Automatic evaluation of end-to-end dialog systems with adequacy-fluency metrics

Luis Fernando D'Haro , Rafael E. Banchs , Chiori Hori , Haizhou Li

Abstract

End-to-end dialog systems are gaining interest due to the recent advances of deep neural networks and the availability of large human–human dialog corpora. However, in spite of being of fundamental importance to systematically improve the performance of this kind of systems, automatic evaluation of the generated dialog utterances is still an unsolved problem. Indeed, most of the proposed objective metrics shown low correlation with human evaluations. In this paper, we evaluate a two-dimensional evaluation metric that is designed to operate at sentence level, which considers the syntactic and semantic information carried along the answers generated by an end-to-end dialog system with respect to a set of references. The proposed metric, when applied to outputs generated by the systems participating in track 2 of the DSTC-6 challenge, shows a higher correlation with human evaluations (up to 12.8% relative improvement at the system level) than the best of the alternative state-of-the-art automatic metrics currently available.

Keywords: Automatic evaluation metrics; dialog systems; DSTC; adequacy and fluency

1. Introduction

Nowadays, there is a huge research interest on end-to-end dialog systems based on deep neural approaches (Vinyals and Le, 2015; Sordoni et al., 2015; Shang et al., 2015; Bordes et al., 2016; Serban et al., 2016). Such systems allow to quickly deploy dialog engines that can interact with human users in an intelligent and contextually relevant manner, while integrating all intermedium modules required in traditional dialog systems (e.g. natural language understanding grammars, dialog management, or state and topic tracking) into a single engine. This is possible by taking advantage of large volume of human-human dialog data that is available online from forums

(Lowe et al., 2015), movie scripts (Banchs, 2012; Danescu-Niculescu-Mizil and Lee, 2011), and Twitter (Petrović et al., 2010) among many others (Serban et al., 2015). However, while it is of fundamental importance to systematically improve system performance, automatic evaluation of system generated responses is still an unsolved problem. Indeed, it has been shown in the literature (Liu et al., 2016; Galley et al., 2015) that the currently used metrics do not generally correlate well with human assessment of dialog quality.

In this paper, we extend adequacy-fluency metrics (AM–FM), originally proposed as a machine translation evaluation framework (Banchs et al., 2015; Banchs and Li, 2011), as an automatic dialogue evaluation metric. The paper presents a comparative analysis against other common proposed metrics and evaluates their correlations with human generated assessments of dialog quality. More specifically, our experiments are conducted on the set of outputs generated by the systems participating in the "End-to-End Conversation Modeling track" of DSTC-6[1] (Hori et al., 2018). The result of this comparative evaluation suggests that AM–FM significantly correlates better than other metrics w.r.t human evaluations for the considered challenge data. In extending the original metric, we have also introduced a modification to the original fluency metric (FM) to adapt it to the evaluation of dialogue systems, which is described and justified in detail. Finally, the paper also provides insights on the situations where AM–FM is able to achieve a high performance and also where it fails, as well as providing supportive reasons (i.e. interpretability) which are useful features to help evaluators to optimize the performance of the metric.

The rest of the paper is organized as follows. Section 2 presents related work on automatic evaluation metrics for dialog systems. Section 3 provides details about the AM–FM evaluation framework. Section 4 presents the comparative evaluation between AM–FM and other state-of-the-art metrics in terms of correlations with human generated evaluations over the challenge data. Section 5 describes and discusses additional experiments carried out to explore the properties of AM–FM in the context of dialog. Finally, Section 6 presents our conclusions and plans for future research.

2. Related work

Dialog systems have been a topic of interest for several years given its potential to automate many services such as call centers, information retrieval systems, control of domestic appliances, entertainment, etc. A dialog system is expected to understand user's requests, ask follow-up questions in case that clarifications are required, and successfully provide or perform the required information or action (i.e. task-oriented dialog), but also to keep the user engaged by providing meaningful answers and behavior based on the history of interactions and current context (i.e. non-task-oriented dialogs or chitchat). In both cases, it is important to be able to evaluate the performance of such dialog systems in the research and system deployment stages. Therefore, a good number of dialog evaluation frameworks, including subjective and objective metrics, have been proposed in order to measure the quality and success of the interaction between the users and dialog systems, especially for task-oriented dialog systems (Hastie, 2012). Some of the subjective metrics include qualitative evaluations for different aspects of the dialog (e.g. handling of errors, easiness to request and obtain help, naturalness of the voice and prompts), user satisfaction surveys, perceived intelligence and completeness of the answers provided by the system, etc. This information is usually collected from surveys presented to the user at the end of the interaction (Hone and Graham, 2000). On the other hand, regarding objective metrics, quantifiable observations such as dialog length, number of successfully completed tasks, word error rate, accuracy of state tracking values and slots, comparison between ground truth prompts and dialog prompts and answers at each turn, etc. are typically considered.

One well known framework for the evaluation of dialog systems is PARADISE (Walker et al., 2000), which seeks to optimize a given quality metric by formulating it as a linear combination of other metrics, e.g. optimizing the user's satisfaction based on increasing task success while reducing dialog length. Another important approach has been proposed in the context of Reinforcement Learning where dialog systems and their answers are automatically trained and evaluated based on learning dialog policies from data or simulated users (Li et al., 2016; Williams and Zweig, 2016). Here, the policy is formulated using a reward function where the system earns points for a full or partially successful dialog (i.e. some part of the requested information is given to the user or confirmed) and losses points if the dialog is too long (i.e. the system requires several turns to get or confirm the information making it less

attractive to the final user) (Schatzmann et al., 2006; Henderson et al., 2005; Khouzaimi et al., 2018). However, in some scenarios or applications, the system could earn points, instead of losing them, when the dialog length is longer as this may indicate user engagement (Foster et al., 2009).

Unfortunately, these frameworks have two main limitations: first, they are mainly applicable to task-oriented dialog where success might be clearly defined and, second, they will fail when the action or answer generated by the system is not the same as the referenced user action. In this latter case, when system performs in a way different from what is given in the ground truth, it is hard to perform comparisons for the following turns, even if the proposed dialog or action is better than the one included in the reference.

In the case of non-task-oriented dialogs, on the other hand, research has been mainly focused on evaluating responses based on subjective metrics such as appropriateness (Robinson et al., 2010) and coherence between human judgements, as well as objective metrics such as word-overlap (e.g. BLEU or ROUGE) or semantic similarity over word-embedding (i.e. extensions or derivations from word-based embedding such as Word2Vec (Mikolov et al., 2013) or Glove (Pennington et al., 2014). In any case, the performance of these metrics, whose correlations with human judgements differs on different studies, are still to be explored and studied in a comprehensive manner. For instance, in Liu et al. (2016) low or almost none correlation is found, while low to moderate correlation are reported by Galley et al. (2015) and, showing an opposite trend, high correlations in constrained domains applications or where low diversity in the answers is required, are reported in Wen et al. (2015).

3. The AM–FM evaluation framework

The AM–FM evaluation framework (Banchs et al., 2015; Banchs and Li, 2011) was originally proposed for the automatic evaluation of machine translation.[2] It is intended to access quality across two linguistic dimensions: semantics and syntax, inspired in the quality-assessment concepts of adequacy and fluency (White et al., 1994). In the machine translation domain, adequacy is understood as the amount of information (meaning) preserved between the reference and the candidate translation, accounting in this manner for the semantic validity of the translation. On the other hand, fluency is understood as the quality of the construction in the target language, accounting in this manner for the syntactic validity of the translation. Most of state-of-the-art evaluation metrics assess the translation based on lexical similarity. The AM–FM evaluation framework marks a departure from the traditional methods by training a scoring mechanism that incorporates both, syntactic and semantic knowledge, about the language. In order to move the AM–FM framework into the evaluation of generative answers in the context of dialog systems, we borrow the same two concepts from the machine translation domain as necessary conditions of quality for good response candidates.

