論文要約 575: 575 の音韻的読みやすさを付与した学術論文の要約文生成

安部 文紀 * 寺田 実 *

概要. 学術論文の投稿と検索に特化した Web サービスの台頭によって, 所望の論文を容易に入手できるようになったが, 論文の概要を把握する手間と概要を忘却する懸念は依然として存在する. そこで本研究では, 学術論文の概要を把握・記憶することの支援を目的として, 学術論文の要約文を 575 形式のキャッチフレーズで生成することに取り組む. これは, 古来より親しまれてきた俳句や川柳といった 575 の持つ音韻的な読みやすさを要約文に適用すれば, 論文が印象に残りやすく内容も把握しやすくなるだろうという仮説に基づいている. 本論文では, 仮説を実証するために実装した提案システム：論文要約 575 について述べる.

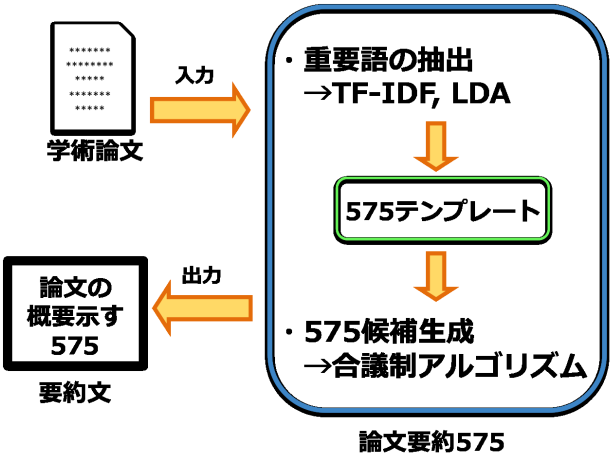

図 1. 論文要約 575 概要図

1 はじめに

Google Scholar[1] のような論文検索エンジンや, arXiV[2] のような論文投稿 Web サイトの台頭で所望の論文を容易に入手できるようになったことに伴い, 膨大な量の論文から自分が必要とする論文の選別や, 論文の概要を素早く把握する能力, 記憶しておく能力がこれまで以上に求められるようになった.

本研究では, 学術論文の概要を把握・記憶することの支援を目的とし, 古来より親しまれてきた俳句や川柳といった 575 の持つ音韻的な読みやすさを付与した論文の要約文生成に取り組む. 具体的には, 内容を把握しやすく, 印象に残りやすい 575 形式のキャッチフレーズを学術論文に対して自動生成する. 例えば, 本論文を提案システムに入力した場合, 図 1 のような要約文が出力されることを目指す.

* 電気通信大学 情報理工学研究科 情報・ネットワーク工学専攻

[1] https://scholar.google.co.jp/

[2] https://arxiv.org/

本研究は, 論文の概要を把握しやすく, 印象に残りやすくすることで読者に貢献する. 著者に対しては論文に付けるキャッチフレーズやキーワードの生成を代行, 考案を支援することで貢献する. 近年の論文では研究のダイジェストとなる画像やイメージ図, もしくは研究のキーワードが論文冒頭に付与されることがあるため, 将来的にはそのような研究のダイジェストにおいて 575 が用いられることを目指す.

2 関連研究

要約文生成の研究には抽出型要約と生成型要約が存在する. 抽出型要約では LaTeX 文書のセグメント構造に基づく手法等を用いて各文を重要度の高い順に並び替え, いくつか選んで要約文としている [1]. 生成型要約では人間が要約する際のパラフレーズのような再構成を Encoder-Decoder 等の機械学習を用いて行い, 生成した単語列を要約文としている [2].

他にも一種の要約文生成とみなした論文タイトルを生成する研究も存在する [3]. タイトル生成は論文概要を短い文で表現する点で本研究と類似するが, タイトルは文章の内容を表現するのに対し, 575 は文章のイメージを喚起させる点で相違する. 言い換えると内容指向かイメージ指向かの違いがある.

古来より俳句は人手で書かれ詠まれ親しまれてきたが, 自然言語処理技術を用いて機械的に生成しようと試みた研究が存在する [4]. これは自動生成する俳句が人手で作ったものにどれだけ近づけられるかを目的に, そして俳句の美しさをどれだけ持てるかを課題に行われてきたため, 俳句生成を要約文生成と見なして取り組んだ研究が存在しない.

3 提案システム：論文要約 575

論文要約 575 は図 1 の通り, 3 つのステップを踏む.

