抽出型自動要約における 低リソース環境下での他言語データ活用方法の検証

桑原 亮介[1*] 斉藤 いつみ[2] 西田 京介[2] 富田 準二[2] 中山 英樹[1]
[1] 東京大学 大学院情報理工学系研究科
[2] 日本電信電話株式会社 NTT メディアインテリジェンス研究所

1 はじめに

自動要約タスクは，今日のニューラルネットワークモデルの発展および大規模データの整備により精度の向上が著しい．例えば，BertSumExt [8] は，BERT [2] を用いて CNN DailyMail [9], NYT [10], XSUM [11] のデータセットにおいてそれぞれ約 29 万，10 万，20 万文書のデータを使って学習を行い，最高精度を記録している．

しかし，こうした大規模データを全ての言語において整備することは難しい．実際，当該タスクにおける対象言語は英語が主であり，日本語などの他言語においては，学習に十分なデータが用意されていないことが多い．

クロスリンガル転移学習 [7] は，高リソース言語で学習させたモデルを低リソース言語において活用する手法である．本研究では, 抽出型自動要約タスクにおいて，日本語の要約モデルを学習させる際に，英語データと同時学習させることで精度が向上するのか検証を行う. また，効果的な同時学習の方法について検証を行う．

Livedoor News コーパス [1]における実験により，入力言語を明示的に指定した上で複数言語データを直接入力する方法において最も高い精度向上を確認した．

2 問題定義

抽出型自動要約は元文書 p のうち重要と思われる複数文 s_i を抽出する．ニューラルネットワークを用いた場合，それぞれの文における特徴表現に対して最終的な要約文とする/しないを判別する分類問題として考える．

本研究で取り組む抽出型自動要約を定式化する．

問題 1. 抽出型自動要約器は，元文書 p における各文 $s = \{s_1^q, \ldots, s_I^q\}$ について，最終的な要約文として出力する文に 1 となるバイナリ変数 $q = \{0, 1\}$ を割り当てる．

3 基本モデル

本研究では，抽出型自動要約の基本モデルとして当該タスクにおいて最高性能を達成した手法を採用する [8]．本手法は元文書 p のエンコーディングに BERT を使用し，複数文の特徴表現を得るために文の数に応じた [CLS] トークンを文の先頭に用いる．また，要約抽出に必要な文書全体の特徴表現獲得のため，2 層 Transformer [5] を BERT の出力した特徴表現上に作用させている．上記を定式化すると次のように表すことができる．

$$\tilde{h}^l = \mathrm{LN}\left(h^{l-1} + \mathrm{MHAtt}\left(h^{l-1}\right)\right), \quad (1)$$

$$h^l = \mathrm{LN}\left(\tilde{h}^l + \mathrm{FFN}\left(\tilde{h}^l\right)\right). \quad (2)$$

ここで，h^0 は，BERT によって得られた文ベクトルであり，MHAtt, LN, FFN はそれぞれマルチヘッドアテンション，線形変換，フィードフォワードニューラルネットワークの関数を表す．2 層 Transformer を適用したそれぞれの文 [CLS] トークンに対応する特徴表現 h_i^2 はシグモイド関数 σ を通して次のように変換される．

$$\hat{y}_i = \sigma\left(W h_i^2 + b\right). \quad (3)$$

W, b はそれぞれ学習可能なパラメーターである．目的関数 L は，モデルの出力 $\hat{y}_i$ と正解ラベル y_i のバイナリクロスエントロピーであり，下記のように表せられ，これを最小化する．

$$L = -\sum_{i=1}^{I} y_i \log \hat{y}_i. \quad (4)$$

*NTT におけるインターンシップ期間中の貢献.
[1]https://github.com/KodairaTomonori/ThreeLineSummaryDataset からクロールしたものより作成

手法	日本語データ数	英語データ数	ROUGE-1	ROUGE-2	ROUGE-L
未学習	0	0	18.21	7.36	17.95
日本語データのみ	152,293	0	26.10	11.47	25.71
直接入力法	152,293	287,226	26.70	12.13	26.34
直接入力法（+言語タグ）	152,293	287,226	**27.26**	**12.41**	**26.91**
翻訳後入力法（英→日）	439,377*	0	26.68	11.66	26.40

