メタ情報に基づいた論文タイトル自動生成

黒田 和矢 [1] 狩野 芳伸 [2]

静岡大学大学院総合科学技術研究科

kkuroda@kanolab.net[1], kano@inf.shizuoka.ac.jp[2]

1. はじめに

研究に取り組むにあたって、研究者は膨大な数の論文の中から自身の研究に関連する情報を取捨選択する必要がある。しかし、日々増え続ける論文のすべてを限られた時間の中で読むことは現実的ではない。そこで重要になるのが論文タイトルと論文概要である。これらは論文の内容を端的に表したものであり、研究者が短い時間で論文の内容を把握するために役立てられる。特に、論文タイトルは論文検索エンジン等で論文を探した際に一番初めに目にする情報であり、その論文が必要かどうかの判断に大きな影響を与えるものである。実際、論文タイトルが論文の引用数やダウンロード数に影響を与えるということが調査されており (Jamali ら 2011, Letchford ら 2015) わかりやすい論文タイトルを付けることは書き手、読み手双方にとって重要であると考えられる。しかし、論文タイトルの付け方を解説した Web ページが多数が存在するように、論文執筆に慣れていない人にとってわかりやすいタイトルを付けることは難しい行為である。そのため、論文タイトル生成の支援として、論文概要を要約して論文タイトルを生成する研究 (安藤ら 2004, Putra ら 2017, 大部ら 2018) がいくつか行われている。

本研究では初学者および非英語話者のための論文タイトル生成の支援として、先行研究と同様に論文概要を要約する形で論文タイトルを生成する手法を提案する。このとき著者やジャーナルといったメタ情報が論文タイトルの表現に与える影響を考慮することで、論文タイトル生成の精度の向上を試みる。実験の結果、ジャーナル名を用いて論文タイトルを生成した際にベースラインと比較して ROUGE スコアが上昇した。

2. 関連研究

2.1. 論文タイトル生成

論文概要を入力とした論文タイトルの自動生成に関する既存研究について述べる。Putra ら (2017) は文の修辞的カテゴリーを考慮したテンプレートベースおよび K 近傍法を用いた論文タイトルの自動生成を試みた。論文概要中の各文を論文の目標や提案手法等の修辞的カテゴリーに基づいて分類した後、修辞的カテゴリーに基づいた文のフィルタリングを行うことで論文概要中の特定の文のみを論文タイトル生成に用いている。安藤ら (2004) は論文タイトルを構成する上で必要となる主題要素を手がかりとした、限定されたパターンで構成される論文タイトルの自動生成を試みた。論文概要から 2 文以内の重要文を抽出した後、抽出された文の文節を一定の規則で探索することで主題要素を獲得し、獲得した主題要素を整形することで論文タイトルを生成した。大部ら (2018) は RNN を用いた抽出型および抽象型の論文タイトルの自動生成を試みた。抽出型の要約では論文概要中の各単語についてその単語がタイトルに出現するかの二値分類を RNN を用いて行った後、テンプレートを用いて論文タイトルを生成した。抽象型の要約では未知語に対応するために Encoder-Decoder 型のモデルに Pointer Networks の機構を組み合わせたモデルを用いて論文タイトルを生成した。

2.2. メタ情報を使った文生成

文生成のタスクにおいて、データに付随するメタ情報を用いることで生成される文を制御する研究が行われている。Dziri ら (2019) は対話応答生成に会話のトピックを組み込んだ THRED（Topical Hierarchical Recurrent Encoder Decoder）を提案し、文脈に応じた対話を生成することを試みている。Mathur ら (2017) は e コマースの Web ページタイトルを自動生成する際に商品のカテゴリーやブランドを用いることで、Web ページの内容に則したタイトルの生成を試みた。Iwama ら (2019) はニュース記事の見出しを生成する際に見出しの順序や文字サイズを用いることで、ひとつのニュース記事に対して複数の見出しを生成することを試みた。

論文タイトルに関する分析として、佐々木 (2017) は国内心理学会機関紙において機関紙の種類によって論文タイトルの長さや用いられる単語の傾向に違いがあることを明らかにしている。機関紙名のようなメタ情報に基づく論文タイトルの傾向を学習することができれば、論文概要から論文タイトルへの要約の精度が向上すると考えられる。そこで、本研究では論文概要に加えて論文に付随するメタ情報を用いて論文タイトルを生成することを試みる。

