Data-Driven Response Generation in Social Media

Alan Ritter
Computer Sci. & Eng.
University of Washington
Seattle, WA 98195
aritter@cs.washington.edu

Colin Cherry
National Research Council Canada
Ottawa, Ontario, K1A 0R6
Colin.Cherry@nrc-cnrc.gc.ca

William B. Dolan
Microsoft Research
Redmond, WA 98052
billdol@microsoft.com

Abstract

We present a data-driven approach to generating responses to Twitter status posts, based on phrase-based Statistical Machine Translation. We find that mapping conversational stimuli onto responses is more difficult than translating between languages, due to the wider range of possible responses, the larger fraction of unaligned words/phrases, and the presence of large phrase pairs whose alignment cannot be further decomposed. After addressing these challenges, we compare approaches based on SMT and Information Retrieval in a human evaluation. We show that SMT outperforms IR on this task, and its output is preferred over actual human responses in 15% of cases. As far as we are aware, this is the first work to investigate the use of phrase-based SMT to directly translate a linguistic stimulus into an appropriate response.

1 Introduction

Recently there has been an explosion in the number of people having informal, public conversations on social media websites such as Facebook and Twitter. This presents a unique opportunity to build collections of naturally occurring conversations that are orders of magnitude larger than those previously available. These corpora, in turn, present new opportunities to apply data-driven techniques to conversational tasks.

We investigate the problem of **response generation**: given a conversational stimulus, generate an appropriate response. Specifically, we employ a large corpus of status-response pairs found on Twitter to create a system that responds to Twitter status posts. Note that we make no mention of context, intent or dialogue state; our goal is to generate any response that fits the provided stimulus; however, we do so without employing rules or templates, with the hope of creating a system that is both flexible and extensible when operating in an open domain.

Success in open domain response generation could be immediately useful to social media platforms, providing a list of suggested responses to a target status, or providing conversation-aware autocomplete for responses in progress. These features are especially important on hand-held devices (Hasselgren et al., 2003). Response generation should also be beneficial in building "chatterbots" (Weizenbaum, 1966) for entertainment purposes or companionship (Wilks, 2006). However, we are most excited by the future potential of data-driven response generation when used inside larger dialogue systems, where direct consideration of the user's utterance could be combined with dialogue state (Wong and Mooney, 2007; Langner et al., 2010) to generate locally coherent, purposeful dialogue.

In this work, we investigate statistical machine translation as an approach for response generation. We are motivated by the following observation: In naturally occurring discourse, there is often a strong structural relationship between adjacent utterances (Hobbs, 1985). For example, consider the stimulus-response pair from the data:

> **Stimulus**: I'm slowly making this soup ..... and it smells gorgeous!

Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, pages 583–593,
Edinburgh, Scotland, UK, July 27–31, 2011.

Response: I'll bet it looks delicious too! Haha

Here "it" in the response refers to "this soup" in the status by co-reference; however, there is also a more subtle relationship between the "smells" and "looks", as well as "gorgeous" and "delicious". Parallelisms such as these are frequent in naturally occurring conversations, leading us to ask whether it might be possible to *translate* a stimulus into an appropriate response. We apply SMT to this problem, treating Twitter as our parallel corpus, with status posts as our source language and their responses as our target language. However, the established SMT pipeline cannot simply be applied out of the box.

We identify two key challenges in adapting SMT to the response generation task. First, unlike bilingual text, stimulus-response pairs are not semantically equivalent, leading to a wider range of possible responses for a given stimulus phrase. Furthermore, both sides of our parallel text are written in the same language. Thus, the most strongly associated word or phrase pairs found by off-the-shelf word alignment and phrase extraction tools are identical pairs. We address this issue with constraints and features to limit lexical overlap. Secondly, in stimulus-response pairs, there are far more unaligned words than in bilingual pairs; it is often the case that large portions of the stimulus are not referenced in the response and vice versa. Also, there are more large phrase-pairs that cannot be easily decomposed (for example see figure 2). These difficult cases confuse the IBM word alignment models. Instead of relying on these alignments to extract phrase-pairs, we consider all possible phrase-pairs in our parallel text, and apply an association-based filter.

We compare our approach to response generation against two Information Retrieval or nearest neighbour approaches, which use the input stimulus to select a response directly from the training data. We analyze the advantages and disadvantages of each approach, and perform an evaluation using human annotators from Amazon's Mechanical Turk. Along the way, we investigate the utility of SMT's BLEU evaluation metric when applied to this domain. We show that SMT-based solutions outperform IR-based solutions, and are chosen over actual human responses in our data in 15% of cases. As far as we are aware, this is the first work to investigate the feasibility of SMT's application to generating responses to open-domain linguistic stimuli.

