2V-04

トピック中心文とキーワードを用いたTextRankによる抽出型要約の冗長性削減の提案

蔡宇鋒 † 望月久稔 †
† 大阪教育大学

1 はじめに

ネットワークの利用で，情報量が指数的に増加している．膨大な情報から重要な部分を抽出するために，自動要約の技術が必要である．TextRank[1] は教師なし学習モデルの一つで，抽出型要約やキーワード抽出などに用いられる．TextRank による抽出文は文書にある文のみで構成されるため記事内容と乖離する内容は生成しない [2]．しかし，同じトピックから複数の文を抽出することを考慮しないため，重複による冗長性が生じえる．本研究の目的はトピックモデル LDA[3] を用いて TextRank の冗長性を削減することである．実験は CNN/DailyMail[4] のデータセットを用いて，抽出文と参考要約の一致度を Rouge スコアで評価する．抽出文の類似度を抽出文の平均 cos 類似度で評価する．

2 トピックの中心文とキーワードによるTextRank の冗長性削減

2.1 記事と LDA で分けたトピック中心文の類似度

記事全体とトピックの中心文の類似度を用いて冗長性の低い文を抽出することを提案する．まず，LDA で記事を 3 つのトピックに分け，それぞれのトピックに属する文から，そのトピックに属する確率の一番高い文を中心文とする．同じトピックの文の場合，共通のキーワードが多く含まれるため，複数の文を抽出するとキーワードが重複し，冗長性が生じる．そこで，要約モデルで 3 文を抽出し，異なる抽出文の間における重複度を調べるために，抽出文の平均単語数 SL と平均共通キーワード数 CK から共起密度 $\rho = \frac{CK}{SL}$ を定義する．抽出文が長くなると共通キーワード数が大きくなるが，ρ が小さければ，抽出文は長くても共通キーワード数が少ないことを示すため重複がなく冗長性が低いことを表す．

それぞれ 3 つのトピック中心文の間には共起密度が低い特徴があると考え，TextRank を用いて計算するスコアに記事の各文とトピック中心文の cos 類似度を用いることで，異なるトピックから文を抽出する．3 つのトピックの中心文から順番に，すべての文との cos 類似度を計算し，中心文との類似度が高い文を抽出しやすくする．最終的に TextRank のスコアと合わせ，記事全文の重みを計算する．

2.2 各文の単語とトピックキーワードの共起回数

LDA で分けたトピックの中心文との類似度を用いるだけでは，中心文に含まれないキーワードと文の長さを考慮しない．よって，2.1 節で提案した手法に各文の単語数，記事各文とトピックキーワードの共起回数を加える．抽出文の単語数が少なく，単一トピックのキーワードが多く含まれると，抽出文に含まれるトピックキーワードの密度が高く，そのトピックに属する確率が高いため，それぞれのトピックから一文ずつ抽出することにより，冗長性を削減できると考える．まず，LDA で記事を 3 つのトピックに分け，それぞれのトピックのキーワードを Python のモジュールである textrank4zh の TextRank4Keyword[5] モデルを用いて 12 個抽出する．次に，抽出したキーワードと記事の文の共起回数が多く，文の単語数が少なければ，その文をトピックの重要文として，単語数あたりの共起回数に高い重みをつける．

3 冗長性削減の検証

実験はデータセット [4] の学習データから 10000 個の記事を抽出して行う．評価は 11490 個記事のテスト用データを用いて評価する．

3.1 記事の各文とトピック中心文の類似度を用いた共起密度減少による冗長性削減

1，TextRank による抽出文，2，トピックの中心文，3，TextRank スコアとトピック中心文との類似度によ

A Proposal to Reduction in Extractive Summarization by Using Topic-Centric Sentences and Keywords with TextRank
†CAI YUFENG and Hisatoshi MOCHIZUKI
‡Osaka Kyoiku University

表 1: 共起密度の比較

	共起密度
textrank 抽出文	0.0413
トピック中心文	0.0108
textrank と中心文との類似度から抽出した文	0.0383

表 2: TextRank と提案手法との比較結果

	rouge-1	rouge-2	rouge-l	cos-sim
textrank	0.3017	0.1030	0.2330	0.6786
textrank+LDA	0.3053	0.1055	0.2403	0.6420
textrank+LDA+keywords	0.3138	0.1092	0.2626	0.6068