3.1. Adequacy metric (AM) and fluency metric (FM)

In the AM–FM evaluation framework, adequacy is measured by means of a similarity metric in a low-dimensional embedding. In the original formulation, which is the same one used here, the continuous space embedding is obtained by applying a dimensionality reduction technique to a discrete bag-of-words representation of the sentences[3]. More specifically, the Latent Semantic Indexing (LSI) approach is used (Landauer et al., 1998).

The AM score is then computed in the low dimensional embedding by calculating the cosine similarity between the system produced response and the target reference response. More specifically, similarities between the system outputs and the available reference answers are computed as follows:

$$AM_{Score} = \frac{r^T U_{MxL} (t^T U_{MxL})^T}{\left| r^T U_{MxL} \right| \left| t^T U_{MxL} \right|} \tag{1}$$

where U_{MxL} is the projection matrix learned by means of LSI, M is the vocabulary size, L is the dimensionality of the embedding, r and t are corresponding vector representations of the candidate response and the target reference

response, and || is the L2-norm operator. As the projected vectors might include small negative values, which can result in angles greater than 90° between vectors, the range of possible values for AM is restricted to the interval [0, 1] by truncating negative similarity values.

In our experiments, we trained one single SVD matrix using only 10,000 randomly selected sentences from Twitter, taking care of not using any of the test data specified in the DSTC-6 challenge. The main motivation for selecting only 10 K sentences is to train the metric with much less data that other existing embedding approaches (see Section 4.1.2). This illustrates the ability of the proposed implementation to build good models with a relatively small in-domain dataset, which can be critical for evaluating performance on new domains or languages where training data is scarce. Additionally, previous reported results for the AM−FM framework in the context of machine translation show that 10 K training sentences was enough to provide a good performance on different corpora and tasks. Finally, the computational time required for estimating the SVD matrix for 10 K is much relatively small. In our implementation, the SVD matrix was computed using the scikit-learn toolkit (Pedregosa et al., 2011).

In the case of fluency, in the original formulation of the AM−FM evaluation framework, it is estimated by using a normalized language model probability. More specifically, the probability for each sentence is calculated by means of an n-gram language model (Manning and Schutze, 1999). For normalization purposes, a compensation factor is introduced by dividing the log-probability space by the number of words contained in the sentence:

$$Prob_{sentence} = exp\left(\frac{1}{N}\sum_{k=1}^{N} log(p(w_k|w_{k-n+1}...w_{k-1}))\right) \tag{2}$$

where N is the length of the sentence in number of words, w_k is the word at position k, and $p(w_k|w_{k\text{-}n+1}...w_{k\text{-}1})$ are the n-gram probabilities.

Different from the original AM−FM implementation, our FM score is computed in a slightly different way (3). In its original definition (2), the FM score consists of an absolute normalized score, while the new proposed formula provides a relative normalized score with respect to the reference score.

$$FM_{Score} = \frac{min(prob\_candidate,\ prob\_reference)}{max(prob\_candidate,\ prob\_reference)} \tag{3}$$

Here, we estimate FM by calculating the ratio between the minimum and the maximum language model probabilities, as defined in (2), of the system generated response and the target reference response instead of using only the probability of the hypothesized sentence as it is commonly done in machine translation applications. The motivations and empirical reasons to propose this extension of the FM score are:

a. Traditionally, in machine translation, fluency is just evaluated over the hypothesis alone (i.e. without reference to the target reference). However, when evaluating dialogue systems, we want to compare the fluency of the hypothesis with respect to the reference. In this way, the metric sets a score of 1.0 when both the reference and hypothesis are the same or near to one when they share the same level of perplexity; On the other hand, it sets a score near to 0.0 when they are completely different in perplexity. Therefore providing additional information than in the original formulation.
b. Because the new formulation keeps the simplicity and normalized range as in the original formula allowing the evaluator to compare the contribution of the FM score against the contribution given by the AM score when interpolating them, and
c. Finally, the proposed formula not only makes the FM score independent of the length of the reference and hypothesis sentences, but also provides a suitable solution when having multiple references.

On the other hand, the goal of this ratio is to measure how similar are both sentences, with respect to their language model probabilities, without requiring them to be exactly the same (as it could be the case when using other traditional machine translation metrics like BLEU, METEOR or ROUGE, see Section 4.1). For instance, consider the following example of a user's utterance, the available human references and some given hypothesis:

U: What is your age?
Ref1: I'm quite young. Thank you.
Ref2: Oh, please don't ask me that kind of questions.

Ref3: I'm 21 years old.
Ref4: Let's talk about another thing.
Hyp1: Why you don't tell me your age first?
Hyp2: I'm 21 years old.

In our formulation, if the hypothesis and one of the multiples references are exactly the same (e.g. Hyp2 and Ref3) then the score will be 1.0. However, the score could be also 1.0 if both sentences have the same language model probability, which results from exhibiting a similar syntactic or grammatical construction or complexity regardless of having the same semantic information or not (i.e. Hyp1 vs Ref1, Ref2 or Ref4). In addition, note that by taking the max probability in the denominator but of either the hypothesis or the reference, we are not just normalizing the scores, we are also doing a relative evaluation of the quality of one sentence over the one selected to be in the denominator. For evaluation dialogue systems this seems a reasonable option since most of the times the denominator will correspond to one of the human references, but when dealing with informal dialogues (i.e. containing slang, infrequent words, misspellings, or very long sentences) it could happen that the generated hypothesis could be more syntactically correct (e.g. a safe answer), then the overall effect is that the FM score will be low therefore measuring the uncertainty on how to consider the hypothesis w.r.t. the references.

Finally, from a practical point of view, the new formulation follows a similar pattern of contribution when compared with the original formulation for machine translation since in both domains the FM component got a similar interpolation weight (see Section 5.3), which seems reasonable since humans tend to give more importance to the semantic content of a sentence (adequacy) than to its syntactic quality (fluency). As future work we plan to extend this new formulation to consider more situations that could be particular when evaluating dialogue systems in contrast to the more formal structure of texts in machine translation.

In our experiments, we trained one single n-gram language model using all the training data collected and preprocessed from Twitter as indicated by the track organizers. The models were computed with the SRILM toolbox (Stolcke, 2002).

3.2. Metric combination and multi-reference AM–FM

In the AM–FM framework, the AM and FM scores can be weighted and combined in different ways. More specifically, it has been shown that either the harmonic, arithmetic or geometric means can be effectively used to combine AM and FM into a single score (Banchs et al., 2015). While in the original formulation AM (1) and FM (2) were used, in our proposed implementation AM (1) and FM (3) will be used instead, and combined by using a weighted arithmetic mean:

$$Final_{Score} = \lambda * AM_{Score} + (1-\lambda) * FM_{Score};\ where\ \lambda \in [0.0 - 1.0] \tag{4}$$

where λ is the score combination factor, which can be optimized, over a development data set, to maximize the correlation between the resulting score and human generated scores.

Another important difference between our implementation here and the original formulation of the AM–FM evaluation framework is that, because of the availability of multiple references in the DSTC-6 dataset under consideration, we explore possible extensions of both metrics, AM and FM, for the case of multiple references, e.g. the example provided in the previous section in which the user asks for the age. The simplest way of implementing the multi-reference AM–FM is by computing scores over the different available references and averaging their values. In the case of the AM metric, it is also possible to compute the centroid for the embedding representations of all the references and compute the cosine similarity between it and the system response. Another possible option, for both AM and FM, is to retain the maximum scores among the different references, instead of averaging them.

