Automatic Transcription of TED Talks

Welly Naptali, Tatsuya Kawahara[†1]

TED.com provides a collection of public speeches on a variety of topics related to Technology, Entertainment and Design (TED). Since 2010, International Workshop on Spoken Language Translation (IWSLT) has held an evaluation campaign on TED talks. In this paper, we describe our ASR system for TED talks in accordance with this campaign. The baseline system is trained on Broadcast News corpus. A lightly-supervised acoustic model training is introduced by retrieving a faithful transcript of TED speech from the corresponding subtitle. Three filtering methods are investigated to select the training data in this work. The resultant acoustic model is effective for improving ASR accuracy, combined with speaker normalization and adaptation techniques.

1. Introduction

TED.com provides a collection of public speeches on a variety of topics related to Technology, Entertainment and Design (TED). TED is a nonprofit organization with a mission of *spreading ideas*. As of January 3, 2012, there are $1,108$ talks of $5-25$ minutes, with more added every week. The videos have subtitles made by professional translators and approved by the speakers. The subtitles are not faithful transcripts, where disfluencies and grammatical errors are edited. Since 2010, International Workshop on Spoken Language Translation (IWSLT) has held an evaluation campaign on TED talks[1)]. In this paper, we describe our ASR system for TED talks in accordance with this campaign. The baseline system is trained with the Broadcast News corpus. A lightly-supervised acoustic model training is introduced by retrieving a faithful transcript of TED speech from the corresponding subtitle. We compare three filtering methods to select the training data for the lightly-supervised training. Finally, a new acoustic model is created based on these data and speaker normalization and adaptation to each talk is performed.

†1 Academic Center for Computing and Media Studies, Kyoto University

The rest of the paper is organized as follows. Section 2 describes the baseline system. Section 3 presents our attempt on lightly-supervised training with three methods. Section 4 presents experimental evaluations.

2. Baseline System

The baseline acoustic model was trained with 145 hours of Broadcast News 1996-1997 (HUB4)[7)] using maximum likelihood estimate (MLE) training. The model is gender-dependent with 46-hour male and 96-hour female training data. The acoustic features are 38-dimensional, comprising of 12 mel-frequency cepstral coefficients (MFCCs) with absolute energy suppressed, the first and second deviation coefficients. Cepstral mean normalization (CMN) and cepstral variance normalization (CVN) are applied. We used the CMU pronunciation dictionary[6)], containing 39 phonemes without lexical stress. HMMs are initialized on the basis of TIMIT phonetic transcriptions (5.4 hours). We also add models to represent silence, short pauses, applause, laugh, cough, lip smack, and music. Cross-word tied-state triphones are made based on the decision tree. There are 7k states with 32 Gaussians per state, except the silent state that has 64 Gaussians. Julius[2)] is used for decoding.

For language model, training data were taken from the TED talks, Europarl, News Commentary, and News Crawl from 2007 to 2011. The training dataset in total contains 2.5 billion words (see Table 1). All the training data were normalized by Non-Standard Words (NSW) tools[10)]. We used 159k vocabulary as a combination of CMU pronunciation dictionary and 100k top words of the training data with Sequitur G2P[12)] and LOGIOS[11)], resulting in the OOV rate of $0.3-0.7\%$ for the training data. We built Kneser-Ney smoothed trigram language model for each corpus using SRILM toolkit[4)], then interpolated them according to the development data.

The development and test data are the official set given by IWSLT 2010 workshop, named as dev2010 and test2010. The development data consist of 92 minutes spoken by 8 speakers or $17,509$ words with the OOV rate of 0.78%, while the test data consist of 149 minutes spoken by 11 speakers or $26,994$ words with OOV rate 0.27%. These data were manually segmented into utterances. Table 1 shows the perplexity of the language model by each training data. TALK corpus

Table 1 Language Model Perplexity

Training data	#Words	Perplexity
TALK	2,063,299	191.16
Europarl	50,023,104	453.38
News Commentary	3,880,801	447.43
News Crawl 2007	305,977,980	728.83
News Crawl 2008	759,301,946	2,042.85
News Crawl 2009	929,178,153	1,595.51
News Crawl 2010	361,016,263	1,016.58
News Crawl 2011	49,258,168	9,577.38
Linear interpolation	2,460,699,714	144.47

Table 2 Baseline WER (%)

