A Dataset for Research on Short-Text Conversation *

Hao Wang[§] **Zhengdong Lu**[‡] **Hang Li**[‡] **Enhong Chen**[§]
[§] xdwangh@mail.ustc.edu.cn [‡]lu.zhengdong@huawei.com
[‡]hangli.hl@huawei.com [§]cheneh@ustc.edu.cn
[§]Univ. of Sci & Tech of China, China [‡]Noah's Ark Lab, Huawei Technologies, Hong Kong

Abstract

Natural language conversation is widely regarded as a highly difficult problem, which is usually attacked with either rule-based or learning-based models. In this paper we propose a retrieval-based automatic response model for short-text conversation, to exploit the vast amount of short conversation instances available on social media. For this purpose we introduce a dataset of short-text conversation based on the real-world instances from Sina Weibo (a popular Chinese microblog service), which will be soon released to public. This dataset provides rich collection of instances for the research on finding natural and relevant short responses to a given short text, and useful for both training and testing of conversation models. This dataset consists of both naturally formed conversations, manually labeled data, and a large repository of candidate responses. Our preliminary experiments demonstrate that the simple retrieval-based conversation model performs reasonably well when combined with the rich instances in our dataset.

1 Introduction

Natural language conversation is one of the holy grail of artificial intelligence, and has been taken as the original form of the celebrated Turing test. Previous effort in this direction has largely focused on analyzing the text and modeling the state of the conversation through dialogue models, while in this paper we take one step back and focus on a much easier task of finding the response for a given short text. This task is in clear contrast with previous effort in dialogue modeling in the following two aspects

- we do not consider the context or history of conversations, and assume that the given short text is self-contained;
- we only require the response to be natural, relevant, and human-like, and do not require it to contain particular opinion, content, or to be of particular style.

This task is much simpler than modeling a complete dialogue session (e.g., as proposed in Turing test), and probably not enough for real conversation scenario which requires often several rounds of interactions (e.g., automatic question answering system as in (Litman et al., 2000)). However it can shed important light on understanding the complicated mechanism of the interaction between an utterance and its response. The research in this direction will not only instantly help the applications of short session dialogue such as automatic message replying on mobile phone and the chatbot employed in voice assistant like Siri[1], but also it will eventually benefit the modeling of dialogues in a more general setting.

Previous effort in modeling lengthy dialogues focused either on rule-based or learning-based models (Carpenter, 1997; Litman et al., 2000; Williams and Young, 2007; Schatzmann et al., 2006; Misu et al., 2012). This category of approaches require relatively less data (e.g. reinforcement learning based) for

*The work is done when the first author worked as intern at Noah's Ark Lab, Huawei Techologies.

[1]http://en.wikipedia.org/wiki/Siri

Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, pages 935–945,
Seattle, Washington, USA, 18-21 October 2013.

training or no training at all, but much manual effort in designing the rules or the particular learning algorithms. In this paper, we propose to attack this problem using an alternative approach, by leveraging the vast amount of training data available from the social media. Similar ideas have appeared in (Jafarpour and Burges, 2010; Leuski and Traum, 2011) as an initial step for training a chatbot.

With the emergence of social media, especially microblogs such as Twitter, in the past decade, they have become an important form of communication for many people. As the result, it has collected conversation history with volume previously unthinkable, which brings opportunity for attacking the conversation problem from a whole new angle. More specifically, instead of generating a response to an utterance, we pick a massive suitable one from the candidate set. The hope is, with a reasonable retrieval model and a *large enough* candidate set, the system can produce fairly natural and appropriate responses.

This retrieval-based model is somewhat like non-parametric model in machine learning communities, which performs well only when we have abundant data. In our model, it needs only a relatively small *labeled* dataset for training the retrieval model, but requires a rather large *unlabeled* set (e.g., one million instances) for candidate responses. To further promote the research in similar direction, we create a dataset for training and testing the retrieval model, with a candidate responses set of reasonable size. Sina Weibo is the most popular Twitter-like microblog service in China, hosting over 500 million registered users and generating over 100 million messages per day [2]. As almost all microblog services, Sina Weibo allows users to comment on a published post[3], which forms a natural one-round conversation. Due to the great abundance of those (post, response) pairs, it provides an ideal data source and test bed for one-round conversation. We will make this dataset publicly available in the near future.

[2] http://en.wikipedia.org/wiki/Sina_Weibo

[3] Actually it also allows users to comment on other users' comments, but we will not consider that in the dataset.

