一般社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
IEICE Technical Report
SP2016-50 (2016-12)

ハードアテンション用いた注意型ニューラルネットワークによる音声翻訳

ク　チュオン　ド† 　サクリアニ　サクティ† 　グラム　ニュービック† 　中村　哲†

† 奈良先端科学技術大学院大学

あらまし　一般的な音声翻訳システムにおいて、発話に含まれる非言語情報は翻訳されない。先行研究において、この問題に対し条件付き確率場（CRF）を用いた手法を提案した。CRF を用いることで、多くの素性を扱い、局所的な文脈情報も考慮できるようになった。一方で、CRF は連続的な変数を扱うことが困難であり、長期的な依存関係を表現することが難しい。本論文では、注意型ニューラルネットワークを用いた強調音声の翻訳手法を提案する。また、人手による主観評価を通して、翻訳言語上で強調か否かを予測する実験において先行研究よりも F 値で 4%性能が向上したことを示す。

キーワード　協調翻訳、音声翻訳、注意型ニューラルネットワーク、ハードアテンション

Hard-Attentional Neural Network Models for Emphasis Speech Translation

Quoc TRUONG DO†, Sakriani SAKTI†, Graham NEUBIG†, and Satoshi NAKAMURA†

† Nara Institute of Science and Technology

Abstract　Traditional speech translation systems are oblivious to paralinguistic information. A recent work has tried to tackle this task by utilizing conditional random fields (CRFs). Although CRFs allow for consideration of rich features and local context, they have difficulty in handling continuous variables, and cannot capture long-distance dependencies easily. In this paper, we propose a new model for emphasis transfer in speech translation using an approach based on neural networks. Our experiments showed a significant improvement of the proposed model over the previous model by 4% target-language emphasis prediction F-measure according to objective evaluation.

Key words　Emphasis translation, speech translation, attentional network model, hard-attentional

1. Introduction

Speech-to-speech (S2ST) translation technologies [1] have been gradually starting to break down the language barriers by translating linguistic information (meaning) of speech across languages. However, conventional S2ST systems ignore emphasis information.

The previous work [2] have proposed an approach to estimate and translate emphasis considering all acoustic features such as power, duration, and F_0 patterns. The emphasis estimation system estimates a real-numbered value representing how emphasized a word is, and emphasis is translated using conditional random fields (CRFs). However, because CRFs require discrete labels, continuous emphasis levels must be quantized into discrete values. Moreover, while CRFs have the ability to capture local dependencies between neighboring labels, they cannot easily handle longer distance dependencies between words in separate parts of the sentence.

In this paper, we propose a model that solves these problems using long short-term memory neural networks (LSTMs) [3]. LSTMs are capable of model long-term dependencies, overcoming the problems of local dependencies in CRFs. In addition, it is possible to define models that can handle continuous variables, and cost functions taking into account label distances, for example, mean squared errors.

2. Emphasis Translation Using Hard-attentional Encoder-Decoders

The proposed emphasis translation system consists of 2 components: an LSTM encoder and an LSTM decoder as illustrated in Fig. 1. The encoder encodes features from the source language, and the decoder takes the encoded features to generate an emphasis sequence in the target language.

The whole encoder-decoder process can be written as a function of input features as follows:

$$\mathbf{o}^{(f)} = f(\mathbf{x}^{(e)}), \tag{1}$$

where $\mathbf{o}^{(f)}$ is the target output sequence, $\mathbf{x}^{(e)}$ is the sequence of the source-language input vector $\mathbf{x}_i^{(e)}$.

2.1 The encoder

The encoder is a standard LSTM model that takes the input vector $\mathbf{x}_i^{(e)}$ consists of words ($w_i^{(e)}$), part-of-speech tags

($p_i^{(e)}$), and emphasis levels ($\lambda_i^{(e)}$), then encodes them into a single vector that is suitable to predict emphasis levels.

The input PoS tags are converted into one-hot vectors with the size is equal to PoS vocabulary size. Also, word embeddings [4] are applied to map words into vectors that capture the similarity between the words. All these input features are concatenated into a single vector and fed to the encoder.

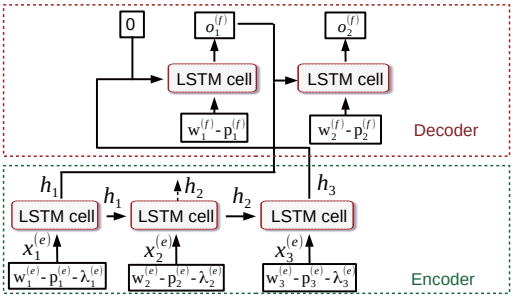

図 1 An unfolded hard-attentional encoder-decoder LSTM model for translating emphasis sequence $\boldsymbol{\lambda}^{(e)}$ into a target output sequence $\boldsymbol{o}^{(f)}$. It takes into account many linguistic features including the word sequence $\mathbf{w}_i^{(e,f)}$ and the part of speech sequence $\mathrm{p}_i^{(e,f)}$ from both source and target languages.

