ライティング支援を想定した情報補完型生成

伊藤 拓海 [1,4*] 栗林 樹生 [1,4*] 小林 隼人 [2,3] 鈴木 潤 [1,2] 乾 健太郎 [1,2]
[1] 東北大学 [2] 理化学研究所 AIP センター [3] ヤフー株式会社 [4]Langsmith 株式会社
{t-ito, kuribayashi, jun.suzuki, inui}@ecei.tohoku.ac.jp hakobaya@yahoo-corp.jp

1 はじめに

自然言語処理の分野では，機械翻訳や要約，キャプション生成など多くの生成問題が注目を浴びてきた．これらの問題の多くは，入力の情報を保持して生成する問題 (情報保存型生成) や，入力の情報を削減する生成問題 (情報削減型生成) と捉えることができる．例えば機械翻訳は，翻訳元の文の意味情報を保持したまま翻訳先の言語に変えて出力するという情報保存型生成，要約は入力情報の重要な箇所以外を除去する情報削減型生成とみなすことができる．本研究ではこれらの生成問題とは対照的な性質を持つ，情報補完型生成問題を考える (図 1)．

情報補完型生成は，画像処理の分野では劣化した画像から元の画像を復元するタスク *image inpainting* [2, etc.] など多数の研究が見られるが，言語処理の分野では該当する研究は多くない．しかしながら，作文途中の不完全な下書きの文から目的に合った完全な文の候補を自動生成する情報補完型生成モデルがあれば，有効な作文支援になることが期待できる．こうした作文支援は，本稿で想定するような第二言語の作文だけでなく，企業でのビジネス文書の作成など，幅広い応用が考えられる．現在盛んに研究されている文法誤り訂正も情報補完型生成の特殊な例とみなせる [9]．こうした応用を想定しながら情報補完型生成という新しい言語生成研究を展開していくためには，少なくとも次の 3 つの問題を検討する必要がある．

第 1 の問いは，情報補完型生成という課題をどのように定義するのが有用かという問題である．本研究では，仮の定義として，特定の目的やスタイルに合った文の集合を目的言語として指定し，目的言語に含まれない文からそれに意味的に近い目的言語内の文を生成する課題を考える．本稿で報告する実験では，ACL Anthology コーパス [1] を目的言語のサンプル文集合として，第二言語学習者が書く不完全な文から，それと意味的に近く，より国際会議論文のスタイルに合った流暢な文を生成する課題に取り組んだ．

第 2 の問いは，出力をどのように評価すべきかである．情報補完型生成では，適切と見なせる生成結果の解空間は広いと考えられるため，生成結果の評価方法を検討する必要がある．本稿では，出力のリファレンスとなる文（正解例）を目的言語からサンプルし，各リファレンス文に対応する想定入力文をクラウドソーシングによって作成することによって評価用データセットを構築する，新しいデータ構築方法を提案する．実験では，実際に ACL Anthology コーパスからクラウドソーシングで評価用データセットを構築し，ベースラインとなるいくつかの生成モデルの出力結果を複数の評価指標に基づいて多角的に評価

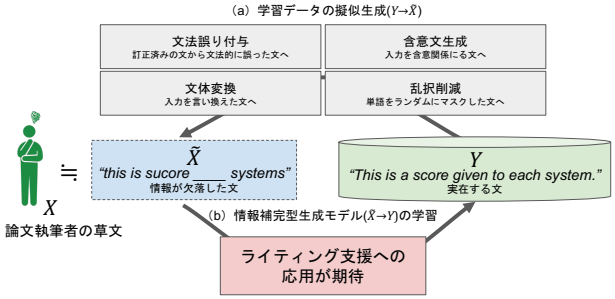

図1: 本研究の概要

することを試みた．

第 3 の問いは，情報補完型生成モデルの構築方法である．情報補完型生成の目標は応用の目的に依存するため，個別の目的ごとに高いコストを支払って教師データを作るアプローチは現実的でない．これに対し本稿では，情報削減型生成や情報保存型生成における近年の成果を活用し，完全な文から疑似的に不完全な文を生成することによって大規模な教師データを自動生成し，情報補完型生成モデルを訓練するアプローチを提案する．

本研究の貢献は以下の通りである．

- 言語処理分野においてこれまで取り組みの少なかった情報補完型生成を新しい言語生成課題として提案する．
- 論文執筆支援という応用を見据え，評価用データの作成方法と評価指標を提案する．
- 学習データの自動生成によって情報補完型生成モデルを訓練できる可能性があることを実験によって示す．

