POSSCORE: A Simple Yet Effective Evaluation of Conversational Search with Part of Speech Labelling

Zeyang Liu
zeyang.liu@nottingham.ac.uk
University of Nottingham
Nottingham, UK

Ke Zhou
ke.zhou@nottingham.ac.uk
University of Nottingham & Nokia Bell Labs
Nottingham, UK

Jiaxin Mao
maojiaxin@ruc.edu.cn
Renmin University of China
Beijing, China

Max L. Wilson
max.wilson@nottingham.ac.uk
University of Nottingham
Nottingham, UK

ABSTRACT

Conversational search systems, such as Google Assistant and Microsoft Cortana, provide a new search paradigm where users are allowed, via natural language dialogues, to communicate with search systems. Evaluating such systems is very challenging since search results are presented in the format of natural language sentences. Given the unlimited number of possible responses, collecting relevance assessments for all the possible responses is infeasible. In this paper, we propose POSSCORE [1], a simple yet effective automatic evaluation method for conversational search. The proposed embedding-based metric takes the influence of part of speech (POS) of the terms in the response into account. To the best knowledge, our work is the first to systematically demonstrate the importance of incorporating *syntactic* information, such as POS labels, for conversational search evaluation. Experimental results demonstrate that our metrics can correlate with human preference, achieving significant improvements over state-of-the-art baseline metrics.

CCS CONCEPTS

• **Information systems → Retrieval effectiveness**; • **Computing methodologies → Discourse, dialogue and pragmatics**.

KEYWORDS

Evaluation, Conversational search, Dialogue, Metric, Part of Speech

ACM Reference Format:
Zeyang Liu, Ke Zhou, Jiaxin Mao, and Max L. Wilson. 2021. POSSCORE: A Simple Yet Effective Evaluation of Conversational Search with Part of Speech Labelling. In *Proceedings of the 30th ACM International Conference on Information and Knowledge Management (CIKM '21), November 1–5, 2021, Virtual Event, QLD, Australia.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3459637.3482463

1 INTRODUCTION

Various existing voice assistants and automatic desk helpers, such as Apple Siri, Google Assistant, and Microsoft Cortana, can be seen as typical platforms for conversational search. The major difference between conversational search and traditional search is that conversational search systems expect users to describe their search tasks using natural language and typically presents the search results through conversations. However, it should be noted that the core of conversational search is still an information retrieval (IR) system. In other words, the goal of conversational search is to seek information and satisfy users' information needs [31, 48].

Evaluation plays a pivotal role in designing and tuning search systems [5]. However, due to the nature of the conversational search, it is difficult to apply traditional evaluation methods, based on relevance assessments, to this new search paradigm. When users directly interact with conversational search systems in natural language, the numbers of possible user utterances and system responses are infinite, which makes it hard to collect a set of reusable assessments for the evaluation of conversational search. To face such challenges, previous work has proposed a number of automatic evaluation metrics to quantify the semantic similarity of utterances to references (ground-truth responses) and leverage this as the proxy for relevance. Examples of those metrics include word-overlap based measures (e.g., BLEU [29], METEOR [2]) , word-embedding based metrics (e.g., Embedding Average [43], Soft Cosine Similarity [44] and BERTScore [53]), and learning-based metrics (e.g., BERT-RUBER [10]). However, most of the above metrics use the entire sentence as the input and treat all the words of the responses equally in the evaluation process, which inevitably brings much noise in estimating relevance. A few prior meta-evaluation studies [22, 23, 28] have revealed the weaknesses of existing automatic metrics. Further, empirical studies have demonstrated that all of these metrics correlate weakly with human preference.

The *syntactic structures* of the utterances might capture additional information for evaluation. The part of speech (POS) defines how a word is used in a sentence and what role the word plays within the grammatical structure of phrases. Usually, words with the same POS tags contain related grammatical functions and display similar semantic behaviour. Novikova et al. [28] have demonstrated that grammar-based metrics correlate better with quality for evaluation in the area of natural language generation. POS words

[1] The codes are available at https://github.com/zy-liu/POSSCORE

CIKM '21, November 1–5, 2021, Virtual Event, QLD, Australia

ACM ISBN 978-1-4503-8446-9/21/11…$15.00
https://doi.org/10.1145/3459637.3482463

Table 1: An example from the PersonaChat dataset [52]. The selected POS tags of our proposed POSSCORE are <'ADJ', 'ADV', 'VERB', 'PROPN', 'NOUN'>. It can be observed that our POSSCORE metric aligns better with human perceived quality score than BLEU4 and METEOR. Higher metric score means the corresponding metric is more likely to prefer this candidate.

