公開

研 究 終 了 報 告 書

「次世代言語生成のための生成文評価基盤」

研究期間: 2018 年 10 月～2022 年 3 月
研 究 者: 須藤 克仁

1. 研究のねらい

深層学習技術に基づく自然言語処理技術の急速な発展により、機械翻訳や自動要約、画像キャプション生成等に代表されるコンピュータによる言語生成の性能が高まり、広く用いられるようになってきている。人間の情報アクセスの手段として、自然言語は特別な訓練を必要とせずに利用可能であり、膨大な情報を検索・閲覧・活用することが求められる今後の社会システムにおいて、言語生成技術は人間の能力の活用や拡大に重要な役割を果たすと考えられる。

現在の言語生成技術により生成されることばは人間と遜色ないと言われることもある一方で、全く意図しない内容や本来言及すべきものと矛盾する内容を含むことがあるなど、深刻な誤解を招くようなことばが生成されることが問題視されている。この問題の背景には、従来の言語生成研究で活用されてきた人手評価・自動評価の方法が、このような問題の識別ができるように設計されておらず、少しのことばの違いによって起こる意味の大きな違いを考慮できていなかったことが挙げられる。特に、生成文を参照文と比較して行う自動評価の仕組みは、システム改良の効果測定が簡便かつ高速に実施できる特徴を活かし、機械翻訳を代表に言語生成技術の着実な発展の礎となってきたが、従来の自動評価では表層的な一致度の影響が大きく、表層上は小さな変化だが意味が大きく変わるようなものに対して非常に脆弱であることが、平均的な言語生成の性能向上により顕在化してきたと見ることができる。

本研究では、次世代の言語生成技術の研究開発を支える、高度にことばの意味を考慮して生成文の評価を行う言語生成評価の新しい枠組みの考案と、それに基づく人手評価・自動評価の方法の確立を目指した。そのために、前記の問題を解決すべく、生成文の言わんとすることが理解できるか否かを問う「文意解釈性」と、生成文の伝達内容が別途評価用に用意した参照文と同一と言えるかを問う「文意正確性」の二つの観点による人手評価・自動評価の検討を行った。これにより、「一見すると正しそうに見えるが重要な箇所で肯定否定の反転があり深刻な誤解を招く生成文」、「文としては整っているが参照文とまったく無関係な内容の生成文」、「内容が解釈できない支離滅裂な文」、などの言語生成において生じうる深刻な誤りを適切に識別し、言語生成技術の改善に繋げることを目標とする。

2. 研究成果

（1）概要

本研究では、①人手評価と②自動評価の大きく二つの課題に取り組んだ。生成文の誤りにより意図していた内容が正確に伝達できない可能性を考慮すべく、生成文がことばとして解釈可能であるかを問う「文意解釈性」と、生成文の伝達内容が参照文のそれと合致するかどうかを問う「文意正確性」の二つの観点で生成文の人手評価・自動評価を行うことを目標に研究を進めた。

①人手評価については、研究開始当初に「文意解釈性」5 種類、「文意正確性」7 種類の分

類を定義し、英日・日英翻訳、様々な言語から英語への翻訳、データから英語の文生成などの各種共通タスクデータを対象に人手評価作業を実施して評価データを蓄積、公開した。当初計画では 10,000 文程度としていたが、最終的には英日・日英翻訳で各 3,000 文程度、他言語から英語への翻訳で 9,000 文程度、英語文生成で 3,000 文程度の評価データ(合計で 18,000 文程度)を構築した(評価者数は原則 3 名)。また、さらなる検討のために人為的に誤りを加えた英語のコーパスを作成し、評価者 1 名ながら 7,000 文×3 パターンの人工負例評価データを構築できた。

②自動評価については、上記人手評価データの評価者間一致度があまり高くなく、安定した性能を得るに至らなかったことから、本研究の着眼点を重視しつつ従来の自動評価の枠組みを拡張する形で展開することとなった。対象としてはデータが十分に入手可能な機械翻訳評価に注力し、まずは機械翻訳評価時に機械翻訳文と参照訳の比較だけでなく、多言語事前学習モデルを用いた原文と機械翻訳文の比較を加えることで評価性能(人手評価との相関)が改善すること、特に訳質の低い機械翻訳文で優れていることを示した。さらに、人工負例を用いたファインチューニングを加えることでさらに評価性能が改善することを示し、提案法は 2021 年の国際会議における機械翻訳評価共通タスクでトップグループに入る性能を達成した。

