論文紹介　ACL2018読み会

Retrieve, Rerank and Rewrite: Soft Template Based Neural Summarization

Ziqiang Cao[1,2] Wenjie Li[1,2] Furu Wei[3] Sujian Li[4]

1 Department of Computing, The Hong Kong Polytechnic University, Hong Kong
2 Hong Kong Polytechnic University Shenzhen Research Institute, China
3 Microsoft Research, Beijing, China
4 Key Laboratory of Computational Linguistics, Peking University, MOE, China

紹介者: 小川泰弘（名古屋大学）

紹介の前に

自動要約研究の推移

- 紹介論文の「要約」は「文圧縮」
- 従来研究:入力文書中から重要文を抽出
  - 要約前と要約後のペアデータは不十分
  - seq2seq アプローチは適用不可
- Annotated English Gigaword Corpus
  - 原文:　ニュース原稿の第1文
  - 目的文: 上記のヘッドライン
  - seq2seq アプローチが適用可能

概要

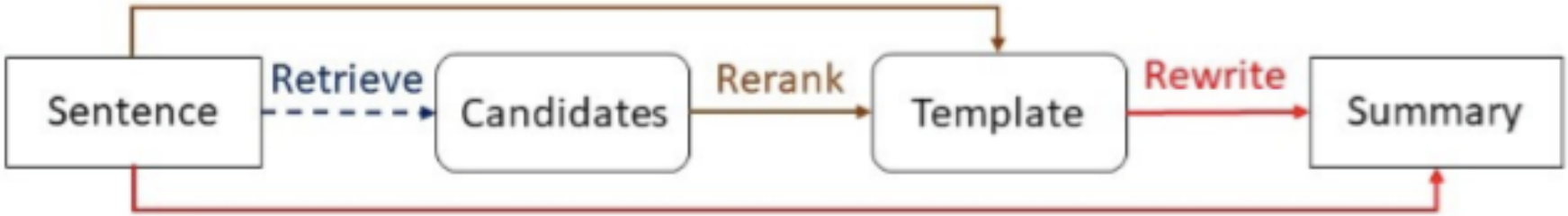
- 問題：要約（文圧縮）
- 従来：ニューラルモデル・seq2seq モデル
- 課題：要約が安定しない
- 手法：入力として、入力文に似た文の要約を
  テンプレートとして追加
- 実験： Annotated English Gigaword
  Corpus を対象に従来手法と比較
- 結果：従来手法を上回る結果

論文のアイデア

- Seq2Seqの要約研究がある
  - でも、入力は原文のみ
- テンプレート要約は使えないか？
  - テンプレートは人手で作るので高コスト
  - 代わりのテンプレート⇒ソフトテンプレート
    - 訓練データ中、入力文に似た原文の要約
- 入力は原文＋ソフトテンプレート

提案手法： Re^{3}Sum

- Retrieve
- Rerank
- Rewrite

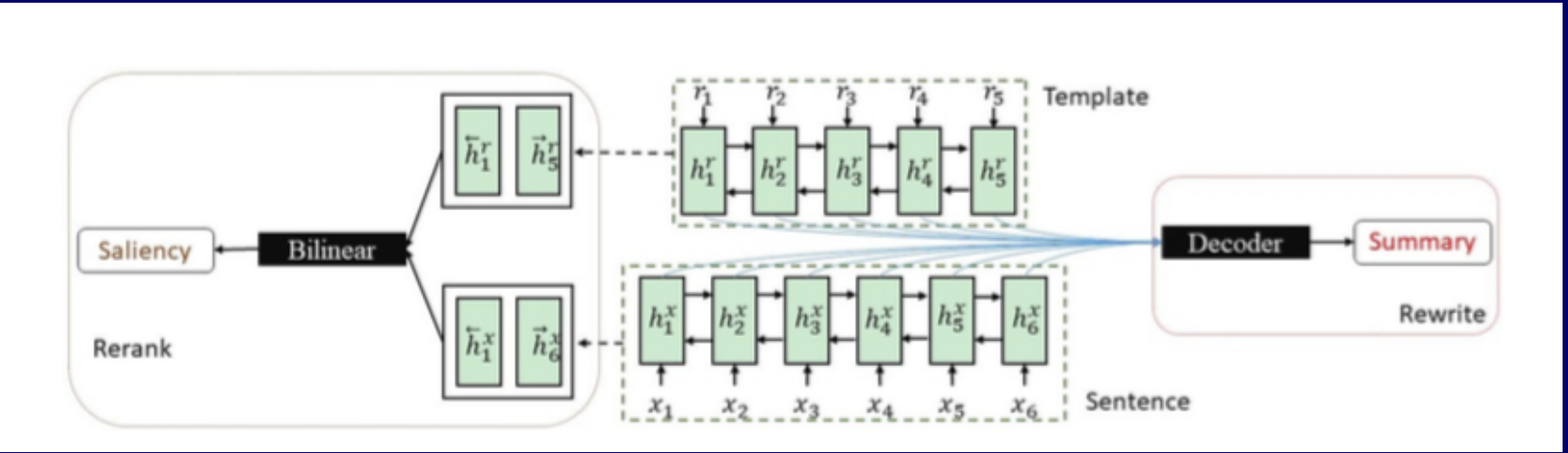


Retrieve

- 情報検索(Lucene) を利用
  - 検索対象が膨大（3M以上）のため
- 入力文と訓練データの原文を比較
  - 上位30位の要約文を候補文とする

Jointly Rerank and Rewrite

双方向RNN



- $\mathbf{h}_i^x = [\overrightarrow{\mathbf{h}_i^x}; \overleftarrow{\mathbf{h}_i^x}]$
- 原文 $\mathbf{x}$: $[\mathbf{h}_1^x; \cdots; \mathbf{h}_{-1}^x]$
- ソフトテンプレート $\mathbf{r}$: $[\mathbf{h}_1^r; \cdots; \mathbf{h}_{-1}^r]$

