Multi-aspects Document Summarization
データセットと要約文における情報の提示順を評価する指標の提案

近藤 崇宏　　宮尾 祐介
総合研究大学院大学 複合化学研究科　東京大学大学院 情報理工学系研究科
{takahiro_kondo, yusuke}@is.s.u-tokyo.ac.jp

1 はじめに

複数文書要約は複数の文書を入力としてその要約を生成するタスクである。要約モデルは入力文書から重要な内容を抽出し、それらの情報を人間が理解しやすいように順番を構成して提示する必要がある。しかし、既存のデータセットはほとんどがニュース記事を対象としていることから、既存の要約モデルはニュース記事の特性である、(1) 記事の作成日時の情報がある、(2) 記事冒頭の内容ほど重要性が高い、(3) 同じ事象についての記事群はしばしば内容が重複する、という性質を利用して情報の提示順を決定しており、これらの特徴を持たない文章に同様の手法を適用することは難しい。

我々は複数文書要約における情報の提示順についての研究を発展させるために、Multi-aspects Document Summarization という新しいタスクを提案する。このタスクは複数文書要約と同様に、複数の文書を入力としてその要約を生成するものである。しかし、1 セットの入力文章群は、各文書がある物事の相異なる側面について記載されている（図 1 参考）。そのため、上記 (1)(2)(3) の性質を利用せずに、要約文においてどの側面をどの順番で提示するかを決定しなければならない。

このタスクのためのデータセットを英語版 Wikipedia から作成する。また、要約文における情報の提示順の適切さを評価するための自動評価指標を提案し、指標が人間による文章の理解しやすさの判断と一致する傾向にあることを評価実験により確認する。最後に、既存の複数文書要約モデルを提案データセットに適用し、既存手法の改善の余地が大きいことを示す。

2 関連研究

2.1 データセット

複数文書要約のデータセットは限られており、ほとんどの既存研究は Document Understanding Conference[*1]（DUC）や Text Analysis Conference[*2]（TAC）で作成されたデータセットを利用している。それらのデータはいくつかのトピックに基づいて集められたニュース記事を入力とし、参照要約（模範的な要約）を専門家が人手により作成している。特定の文脈（ユーザーの検索クエリが指定されているなど）を想定しない複数文書要約のデータについては、DUC において 2001–2004 の 4 年間で計 169 セットの入力文章群と参照要約のペアが作成された。

文章 1：系統・分類について
「バッタ」はバッタ目・バッタ亜目に分類される昆虫の総称。しばしば亜目の共通名称として…

文章 2：身体的特徴について
身体構造は多くの昆虫に典型的なものであり、…特に後脚は強靭に発達しているため、…

文章 3：生態について
雑食性であり、イネ科の植物を好んで食べる。…特定の状況下で集団発生することがあり、…

図 1　（例）バッタの様々な側面についての文章群

Zopf et al. (2016) はニュース記事だけでなく、ブログや QA サイトなど様々なタイプの文章を入力とするデータセットを作成した。参照要約には Wikipedia の冒頭部（リードと呼ばれる）を利用しており、参照要約から人手によってキーワードを抽出し、そのキーワードで検索ヒットした文章を入力とした。この手順を 3 つのカテゴリに属する記事について行い、91 セットの入力文章群と参照要約のペアが作成された。我々のデータセットも同様に Wikipedia のリードを参照要約として利用するが、我々はリード以降の文章を入力として利用する。また、我々は入力文書タイプの多様性ではなく、要約文において様々な側面についての情報をどのような順番で提示するかに着目する。

2.2 要約モデル

要約生成モデルは大きく分けて文抽出型（Extracive）と文生成型（Abstractive）の 2 つの手法に分けられる。

複数文書要約の要約モデルの多くは文抽出型であり、(1) 重要な文の抽出と (2) 文を読みやすいように並び替え、の 2 つの機能がある。(1) 文抽出の主な手法は、文の重要度を示すスコアを推定し、スコアの高い文から

