Diversifying Neural Conversation Model with Maximal Marginal Relevance

Yiping Song,[1] **Zhiliang Tian,**[2] **Dongyan Zhao,**[2] **Ming Zhang,**[1*] **Rui Yan**[2*]
Institute of Network Computing and Information Systems, Peking University, China
Institute of Computer Science and Technology, Peking University, China
{songyiping,zhaody,mzhang_cs,ruiyan}@pku.edu.cn
tianzhilianghit@gmail.com *Corresponding authors

Abstract

Neural conversation systems, typically using sequence-to-sequence (seq2seq) models, are showing promising progress recently. However, traditional seq2seq suffer from a severe weakness: during beam search decoding, they tend to rank universal replies at the top of the candidate list, resulting in the lack of diversity among candidate replies. *Maximum Marginal Relevance* (MMR) is a ranking algorithm that has been widely used for subset selection. In this paper, we propose the MMR-BS decoding method, which incorporates MMR into the beam search (BS) process of seq2seq . The MMR-BS method improves the diversity of generated replies without sacrificing their high relevance with the user-issued query. Experiments show that our proposed model achieves the best performance among other comparison methods.

1 Introduction

Conversation systems are of growing importance since they enable a smooth interaction interface between humans and computers: using natural language (Yan et al., 2016b). Generally speaking, there are two main categories of conversation systems: the retrieval-based (Yan et al., 2016a,b; Song et al., 2016) and the generation-based (Serban et al., 2016b; Shang et al., 2015; Serban et al., 2016a) conversation systems. In this paper, we focus on the generation-based conversation systems, which are more flexible and extensible compared with the retrieval-based ones.

The sequence-to-sequence neural network (seq2seq) (Sutskever et al., 2014) is a prevailing approach in generation-based conversation systems (Shang et al., 2015). It uses a recurrent neural network (RNN) to encode the source sentence into a vector, then uses another RNN to decode the target sentence word by word. Long short term memory (LSTM) (Hochreiter and Schmidhuber, 1997) and gated recurrent units (GRUs) (Cho et al., 2014) could further enhance the RNNs to model longer sentences. In the scenarios of generation-based conversation systems, the training criterion of seq2seq is to maximize the likelihood of the generated replies given the user-issued queries.

As is well known, the generation-based conversation systems suffer from the problem of *universally replies*, which contain less information, such as "I don't know" and "something" (Mou et al., 2016; Mrkšić et al., 2015). According to Li et al., 0.45% generated replies contain the sequence "I don't know." During the interaction between the user and the system, the user may expect more informative and diverse utterances with various expressions. The lack of diversity is one of the bottlenecks of the generation-based conversation systems. Moreover, the quality of generated replies, namely the high relevance between queries and replies, could not be obliterated when trying to improve the diversity.

In this paper, We propose the MMR-BS model to tackle the problem of diversity in the generation-based conversation systems. *Maximum Marginal Relevance* (MMR) (Jaime and Goldstein, 1998; Wang et al., 2009; Yang et al., 2007) has been widely applied in diversity modeling tasks, such as information retrieval (Stewart and Carbonell, 1998), document summarization (Zhou, 2011) and text categorization (He et al., 2012). It scores each candidate by properly measuring them in terms of quality and diversity and selects the current best candidate item at each time step. These properties make it suitable for the sub-

Proceedings of the The 8th International Joint Conference on Natural Language Processing, pages 169–174,
Taipei, Taiwan, November 27 – December 1, 2017