E-024

Web 小説の見出しに有用なキーフレーズの自動抽出の検討
An extraction method of useful keyphrases for web novel headlines

古川　拓実†　　菱田　隆彰†
Takumi Furukawa　Takaaki Hishida

1. はじめに

近年，Web 小説投稿サイトの隆盛が著しく，それに伴い創作される小説の量は膨大である．主要な小説投稿サイトの一つである小説家になろうでは，毎月一万程度の新規作品が発表されている．

小説の数の増加により，読者がより好みの小説を選べるようになったことは好ましい反面，小説を選定するのにかかる時間や労力は増えざるを得ない．選定する時間を短縮する方法にはユーザの読んできた小説の履歴から推薦，小説選定時に閲覧される本文の圧縮などが考えられる．我々は本文の圧縮として見出し生成に着目した．

見出し生成文に求められる性質として，様々なものが考えられる．ニュース記事とその見出しの自動生成を行う自動報道システムについては，Leo ら [1]により，透明性，正確性，修正可能性，ドメイン適用性，流暢さ，データの入手性，話題性が挙げられている．小説分野における見出し生成でも流暢さや正確性は重要だが，それ以上に娯楽性が重要な性質であると我々は考える．

娯楽性を重要視した見出し生成についての先行研究は少ない．我々は娯楽性を重要視した見出しとして，人の興味を引く見出しの自動生成をテーマとして研究を進めてきた．初めに人の興味を引く理論の一つとして広告などの文言に利用されるコピーライティング技術を参考に，人の損得感情を揺さぶる用語，つまり，感情生起要因に着目した．また，感情分析を行うライブラリ kamaboko を作成し [2]，相互情報量を用いて感情生起要因の取得を試みた．

しかし，それだけでは取得できる語彙が少なく，見出しを生成するには不十分であることが判明したため，別の方法で語彙やキーフレーズの抽出を行う必要が生じた．

本論文では，見出しを生成するために有用なキーフレーズを抽出する方法として，Web 小説の本文の一文一文を IOB2 形式の独自のタグでタグづけしたデータセットを用いてモデルのトレーニングをおこない，求めるキーフレーズの抽出精度について評価を行う．

2. キーフレーズ抽出に用いる手法

キーフレーズの抽出にはさまざまな手法が応用される．統計的な手法である TF-IDF や，Google 検索エンジンで使用された実績のある PageRank をテキストに適用できるように拡張した TextRank などがその代表例である．

しかし，いずれの手法もワードを処理単位としており，キーフレーズを抽出するためには十分に近い位置にあるキーワード同士を連結するなどの工夫が必要である．また，これらの手法に基づいた処理によって抽出されたキーフレーズはどのような用途で扱えばいいのかという情報に乏しい．例えば，キーフレーズが名詞であることが分かったとしても，見出しの主語に相応しいのかは見当がつかない．

†愛知工業大学　Aichi Institute of Technology

チャンキングは一連のトークン列一つ一つを分類していくことで標識を与える手法である．チャンキングにより，見出しを生成する際に語彙の用途に関する情報が利用可能となる．また，TextRank や TF-IDF では取得の難しい複雑な表現に対して容易に対応できる．我々は深層学習モデルの一つである BERT をファインチューニングし，チャンキング問題に適合させることでキーフレーズの抽出を実現することとした．

3. 見出し生成のためのフレーズ種別

よく知られた自然言語処理コンペティションである CoNLL-2003 ではチャンキングの一つである固有表現認識が課題として扱われた．抽出する固有表現の種別は人名（PER）・組織（ORG）・場所（LOC）・雑多な名前（MISK）が用いられた．

一般的には，CoNLL-2003 で示される比較的汎用的な意味を持つ種別の定義がそのまま利用されることが多いが，よく利用される定義が見出し生成にとって有用であるかどうかには議論の余地がある．

表 1：フレーズの定義

フレーズ種別	定義
役	主語にできるものであり，そのものが行動を起こし，他の対象に対して影響を与えたもの．また，与えうるもの．かつ，人称代名詞でなく，人名タグまたは種族タグの基準に当てはまらないもの．例：魔女様，兄ちゃん
人名	具体的な人名，または名前の一部． 例：サンライズ・サマー，ガガン
種族	魚やモンスターなどの生物学的な分類の名称． 例：陸ドラゴン，犬
場所	具体的な固有の場所の名前．寝室などは固有ではないので適合しない． 例：サジタリウス，ヤパン
表示	名詞を修飾している部分であり，その他タグの基準に当てはまらないものであること．なおかつ主語になり得ないものであること．助詞を適用対象に含める．主語になり得ない． 例：優秀な人材である，辺境の地へと追放された
出来事	一人以上の集団の物理的・精神的を問わない動作・行為に名称がつけられたもの．それ単体で意味を成す二つ以上の語彙に分割でき，一定期間で終了する一過性を備えていること． 例；政略結婚，お料理対決
能力	現実世界から見て特殊な能力の名称であること．具体的であること． 例：爆裂拳，熱平面

Copyright © 2022 by
The Institute of Electronics, Information and Communication Engineers and
Information Processing Society of Japan All rights reserved.

