Improving Chinese-Japanese Neural Machine Translation with Joint Semantic-Phonetic Word Embedding

以聯合語義語音詞嵌入強化中日文神經機器翻譯

Author: Shih-Chieh Wang
Advisor: Paul Horton

Master Degree Program on Artificial Intelligence,
National Cheng Kung University, Tainan, Taiwan, R.O.C.

Introduction

Neural machine translation (NMT) has achieved sustained success by refining the encoder-decoder model. Several techniques have been proposed to improve the quality of NMT in all languages such as attention-based model (Bahdanau, Cho, and Bengio 2014; Vaswani et al. 2017), byte-pair encoding (Sennrich, Haddow, and Birch 2015b), back-translation for data augmentation (Sennrich, Haddow, and Birch 2015a), etc. The main focus of improving the NMT system between Chinese and Japanese is to explore the features in sub-character level (Zhang and Komachi 2018). In addition, the use of phonetic information as a feature of Western languages and Chinese also enhances the performance of the NMT system (Liu et al. 2018; Khan and Xu 2019). This paper aims to show the improvement of the Chinese-Japanese NMT system after applying phonetic information and the analysis of joint semantic-phonetic embedding from four aspects which shows the advantages of joint embedding.

Objectives

The main purpose of this research is to determine whether the phonetic information is useful for the Chinese-Japanese NMT system. We first attempt to combine the semantic and phonetic embedding; both were trained on a small size parallel corpus (*Asian Scientific Paper Excerpt Corpus, ASPEC*, 672,315 lines of sentences). Next, we examine the translation performance with embeddings on a well-filtered and sampled dataset (50,000 lines of sentences). Finally, we apply four analyses to our joint embedding, namely analogy reasoning, outlier detection, word similarity, and the effects on homonyms and heteronyms.

Methods and Experiments

The tools used for tokenization are *Sentencepiece* (Chinese and Japanese), *Jieba* (Chinese) and *Janome* (Japanese). Phonetic information was extracted using *dragonmapper* and *pykakasi* for Chinese and Japanese respectively. Embeddings were trained and analyzed on *Gensim Word2Vec*, and joint embedding was created by averaging the semantic and phonetic embedding. We filtered the ASPEC dataset with length ratio, overlapping, word alignment score, etc. Following this, the NMT tasks were run on two models, one is an attention-based Bi-GRU encoder-decoder model, the other is a standard Transformer model, both are built in *PyTorch-Lightning*.

Results

NMT Model

The BLEU scores (Papineni et al. 2002) showed the success of joint embedding on the sampled data (50,000 lines of sentences) with different combinations of tokenizers and models.

	Sentencepiece		Jieba & Janome	
	RNN	Transformer	RNN	Transformer
w/o Pre-trained emb	21.63	24.32	25.16	29.31
w/ Semantic emb	21.66	25.72	26.71	31.23
w/ Phonetic emb	21.32	23.48	26.18	30.9
w/ Joint emb	**22.33**	**26.44**	**27.05**	**32.48**

Table 1: BLEU Scores on Sampled Data

Joint embedding also gained the highest score on full well-filtered data (462,582 lines of sentences) with Jieba, Janome as the tokenizers, and Transformer as the model. Moreover, the resulting BLEU score was 6 points higher compared to the baseline of WAT2020 (Nakazawa et al. 2020).

	WAT2020	w/o Pre-trained	w/ Semantic	w/ Phonetic	w/ Joint
Baseline	47.00				
Our Best Model		52.78	52.83	53.04	**53.13**

Table 2: BLEU Scores on Full Well-Filtered Dataset

Case Study

Joint embedding correctly distinguished 試験 (test) from 実験 (experiment), and 測定 (measurement) from 計測 (instrumentation). It was also robust to noise; in the second case shown below, it output the missing symbol H in the target sentence and translated with correct grammar.

Source	微小粒子测量装置的比较试验
Target	微小粒子測定装置の比較試験
Semantic	微小粒子計測装置の比較実験
Semantic + Phonetic	微小粒子測定装置の比較試験
Source	从背景知识B和观测结果O中获得行动规则的集合Y的集合H
Target	背景知識Bと観測結果Oより行動規則の集合Yの集合を獲得する
Semantic	背景背景知識Bと観測結果Oから行動ルールの集合Yの集合H獲得する
Semantic + Phonetic	背景知識Bと観測結果Oから行動規則の集合Yの集合H獲得する

Table 3: Case Study

Embedding Analyses

We observed that the joint embedding could preserve and further improve the semantic meaning of words; it also had positive effects on homonyms (same spelling but different meaning) and heteronyms (same character but different spelling and meaning).

Analogy Reasoning

Joint embedding predicted the correct answer from analogy reasoning with higher cosine similarity.