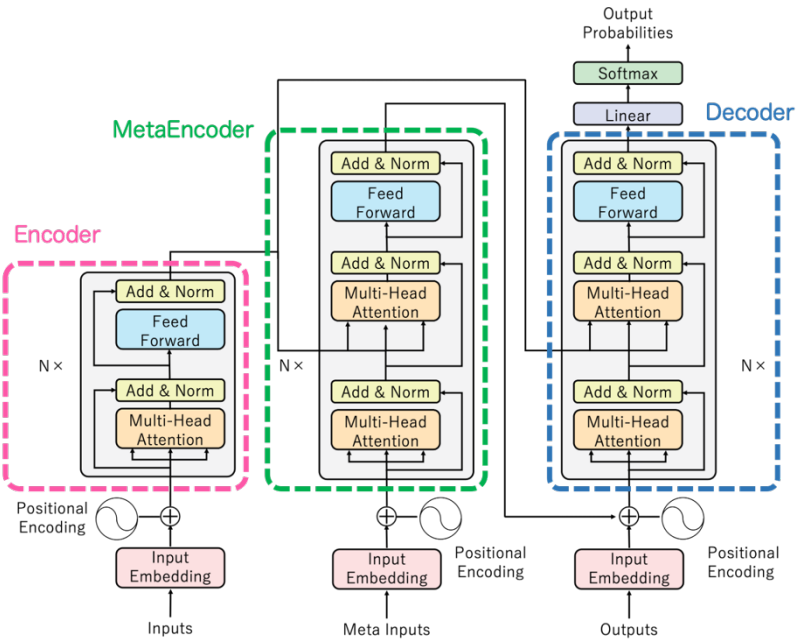

図 1 提案モデル全体図

3. 提案手法

本研究では Transformer (Vaswani ら 2017) にメタ情報を扱うための MetaEncoder を組み込んだモデルを提案する。モデルの全体図を図 1 に示す。MetaEncoder は Transformer の Decoder と同じ構造をしており、論文概要をエンコードした結果とメタ情報を用いてメタ情報をエンコードする。メタ情報をエンコードする際に論文概要をエンコードした結果を用いることで、メタ情報の持つ傾向が Attention Mechanism において論文概要のどの単語に注意を向けるかという形で学習されることを期待している。MetaEncoder に複数のメタ情報を入力する際には、メタ情報を並べて単語列のように扱う。MetaEncoder の出力の平均を Decoder の入力に加えることでメタ情報の持つ傾向を論文タイトルに組み込むことを試みている。

4. 実験

4.1. データセット

本研究では PubMed[1] データベースからダウンロードした論文情報を用いた。PubMed は NLM の NCBI が運営する生物医学および生命科学に関する文献の無料検索エンジンであり、データベースには 3000 万件を超える生物医学文献の要約が含まれている。論文概要、論文タイトルに加えメタ情報が記述されおり、深層学習を行うために十分な量のデータが存在するという点から本データを採用した。ただし、論文情報の中には一部の情報が欠損しているものや、論文概要や論文タイトルの長さが極端なものが存在する。そこで、本実験では英語論文かつ最低限必要な論文概要と論文タイトルを含む論文情報に対して単語数および文数に関する分析を行い、表 1 に示すように使用するデータに制限をかけた。

表 1 データ制限に用いた項目と最大／最小値

	タイトル単語数 (単語)	概要単語数 (単語)	概要文数 (文)	概要によるタイトル単語含有率(%)
最小	5	100	3	70
最大	20	200	10	100

本研究では論文概要および論文タイトルのトークン分析に Genia Tagger[2] を用いた。Genia Tagger は MEDLINE 等の生物医学テキスト用に調整されたトークン分析およびタグ付けを行うライブラリである。タグには単語の基本形、品詞、チャンク、固有表現を表すものがある。本研究では実験データ全体を通して生物医学に関する専門用語が一般用語に対して出現回数が少なく未知語として扱われると考え、論文概要および