[*1] https://duc.nist.gov/
[*2] https://tac.nist.gov/

冗長性を除きつつ選択するものである。スコア推定には、文をノードとし文の類似度をエッジの重みとするグラフを構築する手法 (Erkan and Radev, 2004) や、文書中の文の位置などを特徴量として機械学習によるスコア推定をする手法 (Ng et al., 2012) などがある。(2) 文の並び替えは、単に (1) のスコアが高い順に文を提示するだけのものが多い。Barzilay et al. (2001) は複数文書要約において生成要約の文の順番がほとんど考慮されていないことを課題として指摘し、文のセットを入力としてそれらの文を正しい順番に並び替える Sentence Ordering タスクを定義し、データセットを作成した。Sentence Ordering の手法として、要約文と入力文書でトピックやエンティティの隣接関係が一致する傾向を利用するもの (Barzilay and Lapata, 2008; Ji and Nie, 2008) が提案されているが、これらの手法は入力文書間に内容の重複がなければ、要約全体の文の順番を決めることはできない。

一方、文生成型の要約モデルについては、単文書要約タスクで近年盛んに提案されているが、複数文書要約では非常に少ない。これは既存の複数文書要約のデータセットのサイズが小さいため、近年の文章生成モデルの多くが採用するニューラルネットワークの訓練に適していないことが主な理由であろう。データサイズの小ささを補うために単文書要約で訓練したモデルを転移学習させる手法 (Lebanoff et al., 2018) などが提案されている。

2.3 評価指標

要約タスクで利用される既存の自動評価指標 (Lin, 2004; Papineni et al., 2002; Hirao et al., 2018) はモデルが生成した要約と参照要約のフレーズの一致度を測るものであり、情報の提示順は全く考慮しない。最もよく使用される ROUGE-N (Lin, 2004) の Recall の計算式は以下の通りであり、N グラムの一致度をカウントすることによって計算する。

$$\mathrm{ROUGE_N} = \frac{\sum_{S \in \mathrm{refs}} \sum_{\mathrm{gram_n} \in S} \mathrm{Count_{match}}(\mathrm{gram_n})}{\sum_{S \in \mathrm{refs}} \sum_{\mathrm{gram_n} \in S} \mathrm{Count}(\mathrm{gram_n})} \quad (1)$$

ここで、refs は参照要約（複数あり得る）、$\mathrm{gram_n}$ は N グラム、$\mathrm{Count_{match}}$ は参照要約と生成した要約の間で一致する N グラムを数える関数である。機械翻訳では文中の語順を考慮した指標 (Isozaki et al., 2010) が提案されているが、単一の文内での考慮であり、文章全体における順番の考慮ではない。既存研究では、内容の一致度以外の読みやすさなどの評価については主に人手で行われてきた。DUC では *Grammaticality, Non-redundancy, Referential clarity, Focus, Structure and coherence* の項目について専門家が評価している。しかし、このような人手による評価はコストが大きい。本研究では提案タスクの評価のために、文章における情報の提示順を考慮した自動評価指標を提案する。

3 タスク定義とデータセット作成

我々は複数文書要約における情報の提示順についての研究を発展させるために、Multi-aspects Document Summarization というタスクを提案し、データセットを作成する。我々のタスクは、複数の文書を入力としてその要約を生成するという点では通常の複数文書要約と同じである。ただし、1 セットの入力文章群は各文書がある物事についての相異なる側面について記載されているため、通常の複数文書要約と異なり入力文章間で内容の重複がほとんどない。そのため、内容の重複を利用して異なる文の間の隣接性を推定し、要約文における情報の提示順を決定する手法は適用が困難である。

データセットは英語版 Wikipedia の記事から作成する。Wikipedia の記事は冒頭にリードと呼ばれる導入セクションがあり、その後に階層的なセクション構成を持つ記事本体（本研究ではボディセクションと呼ぶ）が続く。Wikipedia のガイドライン[*3] にはリードは記事の最も重要な内容を要約したものと記載されている。ボディセクションを第一階層のセクションで区切った文章群を入力とし、リードを参照要約として利用する。

