論文の要旨からのタイトル生成における キーワードと分野別 fine-tuning の効果

金野佑太
茨城大学　工学部
17t4039x@vc.ibaraki.ac.jp

古宮嘉那子
茨城大学　理工学研究科
kanako.komiya.nlp@vc.ibaraki.ac.jp

1　はじめに

本稿では、Encoder-Decoder モデルにより論文の要旨からタイトルの生成を行う際に、論文の要旨にキーワードを追加する手法と、論文の分野ごとに fine-tuning する手法の効果について検証する。訓練データに、論文の要旨のみを使用したモデル、論文のキーワードのみを使用したモデル、論文の要旨とキーワードを使用したモデルの合計 3 つのモデルを作成し、スコアの比較を行った。また、様々な学会に投稿された論文のデータを用いてモデルを作成し、電子情報通信学会と日本建築学会に投稿された論文のデータを使用して fine-tuning を行った。テストデータには、電子情報通信学会、日本建築学会、情報処理学会、土木学会、高分子学会の 5 つの学会のデータを使用した。評価には要約用の評価指標である ROUGE を用いた。

実験の結果、論文の要旨にキーワードを追加することで ROUGE スコアの改善は見られなかったが、論文の分野ごとに fine-tuning することで ROUGE スコアの向上が見られた。

2　関連研究

文書要約の研究はこれまでに多く行われており、様々なモデルが考案されている。See ら [1] は、Pointer を介して入力系列から単語をコピーする Pointer-Generator と、入力系列中の要約された内容を追跡する Coverage Vector を使用するモデルを提案した。その結果、要約の不正確さと単語の繰り返し出力を低減することを示した。本研究では、キーワードを利用した文書要約の効果を検証している。キーワードを利用した文書生成の研究には、[2] がある。また、本研究では論文のタイトルと要旨を研究に用いているが、これらから特徴語を抽出する研究には [3] がある。

3　提案手法

3.1　Encoder-Decoder モデルによる論文要旨からのタイトル生成

本稿では、Encoder-Decoder モデルによる論文の要旨からのタイトル生成において、論文の要旨にキーワードを追加する手法と、論文の分野ごとに fine-tuning する手法の効果について検証する。

システムの入力には論文の要旨を単語分割した系列データを使用する。単語分割には形態素解析器である MeCab[1)] を使用した。Encoder では入力となる系列データが Embedding 層で単語分散表現へと変換され、RNN 層により隠れ状態ベクトルが算出される。RNN 層では LSTM を利用し、Forward LSTM 層と Backward LSTM の 2 層を持つ BiLSTM を使用した。Decoder では隠れ状態ベクトルと単語分散表現を用いてタイトルとなるターゲット系列が生成される。

3.2　論文のキーワードの追加

本実験では、Encoder-Decoder モデルへの入力として、論文のキーワードの果たす効果について検証する。本研究で使用した入力データの形態は以下の 3 通りである。

- 要旨のみ
- キーワードのみ
- 要旨とキーワード

要旨は MeCab を用いて単語分割した系列データ、キーワードはキーワードのリストを空白区切りの系列データとして扱う。

論文の要旨にキーワードを追加するのは、タイトルにほぼ必須とも言える重要な単語が欠落することを防ぐ効果を期待している。

1）https://taku910.github.io/mecab/

3.3 分野別 fine-tuning

本実験では、論文の分野ごとに fine-tuning することによる論文要旨からのタイトル生成の効果について検証する。

3.2 節で述べた要旨のみを入力データとしたモデルに対して、特定の分野の論文を訓練データとして fine-tuning を行った。要旨のみを入力データとしたモデルは複数の分野の論文を訓練データとして作成するため、これを General モデルとする。実験では、General モデルと fine-tuning したモデルの ROUGE スコアを比較する。

論文の分野ごとに fine-tuning を行うのは、タイトルを生成したい論文の分野が既にわかっているときに同じ分野の論文で訓練されたモデルの方が高い精度を出すことを期待している。