Context	... A: My dad had me hooked on high protein foods like fish and chicken when I was training. B: What are you training for?				
Reference	I am competing (VERB) for a national (ADJ) chess (NOUN) tournament (NOUN). It helps (VERB) me keep (VERB) focus (NOUN).				
Good candidate	Chess (NOUN) tournaments(NOUN) protein (NOUN) makes (VERB) your brain (NOUN) more (ADV) activate (VERB) for intense (ADJ) chess (NOUN) matches (NOUN).				
Bad candidate	I am a professional (ADJ) chess (NOUN) player (NOUN).				
	BLEU4	METEOR	BERT-Score	POSSCORE	Human score (5-point scale: 0 - poor quality, 5 - excellent quality)
Good candidate	0.019	0.155	0.859	**1.942**	**5**
Bad candidate	**0.032**	**0.257**	**0.892**	1.476	2

and labels, which contain grammatical information, might be helpful for utterance evaluation in conversational search. Therefore, in this paper, we comprehensively analyze the effect of POS words and labels in the evaluation process. To our best knowledge, this is the first work to systematically demonstrate the importance of the part of speech labels for conversational search evaluation. An evaluation example can be found in Table 1, which demonstrates the effectiveness of our proposed metric. Compared to BLEU metrics that utilize only semantic matching, our proposed metric exploits the syntactic POS distributions. Given that the POS distributions of the good candidate result align better with that of the reference (ground truth) than a bad candidate, our proposed metric POSSCORE effectively captures such syntactic matching and correlates better with the human perceived quality score.

Our main contributions are two-fold: (1) We are the first to systematically reveal the connection between POS labels and relevance in conversational search evaluation. We empirically demonstrate that conversational search evaluation should also consider syntactic information, such as POS, rather than only the words in responses. (2) We propose a simple yet effective POS-based metric: POSSCORE. Experimental results show our metrics can correlate strongly with human preference, outperforming state-of-the-art baselines.

2 RELATED WORK

Evaluation of Conversational Search With the lack of a uniform structure, as created by traditional search systems, it is challenging to find suitable features to capture the quality of responses in conversational search evaluation [21]. To address this problem, the general idea of prior evaluation studies is to evaluate the appropriateness of responses by comparing the candidate system responses with ideal ones, which are usually generated by humans. With the similarity between ideal responses (i.e., ground truth or reference response) and candidate responses, automatic metrics can estimate a quality score of candidate responses. As far as we know, there are generally three categories of metrics in conversational search evaluation: word overlap-based metrics, word embedding-based metrics, and learning-based metrics.

Word overlap-based metrics. The basic idea of these metrics is to compute the number of overlapping words between references and candidate responses. Since this type of metric usually has simple algorithms and interpretable structures, word overlap-based metrics have become a popular choice for conversational search and open-domain dialogue evaluation. Typical metrics such as BLEU [29] and METEOR [2] have been widely used to evaluate the adequacy of a response, especially in dialogue-related competitions[13]. However, previous studies [22, 28] indicate that these overlap-based metrics weakly correlate with human judgements.

Word embedding-based metrics. The shortfall of word overlap-based metrics is obvious: the exact matching methods are not able to capture the potential connection between words that are similar topically. Therefore, embedding-based metrics are proposed to address this issue. Popular metrics such as Greedy Matching [34], Vector Extrema[9] and BERTScore [53] are also widely applied in dialogue evaluation [17, 26].

Learning-based metrics. The basic idea of these metrics is to train a supervised model to learn the underlying criteria of human judgements. The features adopted in training models can be the semantic features of ground truth or the context. For example, Lowe et al. [24] propose ADEM, which is a recurrent neural network model, to fit the ratings of human judgements. Tao et al. [47] proposed a mixed evaluation method combining referenced and unreferenced metrics. Especially, their unreferenced part is a supervised model, which aims to estimate the appropriateness of response with respect to the context. Although this kind of metric can achieve good performance in some specific scenarios, their training process is inevitably influenced by the given training datasets. In other words, the evaluation score may be different, even when we test the same dataset, if we use different training settings. Further, it is difficult to interpret the results of learning-based metrics.

