

金融ドメインにおける日本語文書要約の単語単位の品質推定

山内 洋輝¹ 田村 光太郎² 梶原 智之¹

¹ 愛媛大学 ² 株式会社ユーザベース

{yamauchi@ai., kajiwara@}cs.ehime-u.ac.jp koutarou.tamura@uzabase.com

1 はじめに

本研究では、大規模言語モデルによって生成された有価証券報告書の要約の品質推定に取り組む。大規模言語モデルは、文書要約をはじめとする様々なテキスト生成タスクにおいて高い性能を達成している[1]ものの、生成文の発信の際には情報の正誤や過不足についての評価が避けられない。特に、有価証券報告書のような金融ドメインに特化した文書の要約においては、高度な専門知識が要求されるため、詳細な人手評価には多くのコストが発生する。

2 提案手法

人手評価のコストを削減するために、本研究では要約文に対する単語単位の品質推定に取り組む。図1に示すように、入力文書および要約文をマスク言語モデル¹⁾に投入し、要約文中の各トークンに対して品質ラベルを推定する。本研究では、品質ラベルとして{OK, 不必要, 不正確}の3種類を用いる。

単語単位の品質推定は、機械翻訳における既存手法[2]に基づくが、文書要約タスクに特化させるために、入力文書のトークン長を制限する。これは、入力文書が長い場合にマスク言語モデルの最大入力長を超える場合があるため、要約文を含めてトークン数 k となるように、入力文書の末尾のトークンを削るものである。

また、本研究で扱うデータセットは、高性能な大規模言語モデルによって要約文を生成する場合、多くのトークンがOKの品質ラベルを持つ不均衡データとなる。これに対処するために、ラベル数に基づく重みを考慮する均衡クロスエントロピー損失[3]を採用する。その上で、不必要または不正確の品質ラベルが過剰に出力されるのを防ぐために、閾値 $0 < \theta < 1$ 未満の確率で品質ラベルが選択された場合には、OKを出力するようOptuna²⁾で調整した。

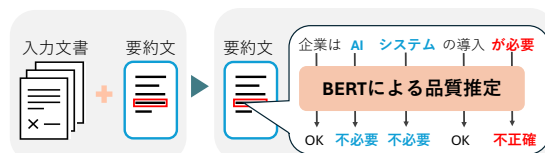


図1 BERTによる要約文の単語単位の品質推定

表1 実験結果

入力長制限	フィルタリング	MCC
1,024	-	0.140
1,536	-	0.142
2,048	-	0.108
1,536	✓	0.149

3 評価実験

有価証券報告書2,048件の大規模言語モデルによる要約文を用いて、提案手法の有効性を検証した。品質推定モデルには、2,048トークンの長文入力に対応した日本語マスク言語モデルのRetrievalBERT¹⁾を使用した。その他、訓練用・検証用・評価用に入力文書を1,657件・193件・198件に分割し、入力文長を $k \in \{1024, 1536, 2048\}$ の中から選択し、評価指標としてMatthews相関(MCC)を使用した。

実験結果を表1に示す。入力長制限が有効であり、最大限に入力文書を使用するよりも、先頭の千トークン程度のみを使用の方が高品質な品質推定を実現できた。また、入力長制限とフィルタリングの組み合わせにより、最高性能を達成した。

参考文献

- [1] 樽本空宙, 畠垣光希, 宮田莉奈, 梶原智之, 二宮崇. Chatgptの日本語生成能力の評価. 自然言語処理, Vol. 31, No. 2, pp. 349–373, 2024.
- [2] Ricardo Rei, Marcos Treviso, Nuno M. Guerreiro, Chrysoula Zerva, Ana C Farinha, Christine Maroti, José G. C. de Souza, Taisiya Glushkova, Duarte Alves, Luisa Coheur, Alon Lavie, and André F. T. Martins. CometKiwi: IST-Unbabel 2022 Submission for the Quality Estimation Shared Task. In **Proc. of WMT**, pp. 634–645, 2022.
- [3] Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge Belongie. Class-Balanced Loss Based on Effective Number of Samples. In **Proc. of CVPR**, pp. 9268–9277, 2019.

1) <https://huggingface.co/retrieval-jp/bert-1.3b>

2) <https://optuna.org/>