その他では、コンピュータの言語生成の評価と、人間の作文の評価の課題の類似性に着目し、様々なことばの評価を包括的に議論するためのワークショップ(国内)を開催した。

(2)詳細

研究テーマ A「深刻な誤訳の識別に向けた段階型機械翻訳評価と評価データセット構築」

本研究では評価の観点として、評価対象文の内容が解釈できるかどうかを評価する「文意解釈性」と、評価対象文の内容が参照文の内容と一致しているかを評価する「文意正確性」の二つを提案した。それぞれの観点では、生成文として許容可能な文をその良し悪しにより 3 種類、許容できない文を文意解釈性では 2 種類、文意正確性では 4 種類に分類した評価基準として設計した。

この評価基準に基づき、以下のデータセットについて人手評価を実施した。

・WAT 2018 英日・日英翻訳データ 各 2,000 文(400 文×5 システム分) 評価者 3 名
・WMT 2015-2017 多言語から英語への翻訳データ 9,280 文 評価者 3 名【公開済み】
・WMT 2020 英日・日英翻訳データ 各 1,000 文 評価者 10 名
・WebNLG 英文生成データ 3,463 文 評価者 3 名

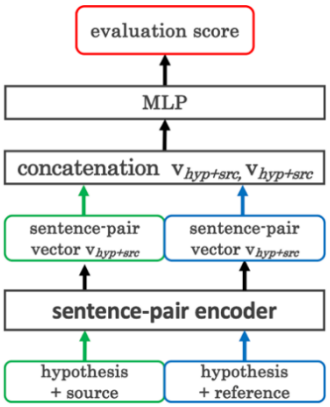
このうち WMT2015-2017 データについて詳細な分析を行った。従来の数値型人手評価との比較では、本研究の文意解釈性と数値型人手評価はほぼ比例関係にあることが示されたが、文意正確性では無関係な訳文に対する平均評価値が最も低く、矛盾を含んでいたり大きな誤解を招きかねない誤訳については、より高い評価値が付されていることが明らかになった。実際に機械翻訳を利用する際には一見して気づきにくい矛盾や誤解を誘発する誤りの存在は深刻な問題であり、従来の機械翻訳評価の仕組みではこうした問題に十分に着目できていないことが示唆される結果となった。しかしながら、この WMT 評価データの評価者間の一致度(Kappa 値)は 0.3 前後と低い値に留まっており、このデータを利用して学習した

自動評価モデルでも許容できる翻訳と許容できない翻訳の識別誤りの影響で正解率は 60%程度であり、非常に難しい問題であることが明らかになった。(論文発表 3)

さらなる検証のために、評価者数を 10 名としてデータを 1,000 文と小規模にした人手評価データ作成を行ったが、評価者間の違いは様々であり、評価を一つの分類に限定するのではなく、こうした評価者間の「揺れ」を踏まえた人手評価の解釈や自動評価への応用の必要性を示唆する結果となった。

研究テーマ B1「参照文に加え原文との意味的類似性を考慮する機械翻訳自動評価法 C-SPEC」

機械翻訳の自動評価では、BERT 等の大規模なテキストデータで事前学習された言語モデルを利用され、高性能な(人手評価との相関が高い)評価ができるようになってきたが、通常機械翻訳の自動評価では参照文 1 文との比較に留まるため、訳語の選択の違いが評価値に与える影響が大きかった。複数の参照文を用意するのは通常困難であるため、本研究では翻訳の原文を訳文と比較して評価の追加材料とすることを考え、多言語に対応した事前学習型言語モデル XLM や XLMRoBERTa を利用した新しい機械翻訳自動評価法 C-SPEC (Cross-lingual Sentence Pair Embedding Concatenation) を提案した。C-SPEC は図に示すように、機械翻訳文と原文のペアを符号化したベクトルと、機械翻訳文と参照訳のペアを符号化したベクトルを結合したベクトルを用いた回帰モデルにより評価値の予測を行う。C-SPEC は表に示す通り機械翻訳文と参照文のペアのみを利用する従来法 (BERT Regressor) よりも高い人手評価との相関を示すことが確認できた。(論文発表 1)



手法＼言語対	独英	露英	土英	中英	平均
文 BLEU	0.432	0.484	0.538	0.512	0.484
従来法(BERT Regressor)	0.729	0.757	0.770	0.702	0.740
提案法 (C-SPEC)	**0.753**	**0.795**	**0.771**	**0.763**	**0.771**

研究テーマ B2「人工負例を用いて追加学習する機械翻訳自動評価法 C-SPECpn」

本研究で着目している言語生成の深刻な誤りは、実際の共通タスクなどのデータにおける出現頻度はさほど多くないことから、前記の実験データにおいては学習・評価の両面で十分に考慮できているとは言い難い。そこで、「単語属性変換」と呼ばれるテキスト変換技術を用いて、参照文の単語のジェンダーを変換したり、反意語への変換を行ったりして、参照文に極めて近いが文意が大きく変わってしまうような文(人工負例)を生成し、C-SPEC の追加学習を行った C-SPECpn (pn は人工負例 pseudo-negative の意) を提案した。