表 1: 日本語データ（Livedoor News）における各手法の ROUGE スコア．直接入力法と翻訳後入力法の両手法においてスコアの向上が確認できた．* 付きのデータ数は翻訳文を含む．

推論時は，$\hat{y} = \{\hat{y}_1, \ldots, \hat{y}_I\}$ について，スコアの高い上位 3 文のインデックスに対して $q = 1$ を割り当てる．

本研究では英語および日本語の複数言語同時学習に対応するために，Multi-Lingual BERT [2] を採用している．Multi-Lingual BERT は，事前学習時に 100ヵ国以上の言語の Wikipedia を使用しているため複数言語の同時学習に対応可能である．

4 同時学習手法

本研究では，高リソース言語である英語データを活用し低リソース言語である日本語データの自動要約の精度向上を図る．そこで，高リソース言語データを活用する方法として，Multi-lingual BERT に対して，**2 言語データの直接入力する方法（直接入力法）と機械翻訳を用いて 1 言語に統一して入力する方法（翻訳後入力法）**の 2 手法について検証を行う．それぞれの手法の概要を図 1 に示す．

4.1 直接入力法

本手法では，Multi-Lingual BERT の多言語性を活用し，英語+日本語の 2 言語データを各言語のままモデルに直接入力し学習を行う．本手法は，モデルやデータに追加作業が発生しないため転移学習のコストが低く，また，複数言語を同時に扱うため，要約文の出現位置パターンなどといった，非言語的な特性の抽出が期待できる．さらに，追加的な手法として，それぞれの言語データの先頭に言語タグを挿入し，明示的に対象言語を指定した場合の精度についても検証を行った．明示的に対象言語を指定することにより，モデル内部での言語表現獲得がより精緻になることが期待できる．

4.2 翻訳後入力法

本手法は，英→日翻訳を用いてデータを 1 言語に統一したのちに Multi-lingual BERT で学習を行う．1 言語に統一することで，モデルがより対象言語の言語性を学習できるとともに，それを加味した要約の出現位置パターンを抽出することが期待できる．機械翻訳に

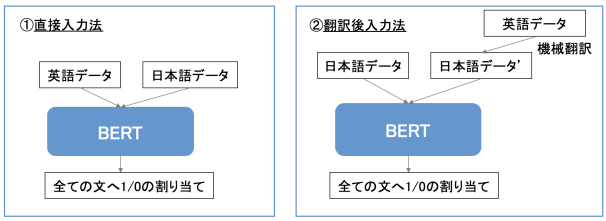

図 1: 本研究における日本語・英語データの同時学習手法．

は Transformer を用いて，約 90 万文対のデータで学習を行った．学習用データには，大規模オープンソース日英対訳コーパス [15]，日英法令コーパス [2]，ロイター対訳コーパス [16]，田中コーパスを用いた [17]．

5 実験

5.1 データセット

本研究では，英語および日本語用の要約データとして，CNN DailyMail，Livedoor News を用いた．CNN DailyMail は自動要約タスクにおいて頻繁に使われるベンチマークデータセットである． CNN DailyMail，Livedoor News データセットのそれぞれにおいて Train/Valid/Test を 287,226/13,368/11,490，152,293/961/963 とした．抽出型自動要約の正解文として必要な Oracle は，Yang ら [8] の手法に沿い，本来の正解文に対する ROUGE-2 が最も高くなるように元文書から文を選択した．

5.2 実験設定

ハイパーパラメーター BERT のハイパーパラメーターは，”bert-base-uncased”の実装に従った．[3] Transformer は 2 層 8 アテンションヘッドのモデルを用い，Embedding を 768 次元，フィードフォワードレイヤーの中間層は 2048 次元とした．それぞれのサブレイヤー

[2] Japanese Law Translation Database System http://www.japaneselawtranslation.go.jp/ より Graham Neubig 氏によるクロールデータ

[3] https://github.com/huggingface/transformers

手法	+日本語データ数	ROUGE-1	ROUGE-2	ROUGE-L
未学習	-	18.21	7.36	17.95
直接入力法	なし（英語データのみ）	22.87	9.40	22.46
直接入力法	10,000	25.07	10.91	24.61
直接入力法	150,000	**26.70**	**12.13**	**26.34**

