Context-Dependent Automatic Response Generation Using Statistical Machine Translation Techniques

Andrew Shin **Ryohei Sasano** **Hiroya Takamura** **Manabu Okumura**
Tokyo Institute of Technology
shin@lr.pi.titech.ac.jp {sasano,takamura,oku}@pi.titech.ac.jp

Abstract

Developing a system that can automatically respond to a user's utterance has recently become a topic of research in natural language processing. However, most works on the topic take into account only a single preceding utterance to generate a response. Recent works demonstrate that the application of statistical machine translation (SMT) techniques towards monolingual dialogue setting, in which a response is treated as a translation of a stimulus, has a great potential, and we exploit the approach to tackle the context-dependent response generation task. We attempt to extract relevant and significant information from the wider contextual scope of the conversation, and incorporate it into the SMT techniques. We also discuss the advantages and limitations of this approach through our experimental results.

1 Introduction

Various approaches have been applied to the response generation task, each with its own merits and drawbacks. While one of the main concerns on the topic has been the semantic relevance of the response, it has mostly been discussed in terms of a limited conversational scope, mostly a single utterance. This provides us with a room for research on a wider scope of conversation, which reflects not only a single preceding utterance, but the overall context of the current conversation.

SMT-based data-driven approach to the response generation task was recently introduced by Ritter et al. (2011). They demonstrated that it was better-suited for response generation than some of the previous approaches, including information retrieval approach. We exploit this model to address the above-mentioned problem of reflecting a wider scope of conversation.

We present a context-dependent model where we attempt to generate more semantically relevant and diverse responses by adding the semantically important words from previous utterances to the most recent one. By doing so, we hope not only to diversify the responses, but also to be able to take semantics from broader scope of the conversation into account.

2 Response Generation using SMT

2.1 Overview

Ritter et al. (2011) remarked that stimulus-response pairs in the same language often have a strong structural resemblance, as shown in the example conversation below, that may be exploited in SMT platforms. In the usual SMT setting, a string f in a source language is translated into a string e in a target language according to probability distribution $p(e|f)$ (Brown et al., 1993). Ritter et al. applied the SMT techniques to monolingual conversation setting, and treated the response as the *translation* of the stimulus.

Stimulus: What is your hobby?
Response: My hobby is hiking.

2.2 Challenges

Although the application of SMT to the response generation task demonstrates potentials, it has a few drawbacks due to its nature.

Human Language Technologies: The 2015 Annual Conference of the North American Chapter of the ACL, pages 1345–1350,
Denver, Colorado, May 31 – June 5, 2015.

First, the lengths of the source and target utterances are not correlated in the conversational setting, and there is hardly any general tendency towards the relative length of the utterances, as shown in the example conversation below. SMT usually works on the data in which the ratio between the lengths of the source and target utterances stays relatively constant (Och and Ney, 2000). However, conversational setting, in which such constant ratio is absent, jeopardizes the functionality of the usual SMT models to make alignments. Although it is highly probable that some of the semantic elements in the source utterance are reflected in the target utterance, it is rarely on a one-to-one basis.

A: Is something going on today? (S_1)

B: Of course, it's dad's birthday. (S_2)

A (most recent stimulus) **:** What?! (S_3)

B (target) **:** Oh, you didn't know? (S_4)

Second, it cannot take into account what was previously discussed in the conversation. Unless the most recent utterance brings a completely new topic, or it has sufficient information in itself, such problem is evident.

Both problems regarding the context and the alignment become more pronounced especially in cases where the source utterance is short as shown in the above example. Clearly, no meaningful response can be derived from the most recent stimulus alone, and it is highly unclear how the alignments should be made. Indeed, the response generated by applying SMT to the most recent stimulus "What?" in this example is "that," which only mimics the syntactic structure but fails to deliver any meaningful content.

3 Context-Dependent Model

3.1 Overview

In order to deal with the issues of context and the lengths of the utterances without correlation, we work on building a context-dependent model, in which we balance the utterance lengths by selecting contextually important words from the previous part of the conversation, and adding them to the source utterance. For example, applying one of our models to the most recent stimulus (S_3) of the previous example conversation results in the following utterance, where the words in the parenthesis are newly added:

A: (today birthday) What?!

The rationale behind this approach is that the topic of a conversation can be characterized by a number of contextually important words, which provide semantic information to be reflected in the response generation process.