2 Related Work

There has been a long history of "chatterbots" (Weizenbaum, 1966; Isbell et al., 2000; Shaikh et al., 2010), which attempt to engage users, typically leading the topic of conversation. They usually limit interactions to a specific scenario (e.g. a Rogerian psychotherapist), and use a set of template rules for generating responses. In contrast, we focus on the simpler task of generating an appropriate response to a single utterance. We leverage large amounts of conversational training data to scale to our Social Media domain, where conversations can be on just about any topic.

Additionally, there has been work on generating more natural utterances in goal-directed dialogue systems (Ratnaparkhi, 2000; Rambow et al., 2001). Currently, most dialogue systems rely on either canned responses or templates for generation, which can result in utterances which sound very unnatural in context (Chambers and Allen, 2004). Recent work has investigated the use of SMT in translating internal dialogue state into natural language (Langner et al., 2010). In addition to dialogue state, we believe it may be beneficial to consider the user's utterance when generating responses in order to generate locally coherent discourse (Barzilay and Lapata, 2005). Data-driven generation based on users' utterances might also be a useful way to fill in knowledge gaps in the system (Galley et al., 2001; Knight and Hatzivassiloglou, 1995).

Statistical machine translation has been applied to a smörgåsbord of NLP problems, including question answering (Echihabi and Marcu, 2003), semantic parsing and generation (Wong and Mooney, 2006; Wong and Mooney, 2007), summarization (Daumé III and Marcu, 2009), generating bid-phrases in online advertising (Ravi et al., 2010), spelling correction (Sun et al., 2010), paraphrase (Dolan et al., 2004; Quirk et al., 2004) and query expansion (Riezler et al., 2007). Most relevant to our efforts is the work by Soricut and Marcu (2006), who applied the IBM word alignment models to a discourse ordering task, exploiting the same intuition investigated

in this paper: certain words (or phrases) tend to trigger the usage of other words in subsequent discourse units. As far as we are aware, ours is the first work to explore the use of phrase-based translation in generating responses to open-domain linguistic stimuli, although the analogy between translation and dialogue has been drawn (Leuski and Traum, 2010).

3 Data

For learning response-generation models, we use a corpus of roughly 1.3 million conversations scraped from Twitter (Ritter et al., 2010; Danescu-Niculescu-Mizil et al., 2011). Twitter conversations don't occur in real-time as in IRC; rather as in email, users typically take turns responding to each other. Twitter's 140 character limit, however, keeps conversations chat-like. In addition, the Twitter API maintains a reference from each reply to the post it responds to, so unlike IRC, there is no need for conversation disentanglement (Elsner and Charniak, 2008; Wang and Oard, 2009). The first message of a conversation is typically unique, not directed at any particular user but instead broadcast to the author's followers (a status message). For the purposes of this paper, we limit the data set to only the first two utterances from each conversation. As a result of this constraint, any system trained with this data will be specialized for responding to Twitter status posts.

4 Response Generation as Translation

When applied to conversations, SMT models the probability of a response r given the input status-post s using a log-linear combination of feature functions. Most prominent among these features are the conditional phrase-translation probabilities in both directions, $P(s|r)$ and $P(r|s)$, which ensure r is an appropriate response to s, and the *language model* $P(r)$, which ensures r is a well-formed response. As in translation, the response models are estimated from counts of phrase pairs observed in the training bitext, and the language model is built using n-gram statistics from a large set of observed responses. To find the best response to a given input status-post, we employ the Moses phrase-based decoder (Koehn et al., 2007), which conducts a beam search for the best response given the input, according to the log-linear model.

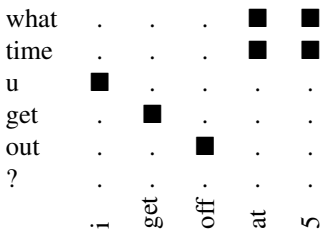

Figure 1: Example from the data where word alignment is easy. There is a clear correspondence between words in the status and the response.

4.1 Challenge: Lexical Repetition

When applied directly to conversation data, off-the-shelf MT systems simply learn to parrot back the input, sometimes with slight modification. For example, directly applying Moses with default settings to the conversation data produces a system which yields the following (typical) output on the above example:

> **Stimulus**: I'm slowly making this soup ..... and it smells gorgeous!
>
> **Response:** i'm slowly making this soup ..... and you smell gorgeous!

This "paraphrasing" phenomenon occurs because identical word pairs are frequently observed together in the training data. Because there is a wide range of acceptable responses to any status, these identical pairs have the strongest associations in the data, and therefore dominate the phrase table. In order to discourage lexically similar translations, we filter out all phrase-pairs where one phrase is a substring of the other, and introduce a novel feature to penalize lexical similarity:

$$\phi_{\text{lex}}(s,t) = J(s,t)$$

Where $J(s,t)$ is the Jaccard similarity between the set of words in s and t.