Besides the above "offline" evaluation methods, online methods, such as satisfaction prediction, are also popular in conversational search evaluation. Different from offline methods, the online evaluation focuses on users' behaviour and feedback when interacting with systems in real-time. Many prior studies have presented methods for satisfaction prediction for intelligent assistants, such as [12, 14, 15]. The basic idea of these methods is to construct a predictive model based on user interaction behaviour signals or semantic features, and estimate a score of an utterance that is close to human judgements. In this paper, we aim to propose a simple POS-based evaluation metric for conversational search. Since the corpora we used do not contain adequate user interaction information, our work only considers offline evaluation methods as baselines.

Predictive Power To examine the efficiency and fidelity of a metric, one common approach is to compare the correlation rate between human annotation scores and the metric scores [18, 22, 53]. However, the correlation score might be invalid when there are many ties (i.e., the values are the same in each pair) in the datasets. In our paper, we choose to adopt *predictive power*, which measures the agreement between user preferences and metrics when presenting a pair of different responses, to evaluate the fidelity of

Table 2: The POS tags in spaCy. 'Adopted POS tags' shows the POS tags used in our experiment and 'NOT adopted POS tags' presents the rest POS tags in spaCy.

Adopted POS tags		NOT adopted POS tags	
POS tag	Description	POS tag	Description
ADJ	adjective	AUX	auxiliary verb
ADV	adverb	CONJ	coordinating conjunction
VERB	verb	DET	determiner
NOUN	noun	INTJ	interjection
PRON	pronoun	NUM	numeral
PROPN	proper noun	PART	particle
		PUNCT	punctuation
		SCONJ	subordinating conjunction
		SYM	symbol
		ADP	adposition
		X	other

proposed metrics. Many prior studies have demonstrated its suitability for measuring the fidelity of metrics in meta-evaluation [4, 23, 37, 39, 41, 42, 54].

3 PROPOSED METRICS

This section focuses on the methodology of our proposed metrics. First, the methods and analysis for Part of Speech (POS) labelling are presented in §3.1 and §3.2. The adaptive extension based on existing metrics, such as BLEU and METEOR, are introduced in §3.3. We describe the design of our POSSCORE metric in §3.4.

3.1 Part of Speech Labelling

The part of speech (POS) explains how a word is used in a sentence and what role it plays within the grammatical structure of the sentence. Words with the same POS tags typically contain identical grammatical information and exhibit similar semantic behaviour. Therefore, using appropriate POS tags allows us to bring grammatical information into the automatic evaluation and to potentially achieve better evaluation performance. In the field of natural language processing, many previous studies have proposed efficient and effective POS tagging methods, including rule-based methods (e.g., [3]) and learning-based methods (e.g., [16, 33]). In our study, we adopt the spaCy toolkit[2], a popular industrial NLP library, to extract the POS tags from the responses. Table 2 shows the categories of POS tags in the spaCy toolkit and which tags were used in our experiment. Here, we select POS tags according to two criteria: (1) *informativeness*: POS tags should have factual information; (2) *interpretability*: selected POS tags can be directly interpreted, which ensures our designed metrics could be interpretable and extensible. Thus, we finally choose ADJ, ADV, VERB, NOUN, PRON, and PROPN as the candidate POS tags. We further refer to the words with these informative and interpretable POS tags as *POS words*.

3.2 Analysis of POS Words in the Responses

Figure 1 shows the average number of POS labels in each response, including references, for 'good' and 'bad' candidates (defined in §4.1). Although the distribution of POS tag words in those selected collections are different, some common trends can be observed across them: the number of POS tag words found in the 'good' candidates is closer to the number found in the references, in comparison to 'bad' candidates. Most of the selected POS words (e.g., 'ADJ', 'ADV', 'PROPN', 'NOUN') follow this trend in all collections.[3] In addition, the similarity between candidates and references in terms of POS tag words might be a useful signal to distinguish the quality of responses. Based on the observed trend, we summarize two basic assumptions in this study:

- **Assumption 1**: The response of good quality should contain similar POS words as the reference response;
- **Assumption 2**: The difference in the distribution of POS tags between the candidate and reference responses is useful for measuring the relevance of the candidate response.

Table 3: An example of POS words and POS tags when the tag set is <NOUN, VERB>.

