組み合わせを考慮した二段階キーフレーズ生成

三沢翔太郎　谷口元樹　三浦康秀　大熊智子

富士ゼロックス株式会社

{misawa.shotaro, motoki.taniguchi, yasuhide.miura, ohkuma.tomoko}@fujixerox.co.jp

1 はじめに

文書の要点を表すキーフレーズは，文書要約や文書検索などに活用することができる．これまで，キーフレーズの付与には文書から適切な部分を抜き出す抽出方法が用いられてきた [1, 2, 3]．しかし，RNN に基づく Encoder-Decoder モデルを用いた Deep Keyphrase Generation [4] (DKG) が提案され，キーフレーズの生成が可能となった．キーフレーズの生成を行うことで文書に現れないキーフレーズも付与できるようになり，文書の内容をより適切に示すキーフレーズを得ることが期待できる．

DKG は文書を入力とした Attention 機構付きの Encoder-Decoder モデルに Copying 機構 [5] を組み合わせた手法である．また，1 つの文書に対して複数のキーフレーズを生成する際にはビームサーチを用い，文書に付与される確率が高いキーフレーズを探索し，それらをキーフレーズ集合として出力する．

しかし，この方法は生成する複数のキーフレーズを探索する際に，各キーフレーズに対する事後確率を独立に計算するため，生成されたキーフレーズの組み合わせが最適でないという課題がある．そのため，生成されやすい頻出キーフレーズを重点的に生成し，その文書特有の内容や専門的な内容などの低頻度なキーフレーズが生成されにくいなどの現象が生じる．

本研究では以下の 2 点を目的とする．(1) 英語を対象とした高性能なキーフレーズ生成手法である DKG を日本語の学術論文データに適用し，その有効性を検証する．(2) 従来手法の生成結果をもとに，同時に生成されたキーフレーズを考慮して最適なキーフレーズ集合を再選択する二段階手法を提案する．

本研究では NTCIR-1 [6] に収録されている，学術論文のタイトル，概要と付与されたキーフレーズのセットを用いた実験により本手法の有効性を確認した．

2 関連研究

キーフレーズ付与　キーフレーズ付与の方法は抽出手法と生成手法に大別される．さらに抽出手法には，教師なし手法と教師あり手法がある．教師なし手法は，キーフレーズ候補の抽出とそれらのランキングから構成され，ランキングには出現頻度やグラフ構造などが用いられる [1]．教師あり手法では CRF や Bidirectional LSTM (Bi-LSTM) などを用いて系列ラベリングとして解く方法が一般的である [2, 3]．また，生成手法に関する研究は DKG のみである．

再選択モデル　再選択は言語処理において構文解析，質問応答や機械翻訳の分野で用いられている [7, 8]．しかし，これまでのキーフレーズ付与の方法においてランキングは一度しか行われず，再選択の有用性を示した研究は存在しない．また，キーフレーズ以外における再選択では，他のモデルの予測結果など外部情報を取り込むことを目的とし，複数の選択肢から 1 つの結果を選択している．しかし，本研究は外部情報を使用せず，組み合わせを考慮して複数の結果を選択するため，これらの研究とは異なる．

3 二段階キーフレーズ生成

3.1 概要

従来手法である DKG は生成時にビームサーチで探索して付与する複数のキーフレーズを決定するため，キーフレーズの組み合わせを考慮できないという課題がある．本手法ではこの課題を解決するため，従来手法で生成されたキーフレーズ集合をキーフレーズ候補として捉えて，その候補から最適なキーフレーズの組み合わせを再選択する．すなわち，本手法は DKG を用いてキーフレーズ候補を生成する生成過程と，その中から最適な組み合わせを選択しなおす再選択過程から構成される．

生成過程では Encoder-Decoder モデルを適用するために，ある文書と割り当てられたキーフレーズのうち 1 つを組みとして扱い，1 文書に対してキーフレーズ個数分のデータ組を用意して学習する．再選択過程は生成過程のモデルを学習した後，そのモデルから生成されるキーフレーズ候補を用いて学習を行う．

3.2 生成過程

生成過程では，文書から候補となるキーフレーズ候補を N 個生成する．このために，生成タスクで多く用いられる Encoder-Decoder モデルと，Decoder の未知語へ対応するために生成対象の文書から単語を抜き出す Copying 機構を組み合わせる．

Encoder-Decoder モデル　Encoder-Decoder モデルは Encoder と Decoder の 2 つの RNN から構成されている．Encoder は文書を単語ごとに入力して，文書全体を 1 つの中間表現に変換する．対して，Decoder は Encoder で獲得した中間表現を元にキーフレーズを単語ごとに生成する．この 2 つのモデルは同時に学習し，生成時はビームサーチによって事後確率の高い単語系列を N 個獲得する．また，この手法で RNN は Attention 機構を組み合わせた GRU を用いる．

