Coverageを考慮したBERTSUMによる生成型自動要約

秋山 和輝[1]　田村 晃裕[2]　二宮 崇[2]　大林 弘明[3]

[1] 愛媛大学 情報工学科，[2] 愛媛大学 大学院理工学研究科 電子情報工学専攻，
[3] トランスコスモス株式会社

[1,2]{k_akiyama@ai., tamura@, ninomiya@}cs.ehime-u.ac.jp,
[3]Oobayashi.Hiroaki@trans-cosmos.co.jp

1 はじめに

入力された文書を短くまとめた要約を自動的に生成する自動要約技術は，自然言語処理の分野で盛んに研究されている．自動要約技術は，要約の作成方法によって「抽出型」と「生成型」の2種類に分けられる．抽出型の自動要約は，入力文書中の重要と思われる文を識別・抽出し，抽出した文を結合させたものを要約とする方法である．一方，生成型の自動要約は，要約の文章を一から生成する方法である．本研究では，生成型の自動要約モデルの技術に着目する．

生成型自動要約は，ニューラルネットワークに基づくエンコーダ・デコーダモデルが出現したことで，要約の品質が大きく向上した [1] ．しかし，古典的な生成型要約モデルでは，入力文書中の内容（各単語等）を複数回繰り返した要約を生成してしまう過剰生成問題が発生する．そこで，生成型要約モデルにおいて，入力文書中の内容（各単語等）が要約文に含まれた度合いを表す，coverageを考慮するモデルが提案されている [2, 3]．coverageに基づき，既に要約文に含まれた入力文書の内容は要約として生成されにくくすることで，過剰生成問題を防いでいる．

また，近年，自然言語処理において汎用的な事前学習モデルであるBERT[4] を抽出型と生成型の両方の要約に拡張させた BertSum[5] という要約モデルが提案され，高い要約性能を実現している．しかし，BertSumでは，coverageが考慮されておらず過剰生成問題が生じやすい可能性がある．

そこで本研究では，BertSumに基づく生成型の自動要約モデルにおいてcoverageを考慮するモデルを提案する．コールセンターにおける通話の自動要約評価実験を行い，coverageを考慮することによりROUGE-LのF値が1.24ポイント改善することを確認した．

2 関連研究

近年の自動要約モデルは，ニューラルネットワークを利用したエンコーダ・デコーダモデルが主流となっている．エンコーダ・デコーダモデルでは，出力単語を決定する際に入力情報のどこに着目するかを捉えるattention機構を組み込むことで性能を大きく向上させている [6].

2.1 Coverage機構

Coverage機構 [3] は，自動要約の分野では，代表的な抽出型エンコーダ･デコーダ要約モデルであるPointer-generator network[2] の中で過剰生成問題を解決するために導入された．Coverage機構の概要図を図1に示す．Coverage機構とは，要約元文書の各単語（各入力単語）が要約として出現した度合い（coverage）を考慮する機構のことである．具体的には，t 番目の出力単語を決めるデコーダでは，各入力単語に対するcoverageをベクトルで表したcoverageベクトル c^t を，式 (1) の通り，これまでのattentionベクトル $a^{t'}$ の合計で算出する．

$$c^t = \sum_{t'=0}^{t-1} a^{t'} \tag{1}$$

そして，attentionベクトル a^t を，式 (2) の通り，coverageベクトル c^t も利用して算出する．

$$e_i^t = v^T \tanh(W_h h_i + W_s s_t + w_c c_i^t + b_{attn}) \tag{2}$$

$$a^t = \mathrm{softmax}(\mathrm{e^t}) \tag{3}$$

ここで，v，W_h，W_s，w_c，b_{attn} は学習可能なパラメータ，h_i はエンコーダの隠れ状態，s_t はデコーダの隠れ状態である．

coverage を考慮したモデルの学習時には，coverage に対する損失（*covloss*）を加えた損失関数を用いる。具体的には，attention ベクトルと coverage ベクトルの各要素の最小値を合計した *covloss* を加える．

$$covloss_t = \sum_i min(a_i^t, c_i^t) \tag{4}$$

以上の coverage 機構により，これまでの要約の生成で attention が向けられた単語に対しては，attention が向けられにくくすることで過剰生成問題を防ぐ．