We hypothesise that given the syntactic nature of FM, averaging scores across the different references might be the best way to proceed, especially in situations where there are more variability in the provided references. On the other hand, different from machine translation evaluation, where the preservation of the meaning (semantics) across the languages is intrinsic to the task, in the case of dialog we can expect different semantic representations for equivalently valid responses. Based on these observations, we assume that retaining the maximum score across the different references might be a better strategy for the case of AM. Our final implementation follows the

retain-the-maximum-score strategy.[4] It is worth noticing that the other objective metrics used for the comparative evaluation are calculated also by selecting the maximum score among all references (Sharma et al., 2017), except for the CIDEr metrics that already considers aggregation across multiple references.

4. Comparative evaluation of automatic evaluation metrics

In order to evaluate the performance of the AM−FM framework within the context of dialog evaluation, a comparative analysis was conducted considering the eight different objective metrics proposed by the track organizers at DSTC-6 (Hori and Hori, 2017). All these metrics were compared with AM−FM in terms of their correlation coefficient with human assessments of response quality.

4.1. Automatic evaluation metrics

The eight evaluation metrics used in the DSTC-6 challenge under consideration include two types of metrics: word-overlap-based metrics (BLEU, METEOR, ROUGE-L, and CIDEr) and embedding-based similarity metrics (Skip-Thoughts Cosine Similarity, Embedding Average Cosine Similarity, Vector Extrema Cosine Similarity, and Greedy Matching Score). These metrics were selected by the DSTC-6 organizers because they constitute the most commonly used automatic evaluation metrics for generative systems in different evaluation campaigns, including machine translation, summarization, and natural language generation. A previous comparative study of these eight metrics, similar to the one presented here, is also available in Liu et al. (2016). Next, we present a brief summary of each of these metrics. For a more comprehensive description and additional information, please refer to Sharma et al. (2017) and the available public implementation nlg-eval.[5]

4.1.1. Word overlap-based metrics

This first set of objective metrics evaluates the amount of word-overlap between the generated hypothesis and the ground-truth responses. All these metrics focus mainly on syntactic similarities which have shown some usability in machine translation and automatic summarization but limited in the context of natural language generation (Liu et al., 2016).

- *BLEU.* Proposed by Papineni et al. (2002), it uses a modified form of precision to compare the occurrence of n-grams of different order that are common between a translation and multiple references. It is used in machine translation (Callison-Burch et al., 2011; Graham et al., 2015), sentence generation (Stent et al., 2005; Espinosa et al., 2010) and natural language inference (Starc and Mladenić, 2017).
- *METEOR.* Proposed by Satanjeev and Lavie (2005), it provides a better correlation at the sentence level with human evaluation since the score is calculated not only on exact matches but also on stem, synonym, and paraphrase matches. However, it is also limited to focusing on local context matches that does not necessarily guarantee the semantic similarity between candidates and reference translations.
- *ROUGE.* Described in Lin (2004), it consists of a set of metrics widely used in summarization and machine translation tasks. The nlg-eval toolkit computes ROUGE-L, which is an F-measure that identifies the longest common subsequence of n-grams between the candidate and the reference utterances as a measure of sentence-level structure similarity.
- *CIDEr.* Proposed by Vedantam et al. (2015), it measures the consensus or similarity of a generated sentence against a set of ground truth references in the context of image description evaluations. The metric encodes the co-occurrences of n-grams that appear in both the hypothesis and reference. The similarity across them is estimated by an average cosine similarity over a TF-IDF vector space model.

[4] After some empirical exploration of different strategies, it seems to be the case that the best one is to retain the maximum score across references. One reason for this could be due to the usage of cosine similarity, which forces the metric to select the closest reference then allowing the hypothesis to be evaluated against the best match, helping the metric to better discriminate among systems. On the contrary, when using the average, if the references are too different among them then the discriminative information could be diluted therefore reducing the correlations. In any case, the observed differences between using max or average were not significant, therefore additional experiments on different datasets will be required to confirm this hypothesis.

[5] https://github.com/Maluuba/nlg-eval

4.1.2. Embedding-based metrics

In order to overcome the limitations of the metrics based on word overlaps, Liu et al. (2016) proposed a set metrics where the cosine similarity between the predicted and the reference sentence is calculated over a continuous space representation.[6]

- *SKIP-THOUGHT.* The Skip-Thought model, proposed by Kiros et al. (2015), uses a recurrent neural network to map a sentence into an embedding representation. Then uses this representation to predict the preceding and following sentences. The generated embeddings have been used to estimate semantic relatedness showing a robust performance among different tasks. For our experiments, nlg-eval uses the Skip-Thought pre-trained model on the BookCorpus dataset (Zhu et al., 2015).
- *EMBEDDING AVERAGE.* This metric computes a sentence-level embedding using additive composition by averaging word embedding representations. Semantic similarity is then calculated by using cosine similarity (Foltz et al., 1998; Wieting et al., 2015). Despite of its simplicity, in our experiments, this was the metric with the second highest correlation with human judgements after our proposed metric.
- *VECTOR EXTREMA.* Proposed by Forgues et al. (2014), it builds the sentence-level embedding by taking the most extreme value for each dimension of the embedding representations of all the words composing the sentence. The semantic similarity between hypothesis and reference is estimated by using cosine similarity.
- *GREEDY MATCHING.* Introduced by Rus and Lintean (2012), it does not generate a sentence-level embedding but each word in the reference sentence is greedily matched to a word in the candidate sentence and the cosine similarity between both word embeddings is computed. All resulting word similarity scores are averaged. A similar score is also computed by reversing the roles of the candidate and reference sentences. The final score is the average of the two scores.

Finally, in order to account for the possible variability of valid responses in the objective evaluations, the DSTC-6 challenge organizers collected 10 additional human generated references for each ground truth response in the dataset. The procedure they followed was to ask 10 different annotators[7] to propose a new sentence as an alternative response for each ground truth response, given the dialog context. The annotators were asked to make their responses different from the corresponding ground truth responses while keeping the dialog topic of the conversation. Therefore, since multiple references are available, both the nlg-eval toolkit and AM—FM compute similarity scores between each system generated responses and all the corresponding references one-by-one, retaining the maximum score.

4.2. Human generated scores

As part of the evaluation campaign for the DSTC-6 challenge, the organizers collected human ratings for each of the system generated responses using a 5 point Likert Scale (5 = Very good, 4 = Good, 3 = Acceptable, 2 = Poor, and 1 = Very poor). The scores were produced by 10 different crowd-sourced workers, who were requested to rate each system response considering the dialog context and the other system responses. Unfortunately, during the human evaluation it was not possible to keep the same evaluators for all systems and sentences, therefore we could not evaluate the inter-annotator agreements nor performing additional experiments to reduce or understand the variability in the responses we can see in the last column of Table 1. However, when presenting the results in Section 5.4, we made sure to compare the same averaged human sentence scores against the scores produced by each metric, and we also provide an upper bound correlation by randomly selecting two different groups of annotators and calculating their correlations.

[6] In all cases, except for Skip-thought, the used word embeddings were extracted from the Glove 300 dimensions embeddings pretrained on Wikipedia and Gigaword datasets (6B tokens, 400 K uncased vocabulary) (Pennington et al., 2014) https://nlp.stanford.edu/projects/glove/.

