SuperChat: Dialogue Generation by Transfer Learning from Vision to Language using Two-dimensional Word Embedding

Baohua Sun, Lin Yang, Michael Lin, Wenhan Zhang, Charles Young, Jason Dong
Gyrfalcon Technology Inc.
1900 McCarthy Blvd, Milpitas, CA, US
baohua.sun@gyrfalcontech.com

Abstract

The recent work of Super Characters method using two-dimensional word embedding achieved state-of-the-art results in text classification tasks, showcasing the promise of this new approach. This paper borrows the idea of Super Characters method and two-dimensional embedding, and proposes a method of generating conversational response for open domain dialogues. The proposed method is language-independent in training and inference, because the text of language is embedded into images. The experimental results on a public dataset shows that the proposed SuperChat method generates high quality responses. An interactive demo is ready to show at the workshop. And code will be available at github soon.

1 Introduction

Dialogue systems are important to enable machine to communicate with human through natural language. Given an input sentence, the dialogue system outputs the response sentence in a natural way which reads like human-talking. Previous work adopts an encoder-decoder, and also the improved architectures with attention scheme added. In architectures with attention, the input sentence are encoded into vectors first, and then the encoded vectors are weighted by the attention score to get the context vector. The concatenation of the context vector and the previous output vector of the decoder, is fed into the decoder to predict the next words iteratively. Generally, the encoded vectors, the context vector, and the decoder output vector are all one-dimensional embedding, i.e. an array of real-valued numbers. The models used in decoder and encoder usually adopt RNN networks, such as bidirectional GRU, and bidirectional LSTM. However, the time complexity of the encoding part is very expensive.

The recent work of Super Characters method has obtained state-of-the-art result for text classification on benchmark datasets in different languages, including English, Chinese, Japanese, and Korean. The Super Characters method is a two-step method. The training method and inference steps are both language-independent. In the first step, the characters of the input text are drawn onto a blank image. And the resulting image is called a Super Characters image. In the second step, Super Characters images are fed into two-dimensional CNN models for classification.

In this paper, we propose the SuperChat method for dialogue generation using two-dimensional embedding. It has no encoding phase, but only has the decoding phase. The decoder is fine-tuned from the pre-trained CNN models in the ImageNet competition. For each iteration of the decoding, the raw text image from both the input and the partial response are fed into the decoder, without any compression into a concatenated vector as done in the encoding phase in the previous

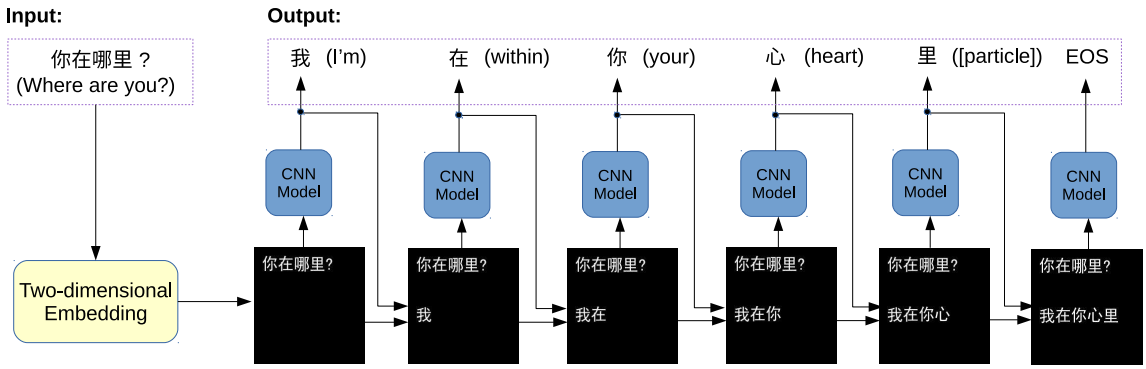

Figure 1: SuperChat method illustration. The input Chinese sentence means "Wher are you?" in English, and the output given by the proposed SuperChat method is "I am within your heart".

work.

2 The Proposed SuperChat Method

The proposed SuperChat method is motivated by the two-dimensional embedding used in the Super Characters method. If the Super Characters method could keep the same good performance when the number of classes in the text classification problem becomes even larger, e.g. the size of dialogue vocabulary, then the Super Characters method should be able to address the task of conversational dialogue generation. This can be done by treating the input sentence and the partial response sentence as one combined text input.

Figure 1 illustrates the proposed SuperChat method. The response sentence is predicted sequentially by predicting the next response word in multiple iterations. During each iteration, the input sentence and the current partial response sentence are embedded into an image through two-dimensional embedding. The resulting image is called as a SuperChat image. And then this SuperChat image is fed into a CNN model to predict the next response word. In each SuperChat image, the upper portion corresponds to the input sentence, and the lower portion corresponds to the partial response sentence. At the beginning of the iteration, the partial response sentence is initialized as null. The prediction of the first response word is based on the SuperChat image with only the input sentence embedded, and then the predicted word is added to the current partial response sentence. This iteration continues until End Of Sentence (EOS) appeared. Then, the final output would be a concatenation of the sequential output.

The CNN model used in this method is fine-tuned from pre-trained ImageNet models to predict the next response word with the generated SuperChat image as input. It can be trained end-to-end using large dialogue corpus. Thus the problem of predicting the next response word in dialogue generation is converted into an image classification problem.

The training data is generated by labeling each SuperChat image as an example of the class indicated by its next response word. EOS is labeled to the SuperChat image if the response sentence is finished. The dataset used is simsimi[1]. This is a Chinese chitchat database. This data set contains 454,561 dialogue pairs.

The experimental results of the proposed SuperChat method show high quality response. An interactive demonstration will be shown at the workshop.

[1] https://github.com/fate233/dgk_lost_conv/blob/master/results/xiaohuangji50w_nofenci.conv.zip