Grounded Response Generation Task at DSTC7

Michel Galley, Chris Brockett, Xiang Gao, Jianfeng Gao, Bill Dolan

Microsoft Research
One Microsoft Way
Redmond, WA 98052, USA
{mgalley,chrisbkt,xiag,jfgao,billdol}@microsoft.com

Abstract

This paper offers and overview of the "Sentence Generation" task (Task 2) at the 7th Dialog System Technology Challenges (DSTC7). In this task, the goal is to generate conversational responses that go beyond chitchat, by producing informational responses that are *grounded* in external knowledge following the framework proposed by Ghazvininejad et al. [1]. We argue that such a grounding can empower end-to-end dialogue modeling research, so far overwhelmingly devoted to chit-chat, to produce informative and ultimately more useful dialogues. We received 26 system submissions from 7 teams, and evaluated them using both automatic and human evaluation. We discuss the outcomes of these two evaluations, noting the merits and drawback of each automatic metric based on their correlation with human judgments. Finally, we briefly outline challenges for future work in grounded response generation.

Index Terms: DSTC, dialog systems, conversational AI, sentence generation, end-to-end conversation modeling, grounding

1. Introduction

Recent work [2, 3, 4, 5, 6, etc.] has shown that conversational models can be trained in a completely end-to-end and data-driven fashion, without any hand-coding. However, such prior work has been mostly applied to chitchat, as this is the salient trait of the social media data (e.g., Twitter [2]) utilized to train these systems. Such end-to-end neural conversation systems have a tendency to produce responses that are conversationally appropriate, but often bland [7], purely chatty, and lacking entities and factual content.

To effectively move beyond chitchat, fully data-driven models would need grounding in the real world and access to external knowledge (textual or structured) in order to produce system responses that are both substantive and "useful". Conventional dialog systems have the ability to inject entities and facts into responses, but often at the cost of significant hand-coding (e.g., slot filling). In DSTC6 [8], the "End-to-End Conversation Modeling" track (Track 2) offered to build fully data-driven systems trained on Twitter customer support data, therefore providing a valuable opportunity to investigate the possibility of fully data-driven conversation in a more goal-oriented scenario than casual chitchat. However, the generation of goal-oriented responses from social media faces several challenges as explained in one of the DSTC6 system descriptions [9]. First, social media is constituted almost entirely of chitchat, which requires very aggressive filtering (i.e., by customer support user IDs) of Twitter data to produce a goal-oriented dataset. This kind of filtering limits the ability of the model to learn the backbone of general conversation, and hinders its ability to mix goal-oriented responses with chatty replies. Second, task-oriented respondents (e.g., customer support) generally avoid responding publicly—often because of customer privacy concerns—and therefore tend to take the exchange offline (e.g., by email or Twitter direct message) early in the conversation, well before the task has been completed or abandoned. Such a behavior makes it difficult to devise reward functions and to measure success in the task.

In justifying this new DSTC task, we argue that dialog shouldn't necessarily be either completely goal-oriented or completely chitchat. This is often reflected in real human-human data, which often combines the two genres. There is also a wide continuum on which a dialog system can have a practical purpose: On one end, task-oriented dialog systems are designed for concrete goals, but to this day still require hand-crafting a fair amount of information specific to the domain or task. On the other end of the continuum, there are chitchat dialog systems that are sometimes seen as less useful, even though they do fulfill a social role and promote user engagement. By adopting the "knowledge-Grounded Neural Conversation Model" framework of [1], we aim to get the benefits of both worlds by augmenting end-to-end conversational modeling with textual data from the user's environment (here, a web page that is talked about). Indeed, such a framework maintains the benefit of fully data-driven conversation while attempting to get closer to task-oriented scenarios, with the goal of informing and helping the users and not just entertaining them.

2. Task

The task follows the data-driven paradigm established in 2011 by Ritter et al. [2], which avoids hand-coding any linguistic, domain, or task-specific information. In the knowledge-grounded approach of [1], that paradigm is extended as each system input consisting of two parts:

- **Conversational input:** Similarly to DSTC6 Track 2 [8], all preceding turns of the conversation are available to the system. For practical purposes, we truncate the context to the K most recent turns.
- **Contextually-relevant "facts":** The system is given snippets of text that are relevant to the context of the conversation. These snippets of text are not drawn from any conversational data, and are instead extracted from external knowledge sources (web pages in the case of this task).