人工負例の文の生成は単語属性変換によって可能であるものの、自動評価の学習データに必要な評価値は不明であるため、本研究では C-SPEC により符号化されたベクトルを用

いた別の補助タスク（評価対象訳文が①原文と一致する、②参照訳と一致する、③人工負例である）による追加学習を行うことで、符号化器のファインチューニングを行った。C-SPECpn は元の C-SPEC より人手評価との高い相関を示すことが示され、また 2021 年の国際会議 WMT における機械翻訳評価尺度の共通タスクにおいて、「人手翻訳を含むニュース分野の翻訳評価サブタスク」で 1 位となった他、有意差のない 1 位グループに加わったサブタスク数で 1 位となり、トップグループに入る結果を残した。（論文発表 2）

ワークショップ開催

本研究ではコンピュータが生成する文の評価に焦点を当てていたが、同じような評価は人間が書いた文章に対しても可能なのではないか、と考え、コンピュータの言語生成と人間の作文の評価は何が違い、何は共有できるのか、ということを広く議論するための場として、2021 年 3 月に「文章の評価と品質推定～人間・機械の「作文」の巧拙をどう見極めるか～」と題したワークショップを言語処理学会年次大会で開催した。さきがけ「社会デザイン」1 期生の高村氏の招待講演等もあり、150 名程度の参加があり大変盛況であった。また、2022 年 3 月の言語処理学会年次大会では、同様のテーマでのテーマセッションを開催し、13 件の発表を集めることができ、関連する研究の幅を拡げることに貢献できたと考える。

3. 今後の展開

研究としての今後の重要な展開としては、参照文に頼らない形での人手評価・自動評価への拡張を考えている。人手評価でのアノテータ間一致度に参照文が与える影響や、参照文の作成が難しい画像キャプショニングの評価への応用を考えたとき、参照文でなく元の文や画像などを確認して行う人手評価が重要である。こうした試みは機械翻訳の人手評価でも注目され始めており、今後の発展が期待される。一方で参照文がない自動評価は原理的に困難であると考えられ、高品質な参照文の整備が当面の方向性であろうと考えられる。さらに、単文の評価でなく文脈を考慮することや、小論文のように文の内容の定めがないものに対する評価など、より挑戦的な研究課題への展開が今後必要になるであろう。

社会実装という面では本研究で提案した機械翻訳の自動評価はかなり実用化に近いフェーズにあると考えることができる。ただし現在の機械学習に基づく自動評価の場合は対象と類似した文体・分野の学習データが必要であることもあり、学習データ整備のための費用やデータ提供者の存在が不可欠と言える。予備検討やデータ整備にかかる期間や費用（1 年程度＋数百万円）の確保ができれば、具体的な応用に繋げられるのではないかと考えている。さらに広範な言語生成への応用についてはまだ基礎研究が不足している段階にあり、実用化までには 4-5 年の期間が必要ではないかと考える。

4. 自己評価

本研究の目的は品質が向上したと言われる言語生成技術が依然として有する重大な問題である、深刻な誤りを的確に識別し評価することであった。自動評価については、既存研究における大規模言語モデルを用いて意味の違いに着目する手法を発展させ、機械翻訳評価において発表当時最高水準の評価性能を示したこと、そして特に訳質の低い機械翻訳結果における有効性が示せたこと、さらに、人工負例を用いた更なる拡張により共通タスクでトップクラス

の評価性能を達成できたことで一定程度の目的達成ができたと考えている。しかしながら、本研究で提案する枠組みに基づく人手評価での評価者間一致度が十分でなかったこと、そのために自動評価の学習への有用性が示せなかったこと、さらに機械翻訳以外の領域への展開と言語生成学習への活用に踏み込めなかったこと、が目標不達の点である。人手評価の困難さは文意に基づく本質的な言語生成評価が今後発展するための大きな障害となっており、人間の自然言語解釈の定量化の困難さを物語るものと言え、今後の発展研究に向けた課題発掘に繋がったと考えることもできるかもしれない。また、変化は小さいが深刻な誤りを生じているような機械翻訳文の評価を適切に行うことの重要性を提起し、そうした評価が国際会議の共通タスクにも導入されたことは、今後の言語生成評価技術にとっての重要な一歩であると考えている。

