

出力長制御を考慮した見出し生成モデルのための
大規模コーパス人見雄太¹ 田口雄哉¹ 田森秀明¹ 西鳥羽二郎² 菊田光²
岡崎直観³ 乾健太郎^{4,5} 奥村学³株式会社朝日新聞社¹ 株式会社レトリバ² 東京工業大学³ 東北大学⁴ 理化学研究所AIPセンター⁵
{hitomi-y1, taguchi-y2, tamori-h}@asahi.com¹

1. はじめに

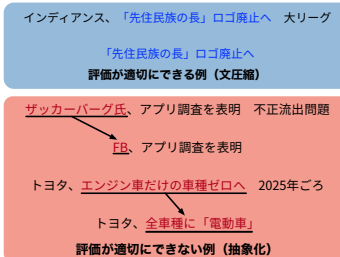
背景

より多くの読者に記事を読んでもらうため、1つの記事を複数のメディアやデバイスに配信している。一方で、配信先のデザインや画面サイズに応じて見出しの文字数に厳密な制約がある。Enc-Decモデルによる自動見出し生成を実用化するには、所望の長さで見出しを生成するタスクに取り組む必要がある。

既存研究

出力長制御に取り組んだ既存研究における評価では、30・50・75文字を上限に指定したシステム出力に対し、75文字を上限に作られた参照要約(DUC 2004)を用いている。

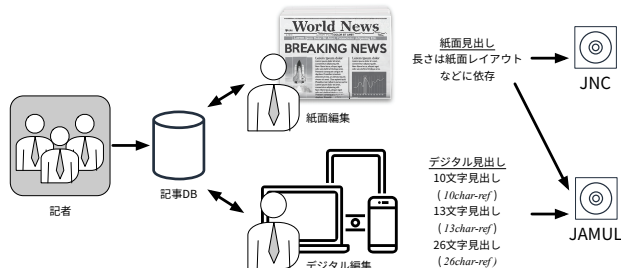
システム出力と参照要約の圧縮率が異なる場合、適切に評価ができていのだろうか？



本研究では、出力長を考慮した見出し生成モデルを評価するためのコーパス「JAMUL」と、日本語の見出し生成のための学習コーパス「JNC」を紹介する。これらのコーパスで既存の評価設定で適切に評価できていたかについて議論する。

2. JNCとJAMUL

2.1 見出し生成のフロー



2.2 JNC

JNCは10年分の記事と紙面見出しのペアを約183万件収録。様々な長さの見出しを含み、一般的な見出し生成モデルのみならず、出力長を考慮したモデルの学習にも利用可能。

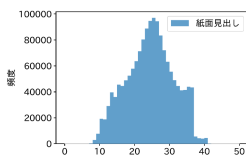
2.3 JAMUL

記事、紙面見出しと、10文字・13文字・26文字を上限として付けられたデジタル見出しの組を1524件収録。7ヶ月間に配信されたもので、JNCとは重複がない。1つの記事に対し、複数の長さの見出しが付いている点が、従来のコーパスと異なる。出力長制御モデルが生成した異なる長さの見出しを評価可能。

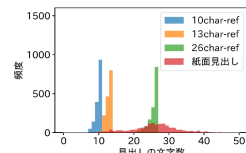
2.4 紙面見出しとデジタル見出しの比較

学習用のJNC（紙面見出し）と評価用のJAMUL（デジタル見出し）では、掲載先が異なるため、デジタル見出しが紙面見出しに対し、どの程度特徴的な違いがあるか検証。

システム出力	参照要約	適合率	再現率
紙面見出し	10char-ref	24.66	64.11
紙面見出し	13char-ref	33.24	66.30
紙面見出し	26char-ref	56.36	55.75



JNCが収録する見出しの長さ分布



JAMULが収録する見出しの長さ分布

デジタル見出しで用いられる語彙の半数以上は紙面見出しに由来するが、約4割は紙面見出しには表れない。
…語彙の違いが大きい

2.5 異なる長さのデジタル見出し間の比較

システム出力	参照要約	適合率	再現率
26char-ref	10char-ref	28.77	78.55
26char-ref	13char-ref	42.53	88.75
26char-refの先頭10文字	10char-ref	38.40	41.65
26char-refの先頭13文字	13char-ref	60.31	65.58
26char-refの末尾10文字	10char-ref	14.55	17.05
26char-refの末尾13文字	13char-ref	23.13	26.56

