TED 英語講演の音声認識・音声翻訳・音声要約の検討

坂野晴彦　桜井陽生　足立十一郎　山本一公　中川聖一

中部大学 工学部 情報工学科

{ep19087-8458@sti, ep18051-1542@sti, ep19002-1545@sti, kazumasayamamoto@isc,

nakagawa@isc}.chubu.ac.jp

概要

我々は、以前から英語の講義や講演を自動音声認識し、要約して日本語で字幕表示する研究を行ってきた。英語講演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、要約システムが必要である。TED 講演音声の音声認識では、DNN-HMM により約 88%の単語正解精度を得た。この書き起こし結果を Transform-er に基づくテキスト翻訳と BERT に基づく重要文抽出型テキスト要約に使用して、音声翻訳と音声要約を評価した。音声翻訳はテキスト翻訳に比べて、BLEU 値が約 14%低下した。一方、音声要約はテキスト要約に比べて、重要文として抽出された区間はほぼ同一であり、音声認識誤りに頑健であった。

1 はじめに

我々は、以前から英語の講義や講演を自動音声認識し、要約して日本語で字幕表示する研究を行ってきた。英語講義・講演音声をすべて翻訳し字幕にすると、ユーザーの読みが追い付かない・読み辛いなどの弊害が生じるため、音声認識システムと翻訳システム以外に要約システムがあると便利である[1]。

音声認識に関しては、DNN-HMM 型のハイブリッド方式と End-to-End 方式があり、後者の方が性能がよいという報告がみられるようになってきた[2,3]。TED 講演音声に関して、英語の音声認識は IWSLT での発表が多いが、英日翻訳に関しては少ない。2021 年には英日の瞬時テキスト翻訳[2]、2022 年には英日の瞬時音声翻訳がタスクに加えられている[3]。上野らは 12 層の Conformer エンコーダ、Attention 付きの単方向 LSTM のデコーダを用いて、TED-LIUM2（211 時間分の学習データ）で、単語誤り率 8.56% という報告をしている[4]。

音声翻訳に関しては、音声認識とテキスト翻訳の接続によるハイブリッド方式と End-to-end 方式があり、最近では後者の進展が目覚ましく、両者の差はなくなってきた[3,5]。我々は講義・講演の翻訳に対して、ハイブリッド方式を採用し、時代に合わせて統計的機械翻訳モデル[6]、RNN ニューラルネットワークの LSTM による系列変換モデル[7]、Transformer モデル[8]を使用してきた。並行して、音声認識誤りに対処するために、N ベストのリスコアリングの利用、認識誤りを含めた対訳ペアによる学習などを試みてきた[6]。本稿では、Transformer にデータ拡張や転移学習を適用した結果を報告する。TED の英日音声翻訳に関しては、福田らは、オフライン方式で 11.6、オンライン・瞬時翻訳方式で 10.6 の BLEU 値を報告している[9]。IWSLT-2022 で最も良かった結果は、大規模な外部音声資源・言語資源を使用した USTC（中国科学技術大学）のシステムで、音声認識誤り率（単語誤り率）約 5%、アンサンブル翻訳ではテキスト入力で 22、音声入力・オフライン方式で約 19 の BLEU 値を報告している[5]。

要約に関しては、抽出型要約と抽象型要約があり、従来は抽出型要約の方が優れていたが、最近はニューラルネットワークの抽象型要約技術が進展し、両者の性能差はなくなってきた[10,11,12]。ただし、人による要約結果の評価では抽出型要約の方がまだ評価が高い[12]。講義・講演を聴講しながら要約を字幕表示する場合は、万遍なく音声に同期した字幕表示が望ましい。そこで我々は、重要な箇所

だけ音声を再生するのにも適している重要文の抽出型要約を採用している。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)[13]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[14]などが提案されている。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、大規模な学習データによるニューラルネットワークを利用した重要文抽出の要約にも利用されている[15,16]。我々は、従来の要約システム MMR に Bidirectional Encoder Representations from Transformers (BERT) [17]を併用する方法を提案してきた[18]。本稿では、本手法と最近の代表的な抽出型要約である BertSumExt 法[16]による音声要約を報告する。

2 TED 講演の音声認識

2.1 Kaldi ツールキットによる DNN-HMM

TED 講演音声をテキストに書き起こすための音声認識ツールキットとして、Kaldi[19]を用いた。Kaldi では様々なコーパスに対応したレシピが公開されており、本稿では TEDLIUMv3 に対応したレシピを用いて音声認識を行った。Kaldi では認識システムとして DNN-HMM を用いており、これは隠れマルコフモデル(Hidden Markov Model)の状態から出力される音素の出力確率を DNN(Deep Neural Network)を用いて算出する手法である。

