Vol.2019-NL-240 No.15
2019/6/14

授業アーカイブの翻訳字幕自動作成システムの試作

須藤 克仁[1,a)]　林 輝昭[1]　西村 優汰[1,†1]　中村 哲[1]

概要：大学・大学院の国際化に伴い英語での授業は増加しているが，依然として多くの授業は日本語でのみ開講されており，日本語を解さない学生が受講できない授業が多く存在するのが実情である．この問題に対応すべく，我々は日本語で行われる授業の音声認識と日英機械翻訳によって録画された授業映像に付与する英語の字幕を自動作成するシステムの開発を行っている．本稿では本開発プロジェクトで構築しているコーパス，システムの構成および要素技術，試作システムにおける予備実験の結果を報告し，今後の展望について述べる．

1. はじめに

我が国での高等教育のグローバル化は徐々に進みつつあり，留学生の受け入れが加速するとともに，英語による教育・指導の実施が広がっている．それに伴い英語で実施される大学・大学院での授業も増加しているが，依然として多くの授業は日本語でのみ開講されており，日本語を解さない学生が実質的に受講できない授業が多く存在するのが実情である．奈良先端大においても全学生の 2 割前後は海外からの留学生であり，基礎科目から専門科目まで多くの英語での授業が行われているものの，比較的比率の高い情報科学系専門科目でおよそ半数に留まっている．大学や大学院等の専門教育に関しては全面的に英語化するべきという意見もある一方で，日本人学生が母語で学ぶ機会を提供するという側面もあり，大学や分野による違いこそあれ，今後も日本語開講科目・英語開講科目が並立する状況は続くと考えられる．

他方，授業映像をインターネットを通じて公開するオープンコースウェアの提供が我が国を含む各国の大学で行われている．奈良先端大では 2004 年度から授業アーカイブとして講義映像・音声の収録と学内公開を開始し，2008 年度からは一部の授業について学外への公開も行っている．[*1] このように蓄積された授業アーカイブは教育・学習のための有用な資源であり，今後さらなる活用が期待される．我々は，先に述べた観点から，日本語を解さない学生の学習の補助のために，アーカイブされた日本語の講義映像に英語の翻訳字幕を自動的に付すためのプロジェクトを開始し，2018 年度までにシステムの試作を行った．本稿では，システム試作に利用したコーパス，システムの構成および要素技術を説明し，試作システムにおける予備実験の結果を報告する．また，最後に本システムにおける課題と今後の展望を述べる．

[1] 奈良先端科学技術大学院大学
NAIST (Nara Institute of Science and Technology), Ikoma, Nara 630–0192, Japan
[†1] 現在，株式会社プレイド
Presently with PLAID, Inc.
[a)] sudoh@is.naist.jp

[*1] http://library.naist.jp/library/archive_top/index-j.html

2. 授業アーカイブシステム

奈良先端大の授業アーカイブシステムでは，収録の許諾が得られた授業や講演会について，学内 6 箇所の講義室・ホールに設置された機器および可搬型機器で収録された講義映像・音声を蓄積し，授業担当教員の許諾の得られた範囲（学内限定もしくは一般向け）で公開している．公開されているものについては PC もしくはスマートフォンの Web ブラウザで閲覧が可能である．図 1 に PC 上の Web ブラウザでの表示例を示す（本稿のシステムによって自動作成された日本語および英語字幕を付与している）．図にもある通り，授業時に映写しているスライド等がある場合は，カメラでの撮影映像と時刻同期して表示されるようになっている．

3. 翻訳字幕自動作成システムの試作

授業映像への字幕の重畳表示が可能なアーカイブシステムの運用が 2018 年に開始されるのに合わせ，音声認識や機械翻訳の技術を利用して翻訳字幕を自動作成するシステムの試作を開始した．まず日本語の講義に対して英語の字幕を付与すべく，日本語の音声認識と日本語から英語への機械翻訳を行う処理エンジンを作成した．本節では 2018