[1] https://www.ncbi.nlm.nih.gov/pubmed/

[2] http://www.nactem.ac.uk/GENIA/tagger/

表 2 実験結果

モデル		ROUGE-1	ROUGE-2	ROUGE-L
Transformer(ベースライン)		43.85	23.29	40.10
提案モデル	ジャーナル名	**44.62**	**23.80**	**40.80**
	公開年	44.43	23.77	40.73
	インパクトファクター	44.07	23.45	40.33
	著者名	42.18	22.12	38.73
	ジャーナル名+公開年	44.39	23.71	40.60
	ジャーナル名+公開年+インパクトファクター	44.22	23.52	40.47
	ジャーナル名+公開年+インパクトファクター+著者名	42.40	22.10	39.00

論文タイトルのトークン分析を行う際に固有表現の置換を行った。置換は論文概要と論文タイトルのペアごとに検出された固有表現のうち文字列長が長いものから順に<NE-X> (X=0,1,...) という形で置き換えた。

4.2. 実験設定

本実験では 4.1 にて用意した 2,170,402 件の論文情のうち、1,736,322 件を学習データ、217,040 件を検証データ、217,040 件を評価データとして使用する。ベースラインは Transformer とし、複数あるメタ情報のうち 1 つのみを用いた場合といくつかのメタ情報を組み合わせた場合について実験を行う。本実験ではメタ情報として、論文が掲載されたジャーナルの名前、ジャーナルの巻号が公開された年、ジャーナルのインパクトファクター、論文の著者名を使用する。なお、著者名については筆頭著者名及び最終著者名のみを使用する。なお、本実験ではメタ情報をモデルへの入力として扱う際、トークン分析は行わず例えばジャーナル名であれば「Current neurology and neuroscience reports」、著者名であれば「Suzuki T」をひとつの入力として扱う。また、インパクトファクターについては取得した値を切り上げて整数値として使用した。

本実験で扱うすべてのモデルのハイパーパラメータについて Encoder、MetaEncoder および Decoder の層数 N = 2、単語の埋め込み表現の次元 d_{model} = 512、全結合層の次元 d_{ff} = 2048、Multi Head Attention のヘッド数 h = 8、ドロップアウト率 P_{drop} = 0.1、オプティマイザを Adam、学習率 lrate = 0.0001 とした。また、単語数は学習データ中の出現回数上位 30,000 件の単語に固有表現を置換した<NE-X> (X=0,1,...,17) を加えた 30,018 件とした。また、メタ情報は学習データ中の出現回数をもとに、ジャーナル名 8,104 件、公開年 105 件、インパクトファクター37 件、著者名 315,755 件を使用し、それ以外は未知語として扱う。

学習の際にはバッチサイズを 64、エポック数を 20 とし、エポックごとの検証データの損失が 3 エポック上昇し続けたら学習を終了する。検証データの損失が最も低かったエポックのモデルを評価データを用いて評価する。評価には ROUGE-1、ROUGE-2、ROUGE-L を用いる。

5. 結果

実験を行った結果を表 2 に示す。表 2 より、提案モデルを用いた結果のうちメタ情報に著者名を含むもの以外のすべてがベースラインと比べて精度が向上していることが見て取れる。このことから、メタ情報のもつ論文タイトルの表現に対する傾向を学習することが論文タイトルの生成に効果的であると考えられる。また、メタ情報のうちジャーナル名のみを使用した際に最もスコアが高くなったことから、ジャーナル名の持つ使用する単語の傾向や論文タイトルの長さの傾向が適切に学習されたのではないかと考えられる。インパクトファクターを使用した際にベースラインと比べて精度があまり変わらなかった原因として、提案モデルがメタ情報の性質を考慮していないことが考えられる。提案モデルでは連続値であるインパクトファクターを一定の範囲で区切り離散値として使用している。その結果、細かい値の差による傾向が学習されずあまり精度が向上しなかったのではないかと考えられる。また、著者名を用いた際に最も精度が低下したことが見て取れる。本実験では学習データ 1,736,322 件に対して 315,755 件の著者名を使用した。ひとつの著者名あたりの学習データの数が少なくなることで、過学習を引き起こし精度が低下したと考えられる。メタ情報を順に組み合わせて使用した際に、使用するメタ情報の種類を増やすたびに精度が低下していることが見て取れる。著者名と同様にひとつのメタ情報の組み合わせ、例えば 2010 年かつ Scientific reports の論文に対する学