Copying 機構　通常の Decoder ではあらかじめ定められた語彙の中から単語を生成する．そのため，低頻度語など Decoder の語彙にない単語が含まれるキーフレーズは生成することができない．この問題を解決する手法として Copying 機構を用いる．Copying 機構は生成対象の文書を構成する単語も生成候補として扱い，Decoder の状態と文書の文脈を元に適切な単語を推定する機構である．本手法では，通常の Decoder による単語選択と Copying 機構による単語選択の事後確率を足し合わせて次の単語を生成するモデルを構築する．この方法を用いることで，Decoder の語彙に含まれない単語でも文書に含まれる語は生成できる．

3.3 再選択過程

再選択過程は，生成過程で生成されたキーフレーズ候補から最適なキーフレーズの組み合わせを選択する．ここで，キーフレーズ候補のうち，キーフレーズとして選択するか判断する対象をターゲット，生成過程においてターゲットよりも事後確率が高いキーフレーズを上位キーフレーズとする．この過程では，上位キーフレーズの情報を考慮して，ターゲットをキーフレーズとするかの判断に用いるスコアを計算する．ターゲットは生成過程における事後確率が高いキーフレーズ候補から順に選び，すべてのターゲットに対するスコアの計算を行った後に，スコアが高いキーフレーズを選択する．これにより，ターゲットの生成過程における事後確率の順位と他のキーフレーズ候補を考慮した再選択を行う．

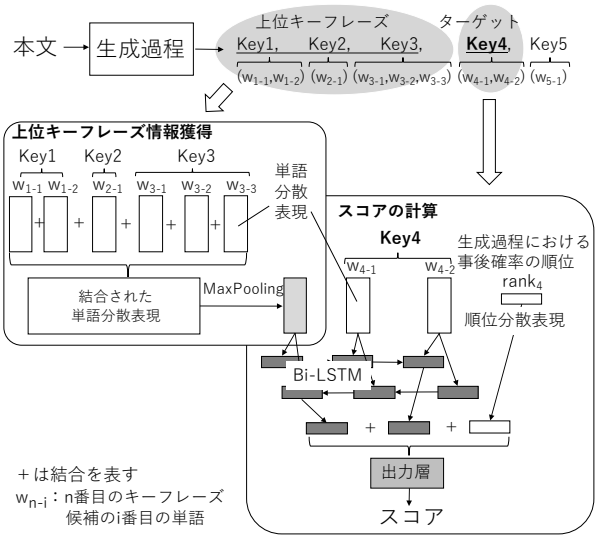

図 1: ターゲット Key4 に対するスコアの計算

再選択過程は，以下の 3 つの手順で構成される．図 1 に生成過程で Key1-5 が生成され，ターゲット Key4 に対するスコアの計算を行う例を示す．

上位キーフレーズ候補情報獲得　上位キーフレーズ候補を構成する単語情報を用いるために，それらの単語分散表現を用意し，分散表現の次元ごとに最大値を獲得する MaxPooling 層で中間表現を獲得する．

スコアの計算　ここでは，ターゲットを構成する単語を系列データとみなして Bi-LSTM で処理した結果を元にスコアの計算を行う．この時，上位キーフレーズ情報を用いるために，前の手順で獲得した中間表現をターゲットの単語系列の前に結合した上で Bi-LSTM で処理する．また，明示的に生成過程におけるターゲットの事後確率の順位を取り込むために，順位を分散表現に変換して用いる．ターゲットのスコアは Bi-LSTM で処理した結果と生成過程における順位の分散表現を入力とした出力層で計算する．

再選択　各ターゲットのスコアを基準に並び替えを行い，その値が高い k 個のキーフレーズ候補をキーフレーズとして選択する．

なお，スコアを計算するネットワークは，学習済みの生成過程で生成されるキーフレーズ候補を用い，ターゲットが正解キーフレーズに含まれるか否かの 2 値分類として学習する．この時，生成過程で候補として

表 1: NTCIR-1 データ概要

	#paper	#keyphrase	unseen rate
train-gen	293,712	1,286,045	33.78%
train-re	10,000	43,711	33.58%
test	10,000	43,650	33.92%