4 実験

4.1 モデル

モデルの構築には、See らの実装を参考にしたモデルである OpenNMT BRNN(1layer, emb128, hid512) モデルのパラメータを使用した。実験では LSTM を用いた Attention 付き Encoder-Decoder モデルを使用した。Encoder-Decoder ともに単語ベクトルサイズを 128、次元数を 512 とする 1 層 LSTM を使用した。Encoder では双方向 LSTM(Bidirectional LSTM) を使用し、各方向の次元数をそれぞれ 256 とした。Attention 層には Bahdanau ら [4] の注意機構 MLP を使用した。最適化アルゴリズムには adagrad を使用し、学習率を 0.15 にした。そして、モデルがソースから単語をコピーするためのコピー機構を導入した。また、実装にはオープンソースのニューラル機械翻訳及びニューラルシーケンス間学習のためのツールである OpenNMT[2] を使用した。

4.2 データセット

実験に使用するコーパスに NII Testbeds and Community for Information access Research（NTCIR[3]）はが提供するデータセット NTCIR-1 と NTCIR-2 を用いた。NTCIR-1 と NTCIR-2 はそれぞれ学会発表データベースからデータを収集したもので、論文の要旨やタイトル、主催学協会名や著者が付与したキーワードのリストなどが含まれている。

表 1 学会別のコーパス (組数)

学会	全てのデータ	キーワードあり
電子	102,385	101,670
建築	74,762	74,655
情報	34,330	33,901
土木	32,851	32,036
高分子	32,829	32,732
その他	171,205	147,063

表 2 キーワードに関する実験に使用するコーパス (組数)

コーパス	学習	開発	テスト
要旨のみ	274,353	66,805	107,899
キーワードのみ	267,802	66,805	107,899
要旨とキーワード	267,802	66,805	107,899

論文のキーワードに関する実験では、学習、開発、テストに使用するデータは学会別に分割した。データ数が多かった上位 5 つの学会である、電子情報通信学会、日本建築学会、情報処理学会、土木学会、高分子学会とその他の学会に対してそれぞれ分割を行った。使用したデータの組数は表 1 に示す。全体ではその他の学会を含め 69 個の学会が存在する。また、キーワードが付与されていない論文データも存在するため、キーワードありのデータは全てのデータよりも数が少なくなっている。

各学会の学習、開発、テストデータの分割の仕方は以下の通りである。電子情報通信学会と日本建築学会のテストデータは fine-tuning 用としても使用するため割合が多くなっている。

- 電子情報通信学会 = (4:1:5)
- 日本建築学会 = (4:1:5)
- 情報処理学会 = (3:1:1)
- 土木学会 = (3:1:1)
- 高分子学会 = (3:1:1)
- その他の学会 = (4:1:0)

要旨のみ、キーワードのみ、要旨とキーワードを入力とするモデルに使用するデータの組数は表 2 に示す。

分野別 fine-tuning の実験に使用する学会はデータ数が多い上位 2 つの学会である電子情報通信学会と日本建築学会を使用する。fine-tuning には、General モデルの学習データと開発データ以外のデータを使用した。fine-tuning 用の電子情報通信学会と日本建築学会のデータは 5 分割し、3:1:1 の割合で学習

2） https://github.com/OpenNMT/OpenNMT-py
3） http://research.nii.ac.jp/ntcir/index-ja.html

表 3 分野別 fine-tuning の実験に使用するコーパス (組数)

学会	学習	開発	テスト
電子	30,501(40,954)	10,167(10,238)	10,167
建築	22,398(29,904)	7,465(7,477)	7,465
情報	0(20,598)	0(6,866)	6,866
土木	0(19,710)	0(6,570)	6,571
高分子	0(19,697)	0(6,566)	6,566
その他	0(136,939)	0(34,266)	0

データ、開発データ、テストデータとして使用し、5 分割交差検定を行った。テストデータには General モデルと同様に 5 つの学会のデータを使用した。実験に使用したデータの組数は表 3 に示す。括弧内は General モデルの作成に使用するデータ数である。