表 3 評価データに対する論文タイトル生成結果

論文概要：pharmacological treatment of colorectal cancer has improved survival rates in recent years . individual genetic variation in genes associated with metabolism and targets of commonly used drugs can be responsible for variability in treatment outcome and toxicity . diverse study designs have been used and heterogeneous end points evaluated by studies assessing the association of genetic markers with treatment outcome . we conducted this systematic review , including 51 studies , to present a comprehensive overview and draw further conclusions . to facilitate comparison of reported study results , risk estimates for observed genetic variants in <NE-0> are presented using defined reference categories and recalculated risk estimates based on data provided in original publications , where necessary . overall , evidence indicates associations of the ugt1a1 (*) 28 variant genotype with toxicity after irinotecan treatment , mutations in gstp1-105 with improved treatment outcome and the xpd-751 variant genotype with poor treatment outcome after oxaliplatin treatment , and amplification of the <NE-1> with improved treatment outcome after therapy with <NE-2> . adequately powered prospective investigations designed specifically for pharmacogenetics are needed .
論文タイトル（正解）：pharmacogenetics in colorectal cancer : a systematic review .
Transformer：pharmacogenetics of treatment outcome after oxaliplatin treatment .
提案モデル：pharmacogenetics of irinotecan treatment outcome in colorectal cancer : a systematic review .

習データの数が減ることが精度低下の原因になっていると考えられる。

Transformer とジャーナル名を使用した提案モデルによる評価データに対する論文タイトルの生成結果を表 3 に示す。生成結果について、提案モデルのみが正解同様「:」で区切られたタイトルを生成していることが見て取れる。学習データについて分析を行った結果、全データ中で「:」が含まれる論文タイトルが 230,143/1,736,322 (13.25%) であったのに対し、表 3 の論文が掲載されている Pharmacogenomics では 53/229 (23.14%)であった。このことから、ジャーナルが持つ「:」を含む論文タイトルが多いという傾向が学習されていると考えられる。

6. おわりに

本研究ではメタ情報を用いた論文タイトルの生成手法を提案した。実験の結果、メタ情報としてジャーナル名を使用した際にベースラインと比べて最も精度が向上し、メタ情報を用いることが効果的であることを確認した。今後の課題として、メタ情報ごとに適切な使い方が可能なモデルの構築やメタ情報に対して十分な量のデータを用いて実験を行うことが挙げられる。また、実際の論文タイトルの傾向と学習済みモデルが生成した論文タイトルの傾向の違いをさらに分析することが精度の向上につながると考えられる。

参考文献

Dziri, N., Kamalloo, E., Mathewson, K., Zaiane, O., 2019. Augmenting Neural Response Generation with Context-Aware Topical Attention. Proceedings of the First Workshop on NLP for Conversational AI , 18-31.

Iwama, K., Kano, Y., 2019. Multiple News Headlines Generation using Page Metadata. Proceedings of 12th International Conference on Natural Language Generation, Vol. 91, pp. 1689-1699.

Jamali, H. R., Nikzad, M., 2011. Article title type and its relation with the number of downloads and citations. Scientometrics, 88(2), 653-661.

Letchford, A., Moat, H. S., Preis, T., 2015. The advantage of short paper titles. Royal Society Open Science, 2(8), 150266.

Mathur, P., Ueffing, N., Leusch, G., 2017. Generating titles for millions of browse pages on an e-Commerce site. Proceedings of the 10th International Conference on Natural Language Generation, 158-167.

Putra, J. W. G., Khodra, M. L., 2017. Automatic Title Generation in Scientific Articles for Authorship Assistance: A Summarization Approach. Journal of ICT Research and Applications, 11, 253.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., Polosukhin, I., 2017. Attention is all you need. Advances in Neural Information Processing Systems, 2017-Decem(Nips), 5999-6009.

佐々木宏之, 2017. 国内心理学会機関誌 7 誌の論文タイトル傾向分析 - KH Coder を用いたテキストマイニングから-. 暁星論叢, 67, 11-47.

大部達也, 大囿忠親, 新谷虎松, 2017. 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. 人工知能学会全国大会論文集, JSAI2017, 3A11.

安藤一秋, 新居雅也, 溝渕昭二, 2004. 論文概要からのタイトル自動生成の試み. 言語処理学会年次大会発表論文集, 10, A10C2-04.