[7] Crowdsourcing services from Amazon Mechanical Turk platform were used to this end.

Table 1
Human evaluation scores, system correlations and p-values for different objective evaluation metrics and ours (using the best parameters) for the 20 evaluated systems at DSTC-6.

System	BLEU_4	METEOR	ROUGE_L	CIDEr	Skip thoughts	Embed avg.	Vector extrema	Greedy matching	AM–FM	Human	
										Mean	Std
S_1	0.1619	0.2041	0.3598	0.0825	0.6379	0.9155	0.6092	0.7543	0.7571	3.3638	1.047
S_2	0.1598	0.2020	0.3608	0.0780	0.6451	0.9113	0.6059	0.7527	0.7669	3.4415	1.024
S_3	0.1623	0.2039	0.3567	0.0828	0.6386	0.9060	0.6091	0.7524	0.7571	3.4298	1.026
S_4	0.1504	0.1826	0.3446	0.0803	0.6446	0.9093	0.5983	0.7488	0.7501	3.4453	1.024
S_5	0.2118	0.2140	0.3953	0.1060	0.7072	0.9281	0.6388	0.7724	0.7651	3.3894	1.045
S_6	0.1851	0.2040	0.3748	0.0965	0.6703	0.9136	0.6167	0.7571	0.7648	3.4778	1.026
S_7	0.1532	0.1833	0.3469	0.0800	0.6458	0.9099	0.5991	0.7499	0.7520	3.4382	1.025
S_8	0.2205	0.2210	0.4102	0.1279	0.6637	0.9277	0.6463	0.7773	0.7701	3.4332	1.014
S_9	0.1602	0.2016	0.3606	0.0782	0.6474	0.9103	0.6050	0.7512	0.7623	3.4504	1.022
S_10	0.1779	0.2085	0.3829	0.0978	0.6257	0.9215	0.6120	0.7647	0.7737	3.5239	1.027
S_11	0.1741	0.2024	0.3703	0.0994	0.6348	0.9021	0.6026	0.7515	0.7606	3.5082	1.011
S_12	0.1342	0.1762	0.3366	0.0947	0.6123	0.8831	0.5931	0.7315	0.7527	3.5107	1.004
S_13	0.1092	0.1731	0.3201	0.0702	0.6129	0.9014	0.5897	0.7344	0.7507	3.3919	1.033
S_14	0.1716	0.2071	0.3671	0.0898	0.6531	0.9127	0.6092	0.7548	0.7579	3.4431	1.024
S_15	0.1480	0.1813	0.3388	0.1025	0.6125	0.9104	0.5935	0.7394	0.7616	3.5209	0.997
S_16	0.0991	0.1687	0.3146	0.0708	0.5944	0.9010	0.5685	0.7204	0.7486	3.3054	1.064
S_17	0.1448	0.1839	0.3375	0.0940	0.6017	0.9103	0.5921	0.7397	0.7639	3.5396	0.998
S_18	0.1261	0.1754	0.3310	0.0945	0.6144	0.8996	0.5820	0.7287	0.7455	3.4546	1.011
S_19	0.1575	0.1918	0.3658	0.1112	0.6453	0.9094	0.6076	0.7490	0.7652	3.5098	0.997
S_20	0.2762	0.1656	0.3482	0.1235	0.6980	0.8054	0.5852	0.7202	0.6512	2.9906	0.935
Reference										3.7245	1.010
Pearson correlation	−0.5108	0.3628	0.1450	−0.1827	−0.4563	0.7768	0.2345	0.4028	**0.8907**	1.0000	
p-value	0.0214	0.1159	0.5420	0.4408	0.0432	0.0001	0.3196	0.0782	**1.41E−7**	0.0000	

For each annotation exercise, a total of 21 responses were presented to the annotators, corresponding to the 19 submitted systems, the baseline system and the ground truth. The organizers instructed the annotators to consider the informativeness, naturalness, and appropriateness of each response, given the provided dialog context. When identical responses were provided by different systems, they were reduced to one single instance with the objective of ensuring consistent ratings for such cases.

4.3. Correlation between automatic metrics and human generated scores

The Pearson correlation coefficients between the different automatic evaluation metrics and the proposed implementation of the AM–FM evaluation framework are presented in Table 1, along with the human generated scores, and the automatically computed scores for each of the 20 systems available: the baseline system and the 19 challenge submissions.

Each figure reported in the upper section of the table represents the average value of the corresponding score (column) over all responses generated by one of the systems (row). In the lower section of the table (last two rows) the Pearson correlation coefficients and their corresponding p-values between each automatic metric (all-but-last columns) and the human generated scores (last column) are presented.

As seen from the table, although the averaged human scores for all systems look very close among them, the scores given by the AM–FM metric significantly (p-value) shows the best correlation with the human evaluations. The hyper-parameter settings for the implementation of AM–FM used here is discussed in further details in Section 5. In the table, it can also be observed that only five metrics, out of the nine considered, exhibit correlation coefficients with significance $p < 0.05$. These are BLEU, CIDEr, Skip Thoughts, Embedding Average and AM–FM. From these, only two exhibit correlation coefficients greater than 0.6 and with significance $p < 0.01$. These are Embedding Average and AM–FM. In the following section we present a more specific comparison on the correlations between human scores and the Embedding Average and AM–FM scores at sentence level to provide better insights on why the AM–FM metric outperforms the average embedding.

5. Further detailed analysis on AM–FM parameterisation

Considering the formulations for the AM–FM scores (i.e. Eqs. (1) and (3)) and the calculation of the final score (i.e. Eq. (4)), there are three hyper-parameters to be adjusted: the dimensionality of the reduced space for the AM metric, the order of the n-gram for the FM metric, and the weighting parameter λ for the linear combination. Given the purpose of the AM–FM metric in the context of the dialog evaluation, these parameter values should be selected to maximize the correlation between the estimated scores and the human generated ones. Different from Callison-Burch et al. (2007) where Spearman's correlations are maximized, we have used the Pearson's correlation coefficient here. Using the Pearson's correlation allows us to focus more on evaluating the significance of the association or lineal correlations between the human evaluations and the scores values given by the different metrics, rather than focusing on the ranking of the systems.

Given that human generated scores and the AM–FM scores are both generated at the sentence level, in order to obtain the scores at the system level (as presented in Table 1) we averaged the sentence scores. Since several references were available (10 generated by the Turkers and the original sentence used in the tweet) sentence level scores were obtained by computing the similarities between the generated hypothesis and each of the references one-by-one and retaining the maximum value (the same procedure was also used for all other objective metrics in nlg-eval except for CIDEr as mentioned before).

In the following three subsections we study the behavior of the AM–FM metrics both, individually and combined, when the three hyper-parameters of the model are varied. By doing this analysis, we can better assess the contribution of each component as well as its sensitivity to the fine tuning of its corresponding hyper-parameter. We also study the behavior of the combined model when the interpolation weight is varied between the two extreme cases of pure AM (λ=1) and pure FM (λ=0).

In the penultimate subsection of this chapter, we study correlations between automatic and human generated sentence-level scores for the case of the two best performing evaluation metrics: Embedding Average and AM–FM. Finally, in the last subsection of this chapter, we conduct a qualitative exploration of the performance of the AM–FM metric at the sentence level and present some specific example of successful and erroneous evaluations.