る抽出文の 3 つモデルの抽出文の共起密度を調べるために，まず，それぞれの抽出文が属するトピックのキーワードとの共起キーワード集合を求める．その後，各モデルの異なる抽出文の平均共通キーワード数を求め，抽出文の平均単語数を合わせ，共起密度を求める．

それぞれモデルの実験結果を表 1 に示す．TextRank による抽出文の共起密度が 0.0413 で最も高く，トピック中心文の場合 0.0108 で最も低かったため，重複するトピックのキーワードは TextRank による抽出文の間に多く，トピック中心文の間には少ないことがわかる．記事の各文とトピック中心文の cos 類似度を用いることにより，TextRank の共起密度を約 7.2 %削減できた．

次に，TextRank と 2.1 節で提案した手法の評価用データにおける実験結果を表 2 の 1,2 行目に示す．トピック中心文との類似度を用いることにより，TextRank のみを用いる場合より，$Rouge-1, Rouge-2, Rouge-l$ がそれぞれ 1.2 %，2.4 %，3.1 %上がり，cos 類似度が 5.4 %下がった．トピック中心文の間に重複するトピックキーワードが少ないため，TextRank に記事各文とトピック中心文の類似度を加え，TextRank の共起密度が下がることで，抽出文に含まれる同一トピックのキーワード数が下がることがわかる．ゆえに，冗長性が下がり，Rouge スコアが上がったと考えられる．したがって，TextRank に記事の各文とトピック中心文の類似度を加えることにより，抽出する文は TextRank のみを用いる場合より参考要約に一致し，冗長性が下がった．

3.2 共起回数を用いた同一トピック内からの抽出文の制限

キーワードと文が同一トピックのときと同一トピックではないときの平均単語共起回数をそれぞれ実験する．同一トピックのキーワードと文の平均共起回数は 1.55 回であり，同一トピックではない場合の平均共起回数は 0.45 回である．同一トピックの場合の平均共起回数は同一ではない場合の 3.44 倍である．よって，各トピックに属する文の単語とそのトピックのキーワードとの平均共起回数が多い．2.2 節で提案した手法の評価用データにおける実験結果を表 2 の 3 行目に示す．中心文の類似度に記事各文とトピックキーワードの共起回数と文の単語数を加えることにより，$Rouge-1, Rouge-2, Rouge-l$ がそれぞれ 4.0 %，6.0 %，12.7 %上がり，cos 類似度が 10.6 %下がった．記事の文に一つトピックのキーワードが多く含まれると，そのトピックの重要文になる確率が高い．それぞれトピックの重要文を抽出することで，TextRank の冗長性が下がり，Rouge スコアが上がったと考えられる．したがって，2.1 節で提案した手法に比べ，抽出文は参考要約にさらに一致し，抽出文の類似度が下がった．

4 おわりに

TextRank にトピック中心文との類似度と記事の文とトピックにおけるキーワードの共起回数や文の単語数を加えることで，冗長性を下げることができた．今後の課題としては，LDA トピックモデルのトピック数を自動的に決めることが挙げられる．

参考文献

[1] 基于 TextRank 算法的自文摘，from<https://blog.csdn.net/qq_51938362/article/details/121985400> (accessed 2022-12-09).

[2] 重松 祐匡, 尼崎 太樹：文章間情報を用いた抽出的要約生成器 EST，人工知能学会全国大会論文集，Vol.36，No.1，p1(2022).

[3] 利用 python 做 LDA 文本分析，从里入手？，from<https://www.zhihu.com/question/39254526/answer/2313310517?utm_id=0>, (accessed 2022-12-09).

[4] CNN-DailyMail News Text Summarization, from<https://www.kaggle.com/datasets/gowrishankarp/newspaper-text-summarization-cnn-dailymail?resource=download>, (accessed 2022-11-09).

[5] TextRank4ZH, from<https://github.com/letiantian/TextRank4ZH/tree/master/textrank4zh> (accessed 2022-12-15).