Wikipedia には質の低い記事も含まれるため、Wikipedia 編集者らにより質の高い記事であると保証された Featured article のみを評価データとして利用し、Featured article 以外の記事は訓練用データにのみ含める。Featured article と認められるための基準[*4]には、リードの適切さも含まれるため参照要約の質を担保できる。また、Featured article ほどではないが比較的質の良い記事である Good article についても区別できるようにデータセットにタグを付与する。

データセット作成手順は次の通りである。文章分割タスクのためのデータセット (Koshorek et al., 2018) の作成スクリプトを利用して Wikipedia のダンプファイル（`enwiki-20180920-pages-articles.xml.bz2`）から記事を抽出し、リードとボディセクションを第一階層で区切った文章に分ける。リードの文数が 5 未満またはボディセクションの第一階層のセクション数が 3 未満の記事を除外し、この過程で 1,261,943 件の記事が除外される。データセットの統計情報を表 1 に示した。Featured article のみでも計 4,793 セットのデータセットであり、既存の複数文書要約データセットに比べてデータサイズが大きいため、学習データの量を必要とするモデルも適用可能である。

[*3] https://en.wikipedia.org/wiki/Wikipedia:Manual_of_Style/Lead_section

[*4] https://en.wikipedia.org/wiki/Wikipedia:Featured_article_criteria

表 1　データセットの統計情報

データタイプ	記事タイプ	記事数	参照要約平均文数	参照要約平均単語数	入力平均セクション数	入力 1 セクションあたり文数	入力 1 セクションあたり単語数
訓練	Featured	3,787	14.1	316	5.4	33.2	762
	Good	20,593	11.3	248	4.7	24.4	544
	Normal	308,642	8.0	170	4.6	15.4	330
検証	Featured	507	13.6	313	5.5	32.2	743
テスト	Featured	499	14.5	325	5.6	33.5	762
合計	All	334,028	8.3	177	4.6	16.2	351

4　自動評価指標の提案と有効性の検証

4.1　ROUGE-N-P

ROUGE (Lin, 2004) などの自動評価指標は要約モデルが生成した文章と参照要約文の内容の一致度を評価するものであり、情報の提示順は全く考慮していない。我々は N グラムによる内容の一致度を計測する ROUGE-N を拡張し、文章中の N グラムの登場位置の一致度を考慮する新しい指標として、ROUGE-N-P（P=Position）を提案する。ROUGE-N-P の Recall の計算式は以下の通りである。

$$\mathrm{ROUGE}_{\mathrm{N\text{-}P}} = \frac{\sum_{S \in \mathrm{refs}} \sum_{\mathrm{gram}_n \in S} \mathrm{Sum}_{\mathrm{pos\_sim}}(\mathrm{gram}_n, S)}{\sum_{S \in \mathrm{refs}} \sum_{\mathrm{gram}_n \in S} \mathrm{Count}(\mathrm{gram}_n)}$$

$$\mathrm{Sum}_{\mathrm{pos\_sim}}(\mathrm{gram}_n, S) = \sum_{\mathrm{pos}_r \in \mathrm{positions\_in}(\mathrm{gram}_n, S)} \Big[1 - \min\{|\mathrm{pos}_r - \mathrm{pos}_s|, \mathrm{pos}_s \in \mathrm{positions\_in}(\mathrm{gram}_n, \mathrm{SysSum})\}\Big]$$

ここで、refs は参照要約（複数あり得る）、SysSum は要約モデルが生成した要約、gram_n は N グラム、$\mathrm{positions\_in}(\mathrm{gram}_n, S)$ は要約文 S における N グラムの相対的位置（[0-1] の値）の集合を返す関数である。上記計算式の refs と SysSum を入れ替えることで Precision を、Recall と Precision を用いて F 値が計算できる。上記の計算式は、ROUGE-N（式 (1)）の $\mathrm{Count}_{\mathrm{match}}(\mathrm{gram}_n)$ の部分のみを変更したものである。N グラムの登場位置が生成要約文と参照要約で完全に一致していれば、ROUGE-N-P と ROUGE-N の値は同じとなる。

