文の分散表現を利用したMMR法と重要文/非重要文分類に基づく講義・講演ドキュメントの要約

早川由倭　　山本一公　　中川 聖一
中部大学　工学部　情報工学科

1. はじめに

英語の講義を自動音声認識し、日本語で字幕表示する研究を行っている。英語講義音声をすべて翻訳し字幕にすると、読みが追い付かない・読み辛いなどの弊害が生じるため講義音声ドキュメントに対応する字幕用の要約システムがあると便利である。しかし講義音声は、長時間であり明確な構造が存在せず冗長性が高い、文の単位や区切りが明確でないなどの特徴があり、このような講義音声ドキュメント特有の現象に対処した要約システムが必要となる。

我々は以前、字幕用の全文表示や重要文表示、重要句表示などの字幕の表示方法と理解度との関係について研究を行い、内容の理解度と表示量の速度や読みやすさから重要文表示が適切であることを示した[1]。過去の重要文抽出の手法として、テキスト・音声要約で有効な手法として広く用いられる Maximal Marginal Relevance (MMR)[2]や、文の重要性を表す素性の抽出と抽出した素性に基づく分類からなる feature-based 法[3]などが提案されている。また原文中の出現頻度が高い重要語 [4]、文の位置情報 [5]、接続詞のような手掛かり語 [5] などを用いる重要文抽出手法も提案されている。実際に FF-SE [6]という要約方法では、文の特徴として、文の長さや文と文の類似度などに加え、文の位置情報も考慮しており、13 種類の要約方法を比較した中で一番良い評価結果が得られている[7]。また、接続詞のような手掛かり語を機械学習によって自動抽出して自動要約の素性とし、重要文の連続性を考慮することで、より講義音声ドキュメントに対応する要約システムに改善する研究[8]もされている。

最近では、文の構文情報・意味表現として、文の分散表現が良いとされている。実際にフレーズや文を固定長のベクトルで表現する研究[9]や、単語の分散表現を得る研究[10]などが行われている。このことから分散表現がいろいろな自然言語処理分野で利用できるとされており、重要文抽出の要約にも利用されている[11]。

本稿では、従来使用されている要約システム MMR にニューラル機械翻訳(NMT)や Bidirectional Encoder Representations from Transformers (BERT) [12]を利用して抽出した文の分散表現を利用する方法、および、重要文/非重要文の分類に基づく方法を提案し、英語や日本語のさまざまな講義・講演音声ドキュメントを要約・評価した結果を報告する。

２．Maximal Marginal Relevance

テキスト要約において有効な手法であるMMRについて説明する。MMR は、ドキュメント（本稿においては講義音声全体）との関連度と、情報の新規性に基づいて抽出する文を順に決定していくことで、全体としてドキュメントとの関連が高くかつ冗長性の低い文集合を抽出することを目指す手法である。いくつかのバリエーションが存在するが、本稿では文献[13]で定義されているものを使用する。

MMR の文抽出アルゴリズムでは、ドキュメント D、文i、抽出文数Rを用いて、文i に含まれる単語からなるベクトルS_iを、単語の出現頻度(Term Frequency)を用いて以下のように定義する。

$$S_i = tf_i = \left(tf_{i,w_1}, tf_{i,w_2}, \dots,\ tf_{i,w_n}\right),$$
$$tf_{i,w} = \ f_w \cdot \log(\frac{f_{\hat{w}}}{f_w}),$$

ここでf_wはドキュメント中の単語wの頻度であり、$\hat{w}$はドキュメント中に最も出現する単語である。ドキュメント全体に分布する単語よりも、特定の箇所に集中して出現する単語の方が重要度が高いと考えられるので、$tf_{i,w}$は Term Frequency の値をドキュメント中の最大単語頻度に基づいて修正している。