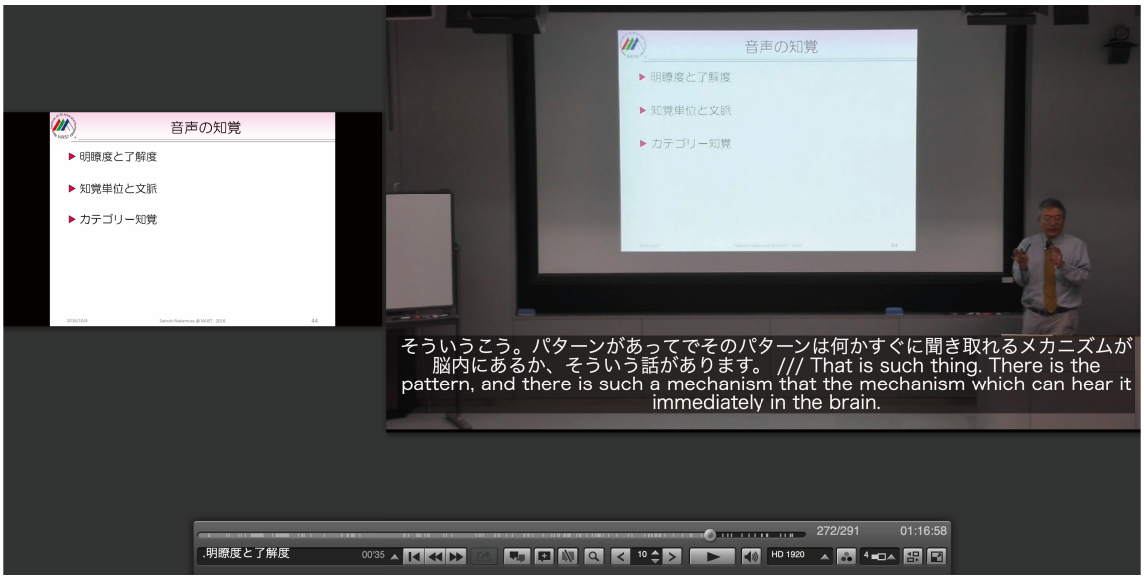

図 1 自動作成された日本語・英語字幕を付与した授業アーカイブ講義映像

年度末時点での試作システムについて述べる．

字幕の付与はアーカイブシステムに時刻情報と字幕テキストを登録する形で行う．現在は授業終了後に手動で動画ファイル (MP4) をダウンロードし，音声認識・機械翻訳を行って得られる字幕情報を手動でアップロードして登録している．理想的には映像・音声の収録と同時に処理する形態も可能であるとは考えられるが，収録後にオフラインで字幕作成を行い必要な処理時間を投じる形態のほうがアーカイブシステムにおいては適切であろう．複数同時収録をしている授業への字幕付与処理を同時に行うのではなく，授業のない夜間等の時間帯も活用して順次行う方式を想定している．

3.1 システム構成

本試作システムを含む授業アーカイブシステムの構成概略を図 2 に示す．

各講義室で収録された講義映像は講義映像蓄積サーバに伝送・保存され，配信サーバを通じて学内および学外に公開される．本施策システムは，講義映像蓄積サーバから講義映像ファイルを取得し，前処理として音声部分の取り出しを行ったのち，音声認識，機械翻訳の処理によって英語の文字列へと変換し，後処理として音声認識による発話時刻情報と機械翻訳結果の英語文字列を統合・整形した字幕登録用のファイルを作成して，最終的に講義映像蓄積サーバ上の当該講義の字幕として登録を行う．

