# Automatic Speech Recognition and Machine Translation System for MIT English Lectures Using MIT and TED Corpus

VERI FERDIANSYAH[†1] SEIICHI NAKAGAWA[†1]

Toyohashi University of Technology, Tenpaku-cho, Toyohashi 441-8580 Japan
E-mail: {veri,nakagawa}@slp.cs.tut.ac.jp

***Abstract:*** This paper presents our attempt to create English automatic speech recognition system (ASR) and English to Japanese machine translation (MT) system. We utilized existing Wall Street Journal corpus for our acoustic model and adapted it with MIT OpenCourseWare lectures while the transciptions of the MIT lectures are utilized to create the needed language model. For the parallel corpus of our statistical machine translation system, we created it by utilizing TED Talks transcripts. Our ASR system can achieve word error rate (WER) 32.1% WER and our MT system can achieve 7.16 BLEU.

***Keywords:*** automatic speech recognition, machine translation system, MIT lectures, TED corpus

## 1. Introduction

Openness in education has made education more accessible to anyone around the globe. Usually licensed under an open copyright license, such as Creative Commons, this type of lecture is free for anyone to access. However, most of the lectures are only available in English and usually they come with separate English transcriptions. This means that a sufficient level of English proficiency is required to understand the contents of the lectures. For students in a non-English native speaking countries like Japan, this will be a problem as there are still many university students that didn't achieve the required proficiency level.

Building an automatic system for simultaneous translation of lectures is very challenging. We are dealing with spontaneous speech, repetitions, pauses, etc. Lectures also tend to have wider topics, virtually unlimited domain, than any other speech translation tasks, such as spoken dialog system for travel assistance [1]. Various speaking styles add more complexity to the task as it is more conversational and informal than in prepared speeches or short utterances.

Combining automatic speech recognition (ASR) and machine translation (MT) system to create a speech translation system is not an easy task. For many years ASR and MT evolved independently from each other but it has been shown in the past that ASR and MT can be coupled in order to directly translate spoken utterances into another language [2]. The simplest approach for the speech translation system was just to translate single best recognizer output. A tighter coupling of ASR and MT is reached when word lattices are translated. Some improvements of translation quality were achieved by using lattices with small densities as explained in [3]. Finally, a fully integrated approach where the whole search space of ASR and MT is integrated can be pursued and this approach was successful only on very small tasks [4].

In this paper, we present our effort in building an ASR and MT system from English to Japanese by utilizing MIT OpenCourseWare lectures and TED corpus. In section 2, we will describe related works to this paper. In Sections 3 and 4, we explain in detail the speech recognition and machine translation components that we build, respectively. Section 5 presented the results of our experiments and in Section 6, a conclusion is made.

## 2. Related Works

Lecture transcription has been the target of bigger research projects like European project CHIL (Computers in the Human Communication Loop) [5] and the American iCampus Spoken Lecture Processing project [6]. Researchers from NTT Communication Science Laboratory have attempted to create a lecture transcription system by utilizing MIT OpenCourseWare lectures. They achieved results as high as 18.8% word error rate (WER) for the ASR system and 27 BLEU score for their MT system by using MIT OpenCourseWare parallel corpus [7].

In [8], the researchers attempted to create English-Spanish and English-German lecture translation system as a part of the TC-STAR project. For their ASR system, they can achieve 9.3% WER and for their final MT system, their system can achieve 18.57 BLEU and 13.22 BLEU for the English-Spanish and English-German lecture translation system respectively.

Researchers in Karlsruhe Institute of Technology (KIT) also attempted to create speech translation system by utilizing the KIT lecture corpus [9]. Their ASR system can achieve WER as low as 29.8% WER and 20.7% WER for Computer Science lectures and miscellaneous talks by the same speaker, respectively while their MT system can get 25.30 BLEU and 34.90 BLEU after being tested on the output of the ASR system and on the reference transcriptions, respectively.

Before [9], [1] already presented their progress in creating a German-English lecture translation system. Their final system has a translation performance of 23.65 BLEU on the ASR output and 29.04 BLEU on the reference transcription.

In this paper, we describe an English ASR system and an English to Japanese MT system by utilizing TED Talks parallel

†1 豊橋技術科学大学
Toyohashi University of Technology

corpus because we could not use MIT OpenCourseWare parallel corpus, unlike NTT Communication Science Laboratory in [7].

