INTERSPEECH 2019: Show & Tell Contribution
September 15–19, 2019, Graz, Austria

Multimedia Simultaneous Translation System for Minority Language Communication with Mandarin

Shen Huang[1], *Bojie Hu*[1], *Shan Huang*[1], *Pengfei Hu*[1], *Jian Kang*[1], *Zhiqiang Lv*[1], *Jinghao Yan*[1], *Qi Ju*[1], *Shiyin Kang*[2], *Deyi Tuo*[2], *Guangzhi Li*[2], *Nurmemet Yolwas*[3]

[1]Tencent Minority-Mandarin Translation
[2]Tencent AI LAB
[3]Xin Jiang University

springhuang@tencent.com

Abstract

Speech recognition for minority language is always behind main stream due to lack of resources. This work presents a system for simultaneous translation between Mandarin and major minority languages such as Uyghur, Tibetan in shape of speech, text and images. The general acoustic model is trained via factorized TDNN with lattice free MMI criteria using mixed-units based lexicon model. For each specific language, acoustic model is trained by multi-task mix-lingual modeling with shared bottleneck layers followed by transfer learning. Besides, the system also supports state-of-the-art OCR, TTS, and machine translation, by which language information will be real-time translated, punctuated and pronounced. The machine translation behind the system gets a high rank in WMT 18 Mandarin-English and CWMT 18 minority language translation task. The system has integrated into a micro-app at WeChat and can facilitate communication between Mandarin and Minority languages.

Index Terms: speech recognition, low resource languages

1. Introduction

Most existing state-of-the-art speech recognition systems were high-resourced languages, such system suffers from accuracy degradation when it comes to low-resource languages. In China, there are more than 50 minority races with over 200 languages. Language differences become an obstacle to communicate, so it is urgent to establish a system with simultaneous translation between Mandarin and these languages. Facing the difficulty of low resources and lack of training data, various solutions have been proposed via our lexicon, language and multi-lingual acoustic models.

In this work, we present our effort to develop a prototype text, speech and image (OCR) translation system that can recognize live minority languages such as Uyghur, Tibetan in shapes of various media forms.

2. System architecture

Our system is built for simultaneous translation in WeChat “mirco app”. In network transmission, we design a set of general real-time slicing & transmission protocols, including disorder processing, jitter buffer, timeout retransmission and other mechanisms to ensure that slicing requests are processed sequentially. WebSocket is selected as frontend and server communication protocol. The frontend divides the input encoded audio into pieces, and on-the fly requests the back-end service, simultaneously the answering pipe is in another stream to display translation results to the front end, which will shorten the latency for the user to get the feedbacks.

Since there are three kinds of medias, namely text, image and speech, all of which share same modules such as word segmentation, post-processing, and name entity recognition (NER) after obtaining target language content, an unified text based post-processing and translation framework is designed. Moreover, all media processes with the same translation direction will be bind to a single GPU card as much as possible in order to save computation consumed by same semantic algorithm. The following is an example architecture. Moreover, inspired by the work in [1], an open vocabulary sub-word OCR decoder is constructed by observing posteriors computed from line image features, whereas the neural network is applied to compute emission probability piggybacked in character based HMM. However, since image and audio features are different in properties such as strides, window lengths, and chunk lengths, various “acoustic model” is trained and composed with the same source language model to form a decoding graph in shape of HCLG WFST [3].

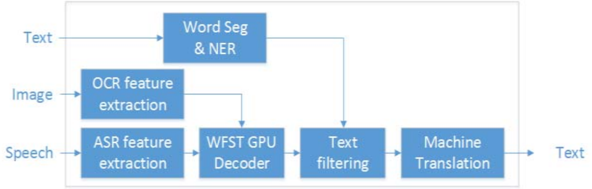

Figure 1: *Unified decoding & translation architecture.*

2.1. Decoder

Computational efficiency is critical for ASR and one of the most time consuming part is decoder. Our decoding algorithm extends the traditional weighted finite-state transducer (WFST) based static decoder. The advantage of using static decoder is that searching on a WFST can be efficient and simple. However, compared to acoustic cost calculation part in GPU, graph search part calculated in CPU is more consuming. As a result, accelerating the speed of graph search part is the key point for a practical online system. Chen in [2] introduces a GPU based WFST decoding algorithm which implements token passing algorithm in parallel as an atomic operation, and propose a dynamic load balancing strategy among GPU threads. The proposed algorithm significantly speeds up the CPU based graph search. Inspired by this work, the backbone decoder is based on WFST GPU decoding and we implement it in form of KALDI nnet3 style [3]. In addition, we accelerate acoustic cost calculation part by using looped parallel acoustic inference. The looped decoding algorithms do not use any previously-computed chunks, which allows us to use shorter chunks effectively. The above methods benefit the overall speech-translation system effectively.

2.2. Lexicon and language model

Lexicon is built across various languages: 1) For Mandarin 300k common words with pronunciations are transcribed with 460 Mandarin tonal phones. 2) For Uyghur, 180k common words are transcribed via 160 phones; 3) For Tibetan, 10k common words are transcribed via 180 phones; all the phone sets are designed to be positional dependent with four types of positions: start, middle, end, and singleton.