2.2 BERTSUM

近年，自然言語処理の様々なタスクにおいて，Transformer[7] モデルをベースとした汎用的な事前学習モデル BERT[4] の有効性が確認されている．自動要約においては，Liu and Lapata[5] が，BERT を抽出型及び生成型の自動要約用に拡張した BertSum を提案している．図 2 に BertSum による要約元文書の Embeddings の概要を示す．通常の BERT では入力の先頭にのみ [CLS] トークンを付けるのに対し，BertSum では入力の各文の先頭に [CLS] トークンを付け，[CLS] トークンにより要約元文書中の各文を表現する．また，Segment Embeddings を導入し，E_A と E_B の二種類の Embeddings を交互に割り当てることで，複数の文を明確に区別できるようにしている．

抽出型要約用の BertSum モデルでは，各 [CLS] トークンに対するエンコーダの最上位層の出力ベクトルをシグモイド分類器にかけることで，各文を要約文に含めるか否かを決定する．このモデルは BertSumExt と呼ばれている．生成型要約用の BertSum モデルでは，エンコーダ・デコーダモデルを適用し，BertSum による要約元文書のエンコード結果から Transformer デコーダを用いて要約文を生成する．エンコーダとして事前学習済みの BertSum を利用するモデルは BertSumAbs と呼ばれており，まず最初に BertSumExt をファインチューニングして，ファインチューニングした BertSumExt をエンコーダとして利用するモデルは BertSumExtAbs と呼ばれている．

3 提案手法

本節では，BertSum[5] の生成型自動要約モデルに coverage を考慮する機構を組み込んだモデルを提案する．文献 [5] では，二種類の BertSum に基づく生成型要約モデルにおいて，BertSumExtAbs モデルの方が BertSumAbs モデルよりも要約性能が高いことが示されていることから，本研究では，BertSumExtAbs モデルに coverage 機構を組み込む．

具体的には，BertSumExtAbs の学習時に，coverage に対する損失（2.1 節の式 (4) 参照）を加えた以下の損失関数を使用する．

$$loss_t = -\log P(w^*) + \lambda\, covloss_t \tag{5}$$

ここで，$P(w^*)$ は正解単語 w^* の予測確率，λ はハイパーパラメータである．

提案手法の概要図を図 3 に示す．提案モデルは Transformer に基づくモデルであり，エンコーダとデコーダは共に複数の層が積み重なり，デコーダの各層でエンコーダとデコーダ間の attention が計算されているが，coverage ベクトルはデコーダの最終層の attention に基づき算出する．また，2.1 節の coverage 機構では推論時にも coverage ベクトルを考慮しているが，本実験では coverage ベクトルはモデル学習時の coverage loss の計算にのみ使用し，推論時には考慮していない．

提案手法では，モデル学習時に coverage loss を考慮することで，既に要約文に含まれた入力文中の単語に対する attention に罰則がつけられ，同じ内容を複数回生成してしまう過剰生成問題を緩和した BertSum による生成型自動要約モデルを実現できる．

4 実験

4.1 データセット

本研究では実験データとして，トランスコスモス株式会社との秘密保持契約の元提供された，コールセンター業務の要約コーパスを使用した．一部のコールセンター業務では，応答品質向上や VoC 分析のために問合せの会話音声を音声認識器を用いてテキストデータに変換し記録している．このデータセットは，コールセンター業務で記録されたテキストデータを要約元文書とし，その内容を人手で要約した要約文とを対にした日本語のデータセットである． データセットの総件数 1,263 件の内，1,063 件を学習データ，100 件を開発データ，100 件をテストデータとした．

