. 349–354, 2019.