本稿では DNN の隠れ層の層数は 13 層とし、隠れ層の各ユニット数は 1024 とした。出力層のユニット数は共有トライフォン数に対応しており、GMM-HMM 学習時の学習データ量に依存して自動決定される。活性化関数には ReLU 関数を用い、出力層のみ softmax 関数を用いた。また、損失関数にはクロスエントロピーを用いた。ネットワークへの入力特徴量は MFCC(メル周波数ケプストラム係数)40 次元である。また、特徴量の前処理として fMLLR, LDA,SAT を行い、不特定話者用の特徴抽出を行っている。

2.2 学習データ

(a) TED 講演データ

TEDLIUMv3[20]は、約 450 時間の TED 講演音声のコーパスである。学習データ、検証データ、評価データでセットが分かれており、それぞれ重複する講演は含まれていない。

(b) 大量読み上げ英語音声データ Librispeech

Librispeech[21]は約 1000 時間の読み上げ英語音声のコーパスである。学習データ、検証データ、評価データでセットが分かれており、それぞれ重複する発話は含まれていない。このうち学習データは約 960 時間である。

TED 講演は話し言葉であるのに対し、Librispeech は読み上げコーパスであるが大量のコーパスであることから、これらのコーパスを合わせて用いることで、より頑健なモデル構築が期待できる。

2.3 テストデータ

音声認識結果の算出には、表 1 に示す IWSLT-2016 の 10 講演と TEDLiumv3 のテストセット 16 講演を使用し、それぞれのセットで認識率を求めた。学習データとテストデータの重複を防ぐためにこの計 26 講演は学習データから除外している。

表 1 : テストデータの詳細

テストセット	TED IWSLT2016	TEDLIUMv3 テストセット
講演数	10 講演	16 講演
総文数	1119 文	2582 文
総単語数	17265 語	45446 語
総時間長	2 時間 2 分 55 秒	5 時間 1 分 16 秒

2.4 TED の音声認識結果と後処理

以下(a), (b),(c)の三通りの手法でモデルを作成し、音声認識を行った際の認識率を比較した。

(a) **TED 講演音声による単独モデル**

TED 講演 450 時間の学習データで DNN-HMM を学習した。言語モデルは TED 講演 450 時間のテキストから作成し語彙サイズは 15 万語程である。

(b) **Librispeech による単独モデル**

Librispeech960 時間の学習データで DNN-HMM を学習した。このモデルのみ、DNN の規模を、層数 17 層、隠れ層のユニット数は 1536 とした。言語モデルは Librispeech960 時間のテキストから作成し、語彙サイズは 20 万語程である。

(c) TED 講演音声と Librispeech の混合学習によるモデル

Librispeech960 時間の学習データと TED 講演 450 時間の計 1410 時間を用いて DNN-HMM を学習した。言語モデルは TED 講演 450 時間と

Librispeech960 時間を合わせたテキストから作成し、語彙サイズは 28 万語程である。

表 2 に各モデルでの認識率を示す。評価として以下の式で定義される単語正解精度を用いた。

$$\text{単語正解精度} = 100 - \frac{\text{置換誤り数} + \text{挿入誤り数} + \text{脱落誤り数}}{\text{発声された入力文中の単語数}} (1)$$

表 2 に示すように、Librispeech 単独モデルでは TED 講演のような話し言葉の認識は難しいことが分かる。また、TED+Librispeech 混合モデルでの認識率は TED 単独モデルよりやや良いことが分かる。これは、Librispeech が朗読音声であるものの 960 時間と大量データであることによる効果である。Librispeech 単独モデルから TED データによる適応を行うとさらに良くなると思われる。なお、(b)のモデルを用いた Librispeech テストデータでの認識実験では単語正解精度は 94.78%であった。

以降、TED2016 に対して一番認識率が高かった TED 単独モデルを用いて音声翻訳と音声要約を行う。このために、音声認識結果の数値の読み表現を数値表現に後処理した(例：thirteen→13)。なお、認識結果には句読点はないので、音声翻訳・音声要約の評価時の参照テキストも句読点を除去した。

表 2： 各モデルでの単語正解精度 [%]

