3-1-8

Word-level Emphasis Transfer in Speech-to-speech Translation [*]

☆ Quoc Truong Do, Shinnosuke Takamichi, Sakriani Sakti, Graham Neubig(NAIST),
Tomoki Toda (NAIST/Nagoya University), Satoshi Nakamura (NAIST)

1 Introduction

Speech-to-speech (S2S) translation systems [1] combine various technologies to help to translate speech across languages. However, most S2S systems ignore paralinguistic information such as emphasis. This paper attempts to solve the problem by proposing two new components: word-level emphasis estimation [2] using linear regression hidden semi-Markov models (LR-HSMM) [3], and emphasis translation that translates the word-level emphasis to a target language with conditional random fields (CRFs) [4]. The result shows that our system can accurately translate emphasis with 91.6% F-measure according to objective evaluation. A listening test with human subjects further showed that they could identify emphasized words with 87.8% F-measure. This paper is a shortened version of our work presented in [5].

2 Word-level Emphasis Translation

Our proposed emphasized speech translation model consists of a conventional S2S system, an emphasis estimation model, and an emphasis translation model as illustrated in Fig. 1. As described in detail in our previous work [2], the emphasis estimation component estimates an emphasis-level sequence $\hat{\boldsymbol{\lambda}}$ for every utterance using LR-HSMM. Each word in the utterance has its own emphasis level representing how emphasized the word is.

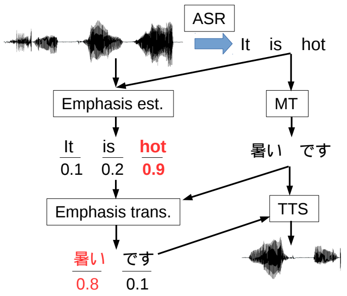

Fig. 1 Emphasis speech-to-speech translation

Adopting the result from [2], our next step is emphasis translation, or to take the estimated emphasis sequence in the source languages $\hat{\boldsymbol{\lambda}}^{(f)}$, and convert them to an emphasis sequence in the target language $\hat{\boldsymbol{\lambda}}^{(e)}$ using CRFs. In order to train the CRF model, we create training data consisting of source and target words $\mathbf{w}^{(f)}$ and $\mathbf{w}^{(e)}$, and the corresponding estimated emphasis values. As $\hat{\boldsymbol{\lambda}}^{(f)}$ is a sequence of continuous values, and CRFs requires discrete state sequences, we first quantize $\hat{\boldsymbol{\lambda}}^{(f)}$ and $\hat{\boldsymbol{\lambda}}^{(e)}$ into buckets, giving us a discrete sequence $\hat{\boldsymbol{\lambda}}^{(f)'}$ and $\hat{\boldsymbol{\lambda}}^{(e)'}$. We then create CRFs training data that consists of N samples $D = [(\mathbf{x}_1, \lambda_1^{(e)'}), \cdots, (\mathbf{x}_n, \lambda_n^{(e)'}), \cdots, (\mathbf{x}_N, \lambda_N^{(e)'})]$, where $\mathbf{x}_n$ is a feature vector consisting of:

- source word-level emphasis $\lambda_j^{(f)}$, and its context,
- source word $w_j^{(f)}$, and word context,
- source part of speech (PoS), and PoS context,
- target word $w_n^{(e)}$, and word context,
- target PoS, and PoS context,

where context means the information of one succeeding and one preceding words.

To decide which source features correspond to a target word $w_n^{(e)}$, we use one-to-one word alignments between $w_j^{(f)}$ and $w_n^{(e)}$. The likelihood of CRFs is given by

$$P(\boldsymbol{\lambda}^{(e)'}|\mathbf{x}) = \frac{\prod_{n=1}^{N} \exp\left\{\sum_{k=1}^{K} \theta_k f_k(\lambda_{n-1}^{(e)'}, \lambda_n^{(e)'}, \mathbf{x}_n^{(k)})\right\}}{\sum_{\tilde{\boldsymbol{\lambda}}^{(e)'}} \prod_{n=1}^{N} \exp\left\{\sum_{k=1}^{K} \theta_k f_k(\tilde{\lambda}_{n-1}^{(e)'}, \tilde{\lambda}_n^{(e)'}, \mathbf{x}_n^{(k)})\right\}}, \tag{1}$$

where θ_k is model parameters, f is feature function, and K is number of feature function. Different combination of input features described above will be used depends on individual feature function f_k.