From that input, the task it to produce a response that is both conversationally appropriate and *informative*. The evaluation setup to rate such a response is presented in Section 5.

Web page	[...] she holds the guinness world record for **surviving** the highest fall without a parachute : **10,160 metres** (**33,330 ft**) . [...] **four years later** , peter hornung-andersen and pavel theiner , two prague-based journalists , claimed that flight 367 had been mistaken for an enemy aircraft and shot down by the czechoslovak air force at an altitude of **800 metres** (2,600 ft) [...]
Turn 1	today i learned a woman fell **30,000 feet** from an airplane and **survived** [URL] .
Turn 2	the page states that a **2009 report** found the plane only fell **several hundred meters** .
Turn 3	well if she only fell a **few hundred meters** and survived then i 'm not impressed at all .
Turn 4	still pretty incredible , but quite a bit different that **10,000 meters** .

Table 1: Sample of the DSTC7 Sentence Generation data, which combines Reddit data (Turns 1-4) along with documents (extracted from Common Crawl) discussed in the conversations. The **emphasis** was added by us. The [URL] links to the web page above.

3. Data

We extracted conversation threads from Reddit data,[1] which is particularly well suited for grounded conversation modeling. Indeed, Reddit conversations are organized around submissions, where each conversation is typically initiated with a URL to a web page (grounding) that defines the topic of the conversation. For this task, we restrict ourselves to submissions that contain exactly one URL and a title. To reduce spamming and offensive language and improve overall the quality of the data, we manually whitelisted the domains of these URLs and the Reddit topic (i.e., "subreddits") in which they appear. This filtering yield about 3 million conversational responses and 20 million facts respectively divided into train, validation and test.[2] For the test set, we selected conversational turns for which 6 or more responses were available, in order to create a multi-reference test set. Given other filtering criteria such as turn length, this yield a 5-reference test set of size 2208 (For each instance, we set aside one of the 6 human responses to assess human performance on this task). More information about the data for this task can be found on the data extraction web site, which makes all the data extraction and evaluation code available, and lets anyone recreate the training, development, validation and test sets.[3] A sample of the data is shown in Table 1.

4. Submitted Systems

The submitted systems include sequence-to-sequence models [3, 4, 5] with memory network and related models [10, 11], copy-based mechanism [12, 13, 14], hierarchical model [6], attention mechanism [15], and variational model [16].

Here is a quick summary of the systems, according to the system descriptions or private communication:

- **TeamA:** This team did not submit a system description and details of their systems are unknown to us.
- **TeamB:** It is a sequence-to-sequence model that introduces a copying mechanism from both the conversation history and facts, so that each output token in a response can be either generated or copied from either sources using pointer-generator networks [12]. To address the problem of generating bland or meaningless responses, they propose a beam search modification that does some semantic clustering.
- **TeamC:** It is a sequence-to-sequence that models the skeleton of dialog response, which is used for pretraining. It is then fine-tuned with a Memory Network encoder [11] that utilizes top-10 related facts (retrieved by highest TF-IDF similarities).
- **TeamD:** It is an ensemble model. It consists of a Memory-augmented Hierarchical Encoder-Decoder (MHRED) that extends [17], a sentence selection module for facts retrieval, and a reranker.
- **TeamF:** It is a variational generative model. It contains a joint attention mechanism conditioning on the contexts and textual facts.
- **TeamG:** It is a variational generative model. Contexts (and response at the training stage) are encoded to extract information from the encoded textual facts using an attention mechanism. The major difference between TeamF and TeamG is that TeamG uses (context + response, facts) for attention during training, while TeamF just uses (context, facts) for attention.

5. Evaluation

We assessed the submitted systems using both human and automatic evaluation. The final ranking of the systems is solely based on human evaluation scores, and automatic evaluation were mainly used for model selection and preliminary results. However, we think the combination of automatic and human scores will help future research in devising metrics that enjoy better correlation with human scores.