4.2 Challenge: Word Alignment

Alignment is more difficult in conversational data than bilingual data (Brown et al., 1990), or textual entailment data (Brockett, 2006; MacCartney et al., 2008). In conversational data, there are some cases in which there is a decomposable alignment between

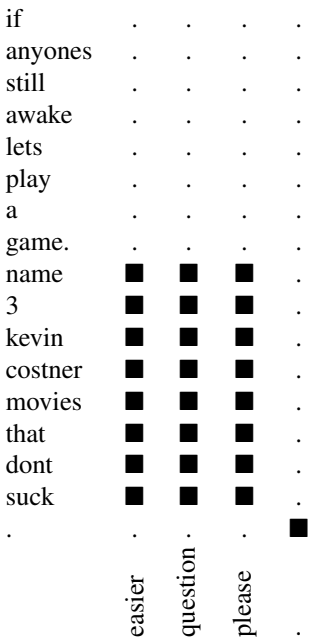

Figure 2: Example from the data where word alignment is difficult (requires alignment between large phrases in the status and response).

words, as seen in figure 1, and some difficult cases where alignment between large phrases is required, for example figure 2. These difficult sentence pairs confuse the IBM word alignment models which have no way to distinguish between the easy and hard cases.

We aligned words in our parallel data using the widely used tool GIZA++ (Och and Ney, 2003); however, the standard growing heuristic resulted in very noisy alignments. Precision could be improved considerably by using the intersection of GIZA++ trained in two directions ($s \rightarrow r$, and $r \rightarrow s$), but the alignment also became extremely sparse. The average number of alignments-per status/response pair in our data was only 1.7, as compared to a dataset of aligned French-English sentence pairs (the WMT 08 news commentary data) where the average number of intersection alignments is 14.

Direct Phrase Pair Extraction

Because word alignment in status/response pairs is a difficult problem, instead of relying on local alignments for extracting phrase pairs, we exploit information from all occurrences of the pair in determining whether its phrases form a valid mapping.

$C(s,t)$	$C(s,\neg t)$	$C(s)$
$C(\neg s,t)$	$C(\neg s,\neg t)$	$N-C(s)$
$C(t)$	$N-C(t)$	N

Figure 3: Contingency table for phrase pair (s,t). Fisher's Exact Test estimates the probability of seeing this event, or one more extreme assuming s and t are independent.

We consider all possible phrase-pairs in the training data,[1] then use Fisher's Exact Test to filter out pairs with low correlation (Johnson et al., 2007). Given a source and target phrase s and t, we consider the contingency table illustrated in figure 3, which includes co-occurrence counts for s and t, the number of sentence-pairs containing s, but not t and vice versa, in addition to the number of pairs containing neither s nor t. Fisher's Exact Test provides us with an estimate of the probability of observing this table, or one more extreme, assuming s and t are independent; in other words it gives us a measure of how strongly associated they are. In contrast to statistical tests such as χ^2, or the G^2 Log Likelihood Ratio, Fisher's Exact Test produces accurate p-values even when the expected counts are small (as is extremely common in our case).

In Fisher's Exact Test, the hypergeometric probability distribution is used to compute the exact probability of a particular joint frequency assuming a model of independence:

$$\frac{C(s)!C(\neg s)!C(t)!C(\neg t)!}{N!C(s,t)!C(\neg s,t)!C(s,\neg t)!C(\neg s,\neg t)!}$$

The statistic is computed by summing the probability for the joint frequency in Table 3, and every more extreme joint frequency consistent with the marginal frequencies. Moore (2004) illustrates several tricks which make this computation feasible in practice.

We found that this approach generates phrase-table entries which appear quite reasonable upon manual inspection. The top 20 phrase-pairs (after filtering out identical source/target phrases, substrings,

[1] We define a possible phrase-pair as any pair of phrases found in a sentence-pair from our training corpus, where both phrases consist of 4 tokens or fewer. The total number of phrase pairs in a sentence pair (s,r) is $O(|s| \times |r|)$.

Source	Target
rt [*retweet*]	thanks for the
potter	harry
ice	cream
how are you	you ?
good	morning
chuck	norris
watching	movie
i miss	miss you too
are you	i 'm
my birthday	happy birthday
wish me luck	good luck
how was	it was
miss you	i miss
swine	flu
i love you	love you too
how are	are you ?
did you	i did
jackson	michael
how are you	i 'm good
michael	mj

Table 1: Top 20 Phrase Pairs ranked by the Fisher Exact Test statistic. Slight variations (substrings or symmetric pairs) were removed to show more variety. See the supplementary materials for the top 10k (unfiltered) pairs.

and symmetric pairs) are listed in Table 1.[2] Our experiments in §6 show that using the phrase table produced by Fisher's Exact Test outperforms one generated based on the poor quality IBM word alignments.