現在の字幕重畳表示は言語の切り替えに対応していないため，図 1 のように日本語と英語を連結した字幕として登録・表示させている．

3.2 コーパスの収集

本試作システムに必要な音声認識や機械翻訳の学習に適

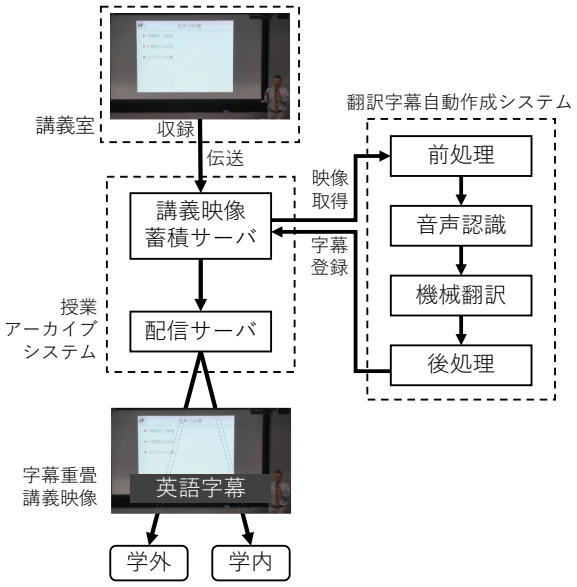

図 2 翻訳字幕自動作成システムを含む授業アーカイブシステムの構成概略

した既存のコーパスは非常に限定されており，特に機械翻訳に関しては講義の対訳コーパスはほぼ存在しないことから，学習・開発・評価のためのコーパス収集を進めている．

2018 年度末までに日本語で行われた 90 分の授業 35 講義分，約 46.5 時間分[*2]について，書き起こしと英語への翻訳を行った．書き起こしにおいてはフィラーやショートポーズ，言い誤り等のラベルを付与した．なお，書き起こしに各発話の時刻情報が付与されているものは 7 講義分のみである．発話単位は書き起こしにおいて句点で区切られる単位とし，英語への翻訳はおおよそ発話を単位として行ったが，翻訳者により発話単位での翻訳が困難であるとされた場合については単位の分割・併合を許容することとした．

*2 収録の不具合で部分的に収録ができていない講義が存在する．

コーパス収集の対象とした授業は，2014 年度に行われたソフトウェア工学，ロボティクス，自然言語処理，音情報処理，ビッグデータ解析等の情報科学分野の授業である．

3.3 音声認識部

音声認識エンジンには Kaldi [1][*3] を利用した．音響モデルは日本語話し言葉コーパス (CSJ) の音声データ (240 時間) を CSJ レシピにより作成した DNN-HMM [2] である．言語モデルは単語 3-gram で，句読点を付加した日本語話し言葉コーパス (20 万文)，ATR コーパス (57 万文)，Web 収集コーパス (10 億文) およびパープレキシティに基づいて選択された過去 3 年間の授業アーカイブ認識結果テキスト (52 万文) から学習した．認識語彙サイズは 26.4 万とした．

3.4 機械翻訳部

機械翻訳エンジンについては Luong らの注視機構つきニューラル機械翻訳の手法 [3] で，注視計算に内積を用いたものを PyTorch[*4]で実装し利用した．エンコーダは双方向エンコーディングを行う LSTM 1 層，デコーダは LSTM 1 層で構成し，単語埋め込みベクトルおよび隠れベクトルの次元数は 512 とした．学習時のドロップアウト率は 0.5，最適化には初期学習率 0.001 の Adam を利用し，対数尤度を目的関数とした．ミニバッチサイズは 64 文とした．また，SentencePiece [4][*5]によるサブワード化を行った．サブワード語彙は日英で共有し，語彙サイズは 16,000 とした．

なお，講義に関する日英対訳コーパスの不足を補うために，論文抄録対訳コーパス ASPEC [5][*6]による事前学習（エポック数 30）を行い，内製講義コーパスで追加学習（エポック数 100）を行った．

4. 予備実験結果

試作システムの音声認識・機械翻訳それぞれの性能評価のための予備実験を行った[*7]．予備実験における評価には，テストセットとして用意した 3 講義（ロボティクス，音情報処理，ソフトウェア工学）を利用した．

音声認識の評価は，3 講義の音声認識結果の，書き起こしに対する単語誤り率により行った（表記揺れの吸収はルールベースで行った）．また，用語やスタイルに対するカバレージの影響を測るために，Kaldi の言語モデルの学習にテストセットの講義の書き起こしを加えたモデルでも合わせて検証を行った．結果を表 1 に示す．