2 The Dialogues on Sina Weibo

Sina Weibo is a Twitter-like microblog service, on which a user can publish short messages (will be referred to as *post* in the remainder of the paper) visible to public or a group specified by the user. Similar to Twitter, Sina Weibo has the word limit of 140 Chinese characters. Other users can comment on a published post, with the same length limit, as shown in the real example given in Figure 6 (in Chinese). Those comments will be referred to as *responses* in the remainder of the paper.

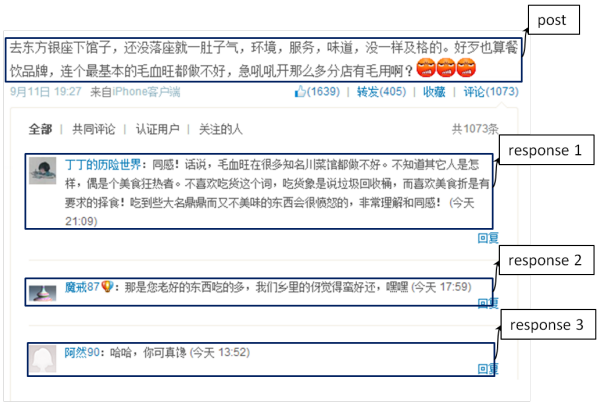

Figure 1: An example of Sina Weibo post and the comments it received.

We argue that the (post, response) pairs on Sina Weibo provide rather valuable resource for studying *one round* dialogue between users. The comments to a post can be of rather flexible forms and diverse topics, as illustrated in the example in Table 1. With a post stating the user's status (traveling to Hawaii), the comments can be of quite different styles and contents, but apparently all appropriate.

In many cases, the (post, response) pair is self-contained, which means one does not need any background and contextual information to get the main point of the conversation (Examples of that include the responses from **B, D, G** and **H**). In some cases, one may need extra knowledge to understand the conversation. For example, the response from user **E** will be fairly elusive if taken out of the context that **A**'s Hawaii trip is for an international conference and he is going to give a talk there. We argue that the number of self-contained (post, response) pairs is vast, and therefore the extracted (post, re-

Post	
User A:	*The first day at Hawaii. Watching sunset at the balcony with a big glass of wine in hand.*
Responses	
User B:	*Enjoy it & don't forget to share your photos!*
User C:	*Please take me with you next time!*
User D:	*How long are you going to stay there?*
User E:	*When will be your talk?*
User F:	*Haha, I am doing the same thing right now. Which hotel are you staying in?*
User G:	*Stop showing-off, buddy. We are still coding crazily right now in the lab.*
User H:	*Lucky you! Our flight to Honolulu is delayed and I am stuck in the airport. Chewing French fries in MacDonald's right now.*

Table 1: A typical example of Sina Weibo post and the comments it received. The original text is in Chinese, and we translated it into English for easy access of readers. We did the same thing for all the examples throughout this paper.

sponse) pairs can serve as a rich resource for exploring rather sophisticated patterns and structures in natural language conversation.

3 Content of the Dataset

The dataset consists of three parts, as illustrated in Figure 2. Part 1 contains the original (post, response) pairs, indicated by the dark-grey section in Figure 2. Part 2, indicated by the light-gray section in Figure 2, consists labeled (post, response) pairs for some Weibo posts, including positive and negative examples. Part 3 collects all the responses, including but not limited to the responses in Part 1 and 2. Some of the basic statistics are summarized in Table 2.

# posts	# responses	vocab.	# labeled pairs
4,6345	1,534,874	105,732	12,427

Table 2: Some statistics of the dataset

Original (Post, Response) Pairs This part of dataset gives (post, response) pairs naturally presented in the microblog service. In other words, we create a (post, response) pair there when the response is actually given to the post in Sina Weibo. The part of data is noisy since the responses given to a Weibo post could still be inappropriate for different reasons, for example, they could be spams or targeting some responses given earlier. We have $628,833$ pairs.

Labeled Pairs This part of data contains the (post, response) pairs that are labeled by human. Note that 1) the labeling is only on a small subset of posts, and 2) for each selected post, the labeled responses are not originally given to it. The labeling is done in an active manner (see Section 4 for more details), so the obtained labels are much more informative than the those on randomly selected pairs (over 98% of which are negative). This part of data can be directly used for training and testing of retrieval-based response models. We have labeled 422 posts and for each of them, about 30 candidate responses.

Responses This part of dataset contains only responses, but they are not necessarily for a certain post. These extra responses are mainly filtered out by our data cleaning strategy (see Section 4.2) for original (post, response) pairs, including those from filtered-out Weibo posts and those addressing other responses. Nevertheless, those responses are still valid candidate for responses. We have about 1.5 million responses in the dataset.