For Uyghur lexicon construction, mixed units by applying a mixture of sub-word BPE and whole-word is applied to build a hybrid lexicon and language models for recognition [4]. Since Uyghur is a highly agglutinative language with a large number of words derived from the same root. For such languages the use of sub-words in ASR is preferred, which can considerably lessen the OOV issues. However, short units in sub-word modeling will weaken the constraint of linguistic context, this compromising solution is inspired by the above observations and advices by native Uyghur researches.

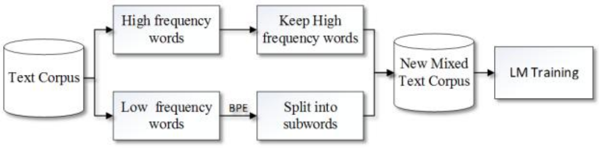

Figure 2: *Mixed word-unit lexicon modeling.*

For Minority languages modeling, the biggest challenge facing is the lack of resources. To deal with this, machine translation is adopted to translate crawled 1GB Mandarin raw transcripts from internet to Uyghur and Tibetan. Translated results with high confidence scores are selected, which are word-classified and NER is recognized. For specific types of word such as name, locations, time, etc, slot expansion using same word classes are enumerated to enlarge the original transcripts, which is sent to a word class based LM at a later stage [5].

2.3. Acoustic model

Our experimental data consists of 10kh Mandarin, 5kh English, 400h Uyghur and 400h Tibetan dictated speeches, respectively. Due to the low resourced data, our method "multilingual modeling + monolingual recognition" is proposed to improve the performance for small language recognition system through mixed training multi-lingual data. Having compared the methods of sharing phone set and sharing bottleneck layers, results suggest that the latter approach has better accuracy due to a more abstract representation of acoustics in intermediate layers. The neural network is initialized via incorporating all multiple languages, and then some layers are shared in the training. The basic principle of this method is that the phoneme distribution for target language is also reflected in the background multilingual data. After initial AM is trained via factorized TDNN with lattice free MMI criterion [6], transfer learning is employed to the target language to improve the target phoneme discrimination, which proves to be more effective than using monolingual data [7]. In this system, since the initial layers in the deep neural networks are "generic" and final layers are language-specific, it's better to tweak the number and position of layers to be shared, basically the closer the languages, the more layers are shared.

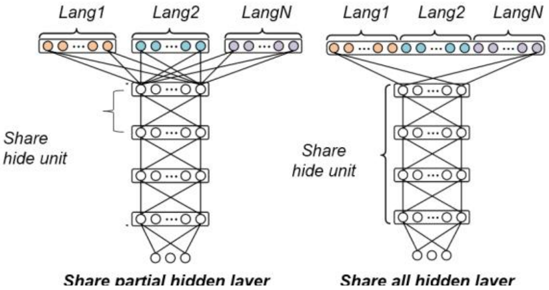

Figure 3: *Mixed word-unit lexicon modeling.*

2.4. Machine translation

Our machine translation system is based on Transformer self-attention mechanism [8]. In addition to the basic settings of transformer training, we uses multi-model fusion techniques, multiple features re-ranking, different segmentation models and joint learning. Finally, data selection strategies are adopted to fine-tune the trained system and stable performance improvement is observed. The prototype system achieved 2rd best BLEU scores in WMT 2018 and ranks 1st in CWMT18 for minority language translation.

3. Conclusions

We have demonstrated the development of our multimedia simultaneous translation system for minority languages. The demo system is the first real application that can real-time recognize, translate and pronounce some of the major minority languages in China.

4. References

[1] A. Arora, CC. Chang, B. Rekabdar, D. Povey, et al, "Using ASR methods for OCR" *in Proc. of ICDAR, 2019.*

[2] Z.H. Chen, J. Luitjens, H. Xu, Y. Wang, D. Povey, S. Khudanpur, "A gpu-based wfst decoder with exact lattice generation," *in Proc. of INTERSPEECH, 2018.*

[3] D. Povey, A. Ghoshal, et al, "The kaldi speech recognition toolkit," *in Proc. of ASRU workshop, 2011.*

[4] P.F. Hu, S. Huang, Z,Q Lv, "Investigating the Use of Mixed-Units based Modeling for Improving Uyghur Speech Recognition," in *Proc. of INTERSPEECH 2018 satellite workshop on under resourced language.*

[5] P. F. Brown, P. V. deSouza, R. L. Mercer, V. J. D. Pietra, and J. C. "Class-based n-gram models of natural language," Comput. Linguist. vol. 18, no. 4, pp. 467–479.

[6] D. Povey, G.F. Cheng, et al, "Semi-Orthogonal Low-Rank Matrix Factorization for Deep Neural Networks, " *in Proc. of INTERSPEECH, 2018.*

[7] J.H Yan, Z.Q Lv, et al, "Low-Resource Tibetan Dialect Acoustic Modeling Based on Transfer Learning," in *INTERSPEECH 2018 satellite workshop on under resourced language.*

[8] B.J. Hu, Ambyer Han, S. Huang, "TencentFmRD Neural Machine Translation System", EMNLP 2018 workshop for Proceedings of the 3rd Conference on Machine Translation.