ISS-SP-061

一般向けと知的障害者向けのニュース記事を利用した外国人向けニュース記事の自動作成

栗野　修平†　　李　淑文†　　韓　東力†
† 日本大学文理学部情報科学科

1. はじめに

グローバル化が進む現代日本には、NewsWeb Easy(NWE)[1]などネット上で閲覧できる外国人向けニュース記事が存在する。しかしそれらの記事はいまだ自動化に至っておらず、一般向けニュース記事等別の書式の記事から自動作成を試みる研究は確認されなかった。そこで、それらの記事を他の書式の記事より自動作成することができれば制作側の負担軽減など一定の価値があると考え、我々は以下の研究を行った。

2. 研究手法

外国人向けニュース記事の自動作成は、既存研究[2]より確認されている一記事当たりの平均文数・平均形態素数・一文当たりの平均形態素数及び日本語能力検定の級位といった数値的特徴をもとに、一般向け・知的障害者向け両記事の数値的特徴を外国人向けニュース記事 NWE の数値的特徴に近づけることで行う。

具体的に近づける手法として、一般向け記事はまず形態素数を絞るため、TF/IDF 法を用いて重要語を抽出する。CaboCha で係り受け解析した一般向け記事を用いて、前行で得られた重要語を含む係り元と主辞に直接かかる名詞節係り元(「○○は　××だ。」における「○○は」等)を記事において重要であるとみなし抽出することで簡約する。最後に簡約した記事内の日本語能力検定 3 級～4 級に該当しない単語を、該当する類義の単語があれば置き換える、といった手法を行う。知的障害者向け記事からは文数を絞るため、同様の方法で重要語を抽出し、重要語を含む文とリード文を重要な文とみなし抽出する。最後に同様の手法で単語の置き換えを行う、といった手法を用いる。図 1 にフローを示す。

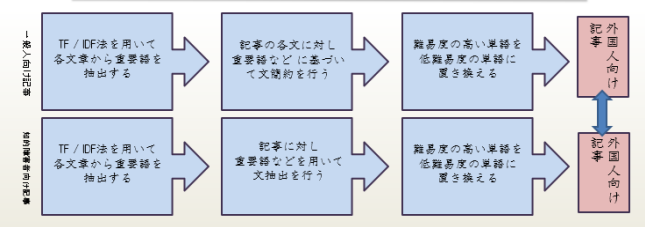

図 1. 外国人向け記事作成フローチャート

3. 記事作成

数値的特徴を近づけるにあたって、一般向け記事として NHK NewsWeb[3]より 50 記事、知的障害者向け記事として知的障害者向け新聞記事「ステージ」のコーパス[4]より 50 記事用意した。これは基準としたい数値的特徴に用いている記事数と異なるため、既存研究[2]の平均値をそのまま使用するのではなく、一般向け記事については一記事平均形態素数を近づけるため、(外国人向け記事の平均形態素数 / 一般向け記事の平均形態素数)で得られた圧縮率 0.646 を基準にして TF/IDF の閾値を定めることとした。

また知的障害者向け記事については一記事平均文数を近づけるため、(外国人向け記事の平均文数 / 知的障害者向け記事の平均文数)で得られた圧縮率 0.679 を基準にして TF/IDF の閾値を定めることとした。その結果、以下の結果を得た。

TF/IDF 閾値	origin	0.015	0.02	0.025	0.028	0.03	0.032	0.035
平均形態素数	341.2	279.0	249.8	234.4	221.1	216.7	206.1	194.4
圧縮率		0.818	0.732	0.687	0.648	0.635	0.604	0.570

図 2.一般向け記事の平均形態素数圧縮率の変遷

TF/IDF 閾値	origin	0.015	…	0.04	0.045	0.05	0.055	0.06
平均文数	12.66	12.36	…	9.84	9.54	9.4	9.12	9.18
圧縮率		0.983	…	0.816	0.794	0.780	0.757	0.761

図 3.知的障害者向け記事の平均文数圧縮率の変遷

以上の結果より、一般向けニュース記事からの作成時には閾値 0.028 を、知的障害者向けニュース記事からの作成時には閾値 0.055 を用いることとした。

4. 結果とまとめ

本大学の外国人留学生 6 名を対象に作成した記事を読んでもらった。その結果より、簡潔に伝わる文章も作成できていたものの、数値的に近づけたために詳細な情報が抜け落ちる、類義語のニュアンスが不適切で意味のずれた単語に置き換わる等のエラーもあり趣旨の伝わりにくい記事もあったようであった。これら可読性を向上させる別の手法も考慮していくことが今後の課題となりそうである。

参考文献

[1]https://www3.nhk.or.jp/news/easy/?utm_int=all_footer_menu_easy
[2]打浪ほか、『知的障害者向け「わかりやすい」情報提供と外国人向け「やさしい日本語」の相違』
言語処理学会 第22回年次大会発表論文集 p1105-p1108
[3]https://www3.nhk.or.jp/news/
[4]GSK2017-E 知的障害者向け新聞『ステージ』テキストデータ
http://www.gsk.or.jp/catalog/gsk2017-e/