新しい文書要約手法の提案

○白川桃子(武蔵野大学 工学部 数理工学科 3年)
指導教員：佐々木多希子(武蔵野大学)，宮田真宏(武蔵野大学)，友枝明保(関西大学)

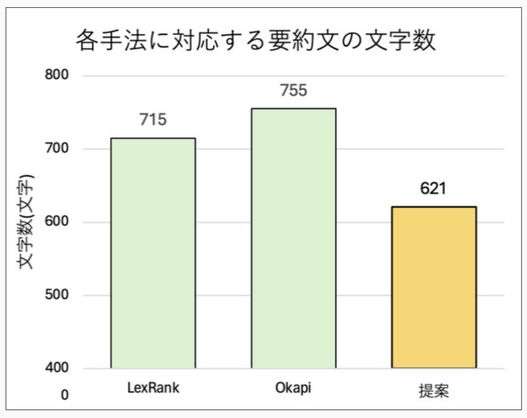
概要

【目的】
大量の文書を**機械的かつ適切に要約**したい！

【提案手法】
国語の教材を実装し既存の分析手法に組み込む

【3つの手法ごとの平均要約正解率】
NHKに公開されている要約との
一致率を用いて評価する.

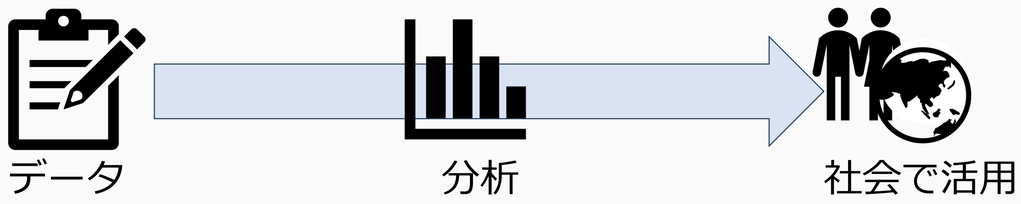
分析手法	正解率(%)
Lexan	88.860
Okapi-bm25	89.115
提案手法	87.815



【結果】
正解率約90%を維持しつつ要約の文量を約**13～18**%削減！

背景

テキストや音声言語を分析し処理する技術．(=自然言語処理)
分析手法も多数存在するが，**絶対的な分析手法の確立**は未だ進化している過程にあるのでは.



↑最適な手法なのか？

本研究の主目的

①既存の文書要約手法 **+ 国語の教材**で文書要約手法を提案する
本来私たちが文章を読むときのプロセスを組み込む！
②３つの手法の要約正解率や違いを明らかにし，考察する

LexRank

①形態素解析（名詞,動詞,副詞,形容詞のみ抽出）
②TF-IDFの計算
tf*idf =

$$\frac{\text{文書}_A\text{における単語}_X\text{の出現頻度}}{\text{文書}_A\text{における全単語の出現頻度の和}} * \log\left(\frac{\text{全文書数}}{\text{単語}_X\text{を含む文書数}}\right)$$

③cos類似度の計算

$$\vec{a} \cdot \vec{b} = |\vec{a}||\vec{b}| \cos\theta$$

④隣接行列に変換（0.5より大きい：1　　0.5以下：0）
⑤確率行列に変換
n*n行列の要素それぞれに対して以下が成立

$$p_{ij} \geq 0 \qquad \sum_{i=1}^{n} p_{ij} = 1$$

⑥固有ベクトルの計算

Okapi-bm25

TF-IDFの問題点
文章内で単語使用数が多いと値が大きくなってしまう
→この問題を解決したのがOkapi-bm25

$$\sum_{i=0}^{n} IDF(q_i) * \frac{f(q_i, D) * (k_1 + 1)}{f(q_i, D) + k_1 * (1 - b + b * \frac{|\boldsymbol{D}|}{\boldsymbol{avgdl}})}$$

$f(q_i, D)$：文書Dにおける単語の出現頻度
$|D|$：文書に含まれる総単語数
$avgdl$：全文章の平均単語数
k_1, b：パラメーター
LexRank-①に対してokapiの合計値を計算

提案手法

①指示語を含む文章の除去
指示語の解釈にはその前後文の参照が必須.
→除去することで文脈の崩れ,要約の冗長を防止.

②順接語，逆接語の重要視
これらの後ろでは結論の内容を示すという性質を活用.
→より文章の本質部分のスコアを高くできる

結果の分析と考察

【評価方法】
正解データ：NHKの作成した要約文
類似度計算方法：**Sentence-BERT**の活用
HuggingFaceに公開されている学習済みBERTモデルを利用