Orignal	it is from our evolution when land animals had both gills and lungs
POS words	<is, evolution, land, animals, had, gills, lungs>
POS tags	<VERB, NOUN, NOUN, NOUN, VERB, NOUN, NOUN>
POS words + POS tags	<is, evolution, land, animals, had, gills, lungs, VERB, NOUN, NOUN, NOUN, VERB, NOUN, NOUN>

3.3 POS-aware Adapted Metrics

To systematically analyze the effect of POS labelling, we firstly use simple methods to extend existing metrics with POS tags. We adopt two adaptive approaches for incorporating POS tag information:

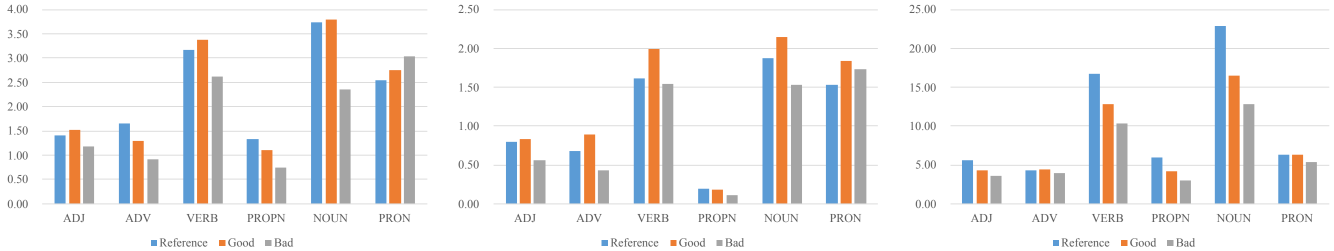
POS Word Extraction (PWE) - This is an intuitive way that only considers the words with specific POS tags. We extract all the words within POS tag sets (i.e., the specific tags which we choose) and filter other words (Table 3). We then combine these POS words into a new sentence, which is used as an input to existing metrics.

POS Tag Linear Combination (PTLC) - We further consider the overlap of POS tag distribution on the basis of PWE. Firstly, we extract all the POS words and combine them into a new sentence like PWE. Then all the corresponding POS tags of these words are also extracted (shown in Table 3). We then use different strategies to combine POS tag words and POS tags: 1) For hard matching metrics (i.e., BLEU), we directly combine POS tag words and POS tags together and put them into one new sentence as the metric input; 2) For synonym-based metrics (e.g., METEOR) and word embedding-based metrics (e.g., EA), we separately calculate the similarity of POS words and POS tags since POS tags do not have synonyms or embedding vectors. The original embedding-based metrics are used for calculating the similarity score of the extracted POS word text. We denote this score as the *POS text score*. The overlap score of POS tags is computed by word overlapping-based metrics. In our paper, we use BLEU1 to compute the overlapping rate of the POS tag sequence. The score of these POS tags is denoted as *POS tag score*. Finally, the POS text score and the POS tag score are linearly combined (added).

Since POS tags generally indicate the grammatical role of a word in a sentence, different types of POS words may contain different information. Therefore, selected POS tags play a vital role in our proposed methods. To comprehensively analyze the effect of POS selection, we list all POS tag combinations that are adopted in the experiment in Table 4. Here a POS tag set means we only select the words with the POS tags in the tag set. For example, 'VERB' means we only extract the 'VERB' tag words from the original sentence. It

[2] https://github.com/explosion/spaCy

[3] Note that 'VERB' and 'PRON' in PersonaChat are two exceptions, due to the unique characteristics of this collection.



(a) Topic -Chat (b)PersonaChat (c) MSDialog

Figure 1: The number distribution of selected POS tags in references, good responses, and bad responses. The x-axis denotes the select POS tags, and the y-axis denotes the average number of the POS tag in the corresponding dataset.

Table 4: The POS tag combinations in the experiment. POS tag set means only the words with the corresponding POS tags (in the POS set) are selected for the evaluation.

POS tag set	
ADJ	ADJ + VERB + PROPN + NOUN
ADV	ADJ + PROPN + NOUN + PRON
VERB	ADV + VERB + PROPN + NOUN
PRON	ADV + ADJ + PROPN + NOUN
PROPN + NOUN	ADV + PROPN + NOUN + PRON
ADV + VERB	VERB + PROPN + NOUN + PRON
VERB + PROPN + NOUN	ADJ + ADV + VERB + PROPN + NOUN
PROPN + NOUN + PRON	ADJ + ADV + VERB + PROPN + NOUN + PRON
ADJ + PROPN + NOUN	

is worth noting that we consider ‘PROPN’ and ‘NOUN’ together since both tags are nominal attributes.