## 3. Automatic Speech Recognition

### 3.1 Acoustic model

Baseline acoustic models are speaker independent, continuous HMM models with 16 Gaussians for non-silence states and 32 for silence states. The feature vector is 39-dimensional, consists of 12 MFCCs plus the $0^{th}$ cepstral, their first- and second- order derivatives, normalized using cepstral mean subtraction. The models were trained for American English using 49,190 utterances from Wall Street Journal (WSJ) corpora. CMU pronouncing dictionary [1] containing 39 phonemes without lexical stress was used. HMMs are initialized on the basis of TIMIT phonetic transcriptions. Cross-word triphones was used. Tied-state triphones was built based on decision tree. The details of our acoustic model can be found in [10]. The system was trained with the help of the HTK toolkit [11].

### 3.2 Acoustic model adaptation

Acoustic model adaptation is a way to increase the recognition rate of ASR system. We created four different acoustic models with different unsupervised adaptation method each. We will then compare the recognition results of those acoustic models. The results of the comparison will help us to determine which adaptation method is the best for our automatic speech recognition system.

Acoustic model adaptation methods that we used in our system are: (1) maximum likelihood linear regression (MLLR) adaptation, (2) maximum a posteriori (MAP) adaptation with updating the mean components of the acoustic model, (3) MAP adaptation with updating mean, variance and mixture weight components, and (4) MAP adaptation with updating mean and variance components of the acoustic model.

For our adaptation data, we gathered more than 1600 hours of lectures and more than 200 male and female lecturers from all engineering departments in MIT OpenCourseWare website. From all these data, only approximately 700 hours of lectures that has separate transcription data. The details of our MIT OpenCourseWare corpus can be seen in Table 1. Because of time constraints, we need to reduce the size of our adaptation data. The final MIT OpenCourseWare corpus that we used for acoustic model adaptation consisted of 30 male speakers with around 300 minutes of speech in total with 1 speaker has around 10 minutes of speech. Table 2 summarizes the number of adaptation data that we used to adapt our baseline acoustic model.

These lectures were then need to be converted into audio files before we used LIUM speaker diarization tool [12] to partition the lectures into homogeneous parts. Then, we manually write the transcription of each speech.

Table 1 MIT OpenCourseWare corpus details

| | |
|---|---|
| Number of speakers: | 209 |
| Number of speakers with transcript: | 56 |
| Total speech duration: | 1600 hours |
| Total speech duration with transcript: | 700 hours |

Table 2 Adaptation data for acoustic model

| | |
|---|---|
| Number of speakers: | 30 |
| Total speech duration: | 5 hours |
| Speech duration for each speaker: | 10 minutes |

### 3.3 Language model

Our baseline language model is a trigram language model trained on the WSJ corpus between the years 1987 and 1989, which containing 36,754,891 words in 85,445 documents. ARPA's official "20o.nvp" (20k most common WSJ words with non-verbalized punctuation) was used as the vocabulary.

Our next language model is a trigram language model trained on the MIT OpenCourseWare lectures transcriptions. Using automated script, we managed to gather approximately 800 PDF files from all engineering department lectures which resulted in around 300k sentences of training data after appropriate filtering was applied and 30k most common words was used as the vocabulary. From this data, we built a trigram language model using the SRI language modeling toolkit [13].

The perplexity and OOV rate of our baseline language model for the MIT OpenCourseWare training data set was 267.9 and 4%, respectively. By using the MIT OpenCourseWare language model, the perplexity was decreased to 146.1 and the OOV rate to 0.57%.

### 3.4 Language model adaptation

To achieve better ASR accuracy rate, we need to adapt our language model. Our adaptation method was to combine the dictionary of the WSJ corpus and the dictionary of the MIT OpenCourseWare corpus. This resulted in a 40k words dictionary. This dictionary was then trained on the MIT OpenCourseWare training data set to make a trigram language model. This new language model further decreased the perplexity to 109.3 and OOV rate to 0.5%. Summary of each language model's perplexity and OOV rate is shown in Table 3.

Table 3 Perplexities (PPL) and OOV-rates of the language models (LM)

| LM | PPL | OOV (%) |
|---|---|---|
| WSJ | 267.9 | 4.00 |
| MIT | 146.1 | 0.57 |
| WSJ+MIT | 109.3 | 0.50 |

[1] http://www.speech.cs.cmu.edu/cgi-bin/cmudict

## 4. Machine Translation

For the machine translation component in our system, we used Moses [14] with GIZA++ for word alignment and SRI language modeling toolkit for language modeling.

### 4.1 Parallel corpus

To train a MT system, we need to provide parallel corpus consisting of sentence pairs of the language that we want to translate from and the target language. We collected several parallel corpuses for our MT system to compare their performance: (1) Japanese-English News Article Alignment Data (JENAAD) [15] and (2) TED Talks transcriptions.

JENAAD corpus is a newspaper corpus. The sources of JENAAD corpus is the Yomiuri Shimbun and the Daily Yomiuri which covers the period from September 1989 to December 2001. This corpus contains around 200k English and Japanese sentence pairs.

