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

sequences choosing in the reply generation process. Hence, we incorporate MMR into the decoding process of *Beam Search* (BS) in `seq2seq`.

To demonstrate the effectiveness of MMR-BS, we evaluate our method in terms of both *quality* and *diversity*. Enhanced with MMR, the MMR-BS model can generate more meaningful replies than other baselines, as we shall show in the experiments.

2 Preliminaries

2.1 `seq2seq` Model

`seq2seq` encodes the user-issued query q using an RNN, and decodes a corresponding reply r with another RNN. At each time step of decoding, the RNN estimates a probabilistic distribution over the vocabulary. The objective function of `seq2seq` is the log-likelihood of reply r given the query q:

$$\hat{r} = \underset{r}{\operatorname{argmax}}\{\log(p(r|q))\} \quad (1)$$

We use the attention mechanism (Bahdanau et al., 2015) to better align input and output sentences and use gated recurrent units (GRUs) to enhance RNNs' ability to handle long sentences.

2.2 Beam Search

Beam search is a prevalent decoding method in `seq2seq` (Vijayakumar et al., 2016), which maintains a set of candidate subsequences at every step of decoding process. At a time step t, we keep N subsequences based on their cumulative probabilities. At the time $t+1$, each subsequence is appended with a word from the entire vocabulary, resulting in a larger candidate set of subsequences. Then we keep the top-N sequences in the same manner. A candidate sequence terminates when RNN predicts EOS, the special symbol indicating the end of a sequence. Let $S(y_1, \cdots, y_t|q)$ be a function that scores a subsequence $\{y_1, \cdots, y_t\}$ given a query q. The original beam search chooses N most probable replies, i.e., $S(\cdot)$ is the logarithmic probability, given by

$$S(y_t|q) = S(y_{t-1}) + \log p(y_t|q, y_1, \cdots, y_{t-1}) \quad (2)$$

3 Diverse Neural Conversation

3.1 MMR-BS Model

The beam search criterion is mainly based on the conditional probabilities of replies given the query. Universal replies, which have relatively higher probabilities, are likely to appear at the top of the candidate list, resulting in the lack of diversity among top replies. To handle the influence of the replies' own probabilities and address the relevance with the query at the same time, we propose MMR-BS, which applies the MMR criterion to every decoding step of beam search to upturn diverse subsequences. The whole architecture of MMR-BS is illustrated in Figure 1.

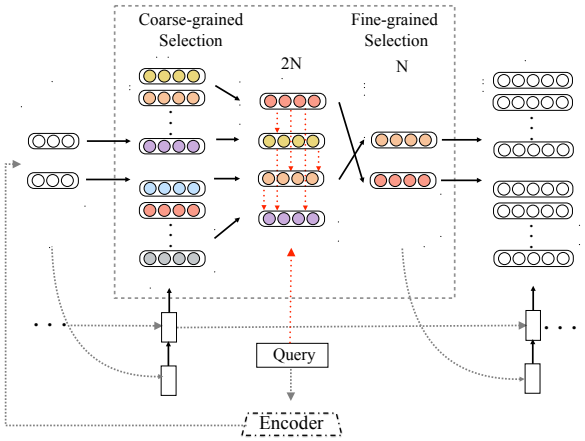

Figure 1: The architecture of MMR-BS.

Specifically, the decoding process maintains a subsequence list S; the number of subsequences in S is N. At each time of decoding, every subsequence is appended with a word from the entire vocabulary V, resulting in $N * |V|$ subsequences. Since only N subsequences would be passed into next time step for further generation, our MMR-BS model uses two granularities of selection, which is performed in two-step strategy. We present the decoding process in Algorithm 1.

• **Coarse-grained Selection.** Coarse-grained selection follows the original scoring function in traditional beam search, which is described in equation 2. This selection strategy conditions on the probabilities of subsequences given the query, which represents the relevance between each subsequence and the query. We use coarse-grained selection to select $2N$ subsequences noted as S^{2N}.

• **Fine-grained Selection.** However, coarse-grained selection focuses on the quality of subsequences but ignores the diversity among them. The fine-grained selection adopts the MMR scoring to balance quality and diversity in the selecting process. It maintains a selected list S^s and continually adds the highest scored subsequence into S^s from the remaining candidates, i.e., $S_t^{2N} \backslash S_t^s$. This process repeats N times, resulting in N best subsequences. The scoring function of MMR con-

Algorithm 1: MMR-BS Decoding

Input: the user-issued query q, the max length of reply l, λ in MMR function
Output: generated reply set R