2 クラウドソーシングを用いた評価用データ作成

情報補完型生成タスクでは，不完全な文を入力として完全な文の生成を目指す．本研究では英語論文執筆支援への応用を見据え，完全な文を ACL Anthology 上の論文に出現する文，不完全な文を非英語ネイティブが書いた草文とする．

不完全な草文についてはクラウドソーシング[*1]を用いて収集した．クラウドソーシングを用いた利点としては，データのスケールのしやすさが挙げられる．

2.1 クラウドソーシングのタスク設定:

英語論文の和訳[*2]から英文に翻訳するタスクを 15 分で 3 問解いてもらった．Aizawa ら [1] の ACL Anthology コーパスから，以下の条件を満たす文を自動で抽出し，PDF からの抽出ミスのある文を人手で取り除いた．

- ACL 2018 に採択された論文に出現する文である．
- 70 文字より長く 120 文字未満である．

[*] 本研究の主要な貢献は第一著者と第二著者によるものであり，両者の貢献は同等である．

[*1] https://crowdsourcing.yahoo.co.jp/

[*2] Google 翻訳（https://cloud.google.com/translate/）を用いた．

表1: クラウドワーカーの評価基準.

評価観点	加減点
作業時間が 2 分以下である	即不採用
すべての回答が 3 単語以下である	即不採用
“.” か “?” で終わっている回答がない	即不採用
全く同じ回答が含まれている	即不採用
回答に日本語が含まれている	即不採用
英語と認識できる回答がない	即不採用
3 単語以下の回答がある	-2
3 種類以下の単語しか使っていない回答がある	-2
自動翻訳サービスの結果と編集距離が 30 以下	-0.5/回答
自動翻訳サービスの結果と編集距離が 20 以下	-1/回答
自動翻訳サービスの結果と編集距離が 10 以下	-3/回答
全ての回答が “.” か “?” で終わっている	+1
マスクを一回以上使った	+1
すべての文が英語と認識できる	+1

- Aizawa ら [1] が予め置換した数学記号や引用のための特殊トークンや，URL，ギリシャ文字，特殊な記号などを含んでいない．

本研究では数式や引用などの生成は研究の対象外とする．翻訳の際にうまく訳せない箇所については，文中に特殊トークン (“___mask___”) を挿入してもらった．

データの質の担保**:** Yahoo! クラウドソーシングではチェック設問を用いてワーカーの質をコントロールすることができる．しかしながら，Yahoo! クラウドソーシングで提供されているチェック方法では予め用意した正解との完全一致による判断しかできない．そのため，今回のような自由記述のタスクではチェック設問を作るのが困難である．そこで，Hanawa ら [4] の方法を参考に，Yahoo! クラウドソーシングから本研究で構築した外部ページに誘導した．ワーカーが訳した 3 文に対して表 1のような観点で評価し，得点が 0 以上である場合その回答を採用した．英語と認識できるかの判断には Spacy の言語判定[*3]を，自動翻訳サービスの結果には Google 翻訳とエキサイト翻訳[*4]の結果を用いた．自動翻訳サービスや自動文法チェックサービスなどが使用されないよう，外部ページではコピー・アンド・ペーストやスペルチェックの機能を制限した．

2.2 クラウドソーシングで獲得したデータの例

本実験では，1125 文中の不完全な草文を収集した．この内 125 文を分析用に分け，どのような例が含まれているかを調査した．また，クラウドワーカーが書いた 45% の文に特殊トークン “___mask___” が含まれていた．実際に獲得されたデータとそこで起きている現象の例を以下に示す．

(1) 文法誤りが生じている．
原文: For **example**, the second candidate **is given** lowest scores by all the three models.
和訳: たとえば、2 番目の候補には 3 つのモデルすべての中で最も低いスコアが与えられます。
草文: For **exemple**, second **gived** lowest score in three models.

(2) 内容を補足する余地がある．
原文: **Table 1: Relative** use of CPU time and peak memory use **(per container)** for various tasks in the NLP **processing** pipeline.
和訳: 表 1：NLP 処理パイプラインのさまざまなタスクにおける CPU 時間とピークメモリ使用量（コンテナごと）の相対的な使用量
草文: The amount of ___**mask**___ use of CPU time and peak memoly in the several task of NLP ___**mask**___ pipe line.