For the TED Talks corpus, we gathered TED Talks transcriptions from the TED website using automated script and we managed to gather around 800 PDF files which contains approximately 100k sentences in total of TED Talks transcriptions in English and Japanese and applied appropriate filtering to get a cleaner corpus.

To find the optimal weights for our translation model, we need to tune our MT system. The tuning set was a randomly selected 2k sentence pairs from the JENAAD corpus.

Before training and tuning the MT system, we tokenized all corpuses using built-in Moses tokenizer for English corpus and using Mecab[2] for Japanese corpus. We also lowercased all words in our English corpus.

### 4.2 Language model

To make English to Japanese translation system we need a language model in the target language, which in this case is Japanese language model. We created five different language models which will be used to compare the performance of the translation system of each language model to determine which language model that we should use for our final system: (1) JENAAD, (2) TED Talks, (3) combination of JENAAD and TED Talks, (4) Corpus of Spontaneous Japanese (CSJ) [16], and (5) combination of CSJ and TED Talks. All language models are 5-gram language model with modified Kneser-Ney smoothing.

For language model training, each language model is trained on its own corpus except for the CSJ and the combination of CSJ and TED Talks. For these two language models, we trained it on the TED Talks corpus of 100k sentences.

## 5. Experiment Results

### 5.1 Test data

Previously, we have conducted an experiment with MIT OpenCourseWare lectures to determine which type of captions are suitable for Japanese to learn lectures being taught in English [17] which resulted in important sentence captions as the most useful caption type to present captions in English lecture videos. We used the lectures in [17] as our test data for our ASR experiments and its transcriptions for our MT experiments. The lectures that we used as our test data are all from Introduction to Computer Science and Programming (Fall 2008) lecture, taught at Electrical Engineering and Computer Science Department in MIT. The full text captions consist of 200k sentences and the important sentence captions consist of 121k sentences.

### 5.2 ASR results

Our baseline system gives a WER of 53.4% WER for full text captions and 56.6% WER for important sentence captions. By changing the language model of our baseline system into adapted MIT OpenCourseWare language model, it can give significant increase in ASR accuracy to around 20% for both captions.

Out of all adaptation methods that we used, only MAP adaptation with only updating mean of the acoustic model that can decrease the WER significantly. The results of our ASR experiments can be seen in Table 4.

Table 4 Experiments—ASR Results: acoustic model (AM), language model (LM), and word error rates (WER) in %

(a) Full text captions

| AM | LM | WER |
|---|---|---|
| WSJ (baseline) | WSJ | 53.4 |
| WSJ | WSJ+MIT | 33.4 |
| MLLR | WSJ+MIT | 36.5 |
| MAP [mean] | WSJ+MIT | 32.1 |
| MAP [mean, variance, mixture weight] | WSJ+MIT | 35.8 |
| MAP [mean, variance] | WSJ+MIT | 35.0 |

(b) Important sentence captions

| AM | LM | WER |
|---|---|---|
| WSJ (baseline) | WSJ | 56.6 |
| WSJ | WSJ+MIT | 37.1 |
| MLLR | WSJ+MIT | 46.3 |
| MAP [mean] | WSJ+MIT | 37.1 |
| MAP [mean, variance, mixture weight] | WSJ+MIT | 46.0 |
| MAP [mean, variance] | WSJ+MIT | 45.0 |

### 5.3 Speech and machine translation results

Two types of machine translation experiments were conducted on the lecture data. For one, the output of the automatic speech recognition system was used as input to the MT system. Before translation, we applied text normalization and converted the texts to all lowercase. A second set of the experiments is the reference transcriptions which we have translated. We used the full text and important sentence captions transcriptions as our test set. The translation quality was presented using the case-insensitive BLEU score [18].

The results of our machine translation experiments can be

[2] http://mecab.googlecode.com/svn/trunk/mecab/doc/index.html

seen in Table 5(a) for the ASR input and in Table 5(b) for the reference transcriptions. For the ASR input, the difference in score between each language model is not very high, except for JENAAD language model. As we can see, the highest score is achieved when we used only TED language model. But when we used TED language model to translate our reference transcriptions, it can give high score for important sentence transcriptions but it fails miserably on translating full text transcriptions. For this case, using only CSJ language model or using combined CSJ and TED language model seems like the best approach. Some samples of the input speech, ASR result, and MT result are provided in Table 6.