3.1 Using the Dataset for Retrieval-based Response Models

Our data can be used for training and testing of retrieval-based response model, or just as a bank of responses. More specifically, it can be used in at least the following three ways.

Training Low-level Matching Features The rather abundant original (post, response) pairs provide rather rich supervision signal for learning different matching patterns between a post and a response. These matching patterns could be of dif-

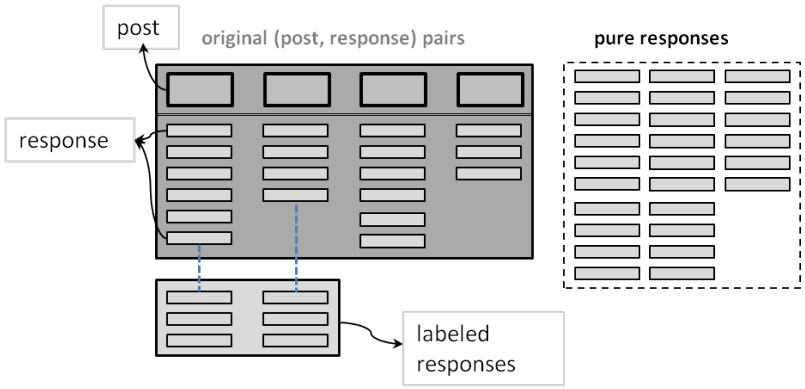

Figure 2: Content of the dataset.

ferent levels. For example, one may discover from the data that when the word "Hawaii" occurs in the post, the response are more likely to contain words like "trip", "flight", or "Honolulu". On a slightly more abstract level, one may learn that when an entity name is mentioned in the post, it tends to be mentioned again in the response. More complicated matching pattern could also be learned. For example, the response to a post asking "how to" is statistically longer than average responses. As a particular case, Ritter et al. (2011) applied translation model (Brown et al., 1993) on similar parallel data extracted from Twitter in order to extract the word-to-word correlation. Please note that with more sophisticated natural language processing, we can go beyond bag-of-words for more complicated correspondence between post and response.

Training Automatic Response Models Although the original (post, response) pairs are rather abundant, they are not enough for discriminative training and testing of retrieval models, for the following reasons. In the labeled pairs, both positive and negative ones are ranked high by some baseline models, and hence more difficult to tell apart. This supervision will naturally tune the model parameters to find the real good responses from the seemingly good ones. Please note that without the labeled negative pairs, we need to generate negative pairs with randomly chosen responses, which in most of the cases are too easy to differentiate by the ranking model and cannot fully tune the model parameters. This intuition has been empirically verified by our experiments.

Testing Automatic Response Models In testing a retrieval-based system, although we can simply use the original responses associated with the query post as positive and treat all the others as negative, this strategy suffers from the problem of spurious negative examples. In other words, with a reasonably good model, the retrieved responses are often good even if they are not the original ones, which brings significant bias to the evaluation. With the labeled pairs, this problem can be solved if we limit the testing only in the small pool of labeled responses.

3.2 Using the Dataset for Other Purposes

Our dataset can also be used for other researches related to short-text conversations, namely anaphora resolution, sentiment analysis, and speech act analysis, based on the large collection of original (post, response) pairs. For example, to determine the sentiment of a response, one needs to consider both the original post as well as the observed interaction between the two. In Figure 3, if we want to understand user's sentiment towards the "invited talk" mentioned in the post, the two responses should be taken as positive, although the sentiment in the mere responses is either negative or neutral.

4 Creation of the Dataset

The (post, comment) pairs are sampled from the Sina Weibo posts published by users in a loosely connected community and the comments they received (may not be from this community). This community is mainly posed of professors, researchers, and students of natural language processing (NLP) and related areas in China, and the users

Query Post:	十点半主楼有XXX博士的关于深度学习的讲座，有兴趣的同学不要错过！ *There is a talk on deep learning given by Dr. XXX on deep learning in the main building. Come if you are interested*
Response1:	太悲催了，我十点钟有节课 *Damn it! I have a class at 10 am*
Response2:	到底主楼哪个房间啊？ *Come on, which room in the main building?*

Figure 3: An example (original Chinese and the English translation) on the difficulty of sentiment analysis on responses.

commonly followed them.

The creation process of the dataset, as illustrated in Figure 4, consists of three consecutive steps: 1) crawling the community of users, 2) crawling their Weibo posts and their responses, 3) cleaning the data, with more details described in the remainder of this section.