文のベクトルの集合$S_{nrk} = \{S_1, S_2, \dots, S_N\}$の$S_i$に対して、$S_1$から順に以下の式を計算し、求めた$S_{max}$を重要文集合 S_{rk}に加える。これを繰り返し行うことで要約文を抽出する。

$$S_{max} = \underset{S_i \in S_{nrk}}{\operatorname{argmax}}\{\lambda\left(Sim(S_i, D)\right) - (1-\lambda)(Sim(S_i, S_{rk}))\}$$

Simは 2 つのベクトル間の類似度を表し、本稿ではコサイン類似度を用いる。式の第一項は文とドキュメントの関連度を表し、第二項は文と重要文集合の類似度の負の値、すなわち情報の新規性を表す。このとき、λはドキュメントとの関連度と冗長性の間のトレードオフである。MMR のツールとして、Text-Summarization-MMR[1]を使用し、本稿では$\lambda = 0.5$と設定する。

MMR では、文を単語の出現頻度を用いたベクトルで表現しているため、文にそれ以上の情報を持たせることが難しいという問題点がある。

３．文の分散表現

3.1 ニューラル機械翻訳の利用

NMT の主流であるエンコーダ－デコーダモデルを用いて文の分散表現を取得する。エンコーダ－デコーダモデルは、原言語の入力文$\{w_1,w_2,\ldots w_n\}$を単語レベルの埋め込みベクトル$\{e_1,e_2,\ldots,e_n\}$に変換してエンコーダへ入力する。エンコーダから出力される分散表現は、入力文の意味や構造を捉えた分散表現$\{h_1,h_2,\ldots,h_n\}$ の h_n（本稿では、文ベクトルと呼ぶ）となる。その文ベクトルと原言語文、直前までに予測された目的言語文をデコーダに入力し、目的言語の次単語の予測を繰り返し、最終的に目的言語文$\{y_1,y_2,\ldots,y_n\}$ を出力する（図 1）。

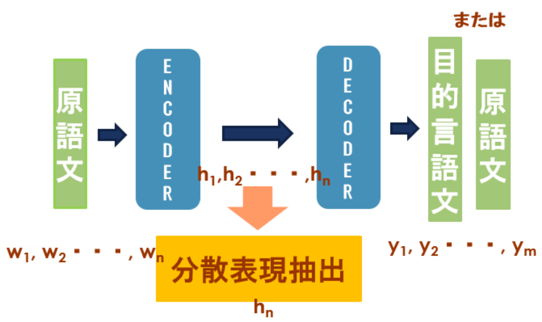

図 1: NMT の流れ図

このエンコーダから出力される文ベクトルは、原言語の入力文の意味や構造を表現した高次元連続空間上の実数値であると考えられる [14,15]。また原言語文の情報を保存しているため、要約に利用できる文の分散表現であるとも考えられる。

本稿では NMT の学習に、エンコーダに双方向 LSTM をもつ OpenNMT[2]ツールを使用した。学習データには英日 100 万文のパラレルコーパスからなる ASPEC コーパス[16]を使用した。隠れユニット数は順方向 250+逆方向 250 とし、250 次元＋250 次元の分散表現を抽出した。本稿では要約に英語と日本語のドキュメントを使用するため、入力する原言語文に英語または日本語のドキュメント、出力する目的言語文に英語または日本語のドキュメントという 4 パターンで学習を行い、4 通りの分散表現を抽出した。入力文と出力文が同一の場合は、オートエンコーダによる分散表現である。

3.2 BERT の利用

自然言語処理モデル BERT を用いて文の分散表現を取得する。BERT とは Google により開発された自然言語モデル[12]であり、Transformer モデルがベースになった双方向性をもつモデルである。転移学習することが可能な汎用性の高いモデルであることから、様々な言語処理に対応できるため、要約の分散表現にも利用できると考えられる。また BERT は、文章中の単語同士のつながりや文脈・文法を理解することができるため、より精度の良い分散表現を得ることが可能である。