[*3] https://kaldi-asr.org/
[*4] https://pytorch.org/
[*5] https://github.com/google/sentencepiece/
[*6] http://lotus.kuee.kyoto-u.ac.jp/ASPEC/
[*7] なお，今回は音声認識において同定される発話区間と評価データにおける書き起こし・翻訳の発話区間が一致しないことから，両処理を行った上でのシステム性能評価を行う実験を実施しなかった．当該検証実験の実施については今後の課題としたい．

表 1 評価データの各講義に対する試作システムの音声認識単語誤り率 (WER)．$\mathrm{WER}_{\mathrm{closed}}$ は言語モデルの学習にテストセットの書き起こしを含めた場合の単語誤り率を示す．

講義名	WER (%)	$\mathrm{WER}_{\mathrm{closed}}$ (%)
ロボティクス	12.48	8.25
音情報処理	12.56	6.65
ソフトウェア工学	17.76	13.31

表 2 評価データの各講義における発話数および言い誤り・言い直しの数．

講義名	発話数	言い誤り・言い直し数
ロボティクス	352	277
音情報処理	458	122
ソフトウェア工学	234	509

講義により単語誤り率の大きな違いが見られるが，1 つの理由として言い誤り、言い直し数の違いが考えられる。表 2 に言い誤り，言い直し数の統計を示す．最も単語誤り率が大きいソフトウェア工学では，音情報処理の 4 倍の言い誤り，言い直しがあり，音声認識に影響を与えた可能性が高い．また，発話数の少なさは 1 発話あたりの継続時間が長いことを示唆しており，これも音声認識に影響を与えた可能性がある．なお，言語モデルを closed にした場合には単語誤り率が 4-6 ポイントも減少していることから，現在はコーパスの不足ゆえの用語や発話スタイルに対するカバレージの不足が示唆される．

機械翻訳の評価は，3 講義から計 500 発話分をテストセットとして抽出したものに対する BLEU [6] により行った．解探索時のビーム幅は開発セットにおける BLEU を最大にするものとして 10 を選択した．BLEU の計算には SacreBLEU [7][*8] を利用した．結果を表 3 に示す．

表には，BLEU (4-gram までを使う BLEU-4) と合わせて簡潔ペナルティ (Brevity Penalty，以下 BP) の値を示してある．BP は参照訳の単語数に対する翻訳結果の単語数の割合を示しており，本予備実験では参照訳よりも 1 割程度短い文を翻訳結果として出力していることが分かる．ニューラル機械翻訳では訳抜けや湧き出しがしばしば生じることが知られており，予備実験結果においても句の訳抜け等で全体的な訳出が短くなる傾向が確認できた．ビーム幅を変化させた実験結果の比較においては，ビーム幅が 10 より小さい場合はビーム幅の拡大に伴い訳文が短くなり BP が強く働く（BP の値が小さくなる）が単語 n-gram 精度の向上により BLEU が向上するが，10 より大きくなると単語 n-gram 精度が向上しなくなり BLEU が低下するという傾向が見られ，点予測に基づく系列モデリングにおける長さバイアスの影響と考えられる．

[*8] https://github.com/mjpost/sacreBLEU/

表 3　評価データの各講義の書き起こしに対する機械翻訳の BLEU と簡潔ペナルティ (BP).