Speaker	HUB4	HUB4-GD	HUB4-MLLR
talkid767	34.6	34.7	29.9
talkid769	35.2	30.5	27.6
talkid779	37.1	36.2	27.9
talkid783	31.4	29.1	27.8
talkid785	28.0	27.9	27.5
talkid790	40.0	37.3	33.1
talkid792	29.5	28.0	26.3
talkid799	30.5	27.7	24.5
talkid805	27.0	25.8	24.7
talkid824	40.0	39.5	28.4
talkid837	32.1	32.1	27.4
Sum/Avg	34.0	32.5	28.0

is created from the subtitle of TED talks, thus it has the best perplexity compared to the others. The worst perplexity is given by the News Crawl corpus because it has a different style and domain compared to the target. However, linearly interpolating them with the in-domain TALK data gave improvement in the perplexity, and the final perplexity is 144.47. Table 2 shows the baseline WER for each speaker of the test set. The average WER is 34.0%. Using the gender-dependent (GD) model improves the WER by 1.5% absolute. On the other side, when we perform maximum likelihood linear regression (MLLR) adapation, the average WER becomes 28.0%.

3. Lightly-Supervised Training of Acoustic Model

Lightly-supervised training[5] exploit subtitles, which are not faithful transcripts, to generate labels for acoustic model training. The most common approach is to create a bias language model that was built from the corresponding subtitle, interpolated with the background model by using a small weight for the background language model. Then, use the resulting language model to automatically transcribe the speech, and finally filter the results by aligning with the subtitle. We compare three methods for filtering the data for lightly-supervised training:

- Word-to-word matching: choose segments or sub-segments of speech in which they are aligned one-to-one with the subtitle. We used time label obtained from the forced alignment to extract the speech segment for training. To avoid unreliable data, we only retrieved segments that consist of two or more aligned words.
- WER filtering: selects segments which have lower WER than a given threshold.
- The combination of WER filtering and word-to-word matching: use the word-to-word matching method for the remaining segments that were not selected by the WER filtering method.

4. Experimental Evaluation

To improve the acoustic model, we collected 780 TED talks of $\sim$ 180 hours made before December 31, 2011, with the corresponding subtitles. The language model used in this experiment is a Kneser-Ney 4-gram based on 794 TED talks subtitles (1.9 million words) and Broadcast News 1996-1997 (1.6 million words). A specific language model was built from the subtitle of the corresponding talk, and linearly interpolated with the background model using a weight of 0.9 : 0.1. The segmentation and gender labeling of TED talks were done based on Bayesian information criterion (BIC) and hierarchical clustering using LIUM speaker diarization tools (SpkDiarization)[13].

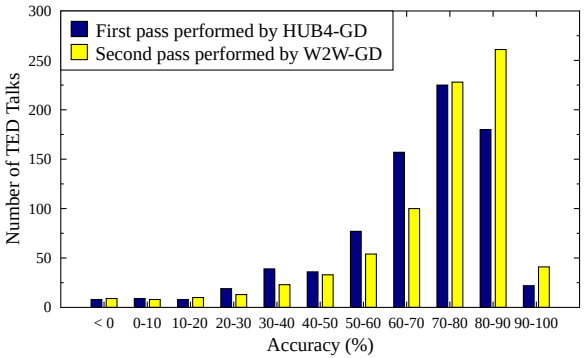

Fig. 1 Distribution of Accuracy in Lightly-Supervised Training.

Table 3 Amount of Training Data Based on WER Filtering.

WER(%) Threshold	#Segments	Duration (hours)
0	5,794	13.2
≤ 10	9,660	23.7
≤ 20	**21,707**	**55.9**
≤ 30	**32,970**	**85.4**
≤ 40	41,040	106.0
≤ 50	46,345	119.2
≤ 60	48,796	125.3
≤ 70	50,446	129.3
≤ 80	51,652	132.1
≤ 90	52,555	134.0
≤ 100	61,391	150.9

4.1 Lightly-Supervised Training

The distribution of accuracy for 780 TED talks by the bias language model is given in Figure 1. The average WER is 32.0%. The amount of data obtained by WER filtering for a given threshold is shown in Table 3. We will focus on threshold of 20% (WER20) and 30% (WER30) which retrieve 55.9 and 85.4 hours, respectively. Word-to-word matching (W2W) method and its combination (WER30-W2W) retrieved 121.9 and 135.7 hours. Using data obtained by these methods, a new acoustic model was trained. WERs of the new models are shown in Table 4. The W2W method performs better than the other two with WER of 30.0%. Adding the HUB4 data improved the WER only by 0.7% absolute. Therefore, the rest of the experiments use only TED talks with the W2W method for training acoustic model.

Table 4 Comparison Filtering Methods for Lightly-Supervised Training.