$R = \varnothing$;
$S_0 = \varnothing$; **for** $t = 1; t \leq l$; **do**
  // obtain the subsequence set at time i
  $S_t = Decoding(q, S_{t-1})$;
  // coarse-grained selection
  $S_t^{2N} = Normal\ Ranking(S_t)$;
  // fine-grained selection
  $S_t^s = \varnothing$; **for** $i = 1; i \leq N$; **do**
    $max = -\infty$;
    forall $s_j \in S_t^{2N} \setminus S_t^s$ **do**
      $score = \lambda \mathrm{sim}_{qua}(s_j, q) - (1-\lambda)\mathrm{sim}_{div}(s_j, S_t^s)$;
      if $score > max$ **then**
        $max = score$;
        $best_s = s_j$;
    $S_t^s = S_t^s \cup best_s$;
$R = S_l^s$;
return R

siders the quality of a candidate as well as its diversity against previously selected ones. In particular, we have two metrics: $\mathrm{sim}_{qua}(s_i, q)$ measures the similarity of a candidate subsequence s_i with respect to the query q, indicating the quality of s_i. $\mathrm{sim}_{div}(s_i, S^s)$ measures the similarity of s_i against other replies in the selected list; $-\mathrm{sim}_{div}(\cdot)$ indicates the diversity.

MMR chooses the next candidate s^* such that

$$s^* = \underset{s_i \in S^{2N} \setminus S^s}{\mathrm{argmax}} \left[\lambda \mathrm{sim}_{qua}(s_i, q) - (1-\lambda)\mathrm{sim}_{div}(s_i, S^s)\right] \quad (3)$$

where λ is a hyper-parameter balancing these two aspects. Thus the fine-grained selection improves the diversity and retains the quality of subsequences at each time of decoding, so the generated replies are of good quality and diversity.

3.2 Quality and Diversity Metrics

The fine-grained selection allows explicit definition of quality and diversity measurements, which are presented in this section.

Quality Metric. The semantic coherence with the query, which is based on the word-level similarity, defines the quality of each candidate subsequence. For each word in the query, we find the best matching word in the subsequence using the cosine similarity of the word embeddings (Mikolov et al., 2015, 2013). Then we sum over all the similarity scores as the final quality score given by

$$\mathrm{sim}_{qua}(s_i, q) = \frac{1}{|q|} \sum_{w_i \in q} \underset{w_j \in s_i}{\mathrm{argmax}} \cos(\boldsymbol{e}_{w_i}, \boldsymbol{e}_{w_j}) \quad (4)$$

where $\boldsymbol{e}_{\boldsymbol{w_i}}$ refers to the embedding of word w_i.

Diversity Metric. The diversity score of a subsequence measures its differences against existing subsequences in the selected set S^s by the word overlapping. We represent a subsequence s_i as a vector and measure the similarity by the cosine score; the average indicates overall diversity,

$$\mathrm{sim}_{div}(s_i, S^s) = \frac{1}{|S^s|} \sum_{s_j \in S^s} \cos(\boldsymbol{s}_i, \boldsymbol{s}_j) \quad (5)$$

where $\boldsymbol{s}_i$ is a binary word vector, each element indicating if a word appears in s_i; the vector length is the number of words in s_i and s_j. Notice that, for diversity, we use binary word vectors instead of embeddings to explicitly avoid word overlap among (top-ranked) candidate subsequences.

4 Experiments

4.1 Dataset

We evaluated each approach on a massive Chinese conversation dataset crawled from Baidu Tieba[1]. There were 1,600,000 query-reply pairs for training, 2000 pairs for validation, and another unseen 2000 pairs for testing. We performed standard Chinese word segmentation.

4.2 Experimental Setups

All the methods are established on the base of the traditional `seq2seq` with same settings. In our study, word embeddings were 610d and hidden layers were 1000d, following the settings in Shang et al. We applied AdaDelta with default hyper-parameters. We kept 100k words (Chinese terms) for both queries and replies, but 30k for the decoder's output due to efficiency concerns. λ in MMR scores was empirically set to 0.5; the beam size was 30.

4.3 Algorithms for Comparison

• **Beam Search (BS).** The standard beam search in `seq2seq` which acts as the baseline.

[1] http://tieba.baidu.com

Method	Top-1		Top-5		Top-10		Quality
	BLEU-1	BLEU-2	BLEU-1	BLEU-2	BLEU-1	BLEU-2	
BS	0.679	0.254	1.803	0.555	2.959	0.980	0.703
DD	0.790	0.192	1.893	0.480	2.991	0.802	0.727
DBS	0.358	0.111	1.123	0.224	2.264	0.401	0.553
MMR-BS	**2.626**	**0.802**	**5.154**	**1.270**	**6.672**	**2.019**	**0.791**

Table 1: Results of quality evaluation. Inter-annotator agreement for human annotation: Fleiss' $\kappa = 0.5698$ (Fleiss, 1971), std $= 0.3453$.

