Learning to Explain Chinese Slang Words

Chuanrun Yi[1,2], Dong Wang[1,2], Chunyu He[1,2], and Ying Sha[1,2*]

[1]Institute of Information Engineering, Chinese Academy of Sciences, Beijing, China

[2]School of Cyber Security, University of Chinese Academy of Sciences, Beijing, China

{yichuanrun, wangdong, hechunyu, shaying}@iie.ac.cn

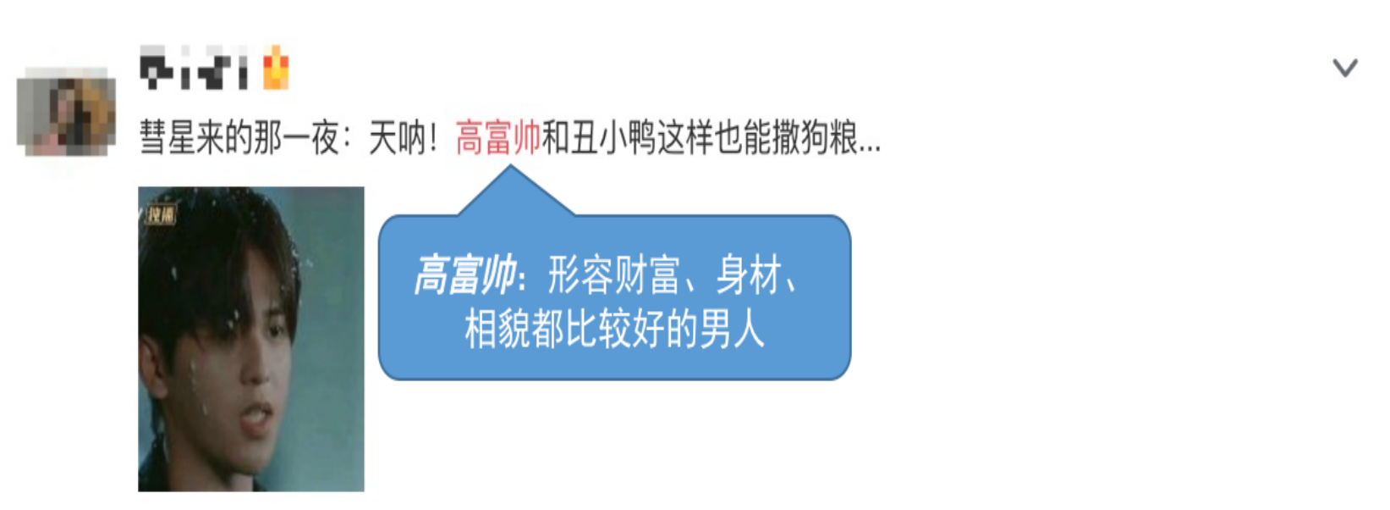
Introduction

■ Problems

- ➢ The explosive development of social media has generated a large number of slang words in Chinese social network.
- ➢ The appearance of Chinese slang words has affected the accuracy of reading comprehension and word segmentation tasks.
- ➢ The most common way to get an explanation of a slang word is to match the word in dictionary.