モデル	TED2016	TEDLIUMv3
(a)TED 講演単独モデル	87.61	88.30
(b)Librispeech 単独モデル	79.11	82.02
(c)TED+Librispeech モデル	87.55	89.24

3 TED の音声翻訳

3.1 Transformer 翻訳モデル[22]

Transformer モデルはエンコーダとデコーダからなり、再帰型ニューラルネットワークの様に時系列データを用いて学習を行う。出力を求める際は自己注意機構を用いる。エンコーダは同じ構成のエンコーダの積み重ねによって構成されており、それらのエンコーダは自己注意とフィードフォワードネットワーク(FFNN)により構成される。デコーダも同じ構成のデコーダが積み重ねられてできている。一つのデコーダは自己注意と FFNN に加えてその間に注意機構が入っている。標準モデルはエンコーダ 6 層、デコーダ 6 層であるが、最適な数は学習データ量に依存する。22 万文の英日パラレルコーパスで学習した場合は、6 層－6 層で 0.80、3 層－4 層で 13.66、3 層－3 層で 13.22 の BLEU 値であった。本実験ではエンコーダとデコーダは共に 3 層とした。

本実験で使用する TED 講演の IWSLT (International Workshop on Spoken Language Translation)のコーパスには英語と日本語の対訳コーパスが 22 万文と少ないため、英語または日本語の単言語コーパスをベースモデルで翻訳/逆翻訳することで英語と日本語の疑似対訳コーパスを作成し、翻訳モデル学習の追加の学習データとすることでデータ拡張を行う。データ拡張には IWSLT2018 英語-スペイン語ペアの英語側コーパスと CSJ (Corpus of Spontaneous Japanese)日本語コーパスの模擬講演を用いた。

100 万文ペアからなる ASPEC (Asian Scientific Paper Excerpt Corpus)コーパスにより英日単/双方向翻訳モデルを学習し、そのパラメータを初期値として IWSLT+データ拡張したデータセットで学習した[8]。なお、ASPEC コーパスによる翻訳モデル学習の際には、TED 語彙に合わせて学習を行った。

3.2 音声翻訳結果

音声翻訳結果の BLEU 値を表 3 に示す。翻訳文の評価には BLUE 値を用いた。音声翻訳はテキスト翻訳と比べて、約 14%BLEU 値が低下した。10%程度の低下に留まるためには、音声認識精度が 90%以上必要と思われる。

表 3： 音声翻訳結果　(TED2016 テストセット)

モデル / 入力	Transformer 単方向モデル	Transformer 双方向モデル
テキスト	15.14	15.52
音声	12.44	13.35

4 TED の音声要約

4.1 MMR+BERT[18]

MMR は、ドキュメント全体との関連度と、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[13][23]。本稿では以下で定義されているものを使用する。最新のものでは文献[24]がある。

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Si を、単語の出現頻度(Term Frequency)に基づいて求め、文のベクトルの集合$S_{nrk} = \{S_1, S_2, \cdots, S_N\}$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。

$$S_{max} = \underset{Si \in S_{nrk}}{\text{argmax}}\{\lambda(Sim(Si, D)) - (1-\lambda)(Sim(Si, S_{rk}))\} \quad (2)$$

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[25]を使用し、本実験ではλ = 0.5 と設定する。

我々は、文脈を学習させた事前学習(Pre-Training)された BERT モデルを利用し、文の分散表現ベクトルを TF に基づくベクトルの代わりに使用する。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)の英語 1024 次元の分散表現を抽出した。また、BERT ベクトルを入力として階層型ニューラルネットワークによる重要文/非重要文識別器を構成し、(3)式のように、その出力の事後確率 p(重要文|入力)の対数値を(2)式に加えた。これを MMR+BERT と呼ぶ。識別器の学習には数千文程度あればよい。なお、同じく MMR+BERT と呼ぶ手法が文献[24]に紹介されているが詳細は不明である。

$$Smax = \underset{S_i \in Snrk}{\text{argmax}}\{\lambda(Sim(S_i, D)) - (1 - \lambda)(Sim(S_i, S_{rk})) + \alpha \times log\ p(\text{重要文}|S_i)\} \quad (3)$$

4.2 BertSumExt[16]

BertSumExt 法は、複数文の文境界記号付き入力単語列（512 トークン以下の制限あり）に対して BERT によって文ベクトル列 $\{T_i\}$ を得て、これを Transformer に入力して、重要文/非重要文のラベル列を抽出する手法である。この手法は標準的な要約タスクである CNN/Daily Mail の重要文抽出要約に適用され、このタスクで標準的な性能を得ており、ベースシステムとしてよく使用されている。MMR と同様に類似な文の抽出を避けるために trigram blocking という機能を取り入れている。