Method	Top-1				Top-5				Top-10			
	distinct-1	distinct-2	distinct-3	distinct-4	distinct-1	distinct-2	distinct-3	distinct-4	distinct-1	distinct-2	distinct-3	distinct-4
BS	0.100	0.261	0.366	0.624	0.038	0.148	0.259	0.346	0.021	0.101	0.200	0.291
DD	0.130	0.333	0.489	0.623	0.047	0.191	0.334	0.448	0.027	0.134	0.263	0.377
DBS	0.113	0.321	0.495	0.649	0.056	0.206	0.371	0.524	0.036	0.171	0.334	0.487
MMR-BS	**0.152**	**0.510**	**0.725**	**0.840**	**0.063**	**0.326**	**0.600**	**0.776**	**0.037**	**0.243**	**0.517**	**0.729**

Table 2: Results of distinct scores.

Method	Top-1	Top-5	Top-10	Rates
BS	0.759	0.765	0.796	56.53%
DD	0.849	0.830	0.846	50.30%
DBS	0.901	0.897	**0.892**	45.53%
MMR-BS	**0.939**	**0.910**	0.878	**15.67%**

Table 3: Results of diverse scores and the rates of the universal replies in Top-10 reply list. Fleiss' $\kappa = 0.2540$ (Fleiss, 1971), std $= 1.563$.

• **Diverse Decoding (DD).** A work proposed by Li et al., which assigns low scores to sibling subsequences.

• **Diverse Beam Search (DBS).** A work proposed by Vijayakumar et al., which adds a similarity punishment to the scoring function.

• **MMR-BS.** The proposed model in this paper, which applies the MMR in the decoding process to select the subsequences.

4.4 Evaluation Metrics

We evaluated each method in terms of two aspects, namely the quality and the diversity. All the subjective evaluation experiments are conducted on 100 randomly sampled cases.

• **Quality Evaluation** We used BLEU scores as objective metrics to measure the coherence between the user-issued query and candidate replies, which is also used in (Li and Jurafsky, 2016; Vijayakumar et al., 2016). We calculated the BLEU scores of Top-1, Top-5 and Top-10 replies, and only display BLEU-1 and BLEU-2 scores due to the flexibility of conversation. We asked three well-educated volunteers to annotate the quality of the generated replies for each comparison method. The volunteers are asked to label each reply with a score: 0 for the improper reply, 1 for the borderline reply and 2 for the proper reply.

• **Diversity Evaluation** We used the distinct scores to measure the diversity in the generated replies, following (Li and Jurafsky, 2016; Vijayakumar et al., 2016). We also conducted the diverse scores, which is used in information retrieval to calculate the differentness between retrieved results (Zhang and Hurley, 2008),

$$\frac{2}{|R|(|R|-1)} \sum_{r_i \in R} \sum_{r_j \in R, r_i \neq r_j} (1 - \cos(\boldsymbol{r}_i, \boldsymbol{r}_j)) \tag{6}$$

where R is whole set of the generated replies and $\boldsymbol{r}_i$ is the binary word vector with same definition in Equation 5. We asked three well-educated volunteers to count the universal replies in the Top-10 reply list and calculated the rates.

4.5 Overall Performance

We presented the quality evaluation results in Table 1. DD achieves almost the same performance with the standard BS. DBS is not as good as the BS and DD. MMR-BS yields the highest BLEU scores and human annotation results, which indicates the effectiveness of the quality measurement.

We presented the diversity evaluation results in Table 2 and Table 3. BS achieves the worst performance as it does not consider about the diversity during the decoding process. DD is better than BS but not as good as DBS. DBS shows a good performance in terms of all the diversity evaluation and even outperforms MMR-BS in the Top 10 diverse score. MMR-BS outperforms all the other methods in most metrics. Compared with BS, it decreases the number of universal replies by 3 times, which is a significant improvement.

It is obvious that MMR-BS yields the highest quality and diverse scores. Compared with BS,

DD does not improve the quality very much but indeed fosters the diversity among the generated replies. DBS achieves a good diversity performance but is still worse than MMR-BS. As DBS does not perform well in quality, we can see that it sacrifices the quality to increase the diversity.

5 Related Work

To tackle diversity problem in generation-based systems, Li et al. propose a diverse decoding method[2] that avoids choosing sibling subsequences during decoding (Li and Jurafsky, 2016). Vijayakumar et al. propose a diverse beam search, which divides the subsequence into several groups during selection. These methods add a diversity punishment term to the scoring function in beam search; it is hard to balance this term with other components in the function.

MMR is widely used in information retrieval (Stewart and Carbonell, 1998), document summarization (Zhou, 2011), and text categorization (He et al., 2012). MMR allows an explicit definition of both quality and diversity, and linearly combines these two aspects. This property fits the requirements in subsequence selection in beam search where the candidate subsequences should be different from each other and retain the high coherence with the user-issued query at the same time.

