2U-02

文章の相対位置関係に基づくユーザ知識レベルに応じた記事要約の提案

阪田晴香[§1]　*Panote Siriaraya*[§1]　王元元[§3]　河合由起子[§1,2]

§1 京都産業大学　§2 大阪大学　§3 山口大学

1 はじめに

本研究では，検索キーワードの出現する文章（センテンス）を起点とした周辺文章との位置関係に基づき文章を抽出し，検索キーワードに対するユーザの知識レベルに応じた記事要約手法を提案する．近年情報過多により，検索や要約技術の重要性が高まっている．その 1 つとして検索結果のスニペットがあるが，それらの要約文ではユーザの知識レベルによっては不明な文章が多くなってしまい，不明な単語や文章による再検索が必要となる．そこで，本研究では，ニュース記事を例にとり，検索キーワード（トピック語）の出現する文章を起点とした文章間の距離ごとに要約することで，知識レベルに応じた要約文生成の実現を目指す．提案手法では，トピック語を含む文章と周辺の文章間の距離に着目し，それら空間長（距離）を変化させることでトピック語に対して関連の高い/低い文章集合として新たに生成し，特徴となる文章を抽出する．これにより多様な要約を生成できる．

具体的には「トピック語を含む文章とその文章の前後に出現する文章との位置関係はトピック語を説明する内容の詳細度と関連する」という仮定に基づき，トピック語を含む記事を検索結果より取得し，トピック語を含む文章の前後の文章を取得し，それら文章集合に対して LDA により類似度の高い文章を抽出することで要約を作成する（図 1）．トピック語を含む文章と前後の文章の間隔を広げていくことで，異なる多様な要約文が作成できる．例えば，トピック語から空間的に遠い文章集合はトピック語との関連が高く，高い知識レベルを必要としない一般読者に向けた要約文を生成できる．一方で，トピック語から空間的に近い文章集合はトピック語との関連が低い文章で構成され，距離の遠いものと比較して専門知識を持つ読者に向けた専門性の高い要約文を生成できる．

本稿では，トピック語の検索結果から複数サイトのニュース記事を取得し，それら全文を用いて抽出したものをベースライン 1，第一パラグラフの集合だけを用いたものをベースライン 2 とし，提案手法による文章間の距離に基づき抽出された各ベクトルの類似度と

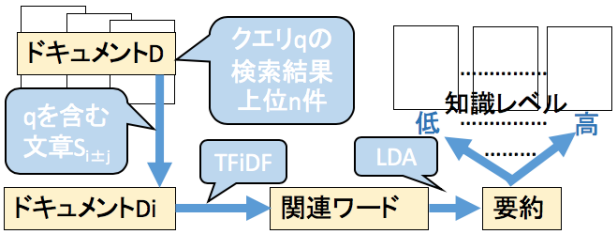

図 1: 要約生成処理の概要

の比較より，提案手法を検討する．

2 関連研究

要約技術は古くから多く研究されており，ドキュメント内から重要な文章を抽出して用いる手法が一般的である．ニュースはリアルタイム性が高いため，要約生成ではタイムラインの時間軸に基づく研究がなされている．柏井ら [1] は，時間経過とともに新情報を持った文書を追加していき，LDA を用いて表層一致と潜在的意味の一致で要約文を生成する手法を提案している．また，Giang ら [2] は，LTR というタイムライン要約のためのトレーニングデータ生成を最適化するフレームワークを提案している．本研究では，LDA を用いる点がこれらとの共通の技術要素となる．

また，空間的位置に基づいた要約では，言語によって異なるがニュース記事では一般的に第一パラグラフあるいは最終パラグラフの重要性が高いことが認知されており，パラグラフの位置に基づいて重要な文章が抽出される．近年では，Gunes ら [3] の LexRank という重要文抽出手法がある．これは PageRank に基づいた手法で，確率グラフベースで相対的重要度を計算し重要な文を選択する．本研究では重要文との関係性が単純に空間的距離の近さとして置き換えられ，多様な話題の要約を生成できる点が従来研究との特異点となる．

3 システム概要

提案する要約生成手法では，トピック語を含む文章の周辺の文章間の距離に着目し，それら空間長（距離）を変化させた文章集合を作成し，LDA を用いてトピックと関連する類似文章を抽出することで異なる要約を生成する．以下に処理の流れを示す（図 1）．

1. トピック語（検索クエリ q）から検索結果上位 n 件のドキュメント D を取得
2. ドキュメント D から TFiDF を算出しコーパス構築

A proposal of document summary according to knowledge levels based on relative positions between sentences

§1 Haruka SAKATA　§1,2 Yukiko KAWAI　§1 Panote SIRIARAYA
§3 Yuanyuan WANG
§1 Kyoto Sangyo University
§2 Osaka University
§3 Yamaguchi University