4.3 System Details

For the phrase-table used in the experiments (§6) we used the 5M phrases with highest association according the Fisher Exact Test statistic.[3] To build the language model, we used all of the 1.3M responses from the training data, along with roughly 1M replies collected using Twitter's streaming API.

We do not use any form of SMT reordering model, as the position of the phrase in the response does not seem to be very correlated with the corresponding position in the status. Instead we let the language model drive reordering.

We used the default feature weights provided by Moses.[4] Because automatic evaluation of response generation is an open problem, we avoided the use of discriminative training algorithms such as Minimum Error-Rate Training (Och, 2003).

5 Information Retrieval

One straightforward data-driven approach to response generation is nearest neighbour, or information retrieval. This general approach has been applied previously by several authors (Isbell et al., 2000; Swanson and Gordon, 2008; Jafarpour and Burges, 2010), and is used as a point of comparison in our experiments. Given a novel status s and a training corpus of status/response pairs, two retrieval strategies can be used to return a best response r':

IR-STATUS $[r_{\text{argmax}_i\, sim(s,s_i)}]$ Retrieve the response r_i whose associated status message s_i is most similar to the user's input s.

IR-RESPONSE $[r_{\text{argmax}_i\, sim(s,r_i)}]$ Retrieve the response r_i which has highest similarity when directly compared to s.

At first glance, IR-STATUS may appear to be the most promising option; intuitively, if an input status is very similar to a training status, we might expect the corresponding training response to pair well with the input. However, as we describe in §6, it turns out that directly retrieving the most similar response (IR-RESPONSE) tends to return acceptable replies more reliably, as judged by human annotators. To implement our two IR response generators, we rely on the default similarity measure implemented in the Lucene[5] Information Retrieval Library, which is an IDF-weighted Vector-Space similarity.

6 Experiments

In order to compare various approaches to automated response generation, we used human evalu-

[2] See the supplementary materials for the top 10k (unfiltered) phrase pairs.

[3] Note that this includes an arbitrary subset of the (1,1,1) pairs (phrase pairs where both phrases were only observed once in the data). Excluding these (1,1,1) pairs yields a rather small phrase table, 201K phrase-pairs after filtering, while including all of them led to a table which was too large for the memory of the machine used to conduct the experiments.

[4] The language model weight was set to 0.5, the translation model weights in both directions were both set to 0.2, the lexical similarity weight was set to -0.2.

[5] http://lucene.apache.org/

ators from Amazon's Mechanical Turk (Snow et al., 2008). Human evaluation also provides us with data for a preliminary investigation into the feasibility of automatic evaluation metrics. While automated evaluation has been investigated in the area of spoken dialogue systems (Jung et al., 2009), it is unclear how well it will correlate with human judgment in open-domain conversations where the range of possible responses is very large.

6.1 Experimental Conditions

We performed pairwise comparisons of several response-generation systems. Similar work on evaluating MT output (Bloodgood and Callison-Burch, 2010) has asked Turkers to rank more than two choices, but in order to keep our evaluation as straightforward as possible, we limited our experiments to pairwise comparisons.

For each experiment comparing 2 systems (*a* and *b*), we built a test set by selecting a random sample of 200 tweets which had received responses, and which had a length between 4 and 20 words. These tweets were selected from conversations collected from a later, non-overlapping time-period from those used in training. Each experiment used a different random sample of 200 tweets. For each of the 200 statuses, we generated a response using method *a* and *b*, then showed the status and both responses to the Turkers, asking them to choose the best response. The order of the systems used to generate a response was randomized, and each of the 200 HITs was submitted to 3 different Turkers. Turkers were paid 1¢ per judgment.

The Turkers were instructed that an appropriate response should be on the same topic as the status, and should also "make sense" in response to it. While this is an inherently subjective task, from inspecting the results, we found Turkers to be quite competent in judging between two responses.

The systems used in these pairwise comparisons are summarized in table 2, and example output generated by each system is presented in Table 3.

System Name	Description
RND-BASELINE	Picks randomly from the set of responses which are observed at least twice in the training data. The assumption is these are likely very general responses
IR-STATUS	$r_{\text{argmax}_i\, sim(s,s_i)}$ as described in §5
IR-RESPONSE	$r_{\text{argmax}_i\, sim(s,r_i)}$ as described in §5
MT-CHAT	Phrase-based translation system as described in §4
MT-BASELINE	Exactly the same as MT-CHAT, except using a phrase table extracted based on word alignments from GIZA++
HUMAN	Actual responses from the test data.