4.1 Sampling Strategy

We take the following sampling strategy for collecting the (post, response) pairs to make the topic relatively focused. We first locate 3,200 users from a loosely connected community of Natural Language Processing (NLP) and Machine Learning (ML) in China. This is done through crawling followees[4] of ten manually selected seed users who are NLP researchers active on Sina Weibo (with no less than 2 posts per day on average) and popular enough (with no less than 100 followers).

We crawl the posts and the responses they received (not necessarily from the crawled community) for two months (from April 5th, 2013, to June 5th, 2013). The topics are relatively limited due to our choice of the users, with the most saliently ones being:

- **Research:** discussion on research ideas, papers, books, tutorials, conferences, and researchers in NLP and machine learning, etc;
- **General Arts and Science:** mathematics, physics, biology, music, painting, etc;
- **IT Technology:** Mobile phones, IT companies, jobs opportunities, etc;
- **Life:** traveling (both touring or conference trips), food, photography, etc.

[4]When user A follows user B, A is called B's follower, and B is called A's followee.

4.2 Processing, Filtering, and Data Cleaning

On the crawled posts and responses, we first perform a four-step filtering on the post and responses

- We first remove the Weibo posts and their responses if the length of post is less than 10 Chinese characters or the length of the response is less than 5 characters. The reason for that is two-fold: 1) if the text is too short, it can barely contain information that can be reliably captured, e.g. the following example

P:	*Three down, two to go.*

  and 2) some of the posts or responses are too general to be interesting for other cases, e.g. the response in the example below,

P:	*Nice restaurant. I'd strong recommend it. Everything here is good except the long waiting line*
R:	*wow.*

- In the remained posts, we only keep the first 100 responses in the original (post, response) pairs, since we observe that after the first 100 responses there will be a non-negligible proportion of responses addressing things other than the original Weibo post (e.g., the responses given earlier). We however will still keep the responses in the bank of responses.
- The last step is to filter out the potential advertisements. We will find the long responses that have been posted more than twice on different posts and scrub them out of both original (post, response) pairs and the response repository.

For the remained posts and responses, we remove the punctuation marks and emoticons, and use ICTCLAS (Zhang et al., 2003) for Chinese word segmentation.

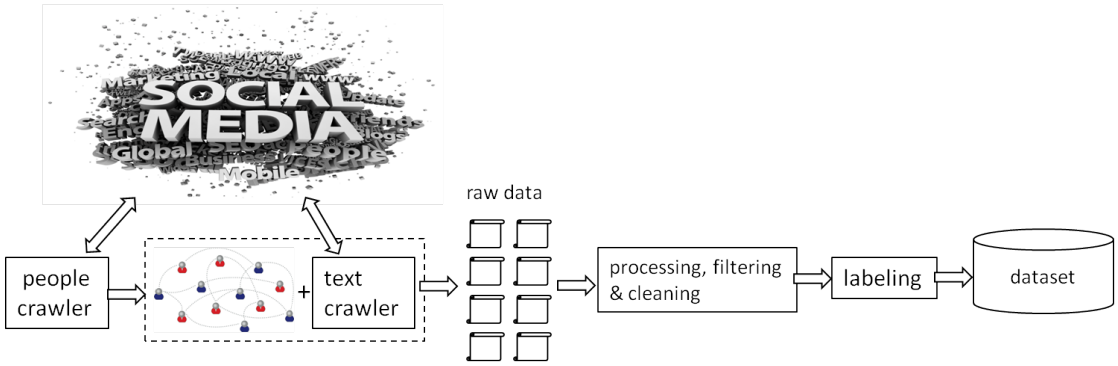

Figure 4: Diagram of the process for creating the dataset.

4.3 Labeling

We employ a pooling strategy widely used in information retrieval for getting the instance to label (Voorhees, 2002). More specifically, for a given post, we use three baseline retrieval models to each select 10 responses (see Section 5 for the description of the baselines), and merge them to form a much reduced candidate set with size ≤ 30. Then we label the reduced candidate set into "suitable" and "unsuitable" categories. Basically we consider a response suitable for a given post if we cannot tell whether it is an original response. More specifically the suitability of a response is judged based on the following three criteria[5]:

Semantic Relevance: This requires the content of the response to be semantically relevant to the post. As shown in the example right below, the post **P** is about soccer, and so is response **R1** (hence semantically relevant), whereas response **R2** is about food (hence semantically irrelevant).