5.1. Dimensionality hyper-parameter and the AM metric

Following the same scripts and instructions provided to the participants in the challenge,[8,9] we downloaded and pre-processed 16,858 dialogs from Twitter (extracted from 1058 different accounts). We also ensured not overlap between the downloaded data and the test data occurred. From all downloaded dialogs, we extracted only the sentences corresponding to agents' turns, which resulted in 18,692 sentences. Then, following Banchs et al. (2015), we randomly selected 10,000 sentences, containing a total of 6059 different words, to train the AM's SVD and set the maximum number of dimension to 2500. After computing the SVD, we performed different estimations of the correlation coefficient between the AM metric and the human generated scores. To this end, we computed the AM scores at embeddings of different dimensionality by projecting the sentence representations with different subsets of singular vectors from the SVD projection matrix.

Fig. 1 shows the resulting correlations between AM and human evaluations by varying the dimensionality of the embedding. It can be seen that as the dimensionality is increased the correlation coefficient diminishes. This can be explained by the few number of topics covered by the messages extracted from the 1058 Twitter accounts, as many of them are related (e.g. Zara and Zara_Care, SubaruCanada and Subaru_USA, etc.); as well as by the reduced vocabulary size and variability of agents' answers (i.e. greetings, apologies, requests for providing additional details using private messages, etc.)

Based on the results depicted in Fig. 1, the AM's dimensionality hyper-parameter was set to 10 for the rest of our experiments.

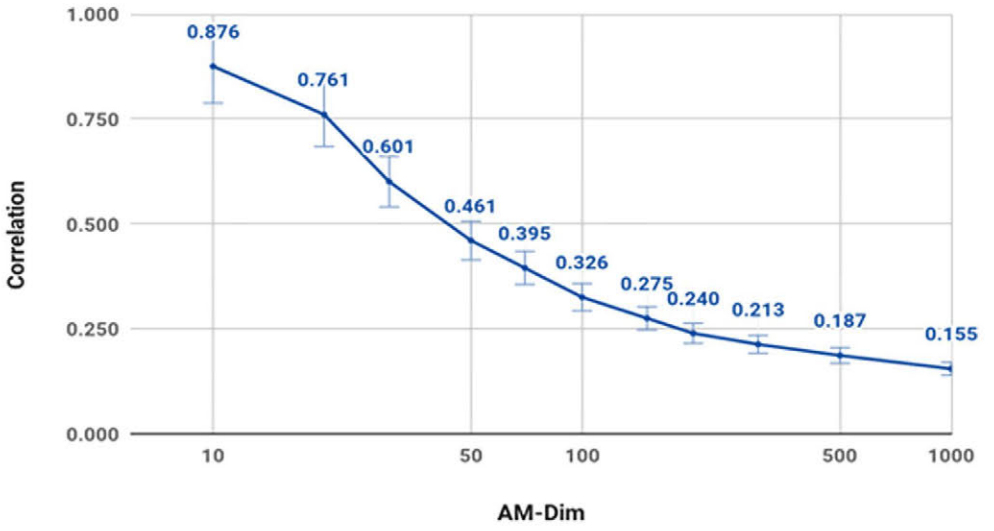


Fig. 1. Correlation values for the AM metric as a function of the dimensionality of the embedding.

Table 2
Correlation values for the FM metric as a function of the order of the n-gram language model.

N-gram Order	Number of n-grams	Avg. Correlation	p-value
Unigram	17,864	0.8128	1.33E−05
Bigram	102,898	0.8596	1.20E−06
Trigram	28,774	0.8272	6.83E−06

5.2. *N-gram order hyper-parameter and the FM metric*

In this section we study the incidence of the n-gram model order on the performance of the FM metric. For this analysis, we used the same 18,692 training sentences collected from Twitter and created Good-Turing backoff models with pruning set to 1e−6. We then computed the correlation coefficients between the obtained FM scores and the human generated ones. The results are presented in Table 2.

From the results in Table 2, we can see that even a unigram language model achieves a higher correlation with human annotations than most of the current automatic metrics presented in Table 1. We theorize that a reason for this might be the fact that many of the systems participating in the challenge were based on the sequence-to-sequence approach. It is well known that this type of system suffers from the safety answer problem (i.e. the system generates short sentences, e.g. yes, no, sure, thanks, or answers that appear very often in the training set, Serban et al., 2016). Therefore even the unigram model allows the metric to assign a low score to this kind of hypothesis when they are compared with the more rich syntactic information found in the reference sentence.[10]

Also, as seen from the table, the best performance of the FM metric, in terms of its correlation with human scores is achieved for the case of bigram. For the rest of our experiments, we used the bigram language model for the FM calculation.

5.3. *Interpolation hyper-parameter and the AM−FM metric*

In Fig. 2 we present the correlation coefficient between the combined AM−FM metric and the human generated scores when varying the interpolation hyper-parameter (λ). As seen in Eq. (4), this hyper-parameter controls the relative weight given to the AM and FM scores. In the proposed implementation, a higher value of λ assigns higher

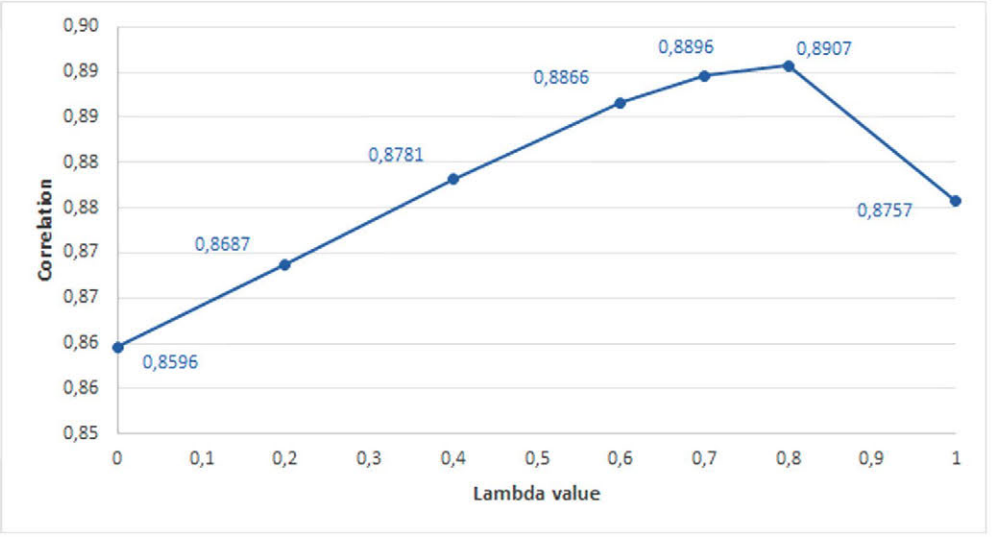


Fig. 2. Correlation values for the interpolated AM–FM metric as a function of the interpolation hyper-parameter (λ). In the figure, a higher value of λ gives more weight to the AM component than to the FM component.

weight to the AM component than to the FM component. The results in Fig. 2 show that the optimal value is λ=0.8, indicating that for this task, similar to the machine translation task, the optimal correlations with humans are found when the combined metric is bias towards AM, giving more importance to the semantic component of the metric. This effect is also manifested in the results obtained by the Average Embedding metric, which gets a high correlation (0.7768 in Table 1) by paying attention to the semantic information in the sentence embeddings. We can see from the figure that the optimized interpolated AM–FM is able to perform even better than each metric by itself, obtaining a higher correlation value with human evaluations (14.6% better) than the best of the 8 objective metrics proposed in the challenge.