4.3 評価

評価には、テキスト要約用の評価指標である ROUGE[5] を使用した。ROUGE には様々なバリエーションがあるが、本実験では以下の 3 つの指標を用いる。

- ROUGE-1
- ROUGE-2
- ROUGE-L

ROUGE-N(N=1,2,...) は N-gram 単位での単語の一致率で評価する手法である。ROUGE-N のスコアは再現率と適合率の調和平均である F 値に等しい。再現率は、正解データの単語数のうち予測データと一致した単語数の割合を表し、式 (1) のように計算できる。また、適合率は、予測データの単語数のうち正解データと一致した単語数の割合を表し、式 (2) のように計算できる。ただし、$Count_{macth}$ は正解データと予測データの n-gram が一致している場合に 1 を返す関数であり、ref は参照要約である正解データ、sum は候補要約である予測データを表す。

$$\frac{\sum_{S \in ref} \sum_{gram_n \in S} Count_{match}(gram_n)}{\sum_{S \in ref} \sum_{gram_n \in S} Count(gram_n)} \quad (1)$$

$$\frac{\sum_{S \in sum} \sum_{gram_n \in S} Count_{match}(gram_n)}{\sum_{S \in sum} \sum_{gram_n \in S} Count(gram_n)} \quad (2)$$

ROUGE-L は、正解データと予測データで一致する最大のシーケンス Longest Common Subsequence（LCS）で評価する手法である。ROUGE-L のスコアは ROUGE-N と同様に再現率と適合率の調和平均に等しい。単語数が m の正解データ X と単語数が n の予測データ Y を用いて、再現率と適合率はそれぞれ式 (3) と式 (4) のように計算できる。ただし、LCS は 2 つの一致する最大のシーケンス長を表す。

$$R_{lcs} = \frac{LCS(X,Y)}{m} \quad (3)$$

$$P_{lcs} = \frac{LCS(X,Y)}{n} \quad (4)$$

表 4 キーワードに関する実験のモデルの ROUGE スコア

モデル	R1F	R2F	RLF
要旨のみ	0.260	0.084	0.243
キーワードのみ	0.197	0.054	0.194
要旨+キーワード	0.244	0.066	0.225

論文のキーワードに関する実験では、要旨のみ、キーワードのみ、要旨とキーワードの 3 つのモデルに対してスコアの比較を行った。電子情報通信学会、日本建築学会、情報処理学会、土木学会、高分子学会をテストデータとし、各テストデータに対して ROUGE スコアを算出してマイクロ平均をとったものを最終的なスコアとして使用した。

分野別 fine-tuning の実験では、General モデルと電子情報通信学会で fine-tuning したモデル、日本建築学会で fine-tuning したモデルの比較を行う。fine-tuning したモデルは 5 分割交差検定を行っているため、5 回分の平均をとったものを最終的なスコアとして使用した。

5 実験結果

表 4 に要旨のみ、キーワードのみ、要旨とキーワードをそれぞれ入力としたモデルの ROUGE スコアを示す。R1F、R2F、RLF はそれぞれ ROUGE-1、ROUGE-2、ROUGE-L の F 値を表す。R1F、R2F、RLF すべてにおいて、要旨のみ、要旨とキーワード、キーワードのみの順に ROUGE スコアが高い結果となった。

次に、表 5 に General モデルと電子情報通信学会で fine-tuning したモデル、日本建築学会で fine-tuning したモデルのテストデータ別の ROUGE スコアを示す。電子情報通信学会で fine-tuning したモデルは、電子情報通信学会をテストデータとしたときが最もスコアが高く、General モデルの電子情報通信学会のスコアを R1F が 0.017、R2F が 0.009、RLF が

表 5 分野別 fine-tuning したモデルの ROUGE スコア

モデル	テスト	R1F	R2F	RLF
General	電子	0.270	0.091	0.253
	建築	0.248	0.076	0.232
	情報	0.250	0.073	0.234
	土木	0.289	0.102	0.271
	高分子	0.225	0.063	0.210
電子 fine-tuning	電子	0.287	0.100	0.273
	建築	0.239	0.071	0.225
	情報	0.261	0.078	0.247
	土木	0.285	0.096	0.267
	高分子	0.228	0.064	0.215
建築 fine-tuning	電子	0.249	0.069	0.231
	建築	0.290	0.090	0.265
	情報	0.236	0.057	0.219
	土木	0.282	0.088	0.262
	高分子	0.225	0.054	0.207