P:	*There are always 8 English players in their own penalty area. Unbelievable!*
R1:	*Haha, it is still 0:0, no goal so far.*
R2:	*The food in England is horrible.*

Another important aspect of semantic relevance is the entity association. This requires the entities in the response to be correctly aligned with those in the post. In other words, if the post is about entity A, while the response is about entity B, they are very likely to be mismatched. As shown in the following example, where the original post is about Paris, and the response **R2** talks about London:

P:	*It is my last day in Paris. So hard to say goodbye.*
R1:	*Enjoy your time in Paris.*
R2:	*Man, I wish I am in London right now.*

This is however not absolute, since a response containing a different entity could still be sound, as demonstrated by the following two responses to the post above

R1:	*Enjoy your time in France.*
R2:	*The fall of London is nice too.*

Logic Consistency: This requires the content of the response to be logically consistent with the post. For example, in the table right below, post **P** states that the Huawei mobile phone "Honor" is already in the market of mainland China. Response **R1** talks about a personal preference over the same phone model (hence logically consistent), whereas **R2** asks the question the answer to which is already clear from **P** (hence logically inconsistent).

P:	*HUAWEI's mobile phone, Honor, sells well in Chinese Mainland.*
R1:	*HUAWEI Honor is my favorite phone*
R2:	*When will HUAWEI Honor get to the market in mainland China?*

Speech Act Alignment: Another important factor in determining the suitability of a response is the

[5]Note that although our criteria in general favor short and general answers like "Well said!" or "Nice", most of these general answers have already been filtered out due to their length (see Section 4.2).

speech act. For example, when a question is posed in the Weibo post, a certain act (e.g., answering or forwarding it) is expected. In the example below, post **P** asks a special question about location. Response **R1** and **R2** either forwards or answers the question, whereas **R3** is a negative sentence and therefore does not align well in speech act.

P:	*Any one knows where KDD will be held the year after next?*
R1:	*co-ask. Hopefully Europe*
R2:	*New York, as I heard*
R3:	*No, it is still in New York City*

5 Retrieval-based Response Model

In a retrieval-based response model, for a given post x we pick from the candidate set the response with the highest ranking score, where the score is the ensemble of several individual matching features

$$\mathsf{score}(x, y) = \sum_{i \in \Omega} w_i \Phi_i(x, y). \quad (1)$$

with y stands for a candidate response.

We perform a two-stage retrieval to handle the scalability associated with the massive candidate set, as illustrated in Figure 5. In **Stage I**, the system employs several fast baseline matching models to retrieve a number of candidate responses for the given post x, forming a much reduced candidate set $\mathcal{C}_{\mathbf{x}}^{(reduced)}$. In **Stage II**, the system uses a ranking function with more and sophisticated features to further evaluate all the responses in $\mathcal{C}_{\mathbf{x}}^{(reduced)}$, returning a matching score for each response. Our response model then decides whether to respond and which candidate response to choose.

In **Stage II**, we use the linear score function defined in Equation 1 with 15 features, trained with RankSVM (Joachims, 2002). The training and testing are both performed on the 422 labeled posts, with about 12,000 labeled (post, response) pairs. We use a 5-fold cross validation with a fixed penalty parameter for slack variable. [6]

5.1 Baseline Matching Models

We use the following matching models as the baseline model for **Stage I** fast retrieval. Moreover, the matching features used in the ranking function in **Stage II** are generated, directly or indirectly, from the those matching models:

POST-RESPONSE SEMANTIC MATCHING: This particular matching function relies on a learned mapping from the original sparse representation for text to a low-dimensional but dense representation for both Weibo posts and responses. The level of matching score between a post and a response can be measured as the inner product between their images in the low-dimensional space

$$\mathsf{SemMatch}(x, y) = \mathbf{x}^\top L_{\mathcal{X}} L_{\mathcal{Y}}^\top \mathbf{y}. \quad (2)$$

where $\mathbf{x}$ and $\mathbf{y}$ are respectively the 1-in-N representations of x and y. This is to capture the semantic matching between a Weibo post and a response, which may not be well captured by a word-by-word matching. More specifically, we find $L_{\mathcal{X}}$ and $L_{\mathcal{Y}}$ through a large margin variant of (Wu et al., 2013)