Rerank

30個の候補の中から予測RG最大を求める

- $\mathbf{h}_x = [\overleftarrow{\mathbf{h}_1^x}; \overrightarrow{\mathbf{h}_{-1}^x}]$
- $\mathbf{h}_r = [\overleftarrow{\mathbf{h}_1^r}; \overrightarrow{\mathbf{h}_{-1}^r}]$
- $s(\mathbf{r}, \mathbf{x}) = \mathrm{sigmoid}\left(\mathbf{h}_r \mathbf{W}_s \mathbf{h}_x^{\mathrm{T}} + b_s\right)$
  - 以下のs^*に近づくよう学習
- $s^*(\mathbf{r}, \mathbf{y}^*) = \mathrm{RG}(\mathbf{r}, \mathbf{y}^*) + \mathrm{RG}(\mathbf{r}, \mathbf{y}^*)$
  - $\mathbf{y}^*$: 正解要約

Rewrite

- Attention RNN
  - ソフトテンプレート中の不要な固有表現を書換
  - 入力は原文＋ソフトテンプレート
    - $\mathbf{H}_c = [\mathbf{h}_1^x; \cdots; \mathbf{h}_{-1}^x; \mathbf{h}_1^r; \cdots; \mathbf{h}_{-1}^r]$
  - 実装はOpenNMT

コーパス

Annotated English Gigaword Corpus

- 原文:　ニュース原稿の第1文
- 目的文: 上記のヘッドライン(ヒューリスティック)

Dataset	Train	Dev.	Test
大きさ	3.8M	189k	1951
平均原文長	31.4	31.7	29.7
平均目的文長	8.3	8.3	8.8
コピー率(%)	45	46	36

Perplexity

Model	Perplexity
ABS	27.1
RAS-Elman	18.9
FTSum	16.4
$\text{OpenNMT}_{\text{I}}$	13.2
PIPELINE	12.5
Re^3Sum	12.9

PIPELINE：Rerank と Rewrite の学習がパイプライン

ROUGEによる評価

Model	RG-1	RG-2	RG-L
ABS	29.55	11.32	26.42
ABS+	29.78	11.89	26.97
Featseq2seq	32.67	15.59	30.64
RAS-Elman	33.78	15.97	31.15
Luong-NMT	33.10	14.45	30.71
FTSum	37.27	17.65	34.24
OpenNMT_{O}	33.13	16.09	32.42
OpenNMT_{I}	35.01	16.55	32.42
PIPELINE	36.49	17.48	33.90
Re^{3}Sum	37.04	19.03	34.46

比較用：5種類のソフトテンプレート

- Random: 訓練データの要約からランダム
- First: Retrieve の時点でのトップ
- Max: 30個の候補中でRG最大
- Optimal: 訓練データの要約中でRG最大
- Rerank: 30個の候補の中で予測RG最大

ソフトレンプレート自体の類似度

Type	RG-1	RG-2	RG-L
Random	2.81	0.00	2.72
First	24.44	9.63	22.05
Max	38.90	19.22	35.54
Optimal	52.91	31.92	48.63
Rerank	28.77	12.49	26.40

- Randomは壊滅的
- First < Rerank　Rerankの有用性
- Optimal はsotaよりはるかに高い

Re^{3}Sum + ソフトテンプレート

Type	RG-1	RG-2	RG-L
+Random	32.60	14.31	30.19
+First	36.01	17.06	33.21
+Max	41.50	21.97	38.80
+Optimal	46.21	26.71	43.19
+Rerank	37.04	19.03	34.46

- Randomもそこそこ。ダメなテンプレートは排除
- Max, Optimal が高い。 Rerankに改善の余地

文の品質評価

項目	Template	OpenNMT	Re³Sum
文長の差	2.6±2.6	3.0±4.4	2.7±2.6
長さ3未満	0	53	1
コピー率(%)	31	80	74
新たな固有表現	0.51	0.34	0.30

原文・正解要約に存在しない

出力例

Source	grid positions after the final qualifying session in the indonesian motorcycle grand prix at the sentul circuit , west java , saturday : UNK
Target	indonesian motorcycle grand prix grid positions
Template	grid positions for **british** grand prix
OpenNMT	circuit
Re^{3}Sum	grid positions for indonesian grand prix
Source	india 's children are getting increasingly overweight and unhealthy and the government is asking schools to ban junk food , officials said thursday .
Target	indian government asks schools to ban junk food
Template	**skorean** schools to ban **soda** junk food
OpenNMT	india 's children getting fatter
Re^{3}Sum	indian schools to ban junk food

まとめ

- 問題：要約（文圧縮）
- 従来：ニューラルモデル・seq2seq モデル
- 課題：要約が安定しない
- 手法：入力として、入力文に似た文の要約を
  テンプレートとして追加
- 実験： Annotated English Gigaword
  Corpus を対象に従来手法と比較
- 結果：従来手法を上回る結果