4.2　指標の有効性の検証

ROUGE-N-P による N グラムの登場位置の重みづけ結果が人間の文章の理解しやすさの評価と一致するかどうかを次の手続きにより検証する。まず、本研究で作成したデータセットから Featured article をランダムに 100 記事選ぶ。次に参照要約をトピックのまとまりによるいくつかの文のまとまりにグルーピングし、グループ単位で順番を入れ替えた文章（以下、入替要約と呼ぶ）を生成する。ここで、グルーピングする理由は入替要約の coherence をなるべく損なわないためである。また、参照要約の最初の文は何についての記事であるかを簡潔に述べる文であり、最初のグループの順番を入れ替えると読みやすさに与える影響が大きいため入れ替えない。入替要約群と参照要約で ROUGE-2-P の F 値を計測し、(1) スコアが最も低い入替要約、(2) スコアが (1) と 1.0 の中間の入替要約、を選ぶ。参照要約と前述の (1)(2) の計 3 つの要約から 2 つを選択したペア（計 3 セット）を作成し、ペアを左右に並べて提示して人間にどちらが理解しやすいかを回答してもらう。その後、ROUGE-2-P の値と人の評価の結果を比較し、両者の傾向が一致するかどうかを確認する。

表 2　人手評価と指標のスコアの対応（数字は該当記事数）

	人間＼指標	左が高スコア	右が高スコア
比較①	左が読みやすい	44	12
	右が読みやすい	8	33
	どちらでもない	1	2
比較②	左が読みやすい	37	15
	右が読みやすい	9	37
	どちらでもない	0	2
比較③	左が読みやすい	34	14
	右が読みやすい	17	33
	どちらでもない	2	0

実験結果は表 2 の通りである。比較パターンは、①参照要約と ROUGE-2-P 最低の入替要約、②参照要約と ROUGE-2-P 中間の入替要約、③ ROUGE-2-P 最低と ROUGE-2-P 中間の入替要約、である。いずれの比較パターンにおいても人の評価と指標のスコアが一致する傾向を示している。Fisher の正確検定を両側検定で適用すると、人の評価と指標のスコアが独立であるという帰無仮説はいずれも有意水準 1% で棄却される。

表 3　文抽出型モデルの適用結果（各 ROUGE は F 値 (%)）

モデル	ROUGE-1	ROUGE-1-P	ROUGE-2	ROUGE-2-P	ROUGE-1-P / ROUGE-1	ROUGE-2-P / ROUGE-2
ランダム	23.98	23.39	4.74	3.85	0.98	0.81
GreedyKL	29.22	28.66	6.76	5.63	0.98	0.83
Oracle	47.58	49.98	18.26	17.87	1.05	0.98
Oracle ランダム順	47.58	42.10	18.13	13.59	0.88	0.75
Oracle 元記事順	47.58	45.88	18.21	15.73	0.96	0.86
G-Flow	25.01	25.72	5.75	4.82	1.03	0.84

5　既存の文抽出型要約モデルの適用

我々のデータセットに対する既存手法の適用可能性を検証するために、既存の文抽出型の要約モデルから coherence を考慮して文の抽出と並び替えを同時に最適化するモデルである G-Flow (Christensen et al., 2013) を適用する。G-Flow は生成する要約文の最大長が 665 バイト固定であるため、プログラムを修正して参照要約のバイト数もしくは 1500 バイトのうち小さい方（実行メモリサイズ制限のため）を最大長として生成する。また比較のために、ランダム抽出モデル、複数文書要約でベースラインとして利用される GreedyKL (Hong et al., 2014)、性能上限を測るための Oracle モデル（後述）、Oracle モデルの出力結果に対して、(1) 文の順番をランダムシャッフルしたもの、(2) 文の順番を元記事の登場順に並び替えたもの、を適用する。Oracle モデルは参照要約の各文との ROUGE-1 F 値が最も高い文を貪欲法により抽出し、抽出した文の順番は抽出時にスコア計算した参照要約の文の順番と合わせる。