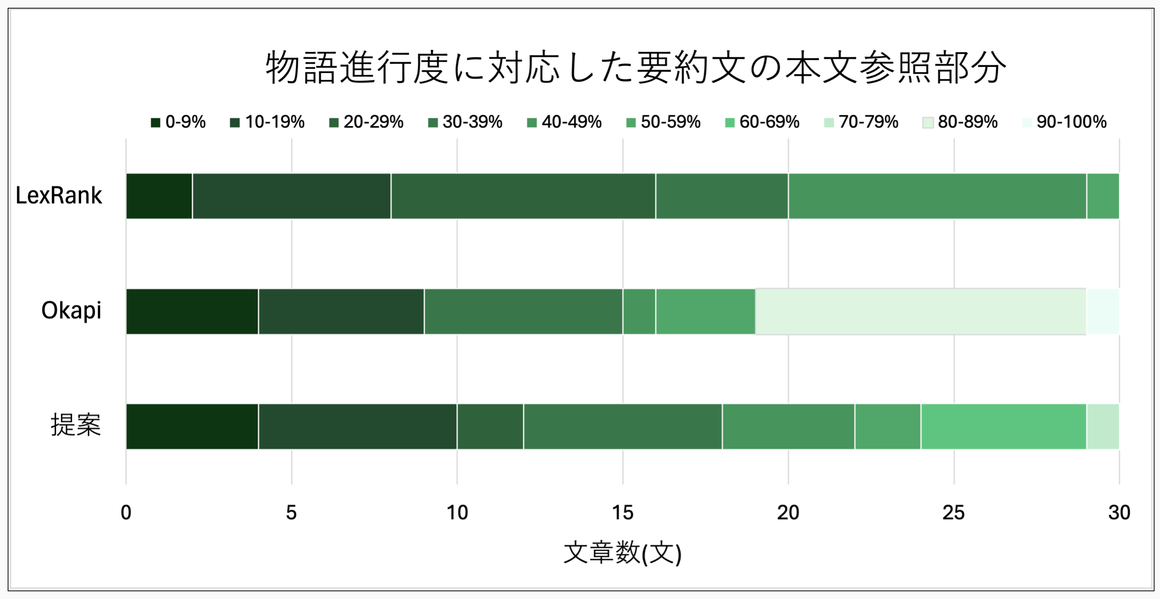
3つの手法における要約文の各分析結果

	文字数(文字)	正解率(%)	指示語(個)	各手法との要約一致率(%)		
				LexRank	Okapi	提案
LexRank	715	88.861	6	100.000	3.448	41.379
Okapi	755	89.115	2	3.448	100.000	3.333
提案	**621**	87.815	0	**41.379**	3.333	100.000

Lexrankと提案手法の要約一致率：41.379%
要約文字数：約13%減　正解率の差：約1%
→**本質部分は残しつつ，文章を短く**できたのでは.

LexRankよりもOkapiの方が正解率が高い.
問題点を解決し精度が向上した事を実際に確認できた.

提案手法の正解率の原因
→指示語や順接語,逆接語を正確に抽出できているのか？
(同じ文字列で役割が違う言葉をスコアアップしている可能性)



LexRank ＜ 提案 ＜ Okapi　の順に文章を幅広く要約に含んでいる.
→LexRankより**提案の方が起承転結が確立されている**のでは.

LexRankの要約文より
「メロスが花婿の肩をたたく」と「セリヌンティウスがメロスのほおを殴る」が約92%の類似率と算出.
部分的に文章が似ているが，人物の関係性は類似していない.
→**主語，述語の関係性**を分析に組み込むと精度が向上する？

今後の展望

①指示語の補完，主語述語関係を文書要約手法に組み込む
②品詞を参照することによる，より高精度の単語抽出の実現
③要約に対する他の評価方法の考案
本研究ではNHKの要約を参照したが，正解の1つに過ぎない.
他の手法での評価により新たな結果が得られるのでは.

〈参考文献〉※順不同
安武 凌, 野中 健一, 岩井 将行「LexRankを用いた小説文章からの自動要約手法の検討」(2024-03-03) https://www.jstage.jst.go.jp/article/wii/16/0/16_38/_article/-char/ja/
NHK「走れメロス（太宰治）|10min.ボックス現代文」(2024-03-03) https://www2.nhk.or.jp/school/watch/outline/?das_id=D0005150184_00000
Crabcamp「LexRank algorithm for text summarization」(2024-03-03) https://github.com/crabcamp/lexrank
Stephen Robertson and Hugo Zaragoza「The Probabilistic Relevance Framework: BM25 and Beyond」(2024-03-03) http://www.staff.city.ac.uk/~sbrp622/papers/foundations_bm25_review.pdf
Yutayamazaki「Okapi-BM25」(2024-03-03) https://github.com/yutayamazaki/okapi-bm25/tree/master
GMOアドパートナーズ株式会社「自然言語処理モデル(BERT)で文の意味上の類似度を計算」(2024-03-03) https://techblog.gmo-ap.jp/2022/12/21/bert_calc_sentence_similarity/
産総研「“自然言語処理”とは？-研究者も驚く急成長！自然言語処理でAIどうしがつながる未来-」(2024-03-03) https://www.aist.go.jp/aist_j/magazine/20230621.html#tid-1
ku-npl「jumanpp」(2024-03-03) https://github.com/ku-nlp/jumanpp
HuggingFace「The AI community building the furue.」(2024-03-03) https://huggingface.co/