(3) 文体や表現が不適切．
原文: If BLEU score was not increased in ten **consecutive** evaluations，then training was **stopped**.
和訳: 10 回の連続評価で BLEU スコアが増加しなかった場合、トレーニングは中止されました。
草文: The cace of blue score don’t increase on 10 **continue** evalution，traning is **canceled**.

(1) の例のように，自動和訳結果がやや不適切になっている文もあった．ワーカーに提示する和訳の質を上げることは質の高いデータを獲得するための課題でもある．

これらの草文のような不完全な文から適切な文にどれだけ補完できたかを評価する指標については次章で考察する．

3 評価指標

不完全な文から情報を補完して文を生成する場合，適切な答えは 1 つに定まらない．したがって，特定のリファレンスと比較した評価では，情報補完型生成モデルの性質を十分に評価できない可能性がある．本実験ではリファレンスとの比較の他に，リファレンスを必要としないコストの低い評価指標も複数用い，多角的な評価を行うことを考える．

2.2節で示したような現象を捉えるため本研究では以下の評価指標を用いた．

1. 妥当性: リファレンスとの BLEU を評価尺度とした．システムへの入力とリファレンス間の BLEU と，モデルの生成結果とリファレンス間の BLEU を求め，生成結果の BLEU が相対的に高くなることで自分たちのシステムが妥当な提案をしていることを示す．

2. 文法性: Napoles ら [8]，浅野ら [15] を参考に，LanguageTool3.2[*5]を用いたリファレンスレス評価指標を用いる．$1 - \frac{\text{誤りの数}}{\text{文のトークン数}}$ によって，文単位の文法性スコアを求めた．システムへの入力と生成結果の文法性を求め，生成結果の文法性スコアが相対的に高くなることで，生成モデルが文法性を改善していることを示す．

3. 論文ドメインでの自然さ: 言語モデルのパープレキシティを用いる．ACL Anthology コーパス [1] から獲得した 5-gram 言語モデルを用いる[*6]．システムへの入力と生成結果のパープレキシティを求め，出力のパープレキシティが相対的に低くなることで，生成モデルが論文ドメインに適した表現を提案していることを示す．

4. 入力の意味内容の保持度: 入力とモデルの生成結果間の METEOR を評価指標とした．入力条件を反映した生成をし

[*3] https://github.com/nickdavidhaynes/spacy-cld
[*4] https://www.excite.co.jp/world/
[*5] https://github.com/languagetool-org/languagetool/releases/tag/v3.2
[*6] https://github.com/kpu/kenlm

表2: 疑似的に生成された不完全な文の例.

使用モデル	原文	疑似生成した不完全な文
文法誤り付与	it is not **surprising** the randompolicy **has** the worst performance .	it is not **surprisingly** this randompolicy **have** the worst performing .
文体変換	we **observe** a similar trend on larger datasets .	we **see** a similar trend on the larger data .
含意文生成	the middle layer includes the core techniques **that are supported by the basic nlp techniques** .	the middle layer is included in the core techniques .
乱択削減	lower perplexity indicates a **better** model .	lower perplexity indicates a **___mask___** model .

表3: 複数の評価尺度による各モデルの性能を示す.

データ	妥当性 (BLEU)	文法性	自然さ (PPL)	条件の反映 (METEOR)	単語数	文字数
入力	12.1	88.0	1435.0	100.0	15.1	80.9
GEC モデル	12.3	91.1	940.6	62.1	15.2	81.3
LSTM	14.9	93.6	444.9	49.2	15.7	85.3
Transformer	16.5	94.5	284.0	43.2	16.4	90.1
リファレンス	100.0	94.9	9.11	23.2	17.1	98.4

た場合，入力条件と生成結果間の METEOR スコアは高くなるという仮定に基づき，この値が高いことを期待する．文法性や自然さの評価指標のみでは，入力情報を無視して ACL Anthology にありそうな文を適当に生成するだけで高い評価を得られてしまう．本評価指標を用いることで，ユーザの条件を反映した出力になっているかを評価する．

5. 情報の増加度: システムへの入力の文長とシステムからの出力の文長を比較し，相対的に出力が長くなっていることで情報の増えた生成が行われていることを示す．

4 学習データの疑似生成

論文に実際に出現する完全な文のみが手に入る設定でタスクを解く．2章で作成したデータは評価時のみに用いる．完全な文は ACL Anthology コーパス [1] などから大量に獲得でき，一方で論文ドメインの不完全なデータを大量に入手することはコストが高く困難であるため，このような設定は現実的であると考えられる．本研究では生成モデルを用いて完全な文から疑似的に不完全な文を生成する (図 1(a))．2.2節での分析を考慮し，不完全な文の自動生成には以下の生成モデルを用いた．