This approach seemingly reduces the grammatical integrity of the source utterance, and it may seem as if we risk confusing the translation model and losing grammaticality of the output. However, grammaticality of the output is handled by the language model, and the language model is constructed upon the target language only, which in our case corresponds to the target utterances that remain untouched. Also, the newly added words are of high relevance to the topic, so the new source utterance frequently demonstrates high semantic coherence both within itself, and in parallel with the target utterance.

The question now is how to determine which words are contextually important throughout the conversation. Since finding such contextually important words is our main concern, we find simple statistical significance test models more suitable than conventional methods from discourse modeling or dialogue systems (Oh et al., 2002). We examine two approaches, namely the pair-based approach, and the token-based approach. The pair-based approach uses Fisher's Exact Test (Moore, 2004), which is reported to give more accurate p-values than χ^2 or G^2 when the counts are small (Ritter et al., 2011). This approach takes advantage of the proximity of utterances, and assumes that a utterance whose distance to the source utterance is shorter is likely to be more contextually related to the source utterance, i.e. S_{n-1} is more likely than S_{n-2} to be semantically relevant to S_n. The token-based approach considers at the entire conversation, and selects the words most characteristic of the conversation, using the most widely used term weighting algorithm, *tf-idf*.

3.2 Pair-Based Model

Given a conversation consisting of utterances S_1,...,S_{n+1}, where S_n is the source utterance and S_{n+1} is the target utterance, we start by computing the p-value from Fisher's Exact Test for every possible word pair between S_n and S_{n-1}. If the p-value

is less than the threshold, implying a significant relevance between the words constituting the pair, we store the words. We then add the stored words to the source utterance, avoiding duplicates with words already in the source utterance, until its length is the same as that of the target utterance. Words are added in a reversed order of their appearance, i.e., we give priority to words that appeared in the later part of the discourse, in light of the previously mentioned assumption. If, after adding all stored words to the source utterance, the length of the source utterance is still less than the length of the target utterance, we repeat the process with word pairs between S_{n-1} and S_{n-2}, and so forth. This "crawling-up" is necessary because S_n is often short or semantically trivial that further comparison of S_n with other previous utterances fails to capture the contextually important words that are continuously discussed in the previous part of the conversation. The procedure ends when there are no more pairs whose p-value is less than the threshold, or the source utterance has the same length as the target utterance.

Note that, for training, we limit the application of our model only to the cases where source utterances are shorter than target utterances, since adding words in the opposite case will exacerbate the difficulty of alignment. In the test setting, however, we do not know the length of the actual target utterance, and thus selectively apply our model based on the absolute length of the source utterance, where the threshold is set to the average length of the source utterance throughout the training data. Also, since it is evident that we are not dealing with grammatically well-formed utterances whose ordering should matter, we opt not to use the reordering table (Bisazza et al., 2011).

3.3 Token-Based Model

The assumption behind the pair-based approach is that a topic of a conversation is something that continues to be discussed throughout the conversation, i.e. something that gets reflected/matched in the later part of the conversation. Finding collocated words using significant test does just that. However, there may be a trade-off here in terms of representing the diversity of context; for example, there may be a characteristic word that is not directly reflected/matched in the later part of the conversation. That provides the motivation for our token-based approach, using tf-idf.

This approach follows a similar manner of adding words to the source utterance until its length is equal to the target utterance, but differs in that it picks contextually important words by examining individual tokens, rather than pairs of words, using *tf-idf*. For *idf*, the total number of documents was set to the number of conversations in our training data.

Also, instead of crawling up the conversation from the source utterance, it scans through the entire conversation and selects characteristic words within the given scope of the conversation. This is intended to reflect that there could be words that are highly relevant to the overall topic of the conversation, yet not very close to the current source utterance. For example, in the following conversation, both (S_3) and (S_4) lack any element characteristic of the conversation that leads to the final response (S_5), while "NBA" or "fans" in (S_1) and (S_2) is indicative of the topic of the conversation, and will be relevant to words like "LeBron" or "dominating" in the target utterance.

A: Well, the NBA season is near again.(S_1)

B: Yeah! So excited for all the NBA fans!(S_2)

A: I'm not. (S_3)

B (source) **:** How come? (S_4)

A (target) **:** It's just gonna be LeBron dominating again. (S_5)

Although we examine the entire conversation, words that are too far from the source utterance (for example, 50 utterances apart) will rarely have much semantic impact to the current topic. Thus, it is necessary to keep the size of the conversation reasonably small, and we restrict it to be at most 8 utterances.