$$\begin{aligned} \arg\min_{L_{\mathcal{X}}, L_{\mathcal{Y}}} \quad & \sum_i \max(1 - \sum_i \mathbf{x}_i^\top L_{\mathcal{X}} L_{\mathcal{Y}}^\top \mathbf{y}_i, 0) \\ s.t. \quad & \|L_{n,\mathcal{X}}\|_1 \le \mu_1,\ n = 1, 2, \cdots, N_x \\ & \|L_{m,\mathcal{Y}}\|_1 \le \mu_1,\ m = 1, 2, \cdots, N_y \\ & \|L_{n,\mathcal{X}}\|_2 = \mu_2,\ n = 1, 2, \cdots, N_x \\ & \|L_{m,\mathcal{Y}}\|_2 = \mu_2\, m = 1, 2, \cdots, N_y. \end{aligned}$$

where i indices the original (post, response) pairs. Our experiments (Section 6) indicate that this simple linear model can learn meaningful patterns, due to the massive training set. For example, the image of the word "Italy" in the post in the latent space matches well word "Sicily", "Mediterranean sea" and "travel". Once the mapping $L_{\mathcal{X}}$ and $L_{\mathcal{Y}}$ are learned, the semantic matching score $\mathbf{x}^\top L_{\mathcal{X}} L_{\mathcal{Y}}^\top \mathbf{y}$ will be treated as a feature for modeling the overall suitability of y as a response to post x.

POST-RESPONSE SIMILARITY: Here we use a simple vector-space model for measuring the similarity between a post and a response

$$\mathsf{sim}_{PR}(\mathbf{x}, \mathbf{y}) = \frac{\mathbf{x}^\top \mathbf{y}}{\|\mathbf{x}\| \|\mathbf{y}\|}. \quad (3)$$

Although it is not necessarily true that a good response has many common words as the post, but this measurement is often helpful in finding relevant responses. For example, when the post and response

[6] The performance is fairly insensitive to the choice of the penalty, so we only report the result with a typical choice of it.

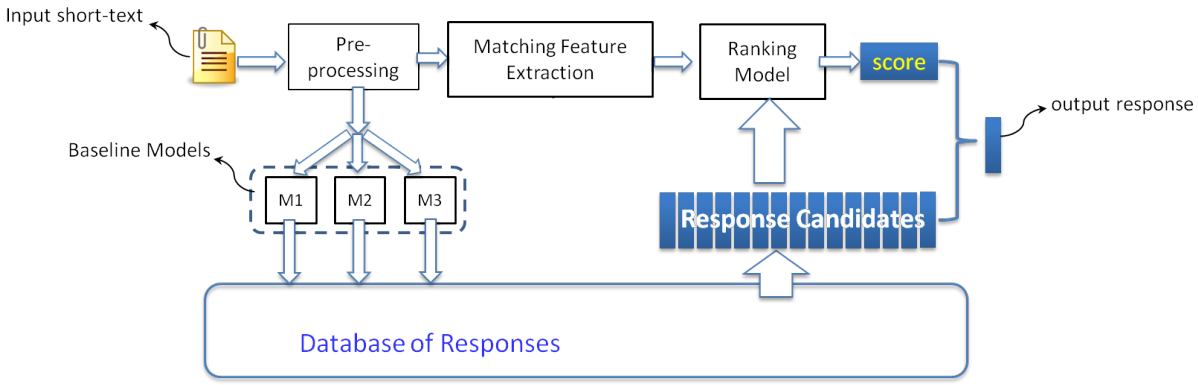

Figure 5: Diagram of the retrieval-based automatic response system.

both have "National Palace Museum in Taipei", it is a strong signal that they are about similar topics. Unlike the semantic matching feature, this simple similarity requires no learning and works on infrequent words. Our empirical results show that it can often capture the Post-Response relation failed with semantic matching feature.

POST-POST SIMILARITY: The basic idea here is to find posts similar to x and use their responses as the candidates. Again we use the vector space model for measuring the post-post similarity

$$\mathsf{sim}_{PP}(x, x') = \frac{\mathbf{x}^\top \mathbf{x}'}{\|\mathbf{x}\|\|\mathbf{x}'\|}. \quad (4)$$

The intuition here is that if a post x' is similar to x its responses might be appropriate for x. It however often fails, especially when a response to x' addresses parts of x not contained by x, which fortunately can be alleviated when combined with other measures.

5.2 Learning to Rank with Labeled Data

With all the matching features, we can learn a ranking model with the labeled (post, response) pairs, e.g., through off-the-shelf ranking algorithms. From the labeled data, we can extract triples (x, y^+, y^-) to ensure that $\mathsf{score}(x, y^+) > \mathsf{score}(x, y^-)$. Apparently y^+ can be selected from labeled positive response of x, while y^- can be sampled either from labeled negative negative or randomly selected ones. Since the manually labeled negative instances are top-ranked candidates according to some individual retrieval model (see Section 5.1) and therefore generally yield slightly better results.