文法誤り付与モデル: Lang-8 [6] と JFLEG [7] を用い，訂正文から誤り文を生成するモデルを学習した．

文体変換モデル: PARANMT-5M [12] を用い，意味は保存しているが言い方が異なる (論文に適切な文体とは限らない) 文を生成するモデルを学習した．言い換えモデルは一般ドメインのデータで学習しているため，本モデルに言い換えを生成させることで論文に則さない表現に変えることを期待している．

含意文生成モデル: SNLI [3] と MultiNLI [13] を用い，含意関係にある文を生成するモデルを学習した．含意文生成モデルは情報を保存または削減する生成をすることが期待できる．このモデルによって2.2節のような “___mask___” がない形での情報削減を期待する．

乱択削減モデル: 単語をランダムに特殊トークン (“___mask___”) に置き換える．

乱択削減モデル以外のモデルには Luong ら [5][*6]の LSTM モデルを用いた． より多様な疑似不完全文を生成するために，デコード時に Ziang ら [14] の提案したランダムノイズと Vijayakumar ら [11] が提案したダイバース・ビームサーチを同時に用いた．また，4 つの不完全文生成モデルを段階的 (文法誤りを付与した後にマスクをするなど) に用いることで，複数の種類の不完全な現象が混在した文を作成した．ACL Anthology コーパスから疑似的な不完全な文を生成し，合計 5000 万文の学習用データを用意した．

疑似的に生成された不完全な文の例を表4に示す．4 つのモデルの生成例から，狙った性質を持つ不完全な文が生成できていると示唆される．

[*6] https://github.com/pytorch/fairseq

5 情報補完型生成モデルの学習

疑似生成した 5000 万文の不完全な文と完全な文のペアを用いて，情報補完型生成モデルを学習した (図 1(b))．モデルとしては，LSTM と Vaswani ら [10] の Tarnsformer[*6]を用いた．また，LSTM を用いて Lang-8 [6] と JFLEG [7] で訓練した文法誤り訂正モデル (GEC) モデルをベースラインとした．GEC モデルについては，“___mask___” を除いて入力した．

6 結果

表 3に，定量評価の結果を示す．妥当性スコアはリファレンスとの BLEU を，条件の反映スコアは入力との METEOR スコアを意味する．全ての評価指標において，GEC，LSTM，Transformer モデルのスコアが入力データのスコアと比較して，リファレンススコアに近づく傾向にあった．このことから，本研究で検討したモデルが論文執筆の目的に則した生成をするモデルであることが分かる．Lang8 コーパスで学習した GEC モデルと擬似不完全文を用いて学習したモデルを比較すると，後者の性能が相対的に高いことが示唆された．前者は論文データを学習に用いていないため，論文ドメインにおいて頑健な生成や訂正ができないと考えられる．

条件の反映スコアに関しては，リファレンスのスコアが非常に低い．浅野ら [15] の分析においても，文意の保存スコアと訂正の良さは逆相関する傾向が出ており，本研究でも同様の現象が起きた．METEOR が条件の反映の評価指標として不適切であること，本タスクでは草文からの大きな書き換えが発生していること，クラウドソーシングで作成したテスト用のデータセット設計に課題があることなどの可能性が示唆される．また，入力とリファレンスに関して文長を比較すると，単語数で 2 単

表4: 情報補完型生成モデルの出力例.

入力		in our knowledge , as we figured out in chapter 4 , our suggestion is that first one .
原文		**to the best of our knowledge, our proposal is the first such proposal, as clarified in Section 4.**
出力	GEC	in our knowledge , as we figured out in chapter 4 , our suggestion is **the** first one .
	LSTM	in our knowledge , as we **find** in chapter 4 , our suggestion is that **the** first one **is the first one** .
	Transformer	**to the best of** our knowledge , as we **discussed** in chapter 4 , our **proposal** is **the** first one .
入力		we propose ___mask___ (sgd) to learn model constracture ___mask___ together .
原文		**we propose a stochastic gradient descent (sgd) algorithm to train the components of our model simultaneously .**
出力	GEC	we propose (sgd) to learn model constructure together .
	LSTM	we propose **a method** (sgd) to learn model **constraints** together .
	Transformer	we propose **stochastic gradient descent** (sgd) to learn model construction **parameters** together .