5.1. Human Evaluation

The overall goal of this task is to move towards dialogues that are more 'useful' and focused on chit-chat. However, this task still remains in an open-ended setting comparable to DSTC6 Task 2 [8], in which there is no pre-specified goal, as opposed to task-oriented dialogue. In real human-to-human conversations, the interlocutor's goals may be diverse or not even known, and these goals may evolve during the course of the conversation. Therefore, we are not able to measure success in terms of percentage of times the task was completed.Instead, we performed a per-response human evaluation judging each system response according to the following criteria:

- **Relevance:** This evaluation criterion asks whether the system response is conversationally appropriate and rel-

[1] We finally used Reddit data for this task. In the original proposal, we planned to use Twitter data, but have decided to use Reddit, owing to the volatility of Twitter data, as well the technical difficulties of aligning Twitter content with data from other sources. Reddit provides an intuitive direct link to external data in the submissions that can be utilized for this task.

[2] We could have easily increased the number of web domains to create a bigger dataset, but we aimed to make the task relatively accessible for participants with limited computing resources.

[3] https://github.com/DSTC-MSR-NLP/DSTC7-End-to-End-Conversation-Modeling

	NIST				BLEU				METEOR	Diversity		Avg.
System	N-1	N-2	N-3	N-4	B-1	B-2	B-3	B-4		1-gram	2-gram	len
Baselines:												
Constant	0.175	0.183	0.184	0.184	39.7%	12.8%	6.06%	2.87%	7.48%	0.05%	0.05%	8.0
Random	1.573	1.633	1.637	1.637	26.4%	6.7%	2.24%	0.86%	5.91%	15.96%	64.66%	19.2
Seq2Seq	0.849	0.910	0.915	0.916	45.2%	14.8%	5.23%	1.82%	6.96%	1.44%	4.75%	10.6
TeamA	0.715	0.748	0.751	0.751	38.8%	11.8%	3.72%	1.48%	5.63%	9.58%	27.55%	10.5
TeamA-c1	0.788	0.831	0.834	0.834	37.1%	11.5%	3.59%	1.38%	5.70%	12.15%	30.23%	10.9
TeamA-c2	1.078	1.121	1.123	1.123	36.1%	9.5%	2.64%	0.82%	5.48%	9.74%	31.92%	12.0
TeamB	2.341	2.510	2.522	**2.523**	41.2%	14.4%	5.01%	1.83%	**8.07**%	10.89%	32.49%	15.1
TeamB-c1	1.648	1.760	1.769	1.771	41.3%	13.7%	4.91%	**1.94%**	7.64%	9.44%	26.74%	12.8
TeamC	1.419	1.509	1.515	1.515	36.8%	10.9%	3.70%	1.32%	6.43%	5.34%	17.09%	12.7
TeamC-c1	1.983	2.113	2.124	2.124	32.4%	9.9%	3.56%	1.32%	6.81%	3.79%	12.42%	16.4
TeamC-c2	1.120	1.193	1.199	1.200	37.9%	11.6%	4.17%	1.66%	6.24%	5.52%	16.87%	11.7
TeamC-c3	1.631	1.730	1.737	1.738	30.0%	8.8%	3.03%	1.21%	5.94%	3.88%	12.16%	14.9
TeamC-c4	1.431	1.532	1.542	1.543	36.3%	11.5%	4.32%	1.77%	6.55%	5.57%	18.04%	12.7
TeamD	1.925	2.039	2.047	2.047	37.1%	11.3%	3.66%	1.35%	6.71%	9.37%	33.37%	14.4
TeamD-c1	0.022	0.023	0.023	0.023	30.6%	6.7%	1.38%	0.34%	3.92%	2.65%	16.12%	6.2
TeamD-c2	0.699	0.729	0.730	0.730	37.0%	9.3%	2.64%	0.58%	5.65%	4.93%	31.29%	10.4
TeamD-c3	0.733	0.765	0.766	0.766	36.9%	9.2%	2.62%	0.68%	5.61%	4.86%	30.94%	10.5
TeamD-c4	0.530	0.554	0.555	0.555	34.9%	8.8%	2.59%	0.76%	5.24%	6.88%	35.24%	9.8
TeamD-c5	1.701	1.797	1.802	1.802	36.9%	10.7%	3.25%	0.92%	6.45%	5.79%	29.20%	13.5
TeamD-c6	1.639	1.742	1.748	1.749	40.3%	12.5%	3.80%	1.10%	6.74%	5.07%	20.71%	13.1
TeamE	1.419	1.509	1.515	1.515	36.8%	10.9%	3.70%	1.32%	6.43%	5.34%	17.09%	12.7
TeamE-c1	1.983	2.113	2.124	2.124	32.4%	9.9%	3.56%	1.32%	6.81%	3.79%	12.42%	16.4
TeamE-c2	1.690	1.812	1.822	1.822	34.8%	11.0%	3.94%	1.55%	6.52%	4.96%	15.63%	14.0
TeamE-c3	1.794	1.917	1.927	1.928	35.0%	10.9%	3.94%	1.51%	6.65%	4.63%	15.21%	14.3
TeamF	0.008	0.009	0.009	0.009	33.9%	10.2%	3.07%	1.01%	4.56%	6.36%	17.65%	5.4
TeamF-c1	0.005	0.005	0.005	0.005	32.5%	9.0%	3.11%	1.26%	4.10%	2.43%	7.17%	5.1
TeamF-c2	0.038	0.041	0.041	0.041	36.4%	11.2%	3.97%	1.45%	5.02%	8.42%	22.40%	6.3
TeamG	2.181	2.312	2.322	2.322	34.9%	10.6%	3.67%	1.21%	7.18%	3.36%	26.47%	16.6
TeamG-c1	1.938	2.034	2.039	2.040	29.2%	8.2%	2.81%	1.05%	7.48%	10.84%	44.86%	22.3
Human	2.424	2.624	2.647	2.650	34.1%	12.4%	5.72%	3.13%	8.31%	16.66%	67.01%	18.8

