Breaking News English Corpus:マルチレベルテキスト平易化に特化した複数文書要約コーパスの提案

永塚 光一　　渥美 雅保
創価大学大学院 工学研究科
e18m5212@soka-u.jp　　matsumi@soka.ac.jp

1 はじめに

複数文書要約とは, ある同一のトピックに関する複数の文書から, その要約文を生成するタスクのことを指す. インターネットを中心として多くの文書データが存在する現代において, 複数文書要約技術の確立は, 様々な実世界アプリケーションの応用に繋がると期待されている. 近年, 自動要約の分野においては, ニューラルネットワークモデルを用いた手法の研究が盛んであり, その多くは大規模データセットに基づいている [1].

こうした中, 複数文書要約のためのデータセットは単一文書要約と比較して開発が遅れており, 複数文書要約の研究を妨げる一つの要因となっている. また, 複数の文書からどのような要約が必要とされるのかといったタスク定義や, 要約の質に対する評価の指標は, 明確に定められていない.

要約タスクの一つとして, 元テキストをある指定された語彙レベルで要約するタスクが設定でき, そこでは, 要約テキストが指定された語彙レベルで要約されているかが評価指標となる. また, それら要約テキストの内容を, より分かりやすい表現に置き換えるテキスト平易化タスクも要約タスクの一つとみなされる. 自動的テキスト平易化の発展により, 外国人や初学者を含む言語学習者や, ある分野でのリテラシーが不十分な人々, 難読症患者への支援システムの開発などが期待されている. また, テキスト平易化技術は, 機械翻訳や情報抽出など他の自然言語処理技術にも応用することが出来る. 複数文書に対して, 様々に指定された語彙レベルの要約を生成するタスクの実用化は, 今後の文書要約研究において, 重要であり, 多くのユーザーにとって有益であると考えられる.

そこで, 本研究では, 複数文書の様々な語彙レベルでの要約, 及び要約テキストの平易化に利用可能な新たなコーパス Breaking News English Corpus(BNEC) を提案する. BNEC は英語学習者向けの教材提供サイト Breaking News English[1] をもとに作成されており, 現在そのサイトからスクレイピングした約 400 の記事クラスターから構成されている. コーパスでは, 人手でのアノテーションにより, 各語彙レベルの要約テキスト間及びそれらとリファレンス記事間でのアラインメントが部分的に取られている.

本稿では, まず複数文書要約に関する先行研究について述べる. その後, 提案するデータセットについて説明したのち, 実際にアノテーションを行ったサンプルデータを用いてコーパスの評価を行う. 最後に全体のまとめと今後の研究に対する課題点について述べる.

2 先行研究

これまで, 複数文書要約の分野では Document

表 1: 主な複数文書要約データセット

データセット	データサイズ 事例数 x 文書数	要約サイズ 単語数
DUC 2001	60x10	50,100,200,400
DUC 2002	60x10	10,50,100,200,400
DUC 2003	60x10,30x25	10,100
DUC 2004	100x10	10,100
DUC 2005	50x32	250
DUC 2006	50x25	250
DUC 2007	25x10	100
TAC 2008	48x28	100
TAC 2009	44x20	100
TAC 2010	46x20	100
TAC 2011	44x20	100
BNE(ours)	408x2,3	100-250

[1]https://breakingnewsenglish.com

表 2: マルチレベルにおける平易化要約記事の例 (easy の場合)

レベル	各レベルにおける平易化後の文 (2 文表示)
easy 3	Japan is changing its immigration policy because it needs workers. Japan is an aging society.
easy 2	Japan is changing because it needs workers. It is an aging society.
easy 1	Japan needs many workers. It is an aging society and does not have enough workers.
easy 0	Japan needs many workers. It does not have enough workers.

表 3: Breaking News English Corpus の統計データ (各平均は記事数で算出)

平易化レベル	記事数	文数	平均文数	単語数	平均単語数	語彙単語数	平均語彙数
easy 0	201	2315	11.51	23419	116.51	2488	12.37
easy 1	201	2755	13.70	33705	167.68	3715	18.48
easy 2	201	3101	15.42	43007	213.96	4503	22.40
easy 3	143	2323	16.24	37388	261.45	4429	30.97
hard 4	207	2537	12.25	31662	152.95	3460	16.71
hard 5	207	2813	13.58	42769	206.61	5330	25.74
hard 6	122	1722	14.11	30413	249.28	4582	37.55
計	1282	17566	-	242363	-	28507	-
リファレンス	1119	33644	30.06	824482	736.80	31847	28.46

Understanding Conference (DUC) と Text Analysis Conference(TAC) データセット群が主なリソースとして使用されてきた [1]. これらのデータセットは, 近年盛んに研究が行われているニューラルネットワークに基づいた自動要約モデルのトレーニング及び評価のベンチマークとして, 重要なコーパスとなっている. しかしながら, 既存の複数文書要約のデータセットは十分とは言えない現状がある. なぜなら, 複数文書からの要約作成は, 単一文書要約と比較して高価であるからである. 加えて, 単一文書要約のデータセットの構築においては, ニュース記事とその見出しをモデルへの入力と教師信号と見なして, コーパスを構築することが可能であるが, 複数文書要約にはこうした半自動的なデータ構築手法が確立されていない. 更に, 先述したように, 人手で要約を作成する際にも, どのような要約を作るべきかといった指標を明確化しなければならないという問題も存在する.