Table 2: Summary of systems compared experimentally

6.2 Results

The results of the experiments are summarized in Table 4. For each experiment we show the fraction of HITs where the majority of annotators agreed system *a* was better. We also show the p-value from an exact Binomial significance test; note that all differences are significant with above 95% confidence. Table 4 also shows the S coefficient (Bennett et al., 1954) as a measure of agreement between annotators for each experiment. S is equivalent the commonly used Kappa, except it assumes the uniform distribution when computing expected or chance agreement (Artstein and Poesio, 2008). The uniform distribution is appropriate in our setup, since annotators are not told which system generated each output, and the order of choices is randomized. If annotators were indeed choosing randomly, they would be expected to agree in 50% of cases. In most comparisons the value of S is in the range of 0.2 to 0.4, which can be considered "Fair agreement" (Landis and Koch, 1977).

Note that agreement between annotators is lower than typically reported in corpus annotation tasks. When annotating which of two automatically generated outputs is better, there is not always a clear answer; both responses might be good or bad. We can expect strong agreement only in cases where one response is clearly better. Strong agreement is not required, however, as we are using many annotations to compare each pair of systems, and the human judgments are not intended to be used as training data. Similar agreement was reported in an evaluation of automatically generated MT output as part

Status	MT-CHAT	MT-BASELINE	IR-STATUS	HUMAN	RND-BASELINE	IR-RESPONSE
riding to the arc on my bike. Hopefully I won't melt into the pavement.	good luck on your bike ride arc	i ca n't wait for it ! i 'm arc is a year .	grab some refreshing drinks ... and a towel ... lol ... then eat ice cream and suika later ...	make sure to ride in the middle of the street	lmfaoooo !	u sure the cup wo n't melt ?
I really hate watching movies in the living room.	me too , what movies are you doing in room ?	me too ! i love those ?) in my mouth .	:(cheer up ! you need to invest in one of those turbo fans	Why? no comfy spots or just too open to public	tas fazendo oq ?	someone was watching in the living room kasi e . :) i was too lazy to go up to my room . :))
Hello Did I miss David Getting his teen choice award? Or did they not show it :(	hello miss ? * he was getting his will probably win .	it 's just like what you just did the exact date hello ?	it did n't show yet -ha- -lol-	nothing yet...he presented though he is so darn cute	you wish .	are they gon na show teen choice awards again ? i missed it
I can think the impossible is possible at any given moment.	i think you can do it , anything is given impossible dreams really specs	i ca n't wait to see if there is no point in the end but if you do a few hours ...	i hope you use those powers for evil . i mean good ! good . that 's what i meant .	for sure Eric, and that makes it not 'impossible' any longer.	no -user- ?	nobody with whom i might apply the blue cobalt ' rabbit ' at the moment . given the economy , perhaps -url- (any of em)
I want to enter my parents house into a home makeover contest. Any help out there?	me too ! thanks for the win , parents are so smart , just got home makeover	oh man , i do n't want to come back to my mom is getting that	i put troy on the potty everyday at lunch and b4 u know it he was going on his own ... only took a week .	check TLC I'm pretty sure it was them who were recently posting about looking for houses to be nominated!	good job	you want to do laundry with me at my parents house after i get off ? maybe get a free meal out of it ? !

Table 3: Example responses from each system. We tried to pick examples where most (or all) systems generate reasonable responses for illustration purposes.

System A	**System B**	**Fraction A**	**p-value**	**Agreement** (S)	**System A BLEU**	**System B BLEU**
MT-CHAT*	IR-STATUS	0.645	5.0e-05	0.347	1.15	0.57
MT-CHAT*	IR-RESPONSE	0.593	1.0e-02	0.333	0.84	1.53
IR-STATUS	IR-RESPONSE*	0.422	3.3e-02	0.330	0.40	1.59
MT-CHAT*	MT-BASELINE	0.577	3.8e-02	0.160	1.23	1.14
MT-CHAT	HUMAN*	0.145	2.2e-16	0.433	N/A	N/A
MT-CHAT*	RND-BASELINE	0.880	2.2e-16	0.383	1.17	0.10

Table 4: Results of pairwise comparisons between various response-generation methods. Each row presents a comparison between systems *a* and *b* on 200 randomly selected tweets. The column **Fraction A** lists the fraction of HITs where the majority of annotators agreed **System A**'s response was better. The winning system is indicated with an asterisk*. All differences are significant.

of the WMT09 shared tasks (Callison-Burch et al., 2009).[6]