Table 2: Automatic evaluation results. Participants submitted primary and contrastive systems, the latter being identified with a -cX suffix in their names. The primary systems (TeamA, TeamB, . . .) were the ones selected by the participants for human evaluation (Table 3). The primary metrics NIST-4, BLEU-4, and METEOR, whose best scores are in bold font.

	Relevance		Interest		Overall	
System	Mean Score	95% CI	Mean Score	95% CI	Mean Score	95 % CI
Baselines:						
Constant	2.60	(2.560, 2.644)	2.32	(2.281, 2.364)	2.46	(2.424, 2.500)
Random	2.32	(2.269, 2.371)	2.35	(2.303, 2.401)	2.34	(2.288, 2.384)
Seq2Seq	2.91	(2.858, 2.963)	2.68	(2.632, 2.730)	2.80	(2.748, 2.844)
TeamA	2.32	(2.267, 2.368)	2.30	(2.252, 2.351)	2.31	(2.262, 2.358)
TeamB	2.99	(2.938, 3.042)	**2.87**	(2.822, 2.922)	**2.93**	(2.882, 2.979)
TeamC	**3.05**	(3.009, 3.093)	2.77	(2.735, 2.812)	2.91	(2.875, 2.950)
TeamD	2.69	(2.635, 2.743)	2.58	(2.527, 2.632)	2.63	(2.583, 2.685)
TeamF	2.52	(2.461, 2.572)	2.40	(2.352, 2.457)	2.46	(2.409, 2.512)
TeamG	2.82	(2.771, 2.870)	2.57	(2.525, 2.619)	2.70	(2.650, 2.742)
Human	3.61	(3.554, 3.658)	3.49	(3.434, 3.539)	3.55	(3.497, 3.596)

Table 3: Human evaluation results. The systems evaluated here are the same as the primary systems in Table 2. Note that we do not report the results of TeamE as their primary system was identical to TeamC's (due to miscommuication at submission time). The best system according to human evaluation (TeamB) also obtained the best NIST-4 and METEOR scores.

evant given the K immediately preceding turns (we set $K = 2$ to reduce the judges' cognitive load).[4] Note that this judgment has nothing to do with grounding in external sources, and is similar to quality human judgments for prior data-driven conversation models (e.g., [3]).