そうした中, 要約作成指標の一つに成り得るタスクとして, テキスト平易化の研究が行われてきた. テキスト平易化においても, データセット構築の研究がなされており,Simple Wikipedia[2] や Newsela[3] といったコーパスが有名である.

表 1 に, 各データセットと今回提案するデータセットにおける入力及び出力要約のサイズを示す. 提案する BNEC は要約元の文書数は少ないものの, 事例数では, 既存のデータセットを大きく上回っている.

3 Breaking News English Corpus

ここでは, 今回提案するコーパスである Breaking News English Corpus(以下,BNEC) の概要及び取集したデータの統計について説明する.

3.1 コーパスの概要

BNEC は英語学習者向けの教材としてニュース記事の要約を無料で公開しているサイトである breakingnewsenglihs.com から収集されたデータに基づき構築されている. このサイトは Sean Baniville 氏によって運営されており,9 つのトピックにおける最新のニュースを対象として, およそ 2 日毎に新たな記事が更新されている.2004 年から 2019 年 1 月までに約 2650 の要約記事が公開されている. このうち, およそ 800 記事ついては, 初級者向けの要約記事である easy(level 0,1,2,3) と, 上級者向けの要約記事である hard(level 4,5,6) という分類で, 合わせて 7 段階の読解レベルに応じて要約

記事が用意されている. 更に, 各要約記事には, 参考文献となった記事の URL が 2 から 4 つ記載されている.

3.2 コーパスの統計データ

コーパスの収集にあたっては, これら 7 段階のレベルで構成された要約記事及びリファレンス記事を対象として, 各記事ページをスクレイピングした. データ収集ツールとして python3.6.2 を使用し, ウェブページへのアクセスと記事本文取得のためのパーサーには,requests2.18.4,newspaper3k0.2.8 の各ライブラリをそれぞれ用いた. 収集したデータの中には, 記事のウェブページが存在しないものや, 異なる HTML 構造に起因するノイズデータを含むものがあったため, これらの記事は例外として除去した. 最終的に,2016 年 5 月から 2019 年 1 月までに作成された記事 408 サンプルを収集した. まず, 一つのリファレンス記事に対し, 文数と単語数の観点から, 要約記事では, 記事のサイズがおよそ半分以下になっている. また,easy, hard の両記事において, 平易化レベルが増加するにつれ, 記事毎の文数, 単語数, 及び語彙数も増加していることが分かる. 一般的に, 記事や文の長さが大きくなるにつれて, 読解の難易度も高くなる. そのため, この統計結果は, 平易化処理においては, 記事と文の長さが調整されたことを反映している. 更に, 平易化レベル間での平均語彙数の変化は最も大きくなっており, これは平易化において語彙の選択が大きな要素の一つであることを表している.

4 アノテーション

本章では,BNE コーパスの特徴である文レベルでのアラインメントに着目し, 実際にサンプリングしたデータのアノテーションを通じて, 複数文書コーパスとしての有用性を示す.

4.1 平易化文の間におけるアラインメント

BNEC では複数の平易化テキスト間で, 同一の意味を含む文の対応関係を人手でアノテーションし, 構造化することを提案している. これまでに, 複数文書要約において, 各要約文とその情報源に当たる元文書内の文とのアラインメントを明示したような大規模なコーパスは開発されていない. そのため, 本研究では, 取集データの中から,12 事例を実際にアノテーションし, 提案コーパスの質的評価を行った. 表 2 に示すように,BNEC では, 平易化レベル毎のテキスト間において sentence レベルでのアラインメントが多く為されている. これは, 既存のテキスト平易化コーパスである Simple Wikipedia や文書レベルのみでアラインメントが為されている Newsela Corpus にはない大きな特徴である.

しかしながら,BNEC は要約記事内において, 基本的に文間のアラインメントが 1 文対 1 文で取られているものの, 中には, 平易化のために、1 対 1 の対応以外のアラインメントペアが存在した. 以下のアノテーションにおいて、確認されたアラインメントタイプを示す.

