5P-01

蓄積情報からの特徴語抽出に基づく自動要約・提示システムに関する研究

松林　圭† 山下 晃弘†　中村　拓哉‡　吉田　優之§ 川村　秀憲††
東京工業高等専門学校†　株式会社調和技研‡　北海道アルバイト情報社§　北海道大学††

1. はじめに

掲示板やSNSが普及してくるにつれ，日報やディスカッションなどの企業業務のナレッジマネジメントに応用をする動きが活発になっている．しかし掲示板やSNSは時系列で投稿が表示される場合が多く重要性の高い情報であってもユーザーが把握できずに埋もれてしまう場合もある．そのため，社内に蓄積された様々な情報を可視化する研究も多く，それらを応用した様々なサービスも存在する．本研究ではそのような大量のテキストデータから重要な情報を自動要約し，社内のナレッジとして有効活用することを目的としている．最終的に社内で活用されるシステムを目指し，複数のテキスト要約手法の比較や実データによる有用性の検討を行った．

2. システム構成と応用例

想定している自動要約システムを図1に示す．システムでは大きく(A)～(D)の4つのプロセスをサイクルさせることによって社内情報の有効活用・ノウハウ蓄積を継続的に行っていく．(A)ユーザが書き込んだ掲示板や社内報等の社内で取り扱われている文書に対して重要な情報の抽出を行う．(B)その文書を本研究で検討するアルゴリズムを用いて自動要約する．その際には文書が単一/複数の場合を考慮する．(C)要約結果を対話用のロボットや人にアウトプット/蓄積を行う．(D)ユーザからの問い合わせや質問に対して応答する形で内容をアウトプットする．

本研究では，アルバイト求人情報を扱う企業の社内スレッド型掲示板の投稿データ約 24,000 投稿を評価用の実データとして用いる．それらのテキストデータを用いた具体的なサービスとして，特定のキーワードをクエリとして検索した場合に求人情報などの要約が取得できるサービスなどを想定する．

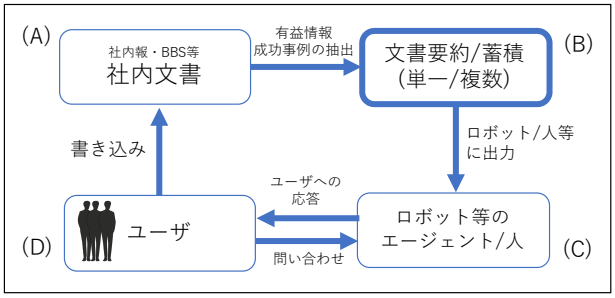

図1 想定しているシステム構成

3. 本研究の位置づけ

まず文書要約の方法を検討する前に要約の目的を明確にする必要がある．一般に要約の目的は「指示的要約」と「報知的要約」に大別される．指示的要約は，元の文書を読むべきかどうかを判断するための要約の事を示す．例としては新聞の見出しの一文が挙げられ，必要に応じて本文を読むことを前提としている．一方で，報知的要約は，元の文書の概要を伝える要約で，テレビニュースの字幕などが例として挙げられる．本研究で想定するサービスでは報知的な場合を取り扱う．次に，入力が単一文書か複数文書かによって分けられる．単一の場合とは，一つの文書から一つの要約を作成することであり，一方で，複数文書から作成する場合とは同じ事柄に対する複数の文書から，一つの要約を生成することをいう．本研究ではある話題について複数の投稿が存在する社内文書を用いるため，複数文書からの要約作成を想定して研究を行う．また，文書の要約には，その文書の概要を示す一般的な要約が必要な場合と，利用者がキーワードなどを指定してそのキーワードに関する要約を文書から生成する場合が存在する．ユーザから何らかの指定を受ける場合をクエリ依存，それ以外の場合をクエリ非依存と呼び，本研究ではクエリ依存型をベースとして要約文書を取得するシステムを想定する．そして，要約手法については要約対象の文書の中から重要と思われる文を抽出する抽出的要約と，意味を汲み取

A research on document summarization and presentation system based on feature word extraction from stored informations
† National Institute of Technology ,Tokyo College Advanced Course of Mechanical and Computer Systems Engineering
‡ CHOWA GIKEN Coporation
§ HAJ Corporation
†† Hokkaido University