本稿では、オープンソースで公開されている、文脈を既に学習させた事前学習(Pre-Training)モデルを利用し、このモデルを基に文の分散表現を生成するツール bert-as-service[3]を使用する。このツールでは、BERT の Pre-Training モデルをもつサーバーにドキュメントを渡すと、文をトークン化し BERT モデルを用いて文の分散表現を出力することができる。本稿では、英語と日本語のドキュメントを要約するため 2 つの事前学習モデルを使用した。英語のドキュメントの文の分散表現を得る際は BERT-Large, Uncased (Whole Word Masking)[4]、日本語のドキュメントの分散表現を得る際は BERT-Base, Multilingual Cased[4] を利用し、英語 1024 次元、日本語 768 次元の分散表現を抽出した。

４．評価実験

4.1 評価データと評価尺度

本稿では、要約に使用する講義・講演音声ドキュメントとして英語と日本語のドキュメントを使用した（表 1)。英語のドキュメントには文数の異なる MIT 講義文 [17]と TED の講演文（英日対訳)、日本語のドキュメントには TED 講演文と CJLC [18]の講義文を使用した。また要約の正解文には人手による要約文を使用した。L11M00 は 6 人の要約結果から要約文集合を作成し[8]、それ以外は著者の一人が作成した。それぞれの元のドキュメントの文に対する要約率、単語に対する要約率を表 2 に示す。Lecture (short)はプログラミングの講義で、要約対象ドキュメントが短く人手でも要約は難しい。また Lecture (long)はアルゴリズムの講義で、これも要約は難しい。

---

[1] https://github.com/fajri91/Text-Summarization-MMR

[2] http://github.com/OpenNMT/OpenNMT-py

[3] http://ghttps://github.com/hanxiao/bert-as-service

[4] https://github.com/google-research/bert

表 3: 要約の評価結果 (ROUGE-3)

要約基準	使用ドキュメント	ベースライン	NMT				BERT
			英日・日英 250 次元	英日・日英 500 次元	英英・日日 250 次元	英英・日日 500 次元	英語（1024 次元） 日本語（768 次元）
正解文数に合わせて要約	Lecture (short)	0.206	0.195	0.173	0.200	**0.212**	**0.216**
	Lecture (long)	0.238	**0.331**	**0.306**	**0.311**	**0.333**	**0.304**
	TED（英語）	0.404	0.392	**0.449**	**0.433**	**0.408**	**0.420**
	TED（日本語）	0.444	**0.446**	**0.526**	**0.526**	**0.491**	**0.495**
	L11M00	0.189	**0.360**	**0.408**	**0.408**	**0.371**	**0.373**
正解単語数に合わせて要約	Lecture (short)	0.202	0.168	0.185	0.167	0.160	0.149
	Lecture (long)	0.350	**0.381**	0.328	**0.373**	**0.395**	**0.363**
	TED（英語）	0.481	0.434	0.437	0.470	**0.490**	**0.485**
	TED（日本語）	0.530	0.501	**0.551**	**0.551**	0.526	**0.547**
	L11M00	0.359	**0.411**	**0.450**	**0.450**	**0.423**	**0.432**

表 1: 評価データ

言語	講義（講演）名・内容	文数（1 講義あたり平均）
英語	Lecture (short) -MIT 講義（1 講義から 5 箇所）	55 文
	Lecture (long) -MIT 講義（2 講義）	580 文
	TED（英語） -IWSLT の TED 講演（10 講演）	101 文
日本語	TED（日本語） -IWSLT の TED 講演（10 講演）	101 文
	L11M00 -日本語講義コーパス CJLC（8 講義）	973 文

評価尺度には Rouge-N [19]を使用する。Rouge-N は評価対象の要約の正解要約に対する N-gram の再現率を表しており、内容の保存に関して評価することが可能である。つまり重要な同一の内容が複数箇所にある場合でも、どれか一箇所が抽出されれば高い評価値となる利点がある。