語，文字数で 18 文字ほど情報が増えていることが分かり，情報を補完する性質を持つタスクであることも示唆された．

7 分析

分析用のクラウドソーシングデータからの生成例を表4に示す．入力がクラウドソーシングによって作成した草文，原文が草文を作る際に用いた ACL Anthology 上の文である．Transformer モデルは “in our knowledge” から “to the best of our knowledge” といった定型的な表現や，“___mask___ (sgd)” から “stochastic gradient descent (sgd)” といった専門用語を生成できていることから，論文ドメインにより適した表現を生成できていることが示唆される．

また，GEC モデルはマスクから復元する生成を学習していないため，マスクを考慮した情報の補完ができていないことが分かる．一方で，疑似生成学習データを用いて学習した LSTM と Transformer はマスクの位置を考慮して情報を補完していることが確認できた．

8 おわりに

言語処理分野においてこれまで取り組みの少なかった情報補完型生成を新しい言語生成課題として提案した．論文執筆支援という応用を見据え，クラウドソーシングを用いて評価用データを作成した．評価方法を提案し，妥当性を確認した．実験結果から，学習データの自動生成によって情報補完型生成モデルを訓練できる可能性があることが示された．今後は，実際のライティング現場における本技術の有効性の調査や，サービスとしての社会実装を目指す．

謝辞

本研究の一部は JST CREST（JPMJCR1301）の支援を受けて行った．

参考文献

[1] Akiko Aizawa et al. “Construction of a New ACL Anthology Corpus for Deeper Analysis of Scientific Papers”. In: *SCIDOCA* (2018).
[2] Marcelo Bertalmio et al. “Image inpainting”. In: *SIGGRAPH*. 2000, pp. 417–424.
[3] Samuel R. Bowman et al. “A large annotated corpus for learning natural language inference”. In: *EMNLP*. 2015, pp. 632–642.
[4] Kazuaki Hanawa et al. “A Crowdsourcing Approach for Annotating Causal Relation Instances in Wikipedia”. In: *PACLIC*. 2017, pp. 336–345.
[5] Thang Luong, Hieu Pham, and Christopher D. Manning. “Effective Approaches to Attention-based Neural Machine Translation”. In: *EMNLP*. 2015, pp. 1412–1421.
[6] Tomoya Mizumoto et al. “Mining Revision Log of Language Learning SNS for Automated Japanese Error Correction of Second Language Learners”. In: *IJCNLP*. 2011, pp. 147–155.
[7] Courtney Napoles, Keisuke Sakaguchi, and Joel Tetreault. “JFLEG: A Fluency Corpus and Benchmark for Grammatical Error Correction”. In: *EACL*. 2017, pp. 229–234.
[8] Courtney Napoles, Keisuke Sakaguchi, and Joel Tetreault. “There’s No Comparison: Reference-less Evaluation Metrics in Grammatical Error Correction”. In: *EMNLP*. 2016, pp. 2109–2115.
[9] Hwee Tou Ng et al. “The CoNLL-2014 shared task on grammatical error correction”. In: *Proceedings of the Eighteenth Conference on Computational Natural Language Learning: Shared Task*. 2014, pp. 1–14.
[10] Ashish Vaswani et al. “Attention is all you need”. In: *NIPS*. 2017, pp. 5998–6008.
[11] Ashwin K. Vijayakumar et al. “Diverse Beam Search: Decoding Diverse Solutions from Neural Sequence Models”. In: *AAAI* (2018).
[12] John Wieting and Kevin Gimpel. “ParaNMT-50M: Pushing the Limits of Paraphrastic Sentence Embeddings with Millions of Machine Translations”. In: *ACL*. 2018, pp. 451–462.
[13] Adina Williams, Nikita Nangia, and Samuel Bowman. “A Broad-Coverage Challenge Corpus for Sentence Understanding through Inference”. In: *NAACL-HLT*. 2018, pp. 1112–1122.
[14] Ziang Xie et al. “Noising and Denoising Natural Language: Diverse Backtranslation for Grammar Correction”. In: *NAACL-HLT*. 2018, pp. 619–628.
[15] 浅野広樹 et al. “文法誤り訂正の文単位評価におけるリファレンスレス手法の評価性能”. In: 研究報告自然言語処理 *(NL)* 2017.3 (2017), pp. 1–8.