The matching features are mostly constructed by combining the individual matching models, for example the following two

- $\Phi_7(x, y)$: this feature measures the length of the longest common string in the post and the response;
- $\Phi_{12}(x, y)$: this feature considers both semantic matching score between query post x and candidate response $\mathbf{y}$, as well as the similarity between x and y's *original* post x':

$$\Phi_{12}(x, y) = \mathsf{SemMatch}(x, y)\mathsf{sim}_{PP}(x, x').$$

In addition to the matching features, we also have simple features describing responses only, such as the length of it.

6 Experimental Evaluation

We perform experiments on the proposed dataset to test our retrieval-based model as an algorithm for automatically generating response.

6.1 Performance of Models

We evaluate the retrieved models based on the following two metrics:

MAP This one measures the mean average precision (MAP)(Manning et al., 2008) associated with the ranked list on $\mathcal{C}_{\mathbf{x}}^{(reduced)}$.

P@1 This one simply measures the precision of the top one response in the ranked list:

$$P@1 = \frac{\#\text{good top-1 responses}}{\#\text{posts}}$$

We perform a 5-fold cross-validation on the 422 labeled posts, with the results reported in Table 1. As it shows, the semantic matching helps slightly improve the overall performance on P@1.

Model	MAP	P@1
P2R	0.565	0.489
P2R + P2P	0.621	0.567
P2R + MATCH	0.575	0.513
P2R + P2P + MATCH	0.621	0.574

Table 3: Comparison of different choices of features, where P2R stands for the features based on post-response similarity, P2P stands for the features based on post-post similarity, and MATCH stands for the semantic match feature.

To mimic a more realistic scenario on automatic response model on Sina Weibo, we allow the system to choose which post to respond to. Here we simply set the response algorithm to respond only when the highest score of the candidate response passes a certain threshold. Our experiments show that when we choose to respond only to 50% of the posts, the P@1 increases to 0.76, while if the system only respond to 25% of the posts, P@1 keeps increasing to 81%.

6.2 Case Study

Although our preliminary retrieval model does not consider more complicated syntax, it is still able to capture some useful coupling structure between the appropriate (post, response) pairs, as well as the similar (post, post) pairs.

Query Post: 创新工场三年庆, 在 我们的智慧树会议室

Today is the 3-year anniversary of Innovation Works. We are in the meeting rooms named Tree of Wisdom

Response: 嗯, 中午去的, 新环境不错, 很宽敞

Yeah, I came in the noon. Nice environment, quite spacious

Figure 6: An actual instance (the original Chinese text and its English translation) of response returned by our retrieval-based system.

Case study shows that our retrieval is fairly effective at capturing the semantic relevance (Section 6.2.1), but relative weak on modeling the logic consistency (Section 6.2.2). Also it is clear that the semantic matching feature (described in Section 5.1) helps find matched responses that do not share any words with the post (Section 6.2.3).

6.2.1 On Semantic Relevance

The features employed in our retrieval model are mostly vector-space based, which are fairly good at capturing the semantic relevance, as illustrated by Example 1 & 2.

EXAMPLE 1:

P:	*It is a small town on an Spanish with 500 population, and guess what, they even have a casino!*
R:	*If you travel to Spain, you need to spend some time there.*

EXAMPLE 2:

P:	*One quote from Benjamin Franklin: "We are all born ignorant, but one must work hard to remain stupid."*
R:	*Benjamin Franklin is a wise man, and one of the founding fathers of USA.*

However our retrieval model also makes bad choice, especially when either the query post or the response is long, as shown in Example 3. Here the response is picked up because 1) the correspondence between the word "IT" in the post and the word "mobile phone" in the candidate, and 2) the Chinese word for "lay off" in the post and the word for "outdated" in the response are the same.