4.2 評価結果と考察

(a) NMT の利用

NMT で得られる英日（原言語文:英語、目的言語文：日本語）・日英（原言語文:日本語、目的言語文：英語）の分散表現を利用した結果および、英英（原言語文:英語、目的言語文：英語）・日日（原言語文:日本語、目的言語文：日本語）の分散表現を利用した結果を表 3 の NMT 欄に示す。今回は要約基準として、正解要約文数に合わせて要約した場合と、正解要約文の単語数に合わせて要約した場合の 2 つの基準で要約した。

表 3 からわかるように、多くの場合でベースラインを上回る結果（太字）を得ることができた。特に、正解

表 2: 正解（人手）要約文の要約率

講義名	文に対する要約率（平均）	単語に対する要約率（平均）
Lecture (short)	45.4%	36.0%
Lecture (long)	35.5%	50.8%
TED（英語）	46.5%	55.2%
TED（日本語）	46.5%	56.1%
L11M00	26.7%	42.6%

要約文数に合わせて L11M00 を要約した場合、日日の 250 次元の分散表現を利用した MMR は、ベースラインを 20%以上上回る結果を得ることができた。逆に Lecture(short)はベースラインを下回ることが多かった。これは、文数が少なく要約しにくいことが原因であると考えられる。また英日・日英の分散表現より、英英・日日のオートエンコーダの分散表現を利用した方が評価結果が良いことが多いことがわかる。これより、NMT は入力する原言語文と出力する目的原語文を一致させることで、より文の意味や構造を理解した分散表現を得ることができると考えられる。

(b)BERT の利用

BERT から得た文の分散表現を利用した結果を表 3 の BERT 欄 に示す。NMT と同様に、正解要約文数に合わせて要約した場合と正解要約文の単語数に合わせて要約した場合の 2 つの基準で要約を行った。表 3 からわかるように、Lecture(short)を正解単語数に合わせて要約した場合以外、ベースラインを上回る結果（太字）を得ることができた。NMT の結果と比べると、英日・日英の 250 次元の分散表現を利用した場合よりは評価結果が良いことが多いが、それ以外の場合では、NMT の評価結果の方が良いことが多いのがわかる。

(c)重要文/非重要文の分類の利用

BERT の文の分散表現を利用して、文毎に重要文か非重要文かを分類する分類器を多層ニューラルネットワ

ークで実現した。文の分散表現を入力し、重要文(1)/非重要文(0)の分類結果を 1 つのノードに出力するように学習する。学習には評価文以外の、TED の 26 講演 2429 文（重要 1095 文、非重要 1334 文）を用いた。学習文の要約文の割合（要約率）は 45.1%で、TED の評価データの文に対する要約率とほぼ同じである。隠れ層は 4〜6 層で試し、一番結果が良いものを使用した。ミニバッチサイズは 200、最大エポック数は 10000、学習率は 0.01 と設定した。評価データには先述の要約実験と同じ TED 講演文（1010 文）を使用し、入力は各文の分散表現の場合と、前後の文の分散表現を連結した場合（±1 文）を比較した。分類器による要約結果を表 4 に示す。

表 4 からわかるように、正解要約文に対する要約率に合わせて要約した場合（ネットワークの出力値の大きさで選択）、ROUGE-3 でベースラインや NMT、BERT の結果を大きく上回る結果を得ることができた。また、文脈なしで学習した場合と前後 1 文を含めた文脈（±1 文）で学習した場合を比べると、±1 文の文脈がある場合の方が結果が良いことがわかる。よって、文の分散表現に文脈情報を加えて学習すると、要約精度が良くなると考えられる。また、英語と日本語を比べると表 2 と同様に日本語の方が若干値が高いことがわかる。

表 4: 文毎の重要文/非重要文分類による要約結果

TED 講演	ROUGE-3	要約率
英語（文脈なし）	**0.600**	46.5%
日本語（文脈なし）	**0.642**	46.9%
英語（±1 文の文脈）	**0.631**	46.7%
日本語（±1 文の文脈）	**0.647**	46.2%