講義名	BLEU (%)	BP (%)
ロボティクス (232 発話)	15.9	90.8
音情報処理 (138 発話)	21.1	89.7
ソフトウェア工学 (130 発話)	12.8	87.2

5. 課題と今後の展望

予備実験で利用した試作システムによって新しく収録された授業について音声認識・機械翻訳が行えるようになったため，試作システムのデモンストレーションのために，いくつかの授業について字幕の作成を行った．書き起こしや参照訳の作成を現在順次進めているために定量的な評価は行えていないが，以下のような課題が見つかっている．

- 言い淀み，言い直し，発声の乱れ等で日本語入力文に乱れが生じた場合に，大きな訳抜けや不必要な句の繰り返しが現れることが多い．
- 固有表現や専門用語に対して音声認識・機械翻訳ともカバレージが不足している．
- 講義の発話スタイルへの適応が十分でない．

こうした問題についてはドメイン適応・データ拡張等の活用が期待され，授業アーカイブコーパスの構築と合わせ外部コーパスの活用方法について引き続き検討を続ける予定である．また，現在は情報科学分野の科目のみを対象としたが，今後は他分野の講義についてもデータ収集や適用可能性検証を行う予定である．奈良先端大ではバイオサイエンス分野，物質科学分野を含めた融合領域の教育・研究を推進していることもあり様々な分野に対応することの意義が大きいとともに，技術的にも重要な挑戦と言える．さらに，英語開講の講義についても同様のアプローチによる日本語への翻訳が可能であると考えられ，非母語話者による講義への対応のための音声認識や文法誤り訂正といった技術の適用についても検討したい．

6. おわりに

本稿では，奈良先端大で開発している授業アーカイブの翻訳字幕自動作成試作システムを紹介した．大学院の講義における専門的な内容を扱うための音声・言語リソース，特に発話スタイルも含めた対象ドメインのコーパスを大量に確保することは困難であり，現在のベースライン技術を適用するのみでは実用的な性能を得るには至っていない．今後はデータの不足を補う技術の利活用と，本システムにおける実践的な課題への対応を推進したい．

謝辞　本研究の一部は JSPS 科研費 JP17H06101 の助成を受けたものである．

参考文献

[1] Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., Schwarz, P., Silovsky, J., Stemmer, G. and Vesely, K.: The Kaldi Speech Recognition Toolkit, *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*, IEEE Signal Processing Society (2011).

[2] Moriya, T., Tanaka, T., Shinozaki, T., Watanabe, S. and Duh, K.: Automation of system building for state-of-the-art large vocabulary speech recognition using evolution strategy, *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, pp. 610–616 (online), DOI: 10.1109/ASRU.2015.7404852 (2015).

[3] Luong, T., Pham, H. and Manning, C. D.: Effective Approaches to Attention-based Neural Machine Translation, *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, Lisbon, Portugal, Association for Computational Linguistics, pp. 1412–1421 (online), DOI: 10.18653/v1/D15-1166 (2015).

[4] Kudo, T. and Richardson, J.: SentencePiece: A simple and language independent subword tokenizer and detokenizer for Neural Text Processing, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, Brussels, Belgium, Association for Computational Linguistics, pp. 66–71 (online), available from ⟨https://www.aclweb.org/anthology/D18-2012⟩ (2018).

[5] Nakazawa, T., Yaguchi, M., Uchimoto, K., Utiyama, M., Sumita, E., Kurohashi, S. and Isahara, H.: ASPEC: Asian Scientific Paper Excerpt Corpus, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC 2016)* (Chair), N. C. C., Choukri, K., Declerck, T., Grobelnik, M., Maegaard, B., Mariani, J., Moreno, A., Odijk, J. and Piperidis, S., eds.), Portoro, Slovenia, European Language Resources Association (ELRA), pp. 2204–2208 (2016).

[6] Papineni, K., Roukos, S., Ward, T. and Zhu, W.-J.: Bleu: a Method for Automatic Evaluation of Machine Translation, *Proceedings of 40th Annual Meeting of the Association for Computational Linguistics*, Philadelphia, Pennsylvania, USA, Association for Computational Linguistics, pp. 311–318 (online), DOI: 10.3115/1073083.1073135 (2002).

[7] Post, M.: A Call for Clarity in Reporting BLEU Scores, *Proceedings of the Third Conference on Machine Translation: Research Papers*, Belgium, Brussels, Association for Computational Linguistics, pp. 186–191 (online), available from ⟨https://www.aclweb.org/anthology/W18-6319⟩ (2018).