E-009

質問応答を用いた日本語要約評価システム
Japanese Summary Evaluation System Using Question and Answer

岡田　直士*　　松澤　智史†
Naoto Okada　　Tomofumi Matsuzawa

1. はじめに

近年，機械学習を用いた自然言語処理の発展は著しく，文章要約の研究も盛んに行われている．しかし要約評価をする上で多く用いられる手法は，文章の意味的な正確さを考慮することが出来ず，表面的な類似性を測ることがある．そのため，真実と異なる要約文に対してまで高い評価値を出力させてしまう場合が存在する．質問と解答のタスクを使用することで一般的に用いられている要約評価よりも意味的な正確さを考慮し，人間との相関が高いものを作成可能であると明らかにした．しかし，これは英語を自然言語として扱った場合であり他の言語については言及されていない．そのため，本研究では日本語を対象にした場合でも質問応答を用いた要約評価は適切であるか検証を行う．

2. QAGS

2.1 概要

QAGS[1] は要約文から質問を生成し原文と要約文に対して同じ質問を投げかけることで，得られた解答との類似度を比較し要約文を評価する．解答の比較では F 値を用いており，全ての質問について F 値を計算し平均を求める．以下に概要図を載せる．(図 1)

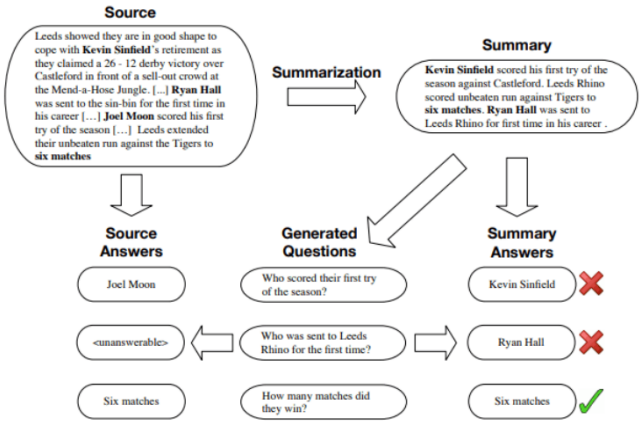

図 1: QAGS の流れ

2.2 結果

要約文について記事の内容と一致するもの，しないもので評価することで各文についての正しさを多数決で求め，結果の平均により要約文全体の正しさを求める．ここで利用される文章は CNN/Daily Mail，XSUM の二つとなる．QAGS は既存手法に比べどちらも高い値を示している．以下に実験結果を載せる．(表 1)

表 1: 関連研究結果

Metric	CNN/DM	XSUM
ROUGE-1	28.74	13.22
ROUGE-2	17.72	8.95
ROUGE-L	24.09	8.86
METEOR	26.65	10.03
BLEU-1	29.68	11.76
BLEU-2	25.65	11.68
BLEU-3	23.96	8.41
BLEU-4	21.45	5.64
BERTScore	27.63	2.51
QAGS	54.53	17.49

3. 提案手法

日本語要約 AI に対応させた評価手法の作成を行う．日本語の原文と機械要約文を質問解答のタスクに入力することで要約評価の F 値を出力させる．この手法を QAGS-Japanese とし，具体的な操作や流れを以下に示す．

1. COTOHA API‡を用いながら要約文の複数の単語の深層格と副品詞を割り出す
2. 単語又は文章を適切な疑問詞に変えることで複数の質問文を作成する
3. 質問文を解答生成の学習済みモデルに入力する
4. 原文をもとにした複数の予想解答が生成される
5. 予想解答と要約文から得られた正しい解答を ROUGE-1 で評価することで要約評価を行う

3.1 質問生成

COTOHA API が分かち書きされた要約文に存在する単語の深層格と副品詞を割り出すことで適切な疑問詞を使用した質問文を作成した．

3.2 解答生成

BERT[2] の日本語版モデル ‘bert-base-japanese-whole-word-masking’ を使用し，SQuAD[3] 形式の運転ドメイン QA データセットでファインチューニングを行うことで解答生成モデルを作成した．