5．まとめ

本稿では、MMR に文の分散表現を利用した講義音声ドキュメントの要約について述べた。従来の MMR アルゴリズムの単語の出現頻度を用いて文をベクトル化する部分に、NMT や BERT を用いて取得した文の分散表現を利用するという方法を提案した。要約した結果、文の分散表現を利用するとベースラインの Rouge-3 の値を多くの場合で上回ることができた。また、文の分散表現を重要/非重要に分類する 2 分類器を用いて要約する方法でも、ベースラインを上回る結果を得ることができた。今後は文の分散表現を利用した MMR に、重要文/非重要文の 2 分類器を併用し、より精度の良い要約結果を得ることを目指す。

謝辞

本研究は、JSPS 科研費 25280062 の助成を受けた。

参考文献

[1] V.Ferdiansyah, S.Nakagawa. Captioning methods of lecture videos for learning in English, Proc. 25th ICCE, pp.902-907, 2017
[2] G. Murray, S. Renal, and J. Carletta. Extractive Summarization of Meeting Recordings, Proc. Interspeech, pp. 593–596, 2005.
[3] 小林,山口,中川.表層的言語情報と韻律情報を用いた講演音声の重要文抽出, 自然言語処理, Vol. 12, No. 6, pp. 3–24, 2005.
[4] H. P. Luhn. The automatic creation of literature abstracts, IBM Journal of research and development, 2(2):159–165, 1958.
[5] H. P. Edmundson. New methods in automatic extracting, Journal of the ACM (JACM), 16(2):264–285, 1969.
[6] R. Abbasi-ghalehtaki, H. Khotanlou, and M. Esmaeilpour. Fuzzy evolutionary cellular learning automata model for text summarization, Swarm Evolution. Comput. 30, pp.11–26. 2016.
[7] P. Verma, S. Pal, H. Om A comparative analysis on Hindi and English extractive text summarization, ACM Transaction on Asian Low-Resource Language Information Processing, Vol.18, No.3, pp.30-39, 2019.
[8] 藤井,山本,中川.重要文抽出に基づく講義音声の要約, 情報処理学会論文誌, Vol. 51, No. 3, pp. 1094–1106, 2010.
[9] Q. Le and T. Mikolov. Distributed Representations of Sentences and Documents, Proc. 31st International Conference on Machine Learning (ICML 2014), pp. 1188–1196, 2014.
[10] T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space, CoRR, Vol.abs/1301.3781, 2013.
[11] M. Kageback and D. Dubhashi O. Mogren, N. Tahmasebi. Extractive Summarization using Continuous Vector Space Models, Proc. 2nd Workshop on Continuous Vector Space Models and their Compositionality (CVSC2014), pp. 31–39, 2014.
[12] J. Devlin, M. Chang, K. Lee, K. Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, arXiv:1810.04805, 2018.
[13] G. Murray, S. Renal, and J. Carletta. Extractive summarization of meeting recording, Proc. Interspeech, pp. 593–596, 2005.
[14] I. Sutskever, O. Vinyals, Q. Le,"Sequence to sequence learning with neural networks", Advanced in neural information processing systems, pp.3104-3112, 2014.
[15] B. Wang, K. Liu, J. Zhao, "Inner attention based recurrent neural networks for answer selection", Proc. ACL, pp.1288-1297, 2016.
[16] 高田,秋葉,塚田. ニューラル機械翻訳における文書トピック情報の利用, 言語処理学会, A2-1, 2019.
[17] 後藤,山本,中川. 音声認識誤りを考慮した英語講義音声の日本語への音声翻訳システムの検討, 言語処理学会, P19-5, 2017.
[18] M. Tsuchiya, S. Kogure, H. Nishizaki, K. Ohta, S. Nakagawa. Developing Corpas of Japanese Classroom Lecture Speech Contents, Proc. 6-th.LREC, 2008.
[19] C. Lin and E. Hovy. Automatic Evaluation of Summaries Using N-gram Co-occurence Statistics, Proc. Human Language Technology Conference, pp. 71–78, 2003.