見出し生成では魔女や，アニキなどの人名ではないものについても人名に近い用途として認識したいという要求や，別のフレーズを修飾するフレーズが取得したいといった要求がある．したがって，我々は興味を引く見出し生成に有効な種別を新たに定義する必要があると考えた．

本研究では，見出し生成に有効なフレーズ抽出に必要な種別を定義するために，小説家になろうに投稿されていた小説三作品の各話の見出しについて，見出しの構成パターンごとに分類し，分析した．各話の見出しが単純過ぎないこと，ハイファンタジー〔ファンタジー〕ジャンルの小説であることを基準として，小説家になろうにおいて小説を一意に特定する ncode と呼ばれる属性が n2759hg，n6251hf，n2614hf の小説を選定した．

我々はその結果を踏まえ，表 1 に示すように役，人名，種族，場所，表示，出来事，能力の計 7 つのフレーズ種別を定義した．

4. 実験

定義したフレーズ種別を元にして，IOB2 タグ付方式に準拠する学習データを作成した．IOB2 タグ付方式は抽出対象である種別のトークン列の先頭には接頭辞として B-“種別名”を付与したラベルを指定し，それ以降には I-“種別名”を付与したラベルを指定する方式である．こうすることで，抽出対象であるフレーズが連続したような場合でも，各まとまりを区別することができる．なお，フレーズ種別に当てはまらないトークンには O のラベルを割り振る．

IOB2 形式の学習データ作成を支援する Web アプリケーション IOB2TagMachinGun を作成することで学習データ作成の効率化を図りつつ，学習データの作成を行った．元データとして，小説家になろうの ncode が n2759hg の本文を冒頭から取得し，約 6000 件のタグ付された学習データを作成した．各フレーズ種別及び O タグの出現する学習データの数にはばらつきがあるため，特定のフレーズ種別または O タグに対して過学習が行われないように学習データ量の調整を行なった．各フレーズ種別が付与された学習データ件数は最低でも 49 件確保し，結果として学習データの総数は 664 件となった．学習データセットは 6：2：2 の割合で分割し，それぞれをトレーニング用，バリデーション用，テスト用として使用した．

ファインチューニングする事前学習済みモデルには東北大学乾研究室が提供する日本語 Wikipedia を事前学習データとした BERT（Bidirectional Encoder Representation Transformer）を利用した．最適化アルゴリズムとしては AdamW を選択した．

5. 評価と考察

作成したトレーニングデータでファインチューニングしたモデルのテストデータに対する正解率は，94.3%であった．なお，正解率は正答したラベルの数をラベルの総数で割った値である．表 2 に各タグの F 値を示す．

各 F 値において 7 割以上の値になることが確認できた．F 値の最も高い平均は人名で 95.2%であった．この結果から今回作成したフレーズ種別の定義がある程度有用に作用したと考えられる．

表 2：各フレーズ種別の F 値

フレーズ種別	F 値		
	B-タグ	I-タグ	平均
役	82.9	70.9	76.9
人名	95.8	94.5	95.2
種族	91.0	79.6	85.3
場所	82.4	88.2	85.3
表示	72.3	73.5	72.9
出来事	77.8	78.8	78.3
能力	90.0	93.3	91.7

F 値の平均が最も低い値となったのは表示の 72.9%であった．正答できなかった具体的な例としては，「商人/と/おも/わし/き/人物」が挙げられる．我々はラベル列には「B-役/O/O/O/O/O」を割り当てた．商人は表示よりも優先したいフレーズ種別である役であり，また，「おもわしき人物」は不完全な修飾のためである．しかし，モデルは「B-役/O/B-表示/I-表示/I-表示/O」と推定した．同様な例として，名詞を修飾する「馬/に/乗っ/た」というトークン列が挙げられる．モデルの推定は「B-表示/I-表示/ I-表示」というラベル列であったが，この場合馬は種族として扱いたいため，我々が望んだ正解ラベル列は，「B-種族/O/O」という形になる．これらの例から分かるように，表示種別のラベルが付与されても良いようなトークン列の場合，優先度の高いその他の種別があることでタグ付けが行なわれていない場合においても学習モデルが反応してしまう．

6. まとめ

本研究では，見出し生成に有用なフレーズ種別を定義し，学習データを作成して，それらによるモデルトレーニングを行った．評価の結果，全てのフレーズ種別において 7 割以上の F 値を獲得することができており，フレーズ種別の定義が有用に作用することが確認できた．

課題点として，今回作成したフレーズ種別には動詞を修飾するフレーズ種別や，特徴的な強調表現を補足するフレーズ種別がないことが挙げられる．これらについて取得可能にすることでより表現豊かな見出しを生成することができると考えられる．

今後の研究では，これまでの研究で利用できるようになった語彙・フレーズを用いて見出しを生成する手法を考案していく予定である．安達ら [3]は，文章を依存木として捉え，そこから文を生成する研究を行った．我々はその内容をより発展させることで興味を引く見出しの生成を行いたいと考えている．

参考文献

[1] Leo Leppänen, Myriam Munezero, Mark Granroth-Wilding, Hannu Toivonen, “Data-Driven News Generation for Automated Journalism”, Proceedings of The 10th International Natural Language Generation conference, pp. 188-197, Santiago de Compostela, Spain, September 4-7 (2017).

[2] 古川拓実，菱田隆彰，“ユーザが変更可能な辞書を持つ感情分析ライブラリ”，令和３年度電気・電子・情報関係学会東海支部連合大会発表論文集，H6-5, (2021).

[3] 足立 顕，牧野 武則，“柔軟な文生成方式”，情報処理学会研究報告，Vol.2004, No.23, pp. 29-36 (2004).

Copyright © 2022 by
The Institute of Electronics, Information and Communication Engineers and
Information Processing Society of Japan All rights reserved.