2.2 The decoder

The decoder is also a standard LSTM model, and the input layer contains both the linguistic information (words, PoS), and vector representations calculated by the encoder.

The name hard-attentional comes from the way the decoder calculates the emphasis representation vectors used as input. The example in Fig. 1 demonstrates this mechanism. Assume that the word pairs $w_1^{(e)}$-$w_2^{(f)}$ and $w_3^{(e)}$-$w_1^{(f)}$ is aligned according to word alignments. To generate the output $o_2^{(f)}$, along with linguistic features $w_2^{(f)}$ and $\mathrm{p}_2^{(f)}$ and the previous output $\lambda_1^{(f)}$, the decoder takes the encoded $\mathbf{h}_1$ from the encoder output, because the word pair $w_1^{(e)}$-$w_2^{(f)}$ are aligned. For unaligned words, we use zero vectors as the emphasis representation vectors.

3. Experiments

In this paper, to evaluate the performance of emphasis translation in isolation, we assume that the MT system produces 100% correct translation outputs. The experiments were conducted using a bilingual English-Japanese emphasized speech corpus [5]. The training and testing data consist of 4330 and 100 utterances, respectively. To evaluate the system, we perform objective evaluation where predicted emphasis levels in the target language are classified into binary values using a threshold of 0.5 and subjective evaluation where native target language listeners decide emphasized words based on their perception.

The encoder's input layer has a size of 138 including 100 dimensions of word embedding, 37 dimensions of one-hot PoS, and emphasis level. Hidden layers have a size of 100 and the output layer predict directly emphasis levels with a size of 1.

3.1 Objective evaluation

Fig. 2 shows the objective F-measure for emphasis prediction. As we can see, in all 3 test sets and in the average, the proposed method performs better than the CRFs. According to the bootstrap resampling significance test [6], both results are significant at the $p < 0.01$ level.

Further analyses have shown that LSTMs perform significantly better than CRF when emphasis levels fall between 0.3-0.6, which is the ambiguous range of emphasis levels. This demonstrates the limitation of CRFs, which require emphasis level quantization to handle continuous variables while LSTMs do not.

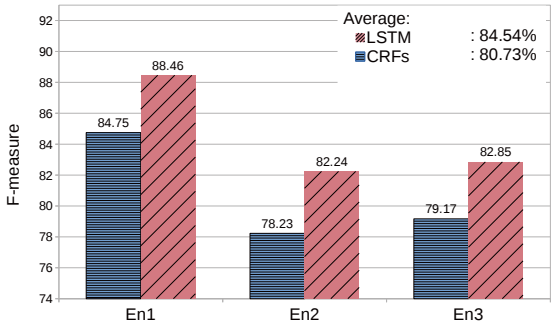

図 2 Objective emphasis prediction F-measure.

3.2 Subjective evaluation on emphasis translation

Finally, we performed the subjective evaluation to verify whether human listeners can perceive the same improvement between *CRFs* and *LSTM* as in the objective evaluation. The test set "En1" is used for the evaluation.

We obtain a result of 83.0% for *LSTM* and 81.0% for *CRFs* indicating that the human perceives a slightly smaller improvement compared to the objective result. Moreover, the performance of the *CRF* system dropped with a smaller margin (3.70%) than proposed method (5.82%).

4. Conclusion

In this paper, we explored encoder-decoder neural net approaches and proposed "hard"-attentional LSTMs for emphasis translation tasks. Compared to previous works, the proposed model has achieved significantly better performance. This is a result of the fact that the model does not require any emphasis quantization and takes into account emphasis label relationships in the loss function.

謝辞 Part of this work was supported by JSPS KAKENHI Grant Number 24240032.

文　　献

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

[2] Q.T. Do, S. Takamichi, S. Sakti, G. Neubig, T. Toda, and S. Nakamura, "Preserving word-level emphasis in speech-to-speech translation using linear regression HSMMs," Proceedings of INTERSPEECH, pp.3665–3669, 2015.

[3] S. Hochreiter and J. Schmidhuber, "Long short-term memory," Neural computation, vol.9, no.8, pp.1735–1780, 1997.

[4] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," arXiv preprint arXiv:1301.3781, pp. • • – • •, 2013.

[5] D.Q. Truong, S. Sakti, G. Neubig, T. Toda, and S. Nakamura, "Collection and analysis of a Japanese-English emphasized speech corpus," Proceedings of Oriental COCOSDA, pp.1–5, Sept. 2014.

[6] P. Koehn, "Statistical significance tests for machine translation evaluation.," Proceedings of EMNLP, pp.388–395, 2004.