The results of the paired evaluations provide a clear ordering on the automatic systems: IR-STATUS is outperformed by IR-RESPONSE, which is in turn outperformed by MT-CHAT. These results are somewhat surprising. We had expected that matching status to status would create a more natural and effective IR system, but in practice, it appears that the additional level of indirection employed by IR-STATUS created only more opportunity for confusion and error. Also, we did not necessarily expect MT-CHAT's output to be preferred by human annotators: the SMT system is the only one that generates a completely novel response, and is therefore the system most likely to make fluency errors. We had expected human annotators to pick up on these fluency errors, giving the the advantage to the IR systems. However, it appears that MT-CHAT's ability to tailor its response to the status on a fine-grained scale overcame the disadvantage of occasionally introducing fluency errors.[7]

Given MT-CHAT's success over the IR systems, we conducted further experiments to validate its output. In order to test how close MT-CHAT's responses come to human-level abilities, we compared its output to actual human responses from our dataset. In some cases the human responses change the topic of conversation, and completely ignore the initial status. For instance, one frequent type of response we noticed in the data was a greeting: "How have you been? I haven't talked to you in a while." For the purposes of this evaluation, we manually filtered out cases where the human response was completely off-topic from the status, selecting 200 pairs at random that met our criteria and using the actual responses as the HUMAN output.

When compared to the actual human-generated response, MT-CHAT loses. However, its output is preferred over the human responses 15% of the time, a fact that is particularly surprising given the very small – by MT standards – amount of data used to train the model. A few examples where MT-CHAT's output were selected over the human response are listed in Table 5.

We also evaluated the effect of filtering all possible phrase pairs using Fisher's Exact Test, which we did instead of conducting phrase extraction according to the very noisy word alignments. We altered our MT-CHAT system to use the standard Moses phrase-extraction pipeline, creating the system denoted as MT-BASELINE. We compared this to the complete MT-CHAT system. Note that both systems account for lexical repetition as described in §4.1. MT-CHAT's output is preferred 58% of the time over MT-BASELINE, indicating that direct phrase extraction is useful in this conversational setting.

Finally, as an additional baseline, we compared MT-CHAT's output to random responses selected from those observed 2 or more times in the training data. One might argue that short, common responses are very general, and that a reply like "lol" could be considered a good response to almost any status. However, the human evaluation shows a clear preference for MT-CHAT's output: raters favour responses that are tailored to the stimulus.

6.3 Automatic Evaluation

The field of SMT has benefited greatly from the existence of an automatic evaluation metric, BLEU (Papineni et al., 2002), which grades an output candidate according to n-gram matches to one or more reference outputs. To evaluate whether BLEU is an appropriate automatic evaluation measure for response generation, we attempted to measure its agreement with the human judgments. We calculate BLEU using a single reference derived from our parallel corpus. We show the smoothed BLEU 1-4 scores for each system on each dataset evaluated in Table 4. Although these scores are extremely low, the overall BLEU scores agree with overall annotator judgments in all cases except when comparing MT-CHAT and IR-RESPONSE. It would seem that BLEU has some agreement with human judgments on this task, but perhaps not enough to be immediately useful.

7 Conclusions

We have presented several data-driven approaches, based on information retrial and phrase-based SMT, to the task of automatically responding to open-

[6] See inter annotator agreement in table 4.

[7] Also, as one can see from the example exchanges in Table 3, fluency errors are rampant across all systems, including the gold-standard human responses.

Status	MT-CHAT	HUMAN
i want to go to a bonfire right about now.	that does sound fun bonfire i wan na go !	did my tweet bring that upon you ? hehe
I'm the invisible man.	man , you could never be invisible	i can still see your face surrounded in blue !
put ya number in my phone ..and ima call you up !	call me when u do that in your phone call :) for what 's up ?	o well hw kum u dnt eva call or txt me :) ? ! ? ! lolz jpn hw r ya tho :) ? ! ? !
Gray Maynard with hair? I did not approve this.	maynard with grey hair !	i take it back, the hair isn't working for maynard.

Table 5: Examples where MT-CHAT output was preferred over HUMAN response by Turker annotators

domain linguistic stimuli.

Our experiments show that SMT techniques are better-suited than IR approaches on the task of response generation. Our system, MT-CHAT, produced responses which were preferred by human annotators over actual human responses 15% of the time. Although this is still far from human-level performance, we believe there is much room for improvement: from designing appropriate word-alignment and decoding algorithms that account for the selective nature of response in dialogue, to simply adding more training data.

We described the many challenges posed by adapting phrase-based SMT to dialogue, and presented initial solutions to several, including direct phrasal alignment, and phrase-table scores discouraging responses that are lexically similar to the status. Finally, we have provided results from an initial experiment to evaluate the BLEU metric when applied to response generation, showing that though the metric as is does not work well, there is sufficient correlation to suggest that a similar, dialogue-focused approach may be feasible.