6 Conclusions

In this paper, we propose an MMR-BS method to tackle the problem of diversity in generative conversation systems. MMR-BS deploys two granularities of subsequence selection during the decoding process. The first one continues to use the original scoring function, and the second one takes advantage of MMR to measure each subsequence, considering both the quality and diversity. The experimental results demonstrate the effectiveness of our method.

Acknowledgments

This paper is partially supported by the National Natural Science Foundation of China (NSFC Grant Nos. 61772039 and 91646202) and CCF-Tencent Open Research Fund.

[2]Originally, the approach is proposed for machine translation, but it applies to conversation systems naturally.

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. *ICLR*.

Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*.

Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378.

Liu He, Xianghong Zhang, Dayou Liu, Yanjun Li, and Lijun Yin. 2012. A feature selection method based on maximal marginal relevance. *Journal of Computer Research and Development*, pages 354–360.

Sepp Hochreiter and Jurgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Carbonell Jaime and Jade Goldstein. 1998. The use of MMR, diversity-based reranking for reordering documents and producing summaries. In *SIGIR*, pages 335–336.

Jiwei Li and Dan Jurafsky. 2016. Mutual information and diverse decoding improve neural machine translation. *arXiv preprint arXiv:1601.00372*.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2015. Distributed representations of words and phrases and their compositionality. *The Workshop on Vector Space Modeling for Natural Language Processing*, pages 8–16.

Lili Mou, Yiping Song, Rui Yan, Ge Li, Lu Zhang, and Zhi Jin. 2016. Sequence to backward and forward sequences: A content-introducing approach to generative short-text conversation. *arXiv preprint arXiv:1607.00970*.

Nikola Mrkšić, Diarmuid Ó Séaghdha, Blaise Thomson, Milica Gašić, Pei Hao Su, David Vandyke, Tsung Hsien Wen, and Steve Young. 2015. Multi-domain dialog state tracking using recurrent neural networks. In *ACL-IJCNLP*, pages 794–799.

Iulian V Serban, Alessandro Sordoni, Yoshua Bengio, Aaron Courville, and Joelle Pineau. 2016a. Building end-to-end dialogue systems using generative hierarchical neural network models. In *AAAI*, pages 3776–3783.

Iulian Vlad Serban, Alessandro Sordoni, Ryan Lowe, Laurent Charlin, Joelle Pineau, Aaron Courville, and Yoshua Bengio. 2016b. A hierarchical latent variable encoder-decoder model for generating dialogues. *arXiv preprint arXiv:1605.06069*.

Lifeng Shang, Zhengdong Lu, and Hang Li. 2015. Neural responding machine for short-text conversation. In *ACL-IJCNLP*, pages 1577–1586.

Yiping Song, Lili Mou, Rui Yan, Li Yi, Zinan Zhu, Xiaohua Hu, and Ming Zhang. 2016. Dialogue session segmentation by embedding-enhanced TextTiling. In *INTERSPEECH*, pages 2706–2710.

Jade Goldstein Stewart and Jaime G. Carbonell. 1998. The use of mmr and diversity-based reranking in document reranking and summarization. *Proceedings of the 14th Twente Workshop on Language Technology in Multimedia Information Retrieval.*, pages 335–336.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *NIPS*, pages 3104–3112.

Ashwin K Vijayakumar, Michael Cogswell, Ramprasath R Selvaraju, Qing Sun, Stefan Lee, David Crandall, and Dhruv Batra. 2016. Diverse beam search: Decoding diverse solutions from neural sequence models. *arXiv preprint arXiv:1610.02424*.

Baoxun Wang, Bingquan Liu, Chengjie SunXiao, and long Wang ad Bo Li. 2009. Adaptive maximum marginal relevance based multi-email summarization. *Artificial Intelligence and Computational Intelligence.*, pages 417–424.

Rui Yan, Yiping Song, and Hua Wu. 2016a. Learning to respond with deep neural networks for retrieval-based human-computer conversation system. In *SIGIR*, pages 55–64.

Rui Yan, Yiping Song, Xiangyang Zhou, and Hua Wu. 2016b. Shall I be your chat companion?: Towards an online human-computer conversation system. *ACM International on Conference on Information and Knowledge Management*, pages 649–658.

Lingpeng Yang, Donghong Ji, and Munkew Leong. 2007. Document reranking by term distribution and maximal marginal relevance for chinese information retrieval. *Information Processing and Management*, 43(2):315–326.

Mi Zhang and Neil Hurley. 2008. Avoiding monotony: Improving the diversity of recommendation lists. In *RecSys*, pages 123–130.

Zhigang Zhou. 2011. Combined features to maximal marginal relevance algorithm for multi-document summarization. *Journal of Convergence Information Technology*, pages 298–304.