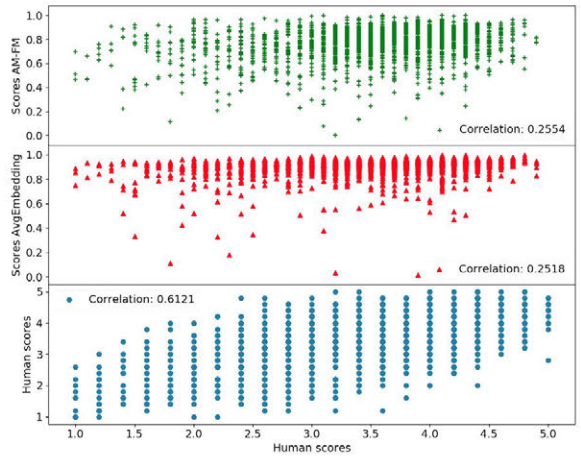
5.4. Analysis of correlations at sentence level

In the previous subsections, we have analysed how the AM–FM metric exhibits higher correlations with human evaluations at a system level, i.e. after averaging the sentence-level scores for each system and looking at their correlation with the average human score for each system (see Table 1). However, in order to better assess the weaknesses and strengths of the AM–FM framework, we need also to study the behavior of the correlations at the sentence level.

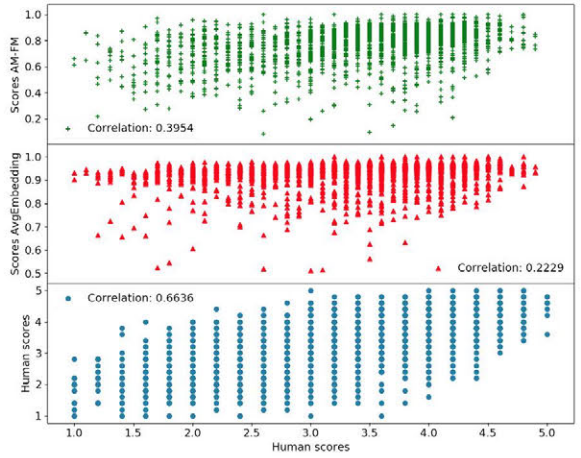
Since conducting an analysis for the 2000 sentences generated by the 20 systems might be too exhaustive, in this section we analyse the correlations at the sentence level for three selected systems: the system with the best human scores, the system with the lowest human scores, and a system with intermediate human scores. We selected these three systems to observe the performance of the metric under the same conditions that appeared in the challenge, or when evaluating a system that is able to generate an n-best list of answers from which the top result must be selected.

Fig. 3 shows the cross-plots (for the three selected systems) of their scores at a sentence-level calculated between the AM–FM score and human scores (in green), then between the Average Embedding scores and the human scores (in red), and finally an upper bound correlation by splitting into two groups the human scores, i.e. 5 raters out of the 10 evaluators were randomly selected without reposition to conform group A and the remaining raters to conform group B. Then, we averaged the scores for each group and generated the cross-plot (in blue).

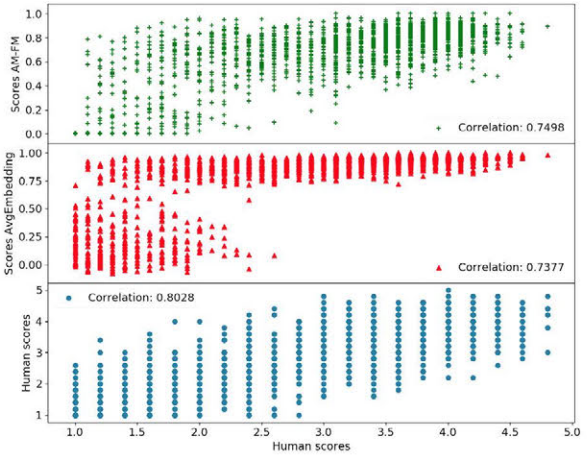
The first thing we can notice from the cross-plots presented in Fig. 3 is the difficulties (low correlation) for the human annotators to agree on the assigned evaluation scores. This is especially true for the good and middle-ranked systems in comparison to the worst system. We think this is an expected situation since the annotators were



(a) Best system (i.e. the one with the higher human evaluation)



(b) Average system (i.e. a system ranked in middle positions)



(c) Worst system (i.e. the system with the lowest human evaluation)

Fig. 3. Cross-plots for three different systems showing the correlations at a sentence level between human evaluators and the AM–FM metric (+ in top), the Average Embedding (▲ in the middle) and humans with humans (● in the bottom).

requested to evaluate the same output for the 20 systems at the same HIT. For them, it is easier to spot quickly which answer is the worst, but more difficult to decide the best with respect to the others. This can also be inferred when observing that most of the systems got a similar average score in Table 1. The difficulty of the task is also evidenced by the fact that they even gave low scores to the ground truth reference (avg. human score: 3.7245 in Table 1).

The reason for the difficulties in distinguishing the ground truth from the generated sentences could be due to most of the systems providing safe answers (e.g. apologizing when facing problems and asking users to send internal messages or just asserting/greeting when detecting the user had a nice experience and was sharing it in Twitter). This conclusion is also confirmed by Hori and Hori (2017), where the best human scores were given to those more "sympathetic" systems.

Fig. 3 clearly shows how the AM–FM evaluation framework follows a more similar pattern to the one observed in the human evaluations (i.e. exhibiting a higher correlation) for all three systems; while, on the other hand, the Average Embedding metric seems to be biased towards providing high similarity scores, and then, making more mistakes for cases when humans will agree the answer is bad (the left upper section of each graph, i.e. low human score but high metric score).

5.5. *Qualitative evaluation of results at the sentence level*

In order to qualitatively appreciate the performance of the AM–FM evaluation metric, in this section we show some examples where the AM–FM metric provides a high or low correlation with the human scores, along with their corresponding dialog contexts, ground truth and generated hypothesis. In these examples, we consider cases where the AM–FM score is higher than 0.8 (i.e. high quality) or below 0.2 (i.e. low quality); while for the human averaged scores we requested values above 4.0 with a standard deviation lower than 1.0 (i.e. high agreement between raters), or averaged score below 2.0 (also with a high agreement in the score, i.e. low standard deviation). The entire hypotheses presented here were generated by the best system submitted to the challenge.

Table 3
Qualitative examples where the metric correctly estimates the score in agreement with the average human scores, and when the metric fails to match the human scores. The provided text is the same that appears in the references, where we can see some misspellings and grammatical errors.

Prediction	Example
	**Dialog Context:** wins ! @infinitiusa ⟨URL⟩ **Ground Truth:** yay we 're happy you 're happy **Hypothesis:** we 're happy to have you as part of the infiniti family, ⟨USER⟩ ! **Human Score:** $\mu = 4.4 \pm 0.6$ **AM–FM Score:** 0.846 (AM: 0.817, FM: 0.962)
	**Dialog Context:** is there a problem with your online booking system this evening ? i 'm trying to book but nothing will load ? **Ground Truth:** i 'm sorry but we are currently facing technical issues with our site we will resolve them as soon as possible **Hypothesis:** thank you . **Human Score:** $\mu = 1.8 \pm 1.0$ **AM–FM Score:** 0.104 (AM: 0.081, FM: 0.194)
	**Dialog Context:** **U:** no one has come to collect car yet . hire car from motability got . not good service **S:** have you contacted jeep assist to collect ? who were you in agreement with ? **U:** jeep friern barnet spoke yesterday to mo **Ground Truth:** it is best to contact the dealership directly **Hypothesis:** if you'd like to speak with someone **Human Score:** $\mu = 1.4 \pm 0.5$ **AM–FM Score:** 0.912 (AM: 0.917, FM: 0.891)
	**Dialog Context:** if you haven't met @dandtwitr yet, you are missing out ! #fes2015 ⟨URL⟩ **Ground Truth:** we agree with you on that get more infor on our website **Hypothesis:** amen to that ! ! **Human Score:** $\mu = 4.3 \pm 0.6$ **AM–FM Score:** 0.215 (AM: 0.154, FM: 0.460)
	**Dialog Context:** do you delivery to serbia **Ground Truth:** i am sorry but we currently do not deliver to serbia **Hypothesis:** hi, yes we do deliver to ireland . **Human Score:** $\mu = 1.6 \pm 1.3$ **AM–FM Score:** 0.816 (AM: 0.792, FM: 0.912)