By generating responses to Tweets out of context, we have demonstrated that the models underlying phrase-based SMT are capable of guiding the construction of appropriate responses. In the future, we are excited about the role these models could potentially play in guiding response construction for conversationally-aware chat input schemes, as well as goal-directed dialogue systems.

Acknowledgments

We would like to thank Oren Etzioni, Michael Gamon, Jerry Hobbs, Dirk Hovy, Yun-Cheng Ju, Kristina Toutanova, Saif Mohammad, Patrick Pantel, and Luke Zettlemoyer, in addition to the anonymous reviewers for helpful discussions and comments on a previous draft. The first author is supported by a National Defense Science and Engineering Graduate (NDSEG) Fellowship 32 CFR 168a.

References

Ron Artstein and Massimo Poesio. 2008. Inter-coder agreement for computational linguistics. *Comput. Linguist.*, 34:555–596, December.

Regina Barzilay and Mirella Lapata. 2005. Modeling local coherence: an entity-based approach. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, ACL '05.

E. M. Bennett, R. Alpert, and A. C. Goldstein. 1954. Communications through limited-response questioning. *Public Opinion Quarterly*, 18(3):303–308.

Michael Bloodgood and Chris Callison-Burch. 2010. Using mechanical turk to build machine translation evaluation sets. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, CSLDAMT '10, pages 208–211, Morristown, NJ, USA. Association for Computational Linguistics.

Chris Brockett. 2006. Aligning the rte 2006 corpus. In *Microsoft Research Techincal report MSR-TR-2007-77*.

Peter F. Brown, John Cocke, Stephen A. Della Pietra, Vincent J. Della Pietra, Fredrick Jelinek, John D. Lafferty, Robert L. Mercer, and Paul S. Roossin. 1990. A statistical approach to machine translation. *Comput. Linguist.*, 16:79–85, June.

Chris Callison-Burch, Philipp Koehn, Christof Monz, and Josh Schroeder. 2009. Findings of the 2009 workshop on statistical machine translation. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, StatMT '09.

Nathanael Chambers and James Allen. 2004. Stochastic language generation in a dialogue system: Toward a domain independent generator. In Michael Strube and Candy Sidner, editors, *Proceedings of the 5th SIGdial Workshop on Discourse and Dialogue*, pages 9–18, Cambridge, Massachusetts, USA, April 30 - May 1. Association for Computational Linguistics.

Cristian Danescu-Niculescu-Mizil, Michael Gamon, and Susan Dumais. 2011. Mark my words! Linguistic style accommodation in social media. In *Proceedings of WWW*.

Hal Daumé III and Daniel Marcu. 2009. Induction of word and phrase alignments for automatic document summarization. *CoRR*, abs/0907.0804.

Bill Dolan, Chris Quirk, and Chris Brockett. 2004. Unsupervised construction of large paraphrase corpora: exploiting massively parallel news sources. In *Proceedings of the 20th international conference on Computational Linguistics*, COLING '04, Morristown, NJ, USA. Association for Computational Linguistics.

Abdessamad Echihabi and Daniel Marcu. 2003. A noisy-channel approach to question answering. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics - Volume 1*, ACL '03, pages 16–23, Morristown, NJ, USA. Association for Computational Linguistics.

Micha Elsner and Eugene Charniak. 2008. You talking to me? a corpus and algorithm for conversation disentanglement. In *Proceedings of ACL-08: HLT*, June.

Michel Galley, Eric Fosler-Lussier, and Alexandros Potamianos. 2001. Hybrid natural language generation for spoken dialogue systems. In *Proceedings of the 7th European Conference on Speech Communication and Technology (EUROSPEECH–01)*, pages 1735–1738, Aalborg, Denmark, September.

Jon Hasselgren, Erik Montnemery, Pierre Nugues, and Markus Svensson. 2003. Hms: a predictive text entry method using bigrams. In *Proceedings of the 2003 EACL Workshop on Language Modeling for Text Entry Methods*, TextEntry '03.

Jerry R. Hobbs. 1985. On the coherence and structure of discourse.

Charles Lee Isbell, Jr., Michael J. Kearns, Dave Kormann, Satinder P. Singh, and Peter Stone. 2000. Cobot in lambdamoo: A social statistics agent. In *Proceedings of the Seventeenth National Conference on Artificial Intelligence and Twelfth Conference on Innovative Applications of Artificial Intelligence*, pages 36–41. AAAI Press.

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

Howard Johnson, Joel Martin, George Foster, and Roland Kuhn. 2007. Improving translation quality by discarding most of the phrasetable. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 967–975, Prague, Czech Republic, June. Association for Computational Linguistics.