タイトル：手話機械辞書作成の基礎検討

概要：手話機械辞書作成の基礎検討聴覚障害者のコミュニケーション手段である手話を機械認識する場合には、音声語の自然言語認識の場合と同じように処理対象に階層構造があり、これに応じたいくつかの機械辞書が必要となる。本研究は、手話動画像から手話単語候補を抽出する際に必要となる機械辞書の構成、すなわち、画像特徴から手話単語を選定する際に使用する辞書構成、に必要な手話の構成要素について基礎的な検討を行った結果を報告する。

正解キーフレーズ：機械辞書 手話 自然言語処理

従来手法生成結果：**手話** 機械翻訳 階層構造 **自然言語処理** 辞書

再選択の結果：**手話** 機械翻訳 階層構造 **自然言語処理** **機械辞書**

図 2: 上段は NTCIR-1 データ構造，下段は生成キーフレーズの具体例を示す．下線は文書中に出現しないキーフレーズを，太字は正解したキーフレーズを表す．

生成されていない正解キーフレーズは考慮せずに，各キーフレーズ候補が正解キーフレーズ集合に含まれるかどうかのみを基準とする．

4 実験

4.1 実験データ

提案手法の効果を確認するために NTCIR-1 のデータで実験を行う．図 2 にデータの例を示す．このデータは日本語学術論文のタイトルと概要および筆者が付与した複数のキーフレーズで構成される．今回は生成過程と再選択過程用に学習データ 2 種類（train-gen, train-re）とテストデータを用意した．表 1 に統計的特徴を示す．なお，unseen rate はキーフレーズのうち文書中に完全に一致する部分がない割合を表す．

4.2 実験設定

生成過程のモデルパラメータは先行研究 [4] と同値に設定し，形態素解析は mecab ipadic を用いた．再選択過程のパラメータは LSTM のユニット数と単語分散表現の次元は 400，順位分散表現の次元は 10，学習には学習率 0.001 の Adam を用い，生成過程で生成するキーフレーズ候補数 N は 20 個とした．なお，train-re は生成過程で Early Stopping の基準としても用い，単語の分散表現は train-gen で事前学習を行った．

表 2: 実験結果

METHOD	F@5	F@10	R_{uns}@5	R_{uns}@10
tf-idf	1.40	2.33	0.00	0.00
DKG	32.94	28.62	4.41	8.19
tar-rank	32.98	28.78	4.49	**8.38**
tar-high-rank	**33.30**	**29.01**	**4.51**	**8.38**

比較手法は，教師なしのベースラインである tf-idf と従来手法の DKG を用いた．また，提案手法の効果確認のため，以下の 2 つの再選択を行う提案手法を用意した．

tar-rank：ターゲットとその事後確率の順位のみを考慮した再選択手法．

tar-high-rank：tar-rank に加えて，上位キーフレーズの情報も考慮した再選択手法．

性能評価はすべてのキーフレーズを対象とした F 値（F@5,10）と文書に出現しないキーフレーズに限定した再現率 (R_{uns}@5,10) で行った．

4.3 実験結果

表 2 に評価実験の結果を示す．なお，太字は評価指標ごとの最良値を表す．この結果から，DKG はベースラインである tf-idf と比較して高性能なキーフレーズ付与が可能となっており，英語と同様に日本語でも効果があることが確認できた．また，tar-rank が従来手法よりも性能向上していることから，再選択を行うことでキーフレーズとして不適切なものを省く学習が行えたと考えられる．さらに，tar-high-rank が最良であることから，上位キーフレーズを考慮することで，より再選択を効果的に行えていることがわかる．また，再選択を行うことで，文書に出現しないキーフレーズの再現率も向上していることがわかる．

4.4 分析と考察

再選択過程の適用による，理論的な性能限界を確認する．図 3 に従来手法の $F@k(k = 1, 2, 3 \ldots 10)$ と，生成過程における上位 20 個の候補を適切に再選択した場合の $F@k(k = 1, 2, 3 \ldots 10)$ を示す．この結果から，再選択手法の改良で更なる性能向上の余地があることがわかる．

tar-rank が性能向上した理由を考察するため，Encoder-Decoder モデルの特徴である同じ単語を繰り返した誤りの数について調査した．その結果，DKG では 274 個あった誤りが tar-rank では 226 個となり，不適切なキーフレーズの除去に成功していることがわかった．