0.020 上回った。また、日本建築学会で fine-tuning したモデルは、日本建築学会をテストデータとしたときが最もスコアが高く、General モデルの日本建築学会のスコアを R1F が 0.042、R2F が 0.014、RLF が 0.033 上回った。

6 考察

入力にキーワードを利用することによる ROUGE スコアの向上は見られなかった。キーワードを利用したモデルの訓練データは少なくなっているが、97.6%以上に当たる量を利用しており、その差は僅かなため大きな影響があるとは考えづらい。実験ではモデルの学習に要約に適したオプションを使用したが、キーワードのみを入力としたモデルを学習する際は、文章生成に適したオプションを使用することで ROUGE スコアが改善されるのではないかと考えられる。また、要旨とキーワードを入力としたモデルは、キーワードを直接入力に組み込むのではなく、出力に組み込めるようにモデルを変更することで ROUGE スコアが改善される可能性がある。その際は、実際にキーワードがタイトルにどの程度含まれているかを調査し、どのくらいの効果があるのかを検証する必要がある。

論文の分野別 fine-tuning をすることによって ROUGE スコアの向上が見られた。つまり、ある学会の論文データで fine-tuning したモデルはその学会の ROUGE スコアを向上させることができる。また、電子情報通信学会で fine-tuning したモデルは、情報処理学会のスコアを R1F が 0.011、RLF が 0.013 上回り、日本建築学会で fine-tuning したモデルは、電子情報通信学会のスコアを R1F が 0.021、R2F が 0.022、RLF が 0.022 下回った。この結果から、論文の分野別 fine-tuning には、他の分野のスコアも大きく変化させることが確認できた。しかし、電子情報通信学会と情報処理学会のように分野が似ている学会であってもスコアの向上に大きく寄与しないこともあり、どの分野にどのような相関があるのかなどを明確に示すには至らなかった。

7 おわりに

本研究では、LSTM を用いて Attention 付き Encoder-Decoder モデルにより論文の要旨からタイトル生成を行い、論文の要旨にキーワードを追加する手法と、論文の分野ごとに fine-tuning する手法の効果を検証した。実験の結果、論文の要旨にキーワードを追加することで ROUGE スコアの改善は見られなかったが、論文の分野ごとに fine-tuning することで ROUGE スコアの向上が見られた。

今後の展望としては、実際にキーワードがタイトルにどの程度含まれているかを調査してどのくらいの効果があるのかを検証し、その上でキーワードを直接出力に組み込めるようにモデルを変更するなどが考えられる。分野別 fine-tuning においてはそれぞれの分野の相関性を求め、類似する分野の論文データを利用することでより大規模なデータで実験を行うことなどが考えられる。

謝辞

本研究は、茨城大学の特色研究加速イニシアティブ個人研究支援型「自然言語処理、データマイニングに関する研究」に対する研究支援　および JSPS 科研費 17KK0002 の助成を受けたものです。

参考文献

[1]Liu P.J. See, A. and C.D. Manning. Get to the point: Summarization with pointer-generator networks. *ACL*, p. 1073–1083, 2017.

[2]張浩達, 上垣外英剛, 高村大也, 奥村学. 複数の言語モデルを考慮したキーワードからの広告文生成. 第 34 回人工知能学会全国大会論文集, pp. 2H6–GS–9–04, 2020.

[3]菊地真人, 山内達登, BUI Tuan Thanh, 梅村恭司. 特徴語抽出の精度改善に向けた反復度と条件付き確率の比較. 第 34 回人工知能学会全国大会論文集, pp. 4Rin1–53, 2020.

[4]Cho K. Bahdanau, D. and Y. Bengio. Neural machine translation by jointly learning to align and translate. *ICLR*, 2014.
[5]Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. *ACL*, p. 74–81, 2004.