Table 3 shows examples for which the proposed metric is successful in providing a correct score to those sentences that human evaluators in average agreed by considering them as good or bad, as well as examples for which the metric is in contradiction with the human raters. If we analyse the successful examples in detail, we can see that the metric is able to assess both the semantic and syntactic similarity information between the hypothesis and the reference. For instance, in the first example it takes into account the usage of similar terms and expressions, as well as that the generated sentence is syntactically correct. In the second example, although the generated sentence is syntactically correct it semantically differs too much from the reference. Therefore the combined metric is in low agreement with the human evaluators this case.

On the other hand, we also present examples for which there are important discrepancies between humans and the AM–FM score. In the first case, we can see that the AM score is high because there is a high semantic similarity between "contacting the dealer" and the verb "speak" in the hypothesis, and that the FM score is also high as both sentences language model scores are similar. Unfortunately, the metric fails in detecting that the hypothesis is incomplete. In the second example, the metric gives a reasonable estimation for the semantic and syntactic similarities between the reference and hypothesis, but fails in not considering the pragmatics of the dialog context. Finally, the last example is interesting because the AM component gives a high score given the high similarity between the reference and the hypothesis, and the FM also provides a high score. However, the metric fails in matching the entity (Serbia instead of Ireland) which could be avoided with the incorporation of a pragmatic component, which should take into account the information given in the dialog context (i.e. reinforcing the entity). Additionally, the metric also fails in not penalizing more that the semantics is not properly preserved. For instance, the generated answer should be negative instead of affirmative. Therefore, we will explore in future works the incorporation of polarity and pragmatic components that could alleviate this kind of problems.

6. Conclusions and future work

In this paper we have extended the use of the AM–FM evaluation framework from the machine translation domain to end-to-end dialog systems. Comparative evaluations were conducted to study how the AM–FM metric correlates with human generated scores for the automatic answers generated by systems participating in the DSTC-6 Track 2 challenge. Additionally, we also provided comparisons between the correlations obtained for AM–FM scores against the correlations obtained for the 8 different objective metrics used in the challenge. The results show that the AM–FM framework provides up to 14.6% better correlations with human generated scores than the other considered metrics.

As future research, we want to study different techniques for constructing the semantic continuous space embeddings and including additional factors, like word dependencies, between references and hypothesis for the AM component (Anderson et al., 2016), as well as using continuous space language models and for the FM metric. The main objective is improving the performance of the metric by taking advantage of multiple references in a similar way as in Ferreira et al. (2018). More specifically, we want to explore implementations based in deep-autoencoders (Hinton and Salakhutdinov, 2006) and recurrent neural networks (Gers and Schmidhuber, 2001), as well as considering joint frameworks as proposed by Guzmán et al. (2017) but applied to evaluating dialog systems instead of machine translation tasks.

Finally, we will explore the idea of introducing a pragmatic metric (PM) that can consider the suitability of the generated response with respect to the dialog context. The incorporation of this component could help to disambiguate those cases where the semantic similarity between hypothesis and reference is low, but the hypothesis can be still considered as a valid answer as it does not break the coherence of the dialog. Some preliminary experimentation and results suggest that this pragmatic component might significantly help to increase the correlations with human evaluations reported in this paper.

References

Anderson, P., Fernando, B., Johnson, M., Gould, S., 2016. Spice: Semantic propositional image caption evaluation. In European Conference on Computer Vision (pp. 382-398). Springer, Cham.

Banchs, R.E., Li, H., 2011. AM-FM: A semantic framework for translation quality assessment. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics, Portland, OR, USA, pp. 153–158.

Banchs, R.E., 2012. Movie-DiC: a movie dialogue corpus for research and development. In: Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers. 2, Association for Computational Linguistics, pp. 203–207.
Banchs, R.E., D'Haro, L.F., Li, H., 2015. Adequacy-fluency metrics: evaluating MT in the continuous space model framework. IEEE/ACM Trans. Audio Speech Lang. Process. 23 (3), 472–482.
Bordes, A., Boureau, Y.L. and Weston, J. 2016. Learning end-to-end goal-oriented dialog arXiv preprint arXiv:1605.07683.
Callison-Burch, C., Fordyce, C., Koehn, P., Monz, C., Schroeder, J., 2007. (Meta-) evaluation of machine translation. In: Proceedings of the Second Workshop on Statistical Machine Translation. Association for Computational Linguistics, pp. 136–158.
Callison-Burch, C., Koehn, P., Monz, C., Zaidan, O.F, 2011. Findings of the 2011 workshop on statistical machine translation. In: Proceedings of the Sixth Workshop on Statistical Machine Translation. Association for Computational Linguistics, pp. 22–64.
Danescu-Niculescu-Mizil, C., Lee, L., 2011. Chameleons in imagined conversations: a new approach to understanding coordination of linguistic style in dialogs. In: Proceedings of the 2nd Workshop on Cognitive Modeling and Computational Linguistics. Association for Computational Linguistics, pp. 76–87.
Espinosa, D., Rajkumar, R., White, M., Berleant, S., 2010. Further meta-evaluation of broad-coverage surface realization. In: Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, pp. 564–574.
Ferreira, R., Cavalcanti, G.D., Freitas, F., Lins, R.D., Simske, S.J., Riss, M., 2018. Combining sentence similarities measures to identify paraphrases. Comput. Speech Lang. 47, 59–73.
Foltz, P.W., Kintsch, W., Landauer, T.K., 1998. The measurement of textual coherence with latent semantic analysis. Discourse Process. 25 (2–3), 285–307.
Forgues, G., Pineau, J., Larchevêque, J.-M., Tremblay, R., 2014. Bootstrapping dialog systems with word embeddings. In: Proceedings of NIPS, Modern Machine Learning and Natural Language Processing Workshop 2.
Foster, M.E., Giuliani, M., Knoll, A., 2009. Comparing objective and subjective measures of usability in a human-robot dialogue system. In: Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP. 2, Association for Computational Linguistics, pp. 879–887.
Galley, M., Brockett, C., Sordoni, A., Ji, Y., Auli, M., Quirk, C., Mitchell, M., Gao, J. and Dolan, B. 2015. deltaBLEU: A discriminative metric for generation tasks with intrinsically diverse targets. arXiv preprint arXiv:1506.06863
Gers, F.A., Schmidhuber, J., 2001. LSTM recurrent networks learn simple context free and context sensitive languages. IEEE Trans. Neural Netw. 12 (6), 1333–1340.
Graham, Y., Baldwin, T., Mathur, N., 2015. Accurate evaluation of segment-level machine translation metrics. In: Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 1183–1191.
Guzmán, F., Joty, S., Màrquez, L., Nakov, P., 2017. Machine translation evaluation with neural networks. Comput. Speech Lang. 45, 180–200.
Hastie, H., 2012. Metrics and evaluation of spoken dialogue systems. Data-Driven Methods for Adaptive Spoken Dialogue Systems. Springer, New York, NY, pp. 131–150.
Henderson, J., Lemon, O., Georgila, K., 2005. Hybrid reinforcement/supervised learning for dialogue policies from communicator data. In: Proceedings of IJCAI Workshop on Knowledge and Reasoning in Practical Dialogue Systems, pp. 68–75.
Hinton, G., Salakhutdinov, R., 2006. Reducing the dimensionality of data with neural networks. Science 313 (5786), 504–507.
Hone, K.S., Graham, R., 2000. Towards a tool for the subjective assessment of speech system interfaces (SASSI). Nat. Lang. Eng. 6 (3–4), 287–303.
Hori, C., Perez, J., Higasinaka, R., Hori, T., Boureau, Y.L., Inaba, M., Tsunomori, Y., Takahashi, T., Yoshino, K., and Kim, S., 2019. Overview of the sixth dialog system technology challenge: Dstc6. Comput. Speech Lang. 55, 1–25.
Hori, C. and Hori, T. 2017. End-to-end conversation modeling track in DSTC6, arXiv preprint arXiv:1706.07440.
Khouzaimi, H., Laroche, R., Lefèvre, F., 2018. A methodology for turn-taking capabilities enhancement in spoken dialogue systems using reinforcement learning. Comput. Speech Lang. 47, 93–111.
Kiros, R., Zhu, Y., Salakhutdinov, R., Zemel, R., Urtasun, R., Torralba, A., Fidler, S., 2015. Skip-thought vectors. In: Cortes, C., Lawrence, N.D., Lee, D.D., Sugiyama, M., Garnett, R. (Eds.), Advances in Neural Information Processing Systems. 28, Curran Associates, Inc., pp. 3294–3302.
Landauer, T.K., Foltz, P.W., Laham, D., 1998. Introduction to latent semantic analysis. Discourse Process. 25, 259–284.
Li, J., Monroe, W., Ritter, A., Galley, M., Gao, J. and Jurafsky, D. 2016 Deep reinforcement learning for dialogue generation. arXiv preprint arXiv:1606.01541.
Lin, C.-Y. Rouge: A Package for Automatic Evaluation of Summaries. Text Summarization Branches Out. 2004.
Liu, C.-W., Lowe, R., Serban, I., Noseworthy, M., Charlin, L., Pineau, P., 2016. How not to evaluate your dialogue system: an empirical study of unsupervised evaluation metrics for dialogue response generation. In: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, pp. 2122–2132.
Lowe, R., Pow, N., Serban, I. and Pineau, J. 2015. The ubuntu dialogue corpus: A large dataset for research in unstructured multi-turn dialogue systems. arXiv preprint arXiv:1506.08909.
Manning, C.D., Schutze, H., 1999. Foundations of Statistical Natural Language Processing. The MIT Press, Cambridge, MA ch. 6.
Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J., 2013. Distributed representations of words and phrases and their compositionality. In: Proceedings of Advances in Neural Information Processing Systems, pp. 3111–3119.
Papineni, K., Salim, R., Todd, W., Wei-Jing, Z., 2002. BLEU: a method for automatic evaluation of machine translation. In: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics. Association for Computational Linguistics, pp. 311–318.
Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., 2011. Scikit-learn: machine learning in python. J. Mach. Learn. Res. 12 (Oct), 2825–2830.
Pennington, J., Socher, R., Manning, C., 2014. Glove: global vectors for word representation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP, pp. 1532–1543.