Table 5 Experiments—MT Results: BLEU score for full text (FT) and important sentence (IS) captions

(a) ASR Input

| Japanese LM | FT | IS |
|---|---|---|
| JENAAD | 2.12 | 1.50 |
| TED | 5.88 | 7.16 |
| JENAAD + TED | 5.01 | 5.27 |
| CSJ | 5.85 | 6.62 |
| CSJ + TED | 5.68 | 6.89 |

(b) Reference transcriptions

| Japanese LM | FT | IS |
|---|---|---|
| JENAAD | 3.53 | 2.35 |
| TED | 0.00 | 6.72 |
| JENAAD + TED | 0.00 | 5.05 |
| CSJ | 5.48 | 5.86 |
| CSJ + TED | 5.60 | 5.55 |

Table 6 Experiments—Samples of ASR and MT results

| Input | ASR Result | MT Result |
|---|---|---|
| ok to work a word of warning | ok to work the word a warning | いいです仕事をするに言葉は警告 |
| i'm going to talk about statements as the key building blocks for writing code | and talk about statements is the key building blocks for writing code | についてお話しの声明重要なは組合わせるを書くのにコード |
| so let me jump straight to it | so let me jump straight to what | で見てみましょうジャンプ直線何を |

## 6. Conclusion

In this paper we presented our efforts in creating ASR and MT system for English lectures by utilizing MIT OpenCourseWare and TED Talks corpus. While our proposed ASR system can achieve moderate results of 32.1% WER, our MT system is not performing as well with only achieving 6.72 BLEU in translation results.

Our planned future work is to couple the ASR and MT system together into one integrated system.

## Reference

1) M. Kolss, M. Wolfel, F. Kraft, J. Niehues, M. Paulik, and A. Waibel: Simultaneous German-English Lecture Translation, Proc. IWSLT, 2008, pp.174-181.
2) E. Matusov, S. Kanthak, and H. Ney: Integrating Speech Recognition and Machine Translation: Where Do We Stand?, Proc. ICASSP, 2006.
3) S. Saleem, S. C. Jou, S. Vogel, and T. Schultz: Using Word Lattice Information for a Tighter Coupling in Speech Translation Systems, Proc. ICSLP, pp.41-44, 2004.
4) E. Vidal: Finite-State Speech-to-Speech Translation, Proc. ICASSP, pp.111-114, 1997.
5) L. Lamel, G. Adda, E. Bilinski, and J.L. Gauvain: Transcribing Lectures and Seminars, Proc. INTERSPEECH, 2005, pp.1657-1660.
6) J. Glass, T.J. Hazen, L. Hetherington, and C. Wang: Analysis and Processing of Lecture Audio Data: Preliminary Investigations, Proc. Human Language Technology NAACL, Speech Indexing Workshop, 2004.
7) T. Hori, K. Sudoh, H. Tsukada, and A. Nakamura: World-Wide Media Browser—Multilingual Audio-visual Content Retrieval and Browsing System, NTT Technical Review, 7(2): 1-7, 2009.
8) C. Fugen, M. Kolss, M. Paulik, S. Stuker, T. Schultz, and A. Waibel: Open Domain Speech Translation: From Seminars and Speeches to Lectures, Proc. ICASSP, 2006.
9) S. Stuker, F. Kraft, C. Mohr, T. Herrmann, E. Cho, and A. Waibel: The KIT Lecture Corpus for Speech Translation, Proc. LREC, pp.3409-3414, 2012.
10) W. Naptali, M. Tsuchiya, and S. Nakagawa: Topic-Dependent-Class-Based n-Gram Language Model, IEEE Trans. On Audio, Speech, and Language Processing, 20(5): 1513-1525, 2012.
11) S. Young, et.al: The HTK Book (for HTK Version 3.4), Cambridge University Engineering Department, 2006.
12) M. Rouvier, G. Dupuy, P.Gay, E. Khoury, T. Merlin, S. Meignier: An Open-source State-of-the-art Toolbox for Broadcast News Diarization, Proc. INTERSPEECH, 2013.
13) A. Stolcke: SRILM – An Extensible Language Modeling Toolkit, Proc. ICSLP, 2002.
14) P. Koehn, et.al: Moses: Open Source Toolkit for Statistical Machine Translation, Proc. ACL, pp.177-180, 2007.
15) M. Utiyama and H. Isahara: Reliable Measures for Aligning Japanese-English News Articles and Sentences, ACL, pp.72-79, 2003.
16) K. Maekawa, H. Kikuchi, and W. Tsukahara: Corpus of Spontaneous Japanese: Design, Annotation and XML Representation, International Symposium on Large-scale Knowledge Resources, pp.19-24, 2004.
17) V. Ferdiansyah and S. Nakagawa: Effect of Captioning Lecture Videos for Learning in Foreign Language, SIG Technical Reports, pp.1-7, 2013.
18) K. Papipeni, S. Roukos, T. Ward, and W. Zhu: Bleu: a Method for Automatic Evaluation of Machine Translation, Technical Report RC22176 (W0109-022), IBM Research Division, 2002.