■ Our goal

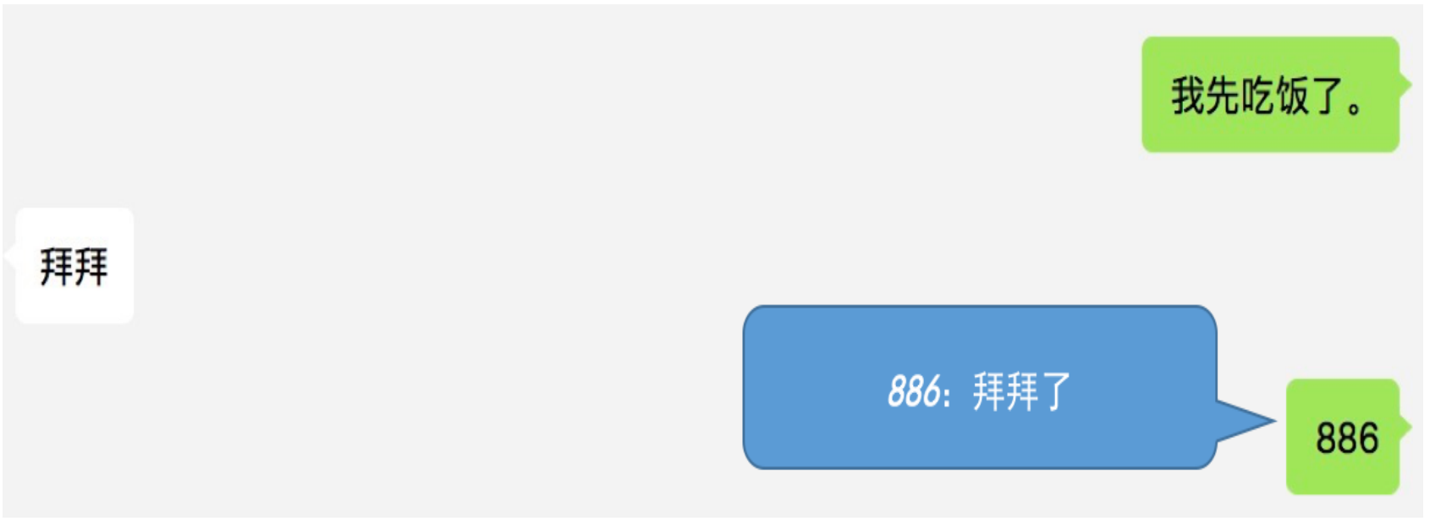
Learning to explain the meaning of Chinese slang words automatically by machiine. Here are two examples. Fig.1 shows the explanation of Chinese slang word "高富帅 (tall, rich and handsome)" in Weibo. Fig.2 shows the explanation of network digital language "886(phonetics: ba ba liu)" in WeChat.



彗星来的那一夜：天呐！高富帅和丑小鸭这样也能撒狗粮...

高富帅：形容财富、身材、相貌都比较好的男人

Fig.1: An example of Chinese slang word in Weibo. We aim at explaining "高富帅 (tall, rich and handsome) " as "形容财富、身材相貌都比较好的男人 (describe a man who is wealthy, fit and good-looking)" automatically.



我先吃饭了。

拜拜

886：拜拜了

886

Fig.2: An example of network digital language in WeChat record. We aim at explaining "886(phonetics: ba ba liu)" as "拜拜了 (bye bye)" automatically.

■ Our contributes

- ➢ To the best of our knowledge, this is the first attempt to explain Chinese slang words automatically by machine, and use DCEAnn(a Dual Character-level Encoder using Attention-based neural network) model for this specific task.
- ➢ Our novel DCEAnn model can generate reasonable explanations for Chinese slang words and get the state-of-the-art results.
- ➢ We constructs a parallel corpus for the explanation of Chinese slang words (continued growth), which can be applied to later academic research.

Our Approach

We define our task as learning to explain Chinese slang words in a given sentence automatically. The input is a sentence containing Chinese slang word, which is used to enrich the semantic information of the slang word. The output is the explanation of the slang word, which is also a sentence.

Our goal is to generate an explanation of Chinese slang word automatically. We select sequence-to-sequence model with attention mechanism as our fundamental framework. Because some words have different meanings in different contexts, we use one character-level BiGRU encoder of example sentence to help machine understand the slang word better. Besides, considering the particularity of Chinese and the fact that many network neologisms are homophones, we use another character-level BiGRU encoder of slang word and its phonetics to learn the representation of neologism. In general, our model is a dual character-level encoder using attention-based neural network, which we call it DCEAnn. An overview of our model is shown in Fig.3 .