研究の進め方については人手評価を研究補助者(アノテータ)による基礎データ作成と外注による大規模データ整備により実施するとともに、学生を研究補助者としてアイデアの実装や実験に従事してもらうことで様々な取り組みを進めることができた。また計算環境整備の費用を十分に確保できたこと、さらに miniRAIDEN 環境を利用できたことで大規模なモデルを用いた自動評価手法の詳細な検証ができたことは、自動評価研究の進捗に非常に有用であった。人手評価データ構築・公開については論文の国際会議採択まで長い時間を要しその後の展開への影響が大きかった。

本研究は今後さらにその重要性を増すと考えられる言語生成技術の基礎となる、ことばの評価に対するものであり、本研究の自動評価技術は、まず機械翻訳の評価において活用可能であると考えられる。日本国内では英日・日英双方向の翻訳が最も需要が多い対象であり、自動評価モデルのファインチューニングを行うため英日・日英翻訳のデータの整備ができれば活用できる。データ整備にあたっては従来型の人手評価を当初は用いることになるが、本研究の知見を活用し将来はより詳細な評価ができるような人手評価データ作成と自動評価の実現、ひいては機械翻訳の改善の加速とビジネスの拡大、そして言語間コミュニケーションの円滑化に繋がると期待する。領域の目標である新しい社会システムデザインへの波及という意味では、機械翻訳というアプリケーションのみならず非言語データからの言語生成への展開が求められるところであり、正解の可能性が多岐にわたる問題への発展が今後求められるところである。

5. 主な研究成果リスト

(1)代表的な論文(原著論文)発表

研究期間累積件数:4件

1. Kosuke Takahashi, Katsuhito Sudoh, Satoshi Nakamura. Automatic Machine Translation Evaluation using a Source and Reference Sentence with a Cross-lingual Language Model. 自然言語処理. 2022, 29 (1), 採録決定
機械翻訳の自動評価において、通常行われる機械翻訳文と参照文の比較だけでなく、多言語事前学習モデルを用いた機械翻訳文と原文の比較も加えた手法 C-SPEC を考案した。特に訳質が低い機械翻訳文の評価において、原文との比較を行わない従来手法より優れた評価性能(人手評価との相関)を示した。

2. Kosuke Takahashi, Yoichi Ishibashi, Katsuhito Sudoh, Satoshi Nakamura. Multilingual Machine Translation Evaluation Metrics Fine-tuned on Pseudo-Negative Examples for WMT 2021 Metrics Task, Proceedings of the 6th Conference on Machine Translation. 2021, pp.1054-1057.
C-SPEC の評価モデルの学習において、通常の人手評価付き機械翻訳評価データの学習に加え、単語のジェンダーの変換や反義語への変換を行う「単語属性変換」の技術を用いて自動的に生成した負例を追加学習させた自動評価手法 C-SPEC_pn を考案し、共通タスクにおいてトップクラスの評価性能を示した。
3. Katsuhito Sudoh, Kosuke Takahashi, Satoshi Nakamura. *Is This Translation Error Critical?*: Classification-Based Human and Automatic Machine Translation Evaluation Focusing on Critical Errors. Proceedings of the Workshop on Human Evaluation of NLP Systems. 2021. pp.46-55.
機械翻訳評価における深刻な誤りの識別に向けて、文意解釈性と文意正確性の二つの観点で、誤りの種類に応じた分類型評価手法を提案し、それに基づく人手評価データの作成し、そのデータを用いた自動評価の結果を示した。また、論文の出版に合わせて人手評価データを研究用に公開した。

（2）特許出願

研究期間全出願件数：0 件（特許公開前のものも含む）

（3）その他の成果（主要な学会発表、受賞、著作物、プレスリリース等）

1. 高橋洸丞, 須藤克仁, 中村哲. 言語横断な言語モデルによる原言語情報を活用した機械翻訳評価. 情報処理学会第 241 回自然言語処理研究会. 2019.08.
2. 高橋洸丞, 須藤克仁, 中村哲. システム訳文のみを用いた自動評価との比較による機械翻訳自動評価の分析. 言語処理学会第 27 回年次大会. 2021.03.
3. 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 27 回年次大会. 2021.03.
4. 須藤克仁, 高橋洸丞, 中村哲. 深刻な誤訳の識別に向けた分類型翻訳評価データセットの構築. 言語処理学会第 27 回年次大会. 2021.03.
5. 高橋洸丞, 石橋陽一, 須藤克仁, 中村哲. 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. 言語処理学会第 28 回年次大会. 2022.03.
6. 石橋陽一, 横井祥, 須藤克仁, 中村哲. 線型部分空間に基づく学習済み単語埋め込み空間上の集合演算. 言語処理学会第 28 回年次大会 2022.03.