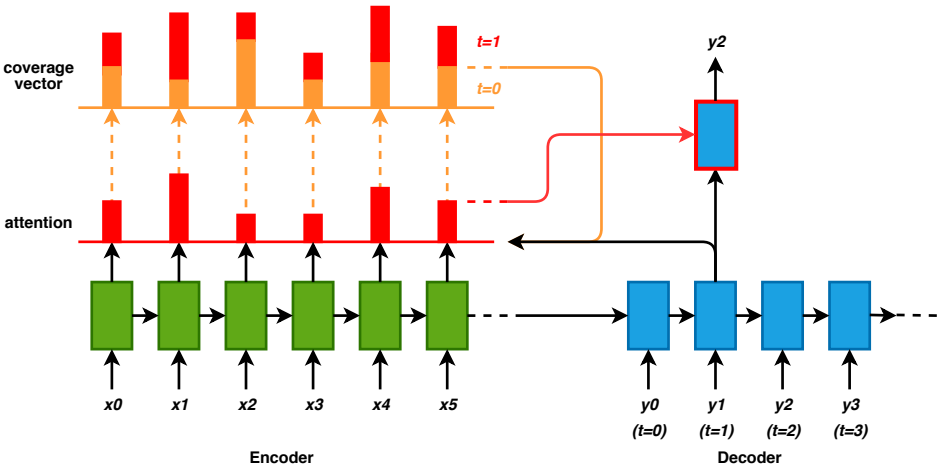

図 1: Coverage 機構の概要図

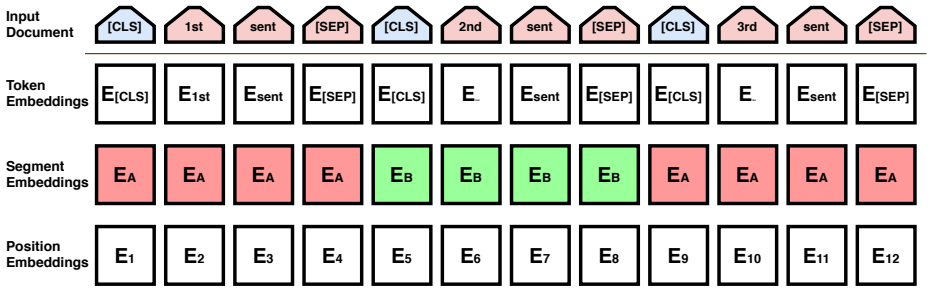

図 2: BertSum の Embeddings

4.2 実験設定

実験では，Coverage 機構を用いない従来の BertSum モデル（BertSumExtAbs）[5] と Coverage 機構を導入した提案の BertSum モデル（BertSumExtAbs+COV）の性能を比較する．4.1 節のデータセットに対して，要約元文書及び要約文書の各文の先頭に [CLS] トークンをつけたものを各モデルの入力とする[1]．要約元文書および要約文書の各文は Juman++ (v2.0.0-rc2)[2]で形態素解析し，BERT の subwords tokenizer[3]でサブワード化した．各モデルは，日本語の事前学習済み BERT モデル[4]を実験データでファインチューニングすることで学習した．まず，抽出型の要約モデルをファインチューニングし，開発データにおいて perplexity が最も低かったモデルを生成型要約モデルのファインチューニング時に利用した．使用するファインチューニング済み抽出型要約モデルは，ベースラインモデルと提案モデルで同一のものである．

生成型要約モデルの学習では，総ステップ数を 25,000，エンコーダ (BERT) の warmup ステップ数を 10,000，デコーダの warmup ステップ数を 5,000，モデルの保存ステップ数を 2,000 とした．また，提案モデルにおける式 (5) のハイパーパラメータ λ は 1.0 とした．

要約生成時は，ビーム探索（size 5）を使用し，1.0 から 0.6 の間の length penalty α をかけた．また，出力文字数の制限である最小文字列長 min_length は 10，最大文字列長 max_length は 512 とし，その他の設定は BertSum の初期設定と日本語の事前学習済み BERT モデルの設定と同じである．

テスト時において，要約性能は，ROUGE-1，ROUGE-2，ROUGE-L の F 値で評価した[5]．開発データに対して perplexity が低かったモデルを上位から 3 つ選択し，各モデルが出力した要約文書に対