- 1 対 1…一文ずつ対応している.
- 削除…記事全体の短くするために, 重要度が低い文を削除する.
- 文分割…比較的長い文を読みやすくするために分割する.
- 文結合…分かれた文を結合し,1 文にまとめる.
- 逆転…2 文間の並びを逆転させる

図 1 に, 既にアラインメントがされている文と例外の割合及び例外における文のタイプの割合を表す.

全体の約 8 割に当たる 449 文がアラインメントされている一方で, 残りの 121 文が例外ケースであった. このうち, 削除が最も発生頻度が高く (55 回), 続いて文分割 (39 回), 文結合 (25 回), 逆転 (2 回) という結果であった.

4.2 リファレンスとのアラインメント

表 4 に実際にアノテーションを施した文の例を示す. 上段の 3 文は hard レベルにおける文であり, 下の 3 文は URL が記載された順で並べたを 3 つのリファレンスの各文である. 表からわかるように, 平易化テキスト間では, レベルが下がる毎に, 文が比較的短く簡単な表現になっているのに対し, アラインメントされたリファレンス記事は長く, 語彙数も複雑な文になっていることがわかる.

図 2 は, アラインメントの結果から, 各リファレンスタイプが含む要約文との対応文の頻度を算出したものである. ここで, 第一リファレンスとは, 要約記事に筆

表 4: 要約文及びリファレンス間のアノテーション

レベル	アラインメントされた文
hard 6	Researchers found that the largest gap between the sexes was in politics.
hard 5	The largest gap between the sexes was in politics.
hard 4	The largest gap between the sexes was in politics.
first	The area with the widest gap between men and women is in politics, the report said.
second	The impact is especially felt in leadership positions.
third	Women are also far behind in politics, and the WEF estimated that at the current pace of change it will take 107 years until there are as many female politicians as male.

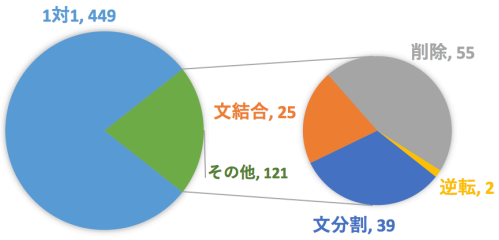

図 1: アラインメントと例外の割合 (各値は文のペア数)

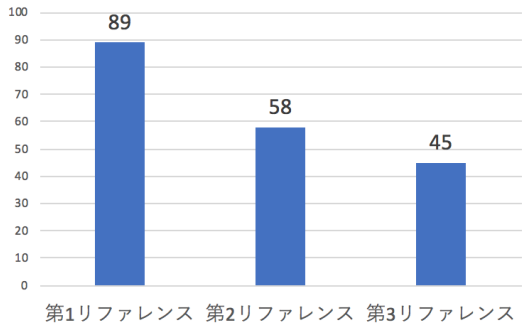

図 2: 各リファレンスタイプが含む要約文との対応文の頻度

頭で記載されたリファレンスのことであり, 同様の順番で各リファレンスを分類している. 図 2 では, リファレンスの記載順とアラインメントの発生頻度が相関関係にあることがわかる. 言い換えれば, 今回の複数文書要約においては, 各文書からの情報を一様に取得するというよりも. 要約の中心となるリファレンスからより多くの情報を獲得していることが伺える. すなわち,BNEC は, 既存の複数文書要約と比べ, リファレンス記事の数が少ないことがボトルネックになる可能性があったが, サンプリング及びアノテーションの結果から, リファレンスの数が少ない場合でも, ある程度の複数文書要約は可能であると言える.

5 おわりに

以上のように, 本稿ではテキスト平易化に着目した新たな複数文書要約コーパス BNEC を提案した. 収集したデータをもとにした統計データからは, テキスト間のレベルに応じて, 記事の長さや文の長さ, 語彙数の観点から, 平易化が段階的に為されていることが分かった. また, 文間のアラインメントに対するアノテーションによる評価では,BNEC がアラインメントに優れたコーパスであること及び, リファレンス記事間で, 取得される情報に差があることが確認出来た. 今後の課題として, データセットの更なる拡張及び実際に, ニューラルモデルを用いて性能評価をする必要がある.

参考文献

[1] Dernoncourt, F., Ghassemi, M., and Chang, W. (2018). A Repository of Corpora for Summarization. Language Resources and Evaluation Conference (LREC).

[2] Zhu, Z., Bernhard, D., and Gurevych, I. (2010). A monolingual tree-based translation model for sentence simplification. In Proceedings of the 23rd International Conference on Computational Linguistics (COLING).

[3] Xu, W., Callison-Burch, C., and Napoles, C. (2015). Problems in current text simplification research: New data can help. Transactions of the Association for Computational Linguistics, 3:283297.