表 3 は各モデルの出力に対する ROUGE-N と ROUGE-N-P の F 値、及び、ROUGE-N-P の ROUGE-N に対する割引率（1 を超えることもある）を示している。なお、ランダムな手法については 100 回の試行の平均値を取った。ベースラインや既存手法と Oracle のスコア差が大きく、手法の改善の余地は大きい。ROUGE-2-P は並び順による割引を表現できているのに対し、ROUGE-1-P の割引率は 1 に近く、情報の提示順を考慮した評価は ROUGE-2-P で実施するべきである。G-Flow とベースラインを比較すると、DUC2004 のデータに対してほぼ同等の ROUGE-1 値を示していたが、本データセットに対しては G-Flow の ROUGE-1 値が低く、ROUGE-2-P の割引率も差がない。G-Flow は文抽出と文の並び替えの両方において、本データセットに対してはうまく機能していないと考えられる。Oracle モデルは ROUGE-2-P の割引率が 1 に近いことから、ROUGE-2-P は文の並び順が適切であれば ROUGE-2 とほぼ同じ値となることがわかる。また、元記事順に文を並べることはある程度有効であると言える。

6　今後の課題

本研究では Multi-aspects Document Summarization という新しいタスクを定義し、そのタスクのデータセットと自動評価指標を提案した。今後は要約文の情報の提示順を明示的に考慮する要約モデルを提案・実装し、本研究で作成したデータセットと指標で評価する。

参考文献

Barzilay, R., N. Elhadad, and K. McKeown (2001) "Sentence Ordering in Multidocument Summarization," in *HLT 2001*.

Barzilay, Regina and Mirella Lapata (2008) "Modeling Local Coherence: An Entity-Based Approach," *Computational Linguistics*, Vol. 34, No. 1.

Christensen, Janara, Mausam, Stephen Soderland, and Oren Etzioni (2013) "Towards Coherent Multi-Document Summarization," in *NAACL 2013*.

Erkan, Günes and Dragomir R. Radev (2004) "LexRank: Graph-based Lexical Centrality As Salience in Text Summarization," *J. Artif. Int. Res.*, Vol. 22, No. 1.

Hirao, Tsutomu, Hidetaka Kamigaito, and Masaaki Nagata (2018) "Automatic Pyramid Evaluation Exploiting EDU-based Extractive Reference Summaries," in *EMNLP 2018*.

Hong, Kai, John Conroy, Benoit Favre, Alex Kulesza, Hui Lin, and Ani Nenkova (2014) "A Repository of State of the Art and Competitive Baseline Summaries for Generic News Summarization," in *LREC 2014*.

Isozaki, Hideki, Tsutomu Hirao, Kevin Duh, Katsuhito Sudoh, and Hajime Tsukada (2010) "Automatic Evaluation of Translation Quality for Distant Language Pairs," in *EMNLP 2010*.

Ji, Donghong and Yu Nie (2008) "Sentence Ordering based on Cluster Adjacency in Multi-Document Summarization," in *IJCNLP 2008*.

Koshorek, Omri, Adir Cohen, Noam Mor, Michael Rotman, and Jonathan Berant (2018) "Text Segmentation as a Supervised Learning Task," in *NAACL 2018*.

Lebanoff, Logan, Kaiqiang Song, and Fei Liu (2018) "Adapting the Neural Encoder-Decoder Framework from Single to Multi-Document Summarization," in *EMNLP 2018*.

Lin, Chin-Yew (2004) "ROUGE: A Package for Automatic Evaluation of Summaries," in *Text Summarization Branches Out*.

Ng, Jun-Ping, P. Bysani, Z. Lin, M. Kan, and C. Tan (2012) "Exploiting Category-Specific Information for Multi-Document Summarization," in *COLING 2012*.

Papineni, Kishore, Salim Roukos, Todd Ward, and Wei-Jing Zhu (2002) "Bleu: a Method for Automatic Evaluation of Machine Translation," in *ACL 2002*.

Zopf, Markus, Maxime Peyrard, and Judith Eckle-Kohler (2016) "The Next Step for Multi-Document Summarization: A Heterogeneous Multi-Genre Corpus Built with a Novel Construction Approach," in *COLING 2016*.