表 2: 日本語データ（Livedoor News）数毎の ROUGE スコア．未学習以外の手法は，CNN DailyMail の全データ（287,226 件）を用いて学習を行なっている． 日本語データを全く使わない場合にも ROUGE スコアが上昇していることが確認できる．また，日本語データを 10,000 程度使用することで更なる精度向上を確認できた．

の最後に $p = 0.1$ のドロップアウトレイヤーを入れている．

学習用設定　最適化手法として Adam [3] を用い，$\beta_1 = 0.9$, $\beta_2 = 0.999$ とした．学習係数は [5] のスケジューリングを活用し，warmup = 10,000 とした．したがって，学習率 lr は下記のように表される．

$$\mathrm{lr} = 2e^{-3} \cdot \min\left(\mathrm{step}^{-0.5}, \mathrm{step} \cdot \mathrm{warmup}^{-1.5}\right). \quad (5)$$

ここで，step はそれぞれパラメーターのアップデート回数である．

全モデルにおいて 1GPU (GTX1080 Ti) で 200,000 ステップの学習を行い，1,000 ステップ毎にパラメーターを保存した．Validation loss の低いパラメーター上位 3 つを用いてテストデータにて評価を行い，その平均を最終スコアとした．

5.3　結果

表 1 に結果を示す通り，2 言語直接入力法，翻訳後入力法のいずれにおいても精度の向上を確認することができた．ここで,「未学習」とは学習を行なっていない当該モデルの ROUGE を計測した結果である．翻訳後入力法では，モデルが単に要約文の出現位置パターンを記憶しているだけでなく，文の意味レベルで重要性の理解をしているため精度が向上したと考えられる．2 言語直接入力法の精度向上理由として，多言語対応モデルを用いているため，対象言語が違う場合にも要約文の出現位置パターンを強く学習していると考えることができる．また，言語タグを用いることで更なる精度の向上が確認できたことから，対象言語を明示的に指定することでより効果的に学習が行われていることがわかる．結果的に当該手法において要約精度が最も高くなった．

6　分析

6.1　低リソース言語のデータ数と要約精度

Livedoor News データセットでは約 15 万のデータで学習を行なっているため，その時点で一定の精度が出ており，高リソース言語（英語）データ活用による効果を測定するのが難しかった．そこで，本分析では Livedoor News のデータ数を減らすことで，より低リソース環境を想定した場合の英語データ活用の効果を検証する．Livedoor News のデータ数を 0, 10,000, 150,000 とした場合の要約の精度を表 2 に示す．日本語データ数が 0，つまり英語データのみを活用した場合にも，要約の精度向上が確認できた．これは，ニュースドメインにおける重要文出現位置パターンが複数言語間で類似していることが理由であると考えられる．また，日本語データ数が 10,000 程度でも未学習状態に比べて ROUGE-1, ROUGE-2, ROUGE-L のそれぞれにおいて 6.86, 3.55, 6.66 の精度向上が確認できた．この検証結果から，同一ドメインの要約においては，英語データが大量にあれば，日本語のデータが 10,000 程度でも一定の精度を保てることが確認できた．

6.2　要約データとモデル出力の分布

ニュース記事の自動要約においては，先頭 5 行程度に重要文が集中していることが知られている．そこで本分析では，学習用の Livedoor News データ数の増加により，節 6.1 で使用したモデル出力文の出現分布が正解文の分布にどの程度近づいていくのかを図 2 で可視化した．出現位置を 10 行目までとした累積分布を示している．未学習の状態では 5~8 文目を予測文として多く出力しているが，学習用の Livedoor News データを増やすごとに正解文の累積分布に近づいていることがわかる．

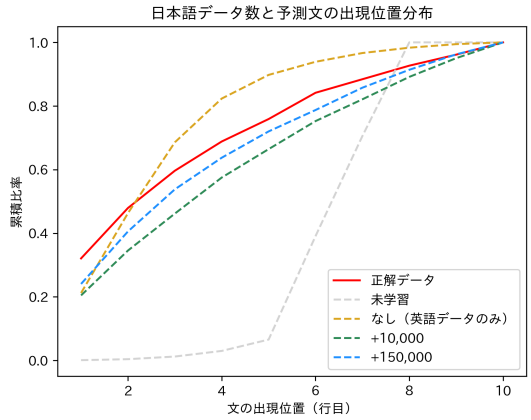

図 2: Livedoor News テストデータにおける正解文と学習データを増やした際のモデル予測文の出現位置分布．学習データを増やすごとに正解文の分布に近づいている．