EXAMPLE 3:

P:	*As to the laying-off, I haven't heard anything about it. "Elimination of the least competent" is kind-off conventional in IT, but the ratio is actually quite small.*
R:	*Please don't speak that way, otherwise you can get outdated. Mobile phones are very expensive when they were just out, but now they are fairly cheap. Look forward, or you will be outdated.*

The entity association is only partially addressed with features like post-response cosine similarity, treating entity name just as a word, which is apparently not enough for preventing the following type

of mistakes (see Example 4 & 5) when the post and response match well on other parts

EXAMPLE 4:

P:	*Professor Wang will give a curse on natural language processing, starting next semester.*
R:	*Jealous.. I wish I can attend Prof. Li's course too some time in the future.*

EXAMPLE 5:

P:	*The fine China from Exhibition at the National Palace Museum in Taipei*
R:	*This drawing looks so nice. National Palace Museum in Taipei is full of national treasures*

6.2.2 On Logic Consistency

Our current model does not explicitly maintain the logic consistency between the response and the post, since Logic consistency requires a deeper analysis of the text, and therefore hard to capture with just a vector space model. Below are two examples which are semantically relevant, and correct with respect to speech act, but logically inappropriate.

EXAMPLE 1:

P:	*I checked. Wang Fengyi is not my great grandfather, although they've done similar deeds and both were called "Wang the Well-doer".*
R:	*wow, Wang Fengyi is your great grand-father*

EXAMPLE 2:

P:	*We are looking for summer interns. We provide books and lunch. If you are in Wu Han and interested, drop us an email. Sorry we don't take any students outside Wu Han.*
R:	*Are you looking for summer intern?*

6.2.3 The Effect of Semantic Matching

The experiments also show that we may find interesting and appropriate responses that have no common words as the post, as shown in the example below. Our bi-linear semantic matching model however performs relatively poorly on long posts, where the topics of the sentence cannot be well captured by the sum of the latent vectors associated with each word.

P:	*Eight England players stand in the penalty area.*
R1:	*What a classic match*
R2:	*Haha, it is still 0:0, no goal so far*

7 Summary

In this paper we propose a retrieval-based response model for short-text based conversation, to leverage the massive instances collected from social media. For research in similar directions, we create a dataset based on the posts and comments from Sina Weibo. Our preliminary experiments show that our retrieval-based response model, when combined with a large candidate set, can achieve fairly good performance. This dataset will be valuable for both training and testing automatic response models for short texts.

References

Peter F. Brown, Vincent J. Della Pietra, Stephen A. Della Pietra, and Robert L. Mercer. 1993. The mathematics of statistical machine translation: parameter estimation. *Comput. Linguist.*, 19(2).

Rollo Carpenter. 1997. Cleverbot.

Sina Jafarpour and Christopher J. C. Burges. 2010. Filter, rank, and transfer the knowledge: Learning to chat.

Thorsten Joachims. 2002. Optimizing search engines using clickthrough data. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '02, pages 133–142, New York, NY, USA. ACM.

Anton Leuski and David R. Traum. 2011. Npceditor: Creating virtual human dialogue using information retrieval techniques. *AI Magazine*, 32(2):42–56.

Diane Litman, Satinder Singh, Michael Kearns, and Marilyn Walker. 2000. Njfun: a reinforcement learning spoken dialogue system. In *Proceedings of the 2000 ANLP/NAACL Workshop on Conversational systems - Volume 3*, ANLP/NAACL-ConvSyst '00, pages 17–20, Stroudsburg, PA, USA. Association for Computational Linguistics.

Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. 2008. *Introduction to Information Retrieval*. Cambridge University Press, New York, NY, USA.

Teruhisa Misu, Kallirroi Georgila, Anton Leuski, and David Traum. 2012. Reinforcement learning of question-answering dialogue policies for virtual museum guides. In *Proceedings of the 13th Annual Meeting*

of the Special Interest Group on Discourse and Dialogue, SIGDIAL '12, pages 84–93.

Alan Ritter, Colin Cherry, and William B. Dolan. 2011. Data-driven response generation in social media. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, pages 583–593, Stroudsburg, PA, USA. Association for Computational Linguistics.

Jost Schatzmann, Karl Weilhammer, Matt Stuttle, and Steve Young. 2006. A survey of statistical user simulation techniques for reinforcement-learning of dialogue management strategies. *Knowl. Eng. Rev.*, pages 97–126.

Ellen M Voorhees. 2002. The philosophy of information retrieval evaluation. In *Evaluation of cross-language information retrieval systems*, pages 355–370. Springer.

Jason D. Williams and Steve Young. 2007. Partially observable markov decision processes for spoken dialog systems. *Comput. Speech Lang.*, 21(2):393–422.

Wei Wu, Zhengdong Lu, and Hang Li. 2013. Learning bilinear model for matching queries and documents. *Journal of Machine Learning Research (2013 to appear)*.

Hua-Ping Zhang, Hong-Kui Yu, De-Yi Xiong, and Qun Liu. 2003. Hhmm-based chinese lexical analyzer ictclas. SIGHAN '03.