- **Interest:** This evaluation criterion measures the degree to which the produced response is interesting and informative in the context of a document provided by the URL. Since it would be impractical to show entire web pages to the crowdworkers, we restricted ourselves at training and test time to URLs with named anchors (i.e., prefixed with '#' in the URL), and the crowdworkers only had to read a snippet of the document immediately following that anchor. Note that models could utilize full web pages as input, and the decision to only show a snippet for each response was again to reduce cognitive load.

We used crowdsourcing to score both evaluation criteria on a 5-point Likert scale, and finally combined the two judgments by weighting them equally.

5.2. Automatic Evaluation

In order to provide participants with preliminary results, we also performed automatic evaluation using standard machine translation metrics, including BLEU [18], METEOR [19], and NIST [20]. NIST is a variant of BLEU that weights n-gram matches by their information gain, i.e., it indirectly penalizes uninformative n-grams such as "I don't" and "don't know".

5.3. Results

The Generation Task received 26 system submissions from a total of 7 teams, whose automatic and human evaluation results are shown respectively in Table 2 and 3. In addition to these systems, we also evaluated a "human"system (one of the six human references set aside for evaluation) and three baselines: a Seq2Seq baseline, a random baseline (which randomly selects responses from the training data), and a constant baseline that always responds "I don't know what you mean". This constant response was automatically selected as follows: we pre-selected a list of high-frequency responses from the training data, and picked the response that had the highest BLEU score on the development set. The reason for including this constant baseline is that such a deflective response generation system can be surprisingly competitive, at least when evaluated on automatic metrics (BLEU).

The findings are as follows for each of the metrics:

BLEU-4: When evaluated on 5 references, the constant baseline, which always responds deflectively, does surprisingly well (BLEU=2.87%) and outperforms all the submitted systems (BLEU4 ranging from 1.01% to 1.83%), and is only outperformed by human. This is due to the deficiency of BLEU when dealing with many references, which was first noted in [21]. In further analysis, we found that reducing the number of references to 1 or 2 solved the problem, causing the constant baseline to be outperformed by most systems. In the case of dialogue, we suspect this deficiency of BLEU exacerbates the blandness problem [7]. For example, if one of the gold responses happens to be also "I don't know what you mean", the constant baseline gets a maximum score for that instance, even if the other references are semantically completely unrelated. Thus, this biases the metric towards bland responses, as often at least one of the 5 reference is somewhat deflective (e.g., contains "I don't know").

NIST-4: The NIST score weights ngram matches by their information gains, and effectively penalizes common n-grams such as "I don't know", which alleviates the problem with multi-reference BLEU mentioned above. None of the baselines is competitive with the top systems according to NIST-4, even when using 5 references. This suggests that NIST might be a more suitable metric than BLEU when dealing with multi-reference test sets, and it penalizes bland responses.

METEOR: This metric suffers from the same problem as BLEU-4, as the constant baseline performs very well on that metric and outperforms all submitted primary systems but one. We suspect this is due to the fact that METEOR (as BLEU) does not consider information gain in its scoring.

Human Evaluation: Owing to the cost of crowd sourcing multiple systems, we limited evaluation to a sample of 1000 conversations from the full test set and used primary systems only. All systems were assigned the same conversations. Three judges were asked to rate outputs for Relevance and Interest using a 5 point Likert scale. Judges were randomly assigned and system outputs were randomly presented to the judges in context. Spammers were replaced when identified. Not unexpectedly, the constant baseline performed moderately well on Relevance (2.60), but poorly on Interest judgments, where it was statistically indistinguishable from the low random baseline (random: 2.35, constant: 2.32). The best system returned a composite score of 2.93 (Relevance: 2.99, Interest: 2.87). This compares with the human baseline of 3.55 (Relevance: 3.61, Interest: 3.49).

In sum, the two primary systems (TeamB and TeamC) that worked best for this grounded generation task use RNN-based architectures that are augmented to incorporate facts, by distinguishing two types of encoders, a dialog encoder and a fact/memory encoder [1]. Both teams do so with an attention mechanism, and the winning system (TeamB) also uses a pointer-generator approach [12].