26char-refには、10/13char-refの単語がほとんど含まれている

先頭からn文字を切り出した評価では、単語が網羅されておらず不十分
必ずしも長い要約の前半に重要なキーワードが集中しているわけではない

3. JAMULによる既存手法の評価

3.1 出力長制御を考慮した見出し生成

6つの手法の組み合わせについて評価

	生成機構	出力長制御機構
(1) Seq2Seq	LenEmb[Kikuchi et al. (2016)]	
(2) Seq2Seq	LenInit[Kikuchi et al. (2016)]	
(3) Seq2Seq	SP-token[Fan et al. (2018)]	
(4) ConvS2S[Gehring et al. (2017)]	SP-token	
(5) ConvS2S	LC[Liu et al. (2018)]	
(6) Transformer[Vaswani et al. (2017)]	SP-token	

- 学習データ：JNC
 - フィルタリングで約157万件を取得
 - 157万件のうち、ランダムに抽出した1%をValidation Dataとした
- 語彙構築：Byte Pair Encoding
- 評価データ：JAMUL
 - JNCと同様のフィルタリングで約1,100件を取得
- 評価指標：ROUGE-1・ROUGE-2・ROUGE-L

3.2 JAMULによる評価

JAMULによる評価												
モデル	10文字			13文字			26文字					
	R-1	R-2	R-L	R-1	R-2	R-L	R-1	R-2	R-L			
(1) Seq2Seq + LenEmb	34.66	15.29	33.56	40.66	19.40	38.23	44.82	20.75	36.62			
(2) Seq2Seq + LenInit	36.50	16.75	35.54	41.40	19.49	38.83	46.77	22.06	38.29			
(3) Seq2Seq + SP-token	38.09	17.43	36.67	42.51	19.79	39.76	47.33	22.12	38.59			
(4) ConvS2S + SP-token	38.90	17.84	37.53	43.32	20.35	40.31	47.10	21.51	37.86			
(5) ConvS2S + LC	37.71	16.89	36.50	42.60	20.11	39.97	45.76	21.93	37.91			
(6) Transformer + SP-token	42.85	19.84	41.40	46.92	22.85	44.09	51.57	24.52	41.05			

いずれもTransformer + SP-tokenのスコアが最も高いが...

26char-ref (長い参照要約)のみを使用した評価

ROUGEの順位が入れ替わる
参照要約が長い
ためスコアの絶対値が小さく、
性能差も小さくなる

短いシステム出力を長い参照要約で評価すると？

参照要約

10文字 米娯楽大手が売却交渉

26文字 21世紀フォックス、ディズニーに事業売却 交渉米報道

システム出力（出力長：10文字）

モデル	見出し	10char-ref (R-1) = Gold	26char-ref (R-1)
(1)	21 世紀フォックス	0.00	25.00
(2)	フォックス、自社売却	16.67	25.00
(3)	フォックス、事業協議	0.00	25.00
(4)	米メディア、売却協議	33.33	25.00
(5)	米メディア 大手売却へ	50.00	16.67
(6)	米娯楽大手が売却協議	83.33	16.67

26char-refで評価したほうが、10char-ref(GOLD)で評価するより高いスコア
＝不当に評価されている可能性

4. まとめ

JAMULによる評価でわかってきたこと：

長さに応じて使用される単語に変化があるため、長い要約にしか含まれないような単語が不当にROUGEスコアを上げている可能性がある

参考文献

- Yuta Kikuchi, Graham Neubig, Ryohei Sasano, Hiroya Takamura, and Manabu Okumura. Controlling output length in neural encoder-decoders. In EMNLP, pp. 1328-1338, 2016.
- Angela Fan, David Grangier, and Michael Auli. Controllable abstractive summarization. In WMT, pp. 45-54, 2018.
- Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N. Dauphin. Convolutional sequence to sequence learning. In ICML, pp. 1243-1252, 2017.
- Yizhu Liu, Zhiyi Luo, and Kenny Zhu. Controlling length in abstractive summarization using a convolutional neural network. In EMNLP, pp. 4110-4119, 2018.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In NIPS, pp. 6000-6010, 2017.



JNC・JAMULの詳細・入手はこちら