り抽象化を行った上で適切な要約を作成する生成的要約が存在する．後者に対しては意味表現の生成文の整合性等の問題があり非常に困難であるため，本研究では抽出的要約を行う．

4. 抽出手法について

要約文を抽出するにあたって一般的に用いられている代表的なアルゴリズムは大きく 3 種類に分類できる．グラフ構造を作成して各ノードの関連性を元に要約を作成するグラフベースの手法[1]，文の特徴語を定義して，その特徴による重み付けを行うことによって文選択を行う特徴ベースの手法[2]，文章のトピックを算出し，そのトピックに沿った文を選択するトピックベースの手法[3]である．本研究ではグラフベースの手法である LexRank[1]を用いる．LexRank とは検索エンジンの重要度を決定する PageRank を応用したアルゴリズムで，ノードを文書，エッジを 2 文書間の類似度としてグラフ構造を作り，重要度を計算する．重要度が高い文は要約文として抽出すべき文と考えられる．類似度は文書を特徴量として用コサイン距離を用いる．特徴量抽出手法として Mecab で名詞のみを抽出した後，LexRank の論文でも用いられている TF-IDF による特徴量抽出と FastText[5]で作成した特徴量を用いる 2 通りの手法を実装した．

5. 冗長性削減

LexRank のアルゴリズムに従って文を抽出していくと，冗長性のある要約文が生成される可能性があり[4]，要約文の長さに限りがあるタスクに置いては効率が悪い．その為，本研究では代表的な文を抽出しつつ冗長でない文を生成可能な MMR を用いる．式は(1)のように表し，検索質問 Q が与えられた時に文書集合 R から次に選択する文書を求めるものである．D は対象文書群，S は既に選ばれた D 内の文集合，λ_w は重みパラメータを示す．

$$MMR = \arg\max_{D_i \in R \setminus S} \left[\lambda_w Sim_1(D_i, Q) - (1 - \lambda_w) \max_{D_j \in S} Sim_2(D_i, D_j) \right] \quad (1)$$

6. クエリ考慮について

クエリを考慮した要約を実現するために，クエリと類似する文に対して LexRank のスコアから加点する．クエリを正解文として考え，その他の文書とのコサイン距離に重み λ_b を掛けることで加点する値を計算する．また，TF-IDF においてキーワード集合の単語が他の文書に存在しなかった場合，値が直交しコサイン距離が 0 となる．λ の値は予備実験により調整した．

$$\text{計算値} = Lexrank(D_i) + \lambda_b CosDist(K, D_i) \quad (2)$$

7. 実験

要約手法の動作確認・検証のため幾つか実験を実施した．パラメータとして MMR は λ=0.5，バイアスは λ=0.5 を掛ける/掛けない状態の TF-IDF, FastText について行った．対象文書に対しては 25～40 行程のアルバイト関連情報の投稿データを用いた．TF-IDF で特徴量を抽出し MMR，バイアスを適用しクエリとして”面接”を指定した計算を行った．MMR を用いた場合長い文書は既に抽出された文書と類似しやすいため順位が低くなる傾向にあった．また，クエリと類似する文に加点した場合，”面接”の単語が含まれる単語が上位に来ていることが確認できた．Fasttext を用いた場合では TF-IDF と比べて順位毎の値の差が小さいものの類似した結果が出力されることが分かった．

8. 結論

与えられた雑多な文書から自動要約を行うシステムを構築するため，LexRank アルゴリズムを用いた．特徴抽出では TF-IDF と FastText による特徴を用いて，MMR による冗長性の排除やクエリと類似する文に対する加点を行う事よってクエリを考慮したモジュールを構築した．

今後は全体のシステムとして実装を行い，実データに基づいて評価を行う．

参考文献

[1] G. Erkan, 他 : LexRank: Graph-based Lexical Centrality as Salience in Text Summarization, Journal of Artificial Intelligence Research 22, pp457-459, 2004.
[2] Jagadeesh J, 他 : Sentence Extraction Based Single Document Summarization, Workshop on Document Summarization, 19th and 20th March, 2005
[3] Makbule Gulcin Ozsoy, 他 : Text summarization using Latent Semantic Analysis, Journal of Information Science, 2011
[4] 北島理沙 ,他 : トピックを考慮した複数文書要約への一考察, 言語処理学会 第 19 回年次大会発表論文集, pp. 504-507, 2013
[5] Armand Joulin, 他 :Bag of Tricks for Efficient Text Classification, 2016