Language	Input ($a : A = b :$)	Output (B) & Cosine Similarity (cs)	
		Semantic Only	Semantic + Phonetic
Chinese	東京:日本=北京:	中國 ($cs = 0.49$)	中國 (**cs = 0.53**)
	長期:三年=短期:	一年 ($cs = 0.38$)	兩周 (**cs = 0.39**)
	進口:買入=出口:	賣出 ($cs = 0.36$)	賣出 (**cs = 0.44**)
Japanese	男性:女性=父親:	母親 ($cs = 0.49$)	母親 (**cs = 0.51**)
	長期:年=短期:	月 ($cs = 0.55$)	月 (**cs = 0.57**)
	左右:前後=水平:	垂直 (**cs = 0.43**)	垂直 ($cs = 0.40$)

Table 4: Analogy Reasoning

Outlier Detection

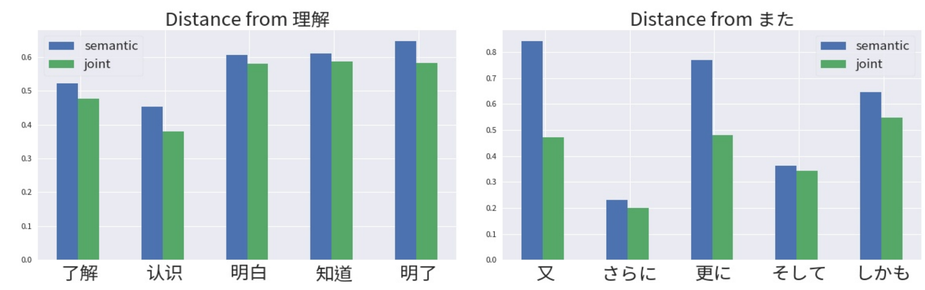
Joint embedding predicted the correct answer from outlier detection with some better answers.

Language	Input	Output (Outlier)	
		Semantic Only	Semantic + Phonetic
Chinese	維持, 保持, 堅持, 建設	建設	建設
	可行, 不行, 可以, 行	可以	不行
Japanese	生み, 創造, 作る, 破壊	破壊	破壊
	普通, 一般, 通常, 異常	異常	異常

Table 5: Outlier Detection

Word Similarity

Joint embedding reduced the cosine distance between synonyms, which can be considered as improving the semantics.

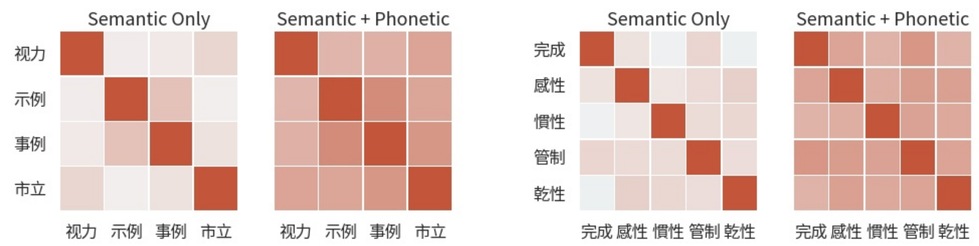


(a) Chinese **(b)** Japanese

Figure 1: Word Similarity

Homonyms and Heteronyms

Joint embedding increased the correlations between homonyms as it reduced the cosine distance between homonyms. This can be useful when typographical errors occurred in the dataset.



(a) Chinese **(b)** Japanese

Figure 2: Correlation between Homonyms

Joint embedding increased the cosine distance between heteronyms, which are the same characters but with different meanings. This can also be seen as an improvement in semantics.

Language	Input	Cosine Distance	
		Semantic Only	Semantic + Phonetic
Chinese	長(ㄔㄤˊ)度, 長(ㄓㄤˇ)大	0.826	**0.853**
	樂(ㄌㄜˋ)趣, 音樂(ㄩㄝˋ)	0.636	**0.682**
	中(ㄓㄨㄥ)午, 中(ㄓㄨㄥˋ)毒	0.841	**0.866**
Japanese	生(なま), 一生(しょう)	0.879	**0.889**
	生(なま), 生(き)地	0.769	**0.867**
	生(なま), 生(う)む	0.829	**0.839**

Table 6: Cosine Distance Between Heteronyms

Conclusion

The purpose of this paper is to utilize phonetic information as a feature of the Chinese-Japanese NMT system. The joint semantic-phonetic embedding trained by Word2Vec in a small corpus had improved the translation qualities in different combinations of tokenization methods and models. Furthermore, the joint embedding also gave positive feedback in the embedding analysis. Taken together, these results indicate that the use of phonetic information can effectively improve the performance of the Chinese-Japanese NMT system.