Alignment Approach	#Segments	Duration (hours)	WER (%)
WER20	21,707	55.9	33.6
WER30	32,970	85.4	32.5
W2W	**166,883**	**121.9**	**30.0**
WER30-W2W	115,759	135.7	31.1
W2W (2nd pass)	**161,232**	**129.8**	**29.6**

Table 5 WER (%) by Lightly-Supervised Training (First Pass).

Model	Corr	Sub	Del	Ins	Err
W2W	73.0	19.2	7.8	3.0	30.0
W2W-VTLN	75.3	17.2	7.4	2.8	27.5
W2W-MLLR	76.9	16.3	6.8	2.5	25.6
W2W-VTLN-MLLR	**78.7**	**15**	**6.2**	**2.4**	**23.7**

We used W2W-GD model (with WER 28.2%) to perform the second pass of lightly-supervised training. The accuracy distribution is seen in Figure 1. Compared to the first pass, the accuracy of TED talks is improved significantly, with average WER 27.9%. The improvements can also be seen from the decreasing number of segments while the amount of data is increasing using word-to-word matching. We obtain additional 7.9-hour data and WER improvements around 0.4% absolute.

4.2 Effect of Adaptation

Then, various normalization and adaptation methods are applied to the model. VTLN and MLLR adaptation significantly improve the WER. The results are given in Table 5 and 6. In the first pass of lightly-supervised training, the adaptation improves WER up to 23.7%. In the second pass, performing speaker adaptive training (SAT) and MLLR on W2W-VTLN model results in the best WER of 22.5%.

4

Table 6 WER (%) by Lightly-Supervised Training (Second Pass).

Model	Corr	Sub	Del	Ins	Err
W2W	73.6	18.8	7.6	3.2	29.6
W2W-VTLN	75.7	17.2	7.1	2.8	27.1
W2W-MLLR	77.2	15.9	6.9	2.5	25.3
W2W-VTLN-MLLR	79.4	14.5	6.0	2.4	23.0
W2W-VTLN-SAT-MLLR	**79.8**	**14.1**	**6.1**	**2.3**	**22.5**

5. Conclusions and Future Works

In this paper, we have presented our system for TED talks task. Lighly-supervised training and adaptation improves the WER from 28.0% to 22.5%. There are a lot of room to improve our system. Adaptation of acoustic and language models, discriminative training, and rescoring should be investigated. We are also planning to perform recognition on unsegmented data, as it is more realistic but challenging.

Acknowledgment

The author would like to thank Mr. Kazuhiko Abe and Dr. Chiori Hori of NICT for data sharing and Mr. Masato Mimura of Kyoto University for discussion on HTK[3] and Julius.

References

1) M. Paul et al., "Overview of the IWSLT 2010 Evaluation Campaign", Proc. of the 7th International Workshop on Spoken Language Translation (IWSLT), pp. 3-27, Paris, France, 2010.
2) A. Lee and T. Kawahara, "Recent Development of Open-Source Speech Recognition Engine Julius" Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), 2009.
3) S. J. Young et al., "The HTK Book Version 3.4", Cambridge University Press, 2006.
4) A. Stolcke et al., "SRILM at Sixteen: Update and Outlook", in Proc. IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), December 2011.
5) L. Lamel, J. Gauvain, and G. Adda, "Investigating lightly-supervised acoustic model training", In Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), volume 1, pp.477–480, 2001.
6) R.L. Wide (1998, 10/8/2011), "The CMU Pronunciation Dictionary", http://www.speech.cs.cmu.edu/cgibin/cmudict.
7) Fiscus J et al., 1997 English broadcast news speech (HUB4), catalog nbr LDC98S71, Linguistic Data Consortium, Philadelphia, PA, 1998.
8) A. R. Aminzadeh et al., "The MIT-LL/AFRL IWSLT2011 MT System", Proc. of the 8th International Workshop on Spoken Language Translation (IWSLT), USA, San Francisco, 2011.
9) N. Ruiz et al., "FBK @ IWSLT 2011", Proc. of the 8th International Workshop on Spoken Language Translation (IWSLT), USA, San Francisco, 2011.
10) A. W. Black et al. (1999, 10/8/2011), "Non-Standard Words", http://festvox.org/nsw.
11) Thomas Harris et al., (2008, 10/8/2011), "LOGIOS Lexicon Tool", http://www.speech.cs.cmu.edu/tools/lextool.html.
12) M. Bisani and H. Ney., "Joint-Sequence Models for Grapheme-to-Phoneme Conversion", Speech Communication, vol. 50, Issue 5, pp. 434-451, May 2008.
13) S. Meignier and T. Merlin, LIUM SpkDiarization: An Open Source Toolkit For Diarization, In: CMU SPUD Workshop, March, Dallas (Texas, USA), 2010.