6. Summary

In this paper, we summarized the Sentence Generation task conducted at the seventh dialog system technology challenge (DSTC7). This task challenged participants to produce *informative* responses in a fully end-to-end manner, and therefore to move beyond chit-chat. This task enabled participants to make response more informative by drawing on textual background knowledge that is external to the dialogue. In this respect, the task was significantly more challenging that the DSTC6 task that was focused on the conversational dimensions of response generation. In general, competing system outputs were judged by humans to be more relevant and interesting than our constant and random baselines. It is also clear, however, that the quality gap between human and system responses is substantial, indicating that there is considerable space for research in future algorithmic improvements.

7. References

[1] M. Ghazvininejad, C. Brockett, M. Chang, B. Dolan, J. Gao, W. Yih, and M. Galley, "A knowledge-grounded neural conversation model," *AAAI*, 2018.

[2] A. Ritter, C. Cherry, and W. Dolan, "Data-driven response generation in social media," in *Proc. of EMNLP*, 2011, pp. 583–593.

[3] A. Sordoni, M. Galley, M. Auli, C. Brockett, Y. Ji, M. Mitchell, J.-Y. Nie, J. Gao, and B. Dolan, "A neural network approach to

[4] We need to limit the number of preceding turns as to not overload MTurk workers with too much information.

context-sensitive generation of conversational responses," in *Proc. of NAACL-HLT*, May–June 2015.

[4] L. Shang, Z. Lu, and H. Li, "Neural responding machine for short-text conversation," *ACL-IJCNLP*, 2015.

[5] O. Vinyals and Q. Le, "A neural conversational model," in *Proc. of ICML Deep Learning Workshop*, 2015.

[6] I. V. Serban, A. Sordoni, Y. Bengio, A. Courville, and J. Pineau, "Building end-to-end dialogue systems using generative hierarchical neural network models," in *Proc. of AAAI*, February 2016.

[7] J. Li, M. Galley, C. Brockett, J. Gao, and B. Dolan, "A diversity-promoting objective function for neural conversation models," in *NAACL-HLT*, 2016.

[8] C. Hori and T. Hori, "End-to-end conversation modeling track in DSTC6," *arXiv:1706.07440*, 2017.

[9] M. Galley, C. Brockett, B. Dolan, and J. Gao, "The MSR-NLP system at dialog system technology challenges 6," in *Dialog System Technology Challenges 6*, 2017.

[10] J. Weston, S. Chopra, and A. Bordes, "Memory networks," *ICLR*, 2015.

[11] S. Sukhbaatar, a. szlam, J. Weston, and R. Fergus, "End-to-end memory networks," in *Advances in Neural Information Processing Systems 28*. Curran Associates, Inc., 2015, pp. 2440–2448.

[12] A. See, P. J. Liu, and C. D. Manning, "Get to the point: Summarization with pointer-generator networks," in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2017, pp. 1073–1083.

[13] J. Gu, Z. Lu, H. Li, and V. O. Li, "Incorporating copying mechanism in sequence-to-sequence learning," in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2016.

[14] S. He, C. Liu, K. Liu, and J. Zhao, "Generating natural answers by incorporating copying and retrieving mechanisms in sequence-to-sequence learning," in *ACL*, vol. 1, 2017, pp. 199–208.

[15] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in *Proc. of the International Conference on Learning Representations (ICLR)*, 2015.

[16] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2013.

[17] I. V. Serban, A. Sordoni, Y. Bengio, A. C. Courville, and J. Pineau, "Building end-to-end dialogue systems using generative hierarchical neural network models," in *Proc. of AAAI*, 2016.

[18] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "BLEU: a method for automatic evaluation of machine translation," in *Proc. of ACL*, 2002.

[19] A. Lavie and A. Agarwal, "METEOR: An automatic metric for mt evaluation with high levels of correlation with human judgments," in *Proc. of the Second Workshop on Statistical Machine Translation*, ser. StatMT '07. Stroudsburg, PA, USA: Association for Computational Linguistics, 2007, pp. 228–231.

[20] G. Doddington, "Automatic evaluation of machine translation quality using n-gram co-occurrence statistics," in *Proceedings of the Second International Conference on Human Language Technology Research*, ser. HLT '02, 2002, pp. 138–145.

[21] R. Vedantam, C. L. Zitnick, and D. Parikh, "CIDEr: Consensus-based image description evaluation," in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, 2015, pp. 4566–4575.