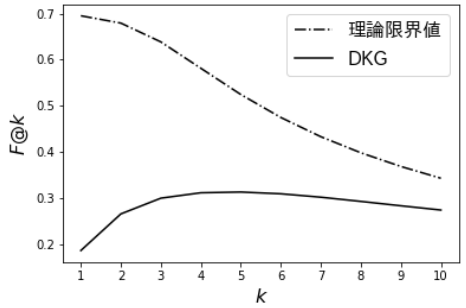

図 3: DKG の $F@k$ と再選択の理論限界値

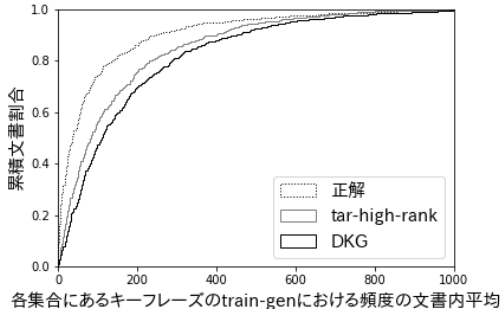

図 4: 各キーフレーズ集合の tarin-gen における出現頻度の文書内平均に関する累積文書割合

DKG と tar-high-rank が上位に生成するキーフレーズの，学習データにおける頻度の傾向を把握するために，2 手法が生成した上位 5 つのキーフレーズと正解キーフレーズの 3 つの集合における各キーフレーズが train-gen に出現する頻度を集計する．図 4 にテストデータの文書ごとに，各キーフレーズ集合の平均頻度を算出し，その平均値ごとの文書数に関する累積文書割合のグラフを示す．この結果から，DKG では正解よりも低頻度語を少なく生成している傾向がわかる．対して，再選択を行うことで低頻度語の生成確率が向上していることがわかる．このことから，tar-high-rank は組み合わせを考慮した結果，低頻度語も生成することができるようになり，全体の性能と未出現キーフレーズの再現率が向上したと考えられる．

最後に，生成結果の定性的評価を行ったところ，「発話」が「漸次的精緻化」に置き換わるなど，平易なキーフレーズが専門的で低頻度なキーフレーズに置き換えられている傾向にあることがわかった．また，図 2 の例にある「辞書」から「機械辞書」など，高頻度な単語を組み合わせて低頻度なキーフレーズとなったものを上位に生成して正解する例もあった．統計的にも単語を組み合わせる傾向は現れており，並び替え前の上位 5 キーフレーズの平均単語数は 1.99 であるのに対して，並び替え後は 2.11 と長くなっている．

5 まとめ

文書のキーフレーズを生成する Deep Keyphrase Generation に対して，生成したキーフレーズ集合から最適なキーフレーズ集合を再選択する手法を提案した．これにより，キーフレーズとして不適切なものを取り除き，キーフレーズの組み合わせを考慮することが可能となり，性能向上が確認できた．現状では再選択による理論限界値と乖離があるため，今後は文書の内容を考慮したキーフレーズの再選択などにより，性能向上を目指す．また，生成過程と再選択過程を統合し，最適なキーフレーズの組み合わせを生成するモデルを構築することで，従来手法では候補として生成しなかったキーフレーズも生成することを目指す．

参考文献

[1] Zhiyuan Liu, Wenyi Huang, Yabin Zheng and Maosong Sun. Automatic Keyphrase Extraction via Topic Decomposition. Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, 2010.

[2] Animesh Prasad and Min-Yen Kan. WING-NUS at SemEval-2017 Task 10: Keyphrase Identification and Classification as Joint Sequence Labeling. Proceedings of the 11th International Workshop on Semantic Evaluations, 2017.

[3] Isabelle Augenstein and Anders Sogaard. Multi-Task Learning of Keyphrase Boundary Classification. Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, 2017.

[4] Rui Meng, Sanqiang Zhao, Shuguang Han, Daqing He, Peter Brusilovsky and Yu Chi. Deep Keyphrase Generation. Proceedings of 55th Annual Meeting of Association for Computational Linguistics, 2017.

[5] Jiatao Gu, Zhengdong Lu, Hang Li, and Victor O. K. Li. Incorporating copying mechanism in sequence-to-sequence learning. arXiv preprint arXiv:1603.06393, 2016.

[6] Noriko Kando, Kazuko Kuriyama, Toshiko Nozue, Koji Eguchi, Hiroyuki Kato and Souichiro Hidaka. Overview of IR Tasks at the First NTCIR Workshop. Proceedings of the first NTCIR workshop on research in Japanese text retrieval and term recognition, 1999.

[7] Chenxi Zhu, Xipeng Qiu, Xinchi Chen and Xuanjing Huang. A Re-ranking Model for Dependency Parser with Recursive Convolutional Neural Network. Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing, 2015.

[8] Jingyi Zhang, Masao Utiyama, Eiichiro Sumita, Graham Neubig and Satoshi Nakamura. Improving Neural Machine Translation through Phrase-based Forced Decoding. Proceedings of the The 8th International Joint Conference on Natural Language Processing, 2017.