図 2: 空間距離に基づくニュース要約システムの出力例

表 1: トピック語とそれを含む文章数と総文章数

トピック語	総文章数	トピック語を含む文章数
クロマグロ	163	24（14.7%）
シャンシャン	214	8（3.7%）
地震	1111	50（4.5%）
大坂なおみ	1,200	10（0.8%）
来訪神	1,430	22（1.5%）
平均	823.6	22.8（2.8%）

3. 取得したドキュメント D のうちトピック語 q を含む文章 S_i を検出（i は文章の出現順番）
4. 文章 S_i から文章間の距離となる文章 $\pm j$ 番目の文章 $S_{i\pm j}$ を抽出し，ドキュメント D_j を生成
5. ドキュメント D_j から TFiDF より関連ワードを抽出
6. ドキュメント D_j の各文章に対して LDA を用いて類似する文章上位 m を抽出し，要約を生成

要約生成の一般的な処理はドキュメント D に対して 4. の TFiDF および LDA などによる重要文抽出となり，提案手法での文章間の距離に基づいた文章抽出は処理コストを削減しつつ，様々な要約生成が可能となる．

4 実装と評価

4.1 インタフェース

図 2 にニュース要約システムのインタフェースを示す．ユーザは検索キーワード（トピック語）を入力すると，知識レベルの低いものから順に複数の要約と関連する単語を取得できる．図 2 では「シャンシャン」の要約結果として，トピック語に近い文章より要約されたものから順に提示しており，ユーザは提示されたトピック語と関連語に基づいて，自身の知識レベルに合わせて要約文を選択できる．

4.2 検証

本稿では，提案手法による文章間の距離に基づき抽出された各ベクトルの類似度を評価し，提案手法の有用性を検討する．

実験は，トピック語より取得する記事数 n=100 とし，「シャンシャン」など 5 件のトピック語を用いた（表 1）．ドキュメント D の平均文章数は 823.6 行となり，そのうちトピック語を含む平均文章数は 22.8 行であった．よって，本実験における 1 記事あたりの文章間の距離

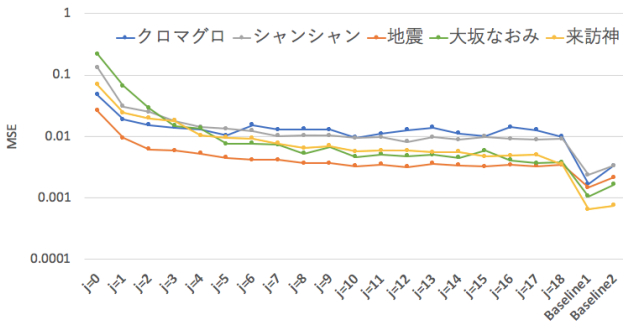

図 3: 類似度評価結果

$\pm j$ は 0 から最大 18 行とした．また，LDA は 30 次元とした．

比較として，D の全文を用いたものをベースライン 1 とし，ニュースで重要とされる第一パラグラフのみを用いた文章集合をベースライン 2 とした．評価方法は，抽出したベクトルの値の平均値を算出し，それらベクトルの平均二乗誤差（MSE）を抽出した．

図 3 に結果を示す．ベースライン 1 と 2 は MSE が 0.01 以下であった．つまり，網羅性が高く概要となる要約が生成されているといえる．提案手法では，全てのトピックに対して j=0 での MSE が最大となっており，局所的な要約となった．j の増加に伴い MSE は減少傾向になっており，$j = 5$ 以降では横ばいとなり，ベースライン 1 と 2 に対する MSE の差は最大でも 0.001 程度となった．以上より，文章の間隔が近いと局所的な話題の要約を生成でき，広げることで概要となる要約が生成されることが明らかとなった．

5 まとめ

本論文では，知識レベルに応じたニュース記事要約の生成を目的に，トピック語を含む文章と周辺の文章間の距離に基づいた要約生成手法を提案し，検証した．実験より，文章間の距離を広げることで概要となる要約が生成されることが明らかとなった．今後，トピック語を増やし要約文に対するユーザ評価を実施する予定である．

謝辞

本研究の一部は，総務省 SCOPE（受付番号 171507010），JSPS 科研費 16H01722，15K00162，17K12686 の助成を受けたものである．ここに記して謝意を表す．

参考文献

[1] 柏井香里, 小林一郎. 文書の潜在情報と表層情報を考慮したタイムライン要約への取り組み.

[2] Nam-Khanh Tran Mohammad Alrifai Giang Binh Tran, Tuan A. Tran and Nattiya Kanhabua. Rouge: a package for automatic evaluation of summaries. 2004.

[3] Gunes Erkan and Dragomir R. Radev. Graphbased lexical centrality as salience in text summarization. 2003.