4 Experiment

4.1 Setting

We first built our baseline model following the procedure proposed by Ritter et al. (2011). In accordance with the paper, we also filtered out the phrase table by Fisher's Exact Test. We then implemented our model using Moses (Koehn et al., 2007) toolkit with KenLM (Heafield, 2011) as the language model in 5-gram setting. In accordance with the baseline, we built our training, tuning, and test data set from

Model A	Model B	A>B	A=B	A<B	p-value	Agreement
Pair	Baseline	287	32	81	4.0e-28	.488
	Actual	58	28	314	1.2e-43	.543
	Token	175	52	173	0.96	.373
Token	Baseline	280	35	85	2.0e-25	.462
	Actual	62	35	303	3.2e-39	.529

Table 1: Performances against Each Model

Twitter, except we collected conversations, consisting of a tweet and successive replies, rather than pairs of tweets. We also restricted each conversation to have 3 to 8 utterances with only two speakers taking turns, to make it more likely that the topic of the conversation is preserved. Although there were some cases in which the topic deviated, our validation of the dataset showed that the amount of such cases was negligible. We ended up having approximately 1.4M pairs of utterances in the training data, which constitute 425,547 conversations. The threshold p-value for Fisher's Exact Test was set to 0.0001, to well-balance the number of selected words with the lengths of the utterances.

4.2 Evaluation

One of the challenging aspects of the researches on conversation is its distinct nature in which there is an extremely wide range of acceptable candidate responses to a stimulus, unlike usual bilingual translation tasks where there are typically pre-set candidates to be referenced with high reliability. Using the automatic evaluation metrics, we obtained slight improvements; for example, BLEU score (Papineni et al., 2002), with the actual responses from Twitter as the gold standard, increased from 0.82 for baseline to 0.89 for the pair-based approach. For the above-mentioned reason, however, we found it dubious whether a higher score in these metrics corresponds to better responses, and we thus resort to human manual evaluation as our primary source of evaluation.

We performed a human evaluation on Amazon Mechanical Turk (Buhrmester et al., 2011). The evaluation task consisted of four different sets of 100 questions, each set of which was handled by 10 workers. Each question was a ranking task, and the workers were shown a part of conversation and were instructed to rank the responses that followed the conversation in consideration of their relevance to the topic of the conversation. For all questions, workers were given four responses; the actual response from Twitter, one generated by the baseline model, and two by each of our context-dependent models.

Model	1st	2nd	3rd	4th	Avg. Rank
Actual	.664	.129	.087	.120	1.66
Baseline	.092	.232	.271	.406	2.99
Pair	.112	.323	.346	.220	2.67
Token	.134	.315	.297	.255	2.67

Table 2: Rankings from Human Evaluation

The order of responses was randomized for each question. In addition, in order to filter out the workers who do not take the tasks seriously, generating noise answers, we selected 10 questions that had obvious answers, and rejected the answers by the workers who failed to achieve 70% or higher accuracy on those questions. As stated in Section 3.1, the threshold for length of source sentence to determine whether to add words or not was set to the average length of source sentences throughout the training data, which in our case was 10. In roughly half of 400 questions, no words were added to the source sentence, and 1 to 6 words were added for roughly 25 to 30 questions respectively. Beyond 6 words, the number of questions begins to decline.

4.3 Results

Table 1 shows how our models performed against the actual responses, the baseline model, and each other, in regards to the number of questions for which our models were ranked higher. Overall, the table shows that our models were preferred over the baseline model, but performed poorly against the actual responses as expected. Yet, it was able to perform better than the actual responses in roughly

Conversation	Response*	Avg. Rank
A: if youre following today maybe follow me **B:** lol ur off to an early start	**1:** im at the bus stop how was your night	2.1
	2: i dont fancy hame pass for	3.6
	3: lmao i got my second home	3.2
	4: updating haha I get u off too	1.9
A: like you're talking about the stupidest things ever. its annoying **B:** who is this about	**1:** the ppl behind you	1.9
	2: I want tie you	3.6
	3: like I said Im talking seriously are you	2.3
	4: one of them is that you	2.9
A: Aww man happy birthday bro!! Lol you know you gotta die right? **B:** What? Lol	**1:** you gotta damn near die from drinking today	1.2
	2: lol yea	3.8
	3: thank you man you know me	2.4
	4: lol I know I know	2.8

* 1 is the actual response on Twitter, while 2,3 and 4 are responses generated by the baseline, pair-based, token-based models respectively.