Petrović, S., Osborne, M., Lavrenko, V., 2010. The edinburgh twitter corpus. In: Proceedings of the NAACL HLT 2010 Workshop on Computational Linguistics in a World of Social Media, pp. 25–26.

Robinson, S., Roque, A., Traum, D.R., 2010. Dialogues in context: an objective user-oriented evaluation approach for virtual human dialogue. In: Proceedings of Language Resources and Evaluation Conference, LREC.

Rus, V., Lintean, M., 2012. A comparison of greedy and optimal assessment of natural language student input using word-to-word similarity metrics. In: Proceedings of the Seventh Workshop on Building Educational Applications Using NLP. Association for Computational Linguistics, pp. 157–162.

Schatzmann, J., Weilhammer, K., Stuttle, M., Young, S., 2006. A survey of statistical user simulation techniques for reinforcement-learning of dialogue management strategies. Knowl. Eng. Rev. 21 (2), 97–126.

Satanjeev, B., Lavie, A., 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In: Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization, pp. 65–72.

Serban, I.V., Lowe, R., Henderson, P., Charlin, L. and Pineau, J. 2018. A Survey of Available Corpora For Building Data-Driven Dialogue Systems: The Journal Version. Dialogue & Discourse, 9(1), pp.1–49.

Serban, I.V., Sordoni, A., Bengio, Y., Courville, A.C., Pineau, J., 2016. Building end-to-end dialogue systems using generative hierarchical neural network models. In: Proceedings of AAAI, 16, pp. 3776–3784.

Shang, L., Lu, Z., and Li, H. 2015. Neural responding machine for short-text conversation. arXiv preprint arXiv:1503.02364.

Sharma, S., Asri, L.E., Schulz, H., and Zumer, J., 2017. Relevance of unsupervised metrics in task-oriented dialogue for evaluating natural language generation. arXiv preprint arXiv:1706.09799.

Starc, J., Mladenić, D., 2017. Constructing a natural language inference dataset using generative neural networks. Comput. Speech Lang. 46, 94–112.

Sordoni, A., Galley, M., Auli, M., Brockett, C., Ji, Y., Mitchell, M., Nie, J.-Y., Gao, J., Dolan, B., 2015. A neural network approach to context sensitive generation of conversational responses. In: Proceedings of NAACL.

Stent, A., Marge, M., Singhai, M., 2005. Evaluating evaluation methods for generation in the presence of variation. In: Proceedings of International Conference on Intelligent Text Processing and Computational Linguistics. Berlin, Heidelberg. Springer, pp. 341–351.

Stolcke, A., Sep.2002. SRILM – an extensible language modeling toolkit. In: Proceedings of the 7th International Conference on Spoken Language Processing, Denver, CO, USA.

Vedantam, R., Lawrence Zitnick, C., Parikh, D., 2015. Cider: consensus-based image description evaluation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4566–4575.

Vinyals, O. and Le, Q. 2015. A neural conversational model. arXiv preprint arXiv:1506.05869.

Walker, M., Kamm, C., Litman, D., 2000. Towards developing general models of usability with PARADISE. Nat. Lang. Eng. 6 (3–4), 363–377.

Wen, T.H., Gasic, M., Mrksic, N., Su, P.H., Vandyke, D. and Young, S. 2015. Semantically conditioned lstm-based natural language generation for spoken dialogue systems. arXiv preprint arXiv:1508.01745.

White, J.S., O'Cornell, T., O'Nava, F., 1994. The ARPA MT evaluation methodologies: evolution, lessons and future approaches. In: Proceedings of Association for Machine Translation in the Americas, pp. 193–205.

Wieting, J., Bansal, M., Gimpel, K. and Livescu, K. 2015. Towards universal paraphrastic sentence embeddings. arXiv preprint arXiv:1511.08198.

Williams, J.D. and Zweig, G. 2016. End-to-end lstm-based dialog control optimized with supervised and reinforcement learning. arXiv preprint arXiv:1606.01269.

Zhu, Y., Kiros, R., Zemel, R., Salakhutdinov, R., Urtasun, R., Torralba, A., Fidler, S., 2015. Aligning books and movies: towards story-like visual explanations by watching movies and reading books. In: Proceedings of the IEEE International Conference on Computer Vision, ICCV.