4. 実験

既存手法との比較実験を 2 つ行う．実験 1 では「事実と一致する要約 AI の場合」，実験 2 では「事実に反する要約 AI の場合」である．

*東京理科大学　理工学研究科　情報科学専攻
†東京理科大学　理工学部　情報科学科

‡自然言語処理 API プラットフォーム

4.1 事実と一致する要約 AI の場合

比較する際に使用する既存手法では，要約評価の際に人間の作成した正しい要約文が必要である．そこで，livedoor ニュースには記事とその三行要約文が記載されているため，この交通系の記事 80 枚を用いる．要約 AI では生成型要約*が可能な COTOHA API の要約 AI を用いる．既存手法の代表として ROUGE-1，ROUGE-2，BLUE，BERTScore[4] を用いる．これらに三行要約文と機械要約文を入力し F 値を出力させる．QAGS-Japanese に原文と機械要約文を入力し F 値を出力させる．(表 2)

4.2 事実に反する要約 AI の場合

上記で要約 AI として用いた COTOHA API は生成型要約を行うものの，事実に反するような要約文をほとんど作成しない．そこで，交通系の記事の機械要約文を別の記事の機械要約文と入れ替えることで，交通系の話題ではあるが違う要約文を作成してしまった状況を作る．その上で，先ほどと同様に既存手法に三行要約文と機械要約文を入力し F 値を出力させる．QAGS-Japanese に原文と機械要約文を入力し F 値を出力させる．(表 3)

5. 結果

5.1 事実と一致する要約 AI の場合

表 2: 要約評価結果

評価手法	平均 F 値
ROUGE-1	0.6728
ROUGE-2	0.4922
BLUE	0.4394
BERTScore	0.7729
QAGS-Japanese	0.8784

5.2 事実に反する要約 AI の場合

表 3: 要約評価結果

評価手法	平均 F 値
ROUGE-1	0.3315
ROUGE-2	0.0758
BLUE	0.0237
BERTScore	0.7485
QAGS-Japanese	0.0326

6. 評価と考察

6.1 全体評価

QAGS-Japanese は事実と一致している際に他の評価手法と比べて一番高い平均 F 値を出している．また，事実に反する場合は他の評価手法に比べて低い平均 F 値を出しているため，一致と不一致における平均 F 値の差が一番大きい．そのため，他の評価手法よりも事実に正確かどうかに重きを置いて評価できている．

*意味を汲み取り新たに自然な文章を作成する手法

6.2 事実と一致する要約 AI の場合

QAGS-Japanese の平均 F 値が 1 にならなかった理由は質問生成と解答生成の精度の低さにある．質問生成では COTOHA API が深層格を間違って判定することで正しくない質問文が作成されることがあった．また，解答生成でも正しい解答を出力出来ないことがあった．

6.3 事実に反する要約 AI の場合

QAGS-Japanese の平均 F 値が 0 にならなかった理由を以下に述べる．この実験では原文と機械要約文の異なるセットを入力値としているため，本来解答が生成されることはない．しかし，全て交通系の記事を使用しているため，似通った単語ベクトルが原因でまれに正しくない解答を出力してしまうことがあった．解答同士を比べる際に ROUGE-1 を用いているため，一致した語に反応してしまい平均 F 値が 0 にならなかった．

7. 今後の課題

結果として，既存手法よりも事実との整合性を評価することが可能だったものの，質問応答タスクの精度に問題があるため改善の余地がある．さらに，この手法は事実と一致しているが要約文にはふさわしくない文章に対してまで高い評価値を与えてしまう問題があるため，その部分を改善するには別の手法が必要になる．

8. おわりに

結果として事実と一致する要約文，事実に反する要約文が作成された状況を仮定した場合，既存手法に比べてより事実との整合性に重きを置くことが可能な評価指標であると確認できた．さらに，この手法は既存手法と違い人間の作成した要約文を必要としない．そのため，既存手法よりも速く要約評価を可能にする．本研究を通じて要約技術のさらなる発展が期待できる．

参考文献

[1] Alex Wang, Kyunghyun Cho, Mike Lewis,“Asking and Answering Questions to Evaluate the Factual Consistency of Summaries”,Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020.

[2] Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova,“ BERT:Pre-training of Deep Bidirectional Transformers for Language Understanding”, Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics, pp 4171-4186, 2019

[3] Pranav Rajpurkar, Robin Jia, Percy Liang,“Know What You Don’t Know:Unanswerable Questions for SQuAD”, Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, pp.784-789, 2018

[4] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, Yoav Artzi,“BERTScore: Evaluating Text Generation with BERT”, ICLR 2020 Conference Blind Submission, 2019