1. 重要語の抽出
2. テンプレートに重要語を割り当てて 575 生成
3. 候補の中から適切な 575 を出力

3.1 重要語の抽出

重要語は, 文の表層情報に基づく手法と文の潜在情報に基づく手法を組み合わせて名詞をスコアの高い順にランク付けして決定する. 表層情報に基づく手法には TF-IDF を用い, TF には入力論文における単語出現頻度を, IDF には論文全文をピリオドで区切って抽出される文の集合を文書集合としたときの単語の文出現頻度を用いた. 潜在情報に基づく手法には, トピックモデリングの潜在的ディリクレ配分法 (Latent Dirichlet Allocation, LDA) を用いた. 潜在トピックのサンプリング手法にはギブスサンプリングを用い, トピック数は階層ディリクレ過程 (Hierarchical Dirichlet Process, HDP) を用いることで無限化した. 文書集合には言語処理学会論文誌 LaTeX コーパス [3] に入力論文を加えたものを用いた. LDA の適用により, (1) 各潜在トピックを表す単語が生起確率とともに得られ, (2) 各論文が持つ潜在トピックが生起確率とともに得られる. 本研究では (2) において入力論文を最も占める割合のトピックを入力論文のトピックとし, そのトピックにおける各単語の生起確率を重要度とした.

文の表層情報と潜在情報を組み合わせるために, unigram-rescaling (式 (1)) を用いた. これは表層情報 u に基づく単語 w の生起確率が, 潜在トピック z が与えられたときの生起確率に調整される. これによって単語の重要度に入力論文の持つトピックも考慮したスコアが得られる.

$$p(w|u,z) \propto p(w|u)\frac{p(w|z)}{p(w)} \tag{1}$$

3.2 テンプレートに重要語を割り当てて 575 生成

575 テンプレートは重要語を当てはめるための 575 の骨組みであり, WISS 2016[4] 論文集の論文 82 本分の手製 575 を用いて作成した. 575 の名詞部分を重要語と置換するための記号に変換しており, 機能語はそのまま配置している. このテンプレートに抽出した重要語上位 10 個を配置し, 候補を生成した.

3.3 候補の中から適切な 575 を出力

候補から最終的な 575 を決めるため, 合議制アルゴリズムで各々に対して評価を行った. 合議制アルゴリズムでは, (1) アブストラクトとの対応, (2) タイトルとの対応, (3) 575 中の係り受け関係 の 3 点について合議制をとる. これは学術論文から 575 を人手で作成した際に, 筆者が重視した点を経験則として反映させたものである. (1) ではアブストラクトに存在する単語が 575 に入っていれば 1 票, (2) ではタイトルに存在する単語が 575 に入っていれば 2 票とし, (3) では係り受け解析器 CaboCha[5] を用いて 575 中の各句の係り受け具合をスコアで取得し, その総和に応じて票数を決定した. 例として, 筆者の過去のデモ発表論文 [6] を入力したときに票数の多かった上位 5 件の 575 を以下に示す.

単語スコア　位置の候補に　手法する
スコアページ　位置の候補に　手法する
ページ単語　位置の候補に　手法する
ページスコア　位置の候補に　手法する
スコアタップ　位置の候補に　手法する

4 まとめと今後の課題

本研究では, 学術論文の要約文を 575 で生成するための重要語抽出, 575 生成手法を提案し, 575 自動生成システムの論文要約 575 として実装した.

本研究のアイデアは, 575 の持つ音韻的な読みやすさを論文の要約文生成に適用すれば, 論文の概要が把握しやすく, 印象に残りやすくなるのではないかという点である. そのため, 論文の全体像を把握しやすく, 印象に残りやすくすることで読者に貢献し, 著者に対しては研究のダイジェストとなるキーワード生成を代行, 考案を支援することで貢献する.

出力された論文要約 575 には音韻的な読みやすさはあるものの, 名詞と動詞の組み合わせが不適切なものや係り受けスコアに全体のスコアが引っ張られている 575 が上位に確認できたため, 合議制アルゴリズム含め, 候補から適切な 575 を選択するアルゴリズムを見直すことを課題とする. また, 学術論文において一般的な単語「手法」が 575 に採用されているため, トピックモデリングにおいて推定した潜在トピックの中からひとつを選ぶ際にどの論文にも当てはまるトピックを避ける等の対策を講ずる. また, システムから生成される 575 の評価実験を通して, 本研究の主張である「論文の概要が把握しやすい」, 「一度読んだ論文が印象に残りやすい」の 2 点を確認することを課題とする.

参考文献

[1] SHIN Wonha 他. “ セグメント構造を考慮した学術論文の包括的要約の自動生成の提案 ”. 言語処理学会第 23 回年次大会発表論文集, pp. 230−233, 2017.

[2] 衣川和尭 他. “ 学術論文の章構造に基づくニューラル自動要約モデル ”. 言語処理学会第 23 回年次大会発表論文集, pp. 150−153, 2017.

[3] 大部達也 他.“ Recurrent Neural Network を用いた抽出型および生成型論文タイトル生成について ”. The 31st Annual Conference of the Japanese Society for Artificial Intelligence, pp. 3A1-1, 2017.

[4] Rafal Rzepka et al. “ Haiku Generator That Reads Blogs and Illustrates Them with Sounds and Images ”. Proceedings of the Twenty−Fourth International Joint Conference on Artificial Intelligence, pp. 2496−2502, 2015.

[3] http://www.anlp.jp/resource/journal_latex/

[4] https://www.wiss.org/WISS2016/

[5] http://taku910.github.io/cabocha/

[6] http://www.wiss.org/WISS2016Proceedings/demo/3-A17.pdf