本手法を TED 講演に利用するために、最大の入力長が 512 トークンという制限から、講演を 512 トークン以下に分割し、それぞれで重要文抽出を行い、最後に講演全体で設定された重要文数なるようにスコアで選定した。また、オリジナルな BertSumExt は、CNN/Daily news 約 30 万記事の 1000 万文から、各記事から 3 文の重要文を抽出するように学習されており、このモデル（**CNN**）と、MRR+BERT と同じく TED の 51 講演の約 4500 文から学習したモデル(**TED**)、CNN のモデルから TED データで適応したモデル(**CNN/TED**)を使用した。

4.3 音声要約結果

(a) テキスト要約

TED10 講演のテキストデータで行った要約実験の結果を表 4 に示す。重要文の割合は約 46%である。TED データだけの学習では MMR+BERT が最も優れているが、BertSumExt の CNN のモデルから TED データで適応した CNN/TED は、これを上回っている。

(b) 音声要約

BertSumExt のみで音声入力の要約実験を行った。講演音声の音声認識結果に対する要約結果を表 4 に示す。音声要約では抽出された文の認識結果に対応する原テキストデータを用いての評価も行い BLEU 値欄の一番右欄に示す。これは重要文箇所を音声で再生した場合は音声認識の誤りに関係なくテキスト内容が正しく聴講できるための尺度である。テキスト入力と比べて音声入力では Rouge-3 で約 20%の性能低下がみられた。しかし、重要文箇所の抽出の尺度では、低下はみられなかった。

表 4：TED 講演の要約結果 (ROUGE-3)

手法	テキスト入力	音声入力	
		認識結果	原文に変換
Lead	50.2		
MMR	40.4		
MMR+BERT	54.8		
BertSum-CNN	53.8	42.9	55.1
BertSum-TED	47.8	35.3	47.5
BertSum-CNN/TED	62.1	48.2	62.1

5 まとめ

TED 英語講演音声の日本語字幕化のための音声認識、音声翻訳、音声要約の結果を述べた。音声認識では、DNN-HMM により約 88%の単語正解精度を得た。この認識結果を Transformer に基づくテキスト翻訳と BERT に基づく重要文抽出型テキスト要約に使用した。音声翻訳はテキスト翻訳に比べて、BLEU 値が約 14%低下した。一方、音声要約はテキスト要約に比べて、重要文として抽出された区間はほぼ同一であり、音声認識誤りに対して頑健であった。

謝辞

本研究の一部は、JSPS 科研費 18H01062、19K12027、22K12084 の研究助成を受けた。

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] A. Anastasopoulos, O. Bojar, et.al : Findings of the IWSLT 2021 evaluation campaign, Proc. IWSLT-2021, pp.1-29, 2021.

[3] A. Anastasopoulos, L. Barrault, et.al : Findings of the IWSLT 2022 evaluation campaign, Proc. IWSLT-2022, pp. 98-158, 2022.

[4] 上野, 李, 河原 : 音声認識のデータ拡張のための話者情報およびマスクを用いた合成音声の周波数スペクトログラム強調, 2022 年日本音響学会秋季講演論文集, pp.1149-1150, 2022.

[5] W. Zhang, Z. Ye, et al : The USTC-NELSLIP offline speech translation systems for IWSLT 2022, Proc. IWSLT, pp.198-207, 2022.

[6] 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, 言語処理学会第 23 回年次大会, pp.1054-1057, 2017.

[7] 佐橋, 秋葉, 中川 : 科学技術論文抄録と講義の英日機械翻訳のリスコアリングの検討, 言語処理学会第 25 回年次大会, pp.1165-1168, 2019.

[8] 足立, 山本, 中川 : TED 講演の英日翻訳と日英翻訳の検討, 言語処理学会第 29 回年次大会, 2023.

[9] R. Fukuda, Y. Ko, Y. Kano, et al : NAIST simultaneous speech-to-text translation for IWSLT 2022, Proc. IWSLT 2022, pp.286-292 , 2022.

[10] T. Nguyen, A. T. Luu, T. Lu, T. Quan : Enriching and controlling global semantics for text suumarization, arXiv:2109.10616v1, 2021.

[11] A. See, P. J. Liu, C. D. Manning: Go to the point summarization with pointer-generative networks, arXiv:1704.04368, 2017.