Sangkeun Jung, Cheongjae Lee, Kyungduk Kim, Minwoo Jeong, and Gary Geunbae Lee. 2009. Data-driven user simulation for automated evaluation of spoken dialog systems. *Comput. Speech Lang.*, 23:479–509, October.

Kevin Knight and Vasileios Hatzivassiloglou. 1995. Two-level, many-paths generation. In *Proceedings of the 33rd annual meeting on Association for Computational Linguistics*, ACL '95.

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 *ACL*. The Association for Computer Linguistics.

J R Landis and G G Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*.

Brian Langner, Stephan Vogel, and Alan W. Black. 2010. Evaluating a dialog language generation system: comparing the mountain system to other nlg approaches. In *INTERSPEECH*.

Anton Leuski and David R. Traum. 2010. Practical language processing for virtual humans. In *Twenty-Second Annual Conference on Innovative Applications of Artificial Intelligence (IAAI-10)*.

Bill MacCartney, Michel Galley, and Christopher D. Manning. 2008. A phrase-based alignment model for natural language inference. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '08, pages 802–811, Morristown, NJ, USA. Association for Computational Linguistics.

Robert C. Moore. 2004. On log-likelihood-ratios and the significance of rare events. In *EMNLP*.

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51.

F. J. Och. 2003. Minimum error rate training for statistical machine translation. In *ACL*, pages 160–167.

K. Papineni, S. Roukos, T. Ward, and W. J. Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *ACL*, pages 311–318.

Chris Quirk, Chris Brockett, and William Dolan. 2004. Monolingual machine translation for paraphrase generation. In *In Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 142–149.

Owen Rambow, Srinivas Bangalore, and Marilyn Walker. 2001. Natural language generation in dialog systems. In *Proceedings of the first international conference on Human language technology research*, HLT '01, pages 1–4, Stroudsburg, PA, USA. Association for Computational Linguistics.

Adwait Ratnaparkhi. 2000. Trainable methods for surface natural language generation. In *Proceedings of the 1st North American chapter of the Association for Computational Linguistics conference*.

Sujith Ravi, Andrei Broder, Evgeniy Gabrilovich, Vanja Josifovski, Sandeep Pandey, and Bo Pang. 2010. Automatic generation of bid phrases for online advertising. In *Proceedings of the third ACM international conference on Web search and data mining*, WSDM '10.

Stefan Riezler, Alexander Vasserman, Ioannis Tsochantaridis, Vibhu Mittal, and Yi Liu. 2007. Statistical machine translation for query expansion in answer retrieval. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 464–471, Prague, Czech Republic, June. Association for Computational Linguistics.

Alan Ritter, Colin Cherry, and Bill Dolan. 2010. Unsupervised modeling of twitter conversations. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, HLT '10, pages 172–180, Morristown, NJ, USA. Association for Computational Linguistics.

Samira Shaikh, Tomek Strzalkowski, Sarah Taylor, and Nick Webb. 2010. Vca: an experiment with a multiparty virtual chat agent. In *Proceedings of the 2010 Workshop on Companionable Dialogue Systems*.

Rion Snow, Brendan O'Connor, Daniel Jurafsky, and Andrew Y. Ng. 2008. Cheap and fast—but is it good?: evaluating non-expert annotations for natural language tasks. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

Radu Soricut and Daniel Marcu. 2006. Discourse generation using utility-trained coherence models. In *Proceedings of the COLING/ACL on Main conference poster sessions*, COLING-ACL '06.

Xu Sun, Jianfeng Gao, Daniel Micol, and Chris Quirk. 2010. Learning phrase-based spelling error models from clickthrough data. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, ACL '10, pages 266–274, Morristown, NJ, USA. Association for Computational Linguistics.

Reid Swanson and Andrew S. Gordon. 2008. Say anything: A massively collaborative open domain story writing companion. In *Proceedings of the 1st Joint International Conference on Interactive Digital Storytelling: Interactive Storytelling*, ICIDS '08, pages 32–40, Berlin, Heidelberg. Springer-Verlag.

Lidan Wang and Douglas W. Oard. 2009. Context-based message expansion for disentanglement of interleaved text conversations. In *HLT-NAACL*.

Joseph Weizenbaum. 1966. Eliza: a computer program for the study of natural language communication between man and machine. *Commun. ACM*, 9:36–45, January.

Yorick Wilks. 2006. Artificial companions as a new kind of interface to the future internet. In *OII Research Report No. 13*.

Yuk Wah Wong and Raymond Mooney. 2006. Learning for semantic parsing with statistical machine translation. In *Proceedings of the Human Language Technology Conference of the NAACL, Main Conference*.

Yuk Wah Wong and Raymond Mooney. 2007. Generation by inverting a semantic parser that uses statistical machine translation. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*.