7 おわりに

本研究では，低リソース言語（日本語）を対象とした抽出型自動要約を行う際の高リソース言語（英語）の活用の有効性とその方法について検証を行った．実験により，複数言語を直接入力する手法，英→日翻訳を行い対象言語を 1 つに統一する手法のいずれにおいても要約精度の向上を確認した．特に，言語タグを入力の先頭に付与し入力を明示的に指定した上で複数言語データを直接入力する手法において高い精度向上を確認した．

関連研究と議論． 抽出型自動要約は，[6] によりニューラルネットワークモデルに適用されて以降，精度の向上が著しい．また，クロスリンガル転移学習 [7] は以前より研究されてきた分野である．Rui ら [12] は，情報検索に 2 つの言語表現を用いてタガログ語などの言語の検索精度を向上させた．Chen ら [13] は，複数言語間で普遍な言語的な特徴を抽出すると同時に，類似性のある言語の特徴を対象とする言語に用いることで，テキスト分類と系列ラベリングタスクにおいて精度を向上させた．自動要約においては，NCLS [14] が，機械翻訳を用いてデータ拡張を行うことにより，自動要約の精度を向上させている．一方で，自動要約における日本語を対象とした研究はあまり見られていない．

本研究の重要性． 自動要約における低リソース問題に日本語データの観点で初めて着眼し，複数の手法の考案比較を行った．本研究から得られた知見は，ニュースドメインに限らず，その他ドメインにおける自動要約の低リソース問題に対して応用可能である．

参考文献

[1] L. J. Ba, R. Kiros, and G. E. Hinton. Layer normalization. *arXiv*, 1607.06450., 2016.
[2] J. Devlin, M. Chang, K. Lee, and K. Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. *arXiv.*, 1810.04805., 2018.
[3] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015.
[4] I. Loshchilov, and F. Hutter. Fixing weight decay regularization in adam. *arXiv.*, 2017.
[5] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention is all you need. In *NeurIPS*, 2017.
[6] R. Nallapati, F. Zhai and B. Zhou. SummaRuNNer: A Recurrent Neural Network based Sequence Model for Extractive Summarization of Documents. arXiv., 1611.04230., 2016.
[7] Y. David, N. Grace and W. Richard. Inducing Multilingual Text Analysis Tools via Robust Projection across Aligned Corpora. In *HLT*, 2001.
[8] L. Yang, L.Mirella Text Summarization with Pretrained Encoders. In *EMNLP*, 2019.
[9] KM. Hermann, T. Kočiský, E. Grefenstette, L. Espeholt W. Kay, M. Suleyman and P. Blunsom. Teaching Machines to Read and Comprehend. In *NeurIPS*, 2019.
[10] E. Sandhaus, The New York Times Annotated Corpus. 2008.
[11] N. Shashi, C. Shay B. and L. Mirella. Don't Give Me the Details, Just the Summary! Topic-Aware Convolutional Neural Networks for Extreme Summarization. In *EMNLP*, 2018.
[12] Z. Rui, W. Caitlin, S. Sungrok, B. Garrett, F. Alexander, H. William, V. Neha and R. Dragomir. Improving Low-Resource Cross-lingual Document Retrieval by Reranking with Deep Bilingual Representations. In *ACL*, 2019.
[13] X. Chen, AH. Awadallah, H. Hassan, W. Wang and C. Cardie. Multi-Source Cross-Lingual Model Transfer: Learning What to Share. In *ACL*, 2019.
[14] Z. Junnan, Z. Junnan, W. Qian, W. Yining, Z. Yu, Z. Jiajun, W. Shaonan and Z. Chengqing. NCLS: Neural Cross-Lingual Summarization. In *EMNLP*, 2019.
[15] 石坂達也，内山将夫，隅田英一郎，山本和英. 大規模オープンソース日英対訳コーパスの構築. 情報処理学会 第 191 回自然言語処理研究会, 2009.
[16] Utiyama. M and Isahara. M. Reliable Measures for Aligning Japanese-English News Articles and Sentences. In *ACL*, 2003.
[17] Tanaka. Y. Compilation of A Multilingual Parallel Corpus. In *Pacling*, 2001.