Table 3: Examples of Responses

15% of the questions, especially when the actual responses were grammatically poor, or irrelevant to the topic of the conversation. There was no significant difference between the performances of our models. It also shows the p-value and mutual agreement between two models. Using S coefficient (Bennett et al., 1954) as a measurement of agreement yields the following result. Most of them fall into "moderate agreement" range of 0.4 to 0.6, except Token-based model against Pair-based model is slightly lower and falls into "fair agreement" range (Landis and Koch, 1977).

Table 2 shows the distribution of each model over each ranking and their average rankings. Our models outperform the baseline model in higher rankings. Table 3 features examples of responses generated by each model and the actual responses on Twitter, along with their average ranking in the final evaluation. In the first conversation, one of our models was ranked higher than both the baseline model and the actual response. In other conversations, our models were ranked higher than the baseline model, but lower than the actual response. Generally, our models have a wider range of topic-relevant vocabularies, and sound comparatively coherent than the baseline model, without too much grammatical violations.

5 Conclusion and Future Work

As we observed in the experimental results, our context-dependent model outperformed the baseline model when examined in a wider scope of conversations. Although its performance against the actual responses was not as satisfactory, it could outperform them when the actual responses diverted from the topic, or had poor coherence and grammaticality.

Possible applications include chatterbots or conversational agents. Most such applications are based on one-turn conversation, where user says something, system gives some response, and that is technically the end of the conversation of current topic, which will not be referred to in later conversations. Our work can, for example, provide the system with possible topics to talk about, especially when the input from the user is short or trivial. Diversity of the responses is obtained because, even when the system is given the same input, it will return completely different responses depending on what was previously talked about, as opposed to the applications where certain responses can be expected given an input.

An improvement is likely to come from attempting different methods to extract the core tokens from the past utterances. We relied on the Fisher's Exact Test and *tf-idf* throughout the research, but other approaches may perform better. Alternatively, we may try different weighting systems depending on whether a token is from the same speaker as the current utterance or a different speaker, since it would generally make more sense for a particular speaker not to repeat him/herself.

References

Arianna Bisazza, Nick Ruiz, and Marcello Federico. 2011. *Fill-up versus Interpolation Methods for Phrase-based SMT Adaptation*. In *International Workshop on Spoken Language Translation*, pages 136–143.

E. M. Bennett, R. Alpert, and A.C. Goldstein 1954. *Communications through limited-response questioning*. In *Public Opinion Quarterly*, pages 303–308.

Peter Brown, Stephen A. Della Pietra, Vincent J. Della Piera, and Robert J. Mercer. 1993. *The Mathematics of Statistical Machine Translation: Parameter Estimation*. In *Computational Linguistics*, pages 263–311.

Michael Burhmester, Tracy Kwang, and Samuel D. Gosling 2011. *Amazon's Mechanical Turk: A New Source of Inexpensive, Yet High-Quality Data?* In *Perspectives on Psychological Science*, 6, pages 3–5.

Kenneth Heafield. 2011. *KenLM: Faster and Smaller Language Model Queries*. In Proceedeings of the *Sixth Workshop on Statistical Machine Trnaslation*, pages 187–197.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. 2007. *Moses: Open Source Toolkit for Statistical Machine Translation*. In Proceedings of the *Association for Computational Linguistics*, pages 177–180.

J. R. Landis and G. G. Koch. 1977. *The Measurement of Observer Agreement for Categorical Data*. *Biometrics*, Vol. 33, No. 1, pages 159–174.

Robert C. Moore. 2004. *On Log-Likelihood Ratios and the Significance of Rare Events*. In Proceedings of the *Conference on Empirical Methods in Natural Language Processing*, pages 333–340.

Franz J. Och, and Hermann Ney. 2000. *A Comparison of Alignment Models for Statistical Machine Translation*. In the *18th International Conference on Computational Linguistics*, pages 1086–1090.

Alice H. Oh, and Alexander I. Rudnicky. 2002. *Stochastic Natural Language Generation for Spoken Dialogue Systems*. In *Computer Speech & Language*, pages 387–407.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. *BLEU: a Method for Automatic Evaluation of Machine Translation*. In Proceedings of the *Association for Computational Linguistics*, pages 311–318.

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*, pages 583–593.