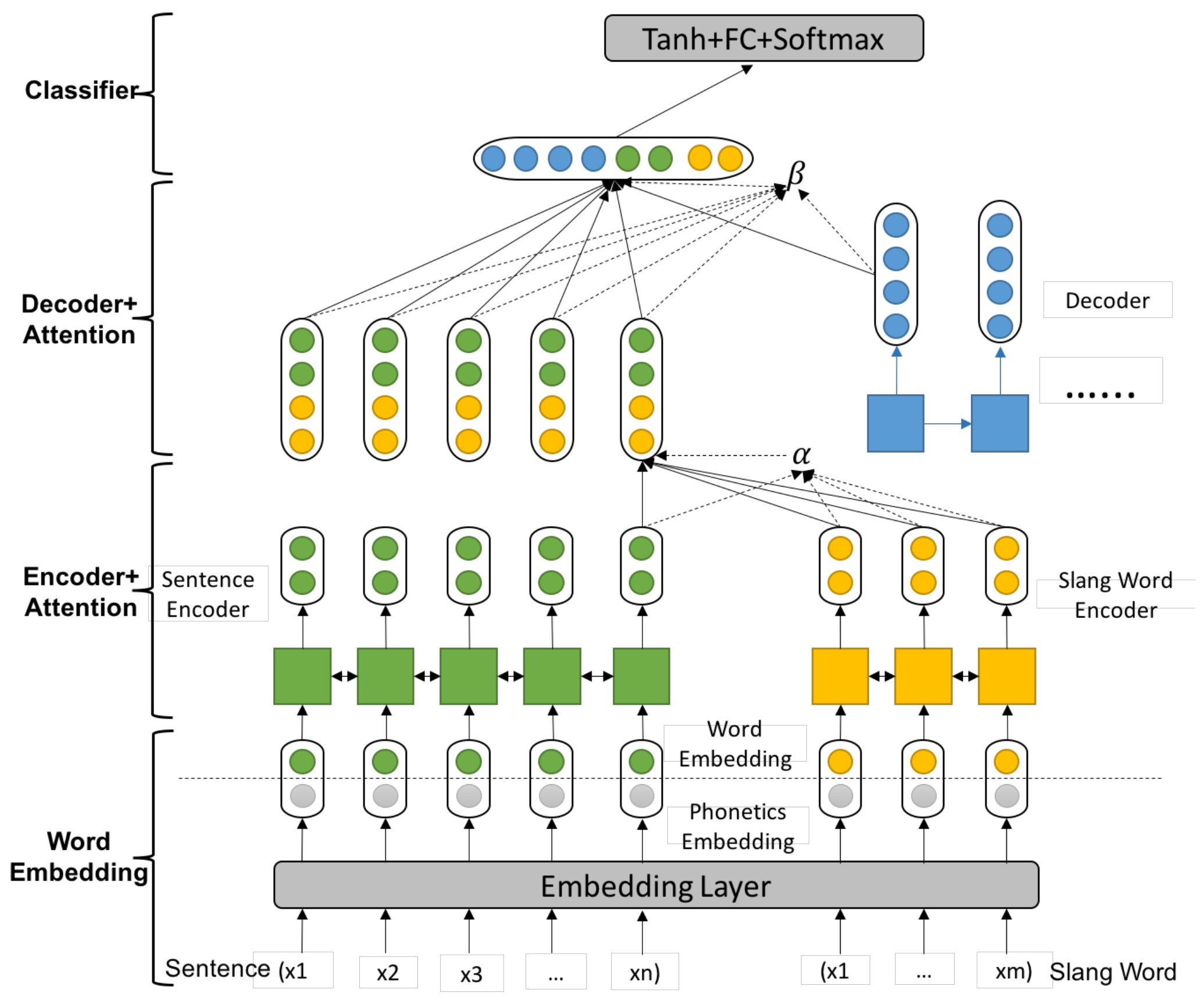


Fig.3: The overview of our model. Two encoders encode the example sentence and Chinese slang word spliced phonetics respectively. Decoder generates the interpretation which relies on both hidden states of decoder and encoder. α and β indicate the attention sequence.