[1] 要約元文書は一発話を一文とみなした．
[2] https://github.com/ku-nlp/jumanpp
[3] https://github.com/huggingface/transformers
[4] http://nlp.ist.i.kyoto-u.ac.jp/
[5] pyrouge（https://github.com/bheinzerling/pyrouge）を用いて各 ROUGE 値を算出した．

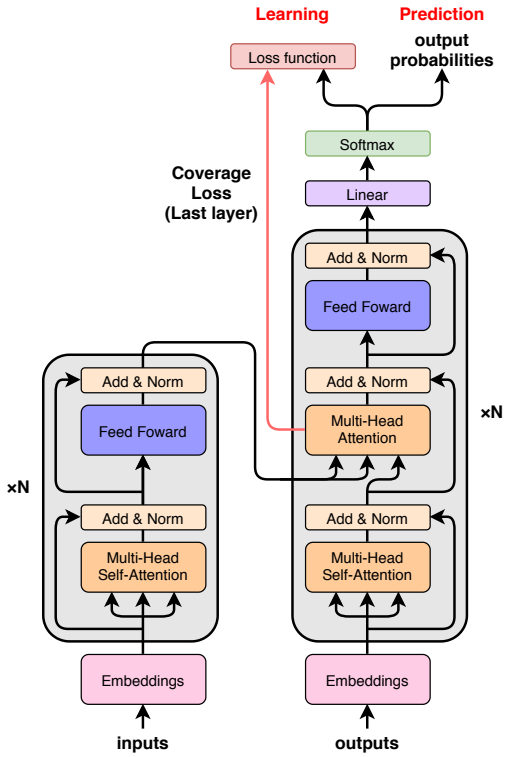

図 3: 提案手法のモデル図

する ROUGE スコアの平均値を評価・比較した．

4.3 結果

評価結果を表 1 に示す．表 1 より，従来モデルである BertSumExtAbs モデルに比べ，coverage 機構を組み込んだ提案モデル (BertSumExtAbs+COV) では，ROUGE において全体的な向上が見られた．特に，ROUGE-L の F 値は提案モデルが従来モデルよりも 1.24 ポイント上回っている．このことから提案モデルの有効性が実験的に確認できた．

表 1: 各モデルの要約性能

Model	ROUGE[%]		
	1	2	L
BertSumExtAbs	18.54	3.62	17.81
BertSumExtAbs+COV	19.80	4.00	19.05

5 おわりに

本研究では，自動要約タスクにおいて高い性能を実現している BERTSUM による生成型自動要約モデル (BertSumExtAbs) に対し，要約元文書と要約文書の coverage を考慮させる coverage 機構を組み込んだモデルを提案した．具体的には，coverage loss を加えた損失関数に基づきモデルを学習することで，同じ内容が複数回生成されることを抑制する要約モデルを獲得する．コールセンター業務の要約コーパスを用いた評価実験により，要約精度の改善を確認した．

今後は，coverage ベクトルを学習時だけではなく推論時にも考慮するモデルに拡張予定である．また，CNN/Daily Mail や XSum などの標準的なデータセットに対しても提案モデルの有効性を検証予定である．

6 謝辞

本研究は，トランスコスモス株式会社により助成を受けたものである．ここに謝意を表する．

参考文献

[1] Alexander M Rush, Sumit Chopra, and Jason Weston. A neural attention model for abstractive sentence summarization. In *Proc. EMNLP*, 2015.

[2] Abigail See, Peter J Liu, and Christopher D. Manning. Get to the point: Summarization with pointer-generator networks. In *Proc. ACL*, 2017.

[3] Zhaopeng Tu, Zhengdong Lu, Yang Liu, Xiaohua Liu, and Hang Li. Modeling coverage for neural machine translation. In *Proc. ACL*, 2016.

[4] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proc. NAACL: Human Language Technologiess*, 2019.

[5] Yang Liu and Mirella Lapata. Text summarization with pretrained encoders. In *Proc. EMNLP*, 2019.

[6] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. In *Proc. ICLR*, 2015.

[7] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Proc. NIPS*, 2017.