[12] O. Ernst, A. Caciularu, et. al : Proposition-level clustering for multi-document summarization, Proc. NACACL, pp.1765-1779, 2022.

[13] J. Carbonell, J. Goldstein : The use of MMR, diversity-based reranking for reordering documents and producing summaries. Proc. ACM SIGIR, pp. 335–336, 1998.

[14] 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.

[15] 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, CVSC, pp. 31–39, 2014.

[16] Yang Liu, Mirella Lapata ; Text summarization with pretrained encoders, Proc. conf. EMNLP, pp. 3730-3740, 2019.

[17] J. Devlin, M. Chang, K. Lee, K. Toutanova : BERT pre-training of deep bidirectional transformers for language understanding, arXiv:1810.04805, 2018.

[18] K. Masuda, Y. Hayakawa, K. Yamamoto, S. Nakagawa : Summarization of spoken lectures based on MMR method and important/unimportant sentence classification using BERT, Proc. GCCE, OS-SLP, 2022.

[19] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Cchwarz, J. Silovski, G. Stemmer and K. Veseh : The Kaldi speech recognition toolkit, Proc. ASRU, 2011.

[20] OpenSLR : TED-LIUM corpus release3, https://www.openslr.org/51/, 2022.

[21] V. Panayotov, G. Chen, D. Povey and S. Khudanpur : Librispeech an ASR corpus based on public domain audio books, 2015 IEEE International Conference on Acoustics, Speech and Signal Processing(ICASSP), pp.5206-5210, 2015.

[22] I. Polosukhin, L. Kaiser : Attention is all you need, Conference on Neural Information Processing Systems, 2017.

[23] G. Murray, S. Renal, and J. Carletta : Extractive summarization of meeting recordings, Proc. Interspeech, pp. 593–596, 2005.

[24] X. Liang, J. Li, S. Wu, M. Li, Z. Li : Improving unsupervised extractive summarization by jointly modeling facet and redundancy, IEEE Trans. ASLP, Vol.30, pp.1546-1556, 2022.

[25] GitHub : Text-Summarization, https://github.com/fajri91/Text-Summarization-MMR, 2023.

A 付録

表：TED 講演音声の音声認識と翻訳の例

例 1	原文	But augmented reality is not just an enhanced playbook.
	音声認識結果	but augmented reality is not just in the hands playbook
	日本語参照文	しかし 拡張 現実 は フォーメーション を 覚え 易く する だけ で は あり ませ ん
	テキスト翻訳結果	しかし 拡張 現実 は 単なる 優れ た 戦略 で は あり ませ ん
	音声翻訳結果	しかし 拡張 現実 は 単なる 戦略 上 の 戦略 上 の 手 だけ で は あり ませ ん
例 2	原文	and it got to the point where every time I walked up to a table that had a kid anywhere between three and 10 years old , I was ready to fight .
	音声認識結果	and we got the point that every time I walked up to a table that had a kid anywhere when three in 10 years old I was ready to fight
	日本語参照文	その うち 3～10 歳 の 子ども が いる ― テーブル へ は 戦闘 態勢 で 臨む よう に なり まし た
	テキスト翻訳結果	そして 子供 が 3 歳 から 10 歳 の 間 に 子供 が いる テーブル に 上がる たび に 喧嘩 を する 準備 が 整い まし た
	音声翻訳結果	そして 私 が 子供 を 持つ テーブル に 上がる たび に 10 歳 の 時 に 喧嘩 を する 準備 が 整い まし た
例 3	原文	it means that a stadium in Brooklyn can be a stadium in Brooklyn , not some red-brick historical pastiche of what we think a stadium ought to be .
	音声認識結果	it means that a stadium in Brooklyn can be a stadium in Brooklyn not some pastiche of what we think a stadium ought to be
	日本語参照文	ブルックリン の 競技 場 は ブルックリン の 競技 場 で あり 赤 レンガ 造り の 歴史 の 寄せ集め の よう な 私 たち が 考える 競技 場 の 姿 で は ない の です
	テキスト翻訳結果	つまり ブルックリン の スタジアム は 歴史 的 な 模倣 版 で は なく ブルックリン の スタジアム に なる こと が できる と いう こと です
	音声翻訳結果	つまり ブルックリン の スタジアム は 私 たち が 思っ て いる こと を 模倣 し て い ない ブルックリン の スタジアム に なる べき で は ない と いう こと です