3 Experiments

Experiments were performed using a bilingual English-Japanese emphasis corpus [6]. The corpus consists of 1015 parallel utterances of English and Japanese. We use the data from 2 speakers, a native English speaker, and a Japanese native speaker as the training and testing data. After filtering out

[*] 音声翻訳システムにおける単語強調の伝達

Table 1 F-measure for different combinations of input features. *e_en* and *e_ja* denote word-level emphasis, *w_en* and *w_ja* denote word information, *t_en* and *t_ja* denote PoS tag of English, and Japanese, respectively, and **_c* denotes context information.

Feature type	***F*-measure**
w_ja, *t_ja*, *t_ja_c*, *e_ja*	81.6
e_en, *e_ja*	82.8
e_en, *e_ja*, *e_en_c*	82.8
e_en, *e_ja*, *w_en*, *w_ja*	84.8
e_en, *e_ja*, *w_en*, *w_ja*, *t_en*, *t_ja*	90.0
e_en, *e_ja*, *w_en*, *w_ja*, *t_en*, *t_ja*, *t_ja_c*	**91.6**

long sentences over 10 words, we obtained 966 utterances, which we divided into 916 sentences with 1,186 emphasized words for training and 50 sentences with 62 emphasized words for testing. The word-level emphasis is quantized to the closest of {0, 0.3, 0.6, 0.9}.

To measure the accuracy, we calculate emphasis F-measure, the harmonic mean of the precision and recall with which the system detects emphasis.

3.1 Emphasis translation evaluation

In our first experiment, we evaluate the ability of the proposed method to reproduce emphasis in the target language. In addition, we also evaluate the effect of the combination of input features described in Section 2 to find out which features give the highest F-measure. The result is shown in Table 1.

We can see that the model that use the features including the source language information (2nd-6th row) is better than the model use only target information (1st row). This demonstrates that our model effectively translates emphasis from the source, as opposed to simply predicting based on the target.

Looking at the second and third rows, we can see that emphasis contexts in the source language does not help word-level emphasis translation, indicating that word-level emphasis in the target language depends mainly on emphasis of the corresponding source word. By adding more linguistic information such as word and PoS, the best system is achieved with 91.6% F-measure.

In addition, we carried out a listening evaluation with 6 native Japanese speakers. We asked to 6 native Japanese speakers to listen to the emphasis translated utterances, and select the words that they think are emphasized in 150 randomized testing utterances from the following 3 systems: **Baseline**–no emphasis translation, **CRF-based**–with CRF-based emphasis translation, and **Natural**–natural spoken speech by a Japanese speaker. The result in Fig. 2 shows that the **CRF-based** system achieved 87.8% F-measure that outperforms the **Baseline** system with 77.0% F-measure.

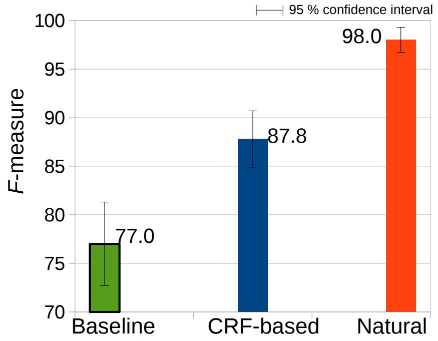

Fig. 2 Prediction F-measure for manual evaluation

4 Conclusion

In this paper, we present a new speech translation architecture that is able to translate emphasis across languages. Experiments showed that our proposed approach can accurately convey emphasis in speech translation models. Future work will improve the emphasized speech synthesis quality, and the emphasis translation model.

Acknowledgements

Part of this research was supported by JSPS KAKENHI Grant Number 24240032 and 26870371, and the Commissioned Research of National Institute of Information and Communications Technology (NICT), Japan.

References

[1] S. Nakamura. Overcoming the language barrier with speech translation technology. *Science & Technology Trends - Quarterly Review No.31*, April 2009.

[2] D. Q. Truong, S. Takamichi, S. Sakti, G. Neubig, T. Toda, and S. Nakamura. Study on word-level emphasis across English and Japanese. In *Proceedings of ASJ, Autumn meeting*, 2015.

[3] T. Nose, J. Yamagishi, T. Masuko, and T. Kobayashi. A style control technique for HMM-based expressive speech synthesis. *IEICE*, E90-D(9):1406–1413, September 2007.

[4] J.D. Lafferty, A. McCallum, and F.C.N. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of ICML*, pages 282–289, 2001.

[5] S. Sakti G. Neubig T. Toda S. Nakamura Q. T. Do, S. Takamichi. Preserving word-level emphasis in speech-to-speech translation using linear regression HSMMs. In *INTERSPEECH*, 2015.

[6] D. Q. Truong, S. Sakti, G. Neubig, T. Toda, and S. Nakamura. Collection and analysis of a Japanese-English emphasized speech corpus. In *Proceedings of Oriental COCOSDA*, September 2014.