Experiment

■ datasets

➢ datasets Construction

we collected 2,831 Chinese slang words and 6,519 example sentences, among which 386 are network numbers. As shown in Fig. 4, the structure of each item is: Chinese slang word, example sentence and explanation.

Chinese Slang : *高富帅*
Example: 几乎每个男人都想成为"高富帅"，而不是"矮丑穷"
Explanation: 形容财富、身材、相貌都比较好的男人

Chinese Slang : *感冒*
Example: 我对帅哥特感冒, 对老男人没有兴趣
Explanation: 喜欢或感兴趣的意思

Chinese Slang : *520*
Example: *520*，不管怎样，我都爱你
Explanation: 我爱你的意思

Fig.4:The example of parallel corpus for the explanation of Chinese slang words.

Seq-to-Seq(with attention)	BLEU-1	BLEU-2
single character-level encoder	20.05	2.89
single character-level encoder(phonetics)	22.25	3.13
dual character-level encoder	21.62	3.12
dual character-level encoder(phonetics)	**23.64**	**3.65**

Table 1: BLEU scores for the explanation of Chinese slang words on test dataset.

➢ datasets Settings

The training dataset and test dataset used in our experiments were all from the dataset we collected. The training dataset contains 2,531 Chinese slang words and 5,689 example sentences. The test dataset contains 300 Chinese slang words, which did not exist in the training dataset, and each test example corresponds to each slang word.

■ Experimental Results and Analysis

We conducted four sets of comparison experiments, the experimental results are showed in Table 1. The first set used a single character-level encoder model as baseline on the basic model of sequence-to-sequence with attention. The second set was added phonetic vector on baseline. The third set encoded both example sentence and slang item character. The forth set used the novel model we proposed, that is, a dual character-level encoder attention-based model with phonetic features. As we can see from the Table 1, our novel dual encoder approach is superior to the other three methods.

■ Network Digital Language Experiments

Network digital language is one kind of Chinese slang, which has special meanings in Chinese. In general, network digits are not the meanings of the digits literally. For example, "520" usually refers to "我爱你 (I love you)" in Chinese. Since they often appear alone, we only use a single-encoder sequence-to-sequence model to generate the meaning of its allusion. The training dataset contains 300 pairs of parallel corpus of network digits, 86 network digits are used for test dataset, such as "520-我爱你". The results of the two groups of experiments are shown in Table 2. We can observe that the BLEU scores with phonetic features are better than none. There are several examples of model-generated explanations shown in Fig.5.

Network Digital Language: *71345*
Reference Explanation: 请你相信我
Generated Explanation(None): 请你相信我
Generated Explanation(Pinyin): 请你相信我

Network Digital Language: *995*
Reference Explanation: 救救我
Generated Explanation(None): 救我
Generated Explanation(Pinyin): 救救我

Network Digital Language: *1314*
Reference Explanation: 一生一世
Generated Explanation(None): 要生一世
Generated Explanation(Pinyin): 一生一世

Fig.5: Examples of automatically generated explanations of network digital language based on our experimental models.

Seq-to-Seq(with attention)	BLEU-1	BLEU-2
single character-level encoder	50.05	35.34
single character-level encoder(phonetics)	**54.23**	**38.17**

Table 2: BLEU scores for the explanation of network digital language on test dataset.

Conclusion

In this paper, we propose to explain Chinese slang words automatically for the first time. We use a novel dual character-level encoder using attention-based neural network to deal with this specific task. The experimental results show that our novel DCEAnn model can generate reasonable interpretations of Chinese slang words that have not appeared before, and perform better in the interpretation of network digital language. Also we construct a public dataset for training model. Now we are building an end-to-end system working from slang words discovery to slang words interpretation. In our future work, we will put the Chinese slang parallel corpus online and use transfer learning to optimize our model.