3.4 Proposed POSSCORE Metric

In this paper, our goal is to propose simple yet effective POS-based metrics for conversational search. The basic idea of our metrics is to increase the importance of selected POS words and give more weight to the POS similarity scores if the POS distribution is similar to the reference.

Given a reference response $r =< r_1, r_2, r_3, ..., r_j >$ and a candidate response $\hat{r} =< \hat{r}_1, \hat{r}_2, \hat{r}_3, ..., \hat{r}_k >$, we use word embeddings to present the tokens. Then each response is split into two group: POS words sequence and Non-POS words sequence. The POS word sequence only contains the words with the selected POS tags, while the Non-POS sequence is the remaining words of the response. Thus, a reference is split as a POS word sequence $r_p =< r_{p1}, r_{p2}, ..., r_{pm} >$ and a Non-POS word sequence $r_q =< r_{q1}, r_{q2}, ..., r_{qm'} >$, and a candidate response is grouped as a POS word sequence $\hat{r}_p =< \hat{r}_{p1}, \hat{r}_{p2}, ..., \hat{r}_{pn} >$ and a Non-POS word sequence $\hat{r}_q =< \hat{r}_{q1}, \hat{r}_{q2}, ..., \hat{r}_{qn'} >$. The POSSCORE could be calculated as below shown in Equation 1:

$$POSSCORE(r, \hat{r}) = w * S(r_p, \hat{r}_p) + S(r_q, \hat{r}_q) \tag{1}$$

where w means the weight function of POS tag rewarding, and $S(x, \hat{x})$ is the cosine similarity of the average embedding between sentence x and $\hat{x}$ (defined in Equation 2).

$$S(x, \hat{x}) = cosine(\bar{E}(x), \bar{E}(\hat{x})) \tag{2}$$

POSSCORE should consider both the quality of POS word content and the difference in the distribution of POS tags between the references and candidate responses. Since the similarity score $S(r_p, \hat{r}_p)$ can capture the content-level similarity of POS words, the design of the weight function w should consider the distribution difference. Therefore, the weight function needs to meet these requirements: 1) If the number of POS words in a candidate is less than that in the reference, the gain from POS words could be small and w could reduce the importance of the POS similarity part; 2) If the number of POS distribution is the same, w could be 1 and keep the original POS similarity scores; 3) If the number of POS words is larger than that in references, POS word part could be important and w could increase the gain of POS similarity part. Therefore, following these criteria, the weight function w is defined as Equation 3. This weight function is inspired by the penalty function of BLEU metrics [29]. The range of w is $0 < w < e$.

$$w = \exp(1 - \frac{n_r}{n_{\hat{r}}}) \tag{3}$$

where n_r is the percentage of POS words in references (i.e., the number of selected POS words in reference r divided by the length of reference r), and $n_{\hat{r}}$ is the percentage of POS words in candidate responses (i.e., the number of selected POS words in candidate $\hat{r}$ divided by the length of candidate $\hat{r}$). This weight function entails three different scenarios:

- if $n_r > n_{\hat{r}}$, then w < 1. This means candidate responses do not have enough POS tag words by comparing to the references. In other words, the candidates may lack the necessary information expressed by the POS words in the reference response. Therefore, the gain from POS similarity scores should decrease.
- if $n_r = n_{\hat{r}}$, then w = 1. This means the POS distribution of candidate responses might be similar to the references. There is no extra gain for the POS word part.
- if $n_r < n_{\hat{r}}$, then w > 1. This means the number of the POS words in the candidate response is more than that in the reference and more likely to cover the POS information of references. In other words, the gain from POS word parts is more important than the rest of a candidate response. Therefore, w increases the importance of the POS part and give more weight to the POS similarity scores.

Thus, we simply use the difference of POS word percentage between references and candidates to capture the distribution of POS information and dynamically control the importance of the POS similarity part. In order to find the influence of selected POS tags on the performance of POSSCORE, we also test different POS tag combinations as shown in Table 4.

4 EXPERIMENTAL SETTINGS

This section describes the key elements of our experimental setup, including dataset selection and pre-processing (§4.1), baselines (§4.2), and metric evaluation methods (§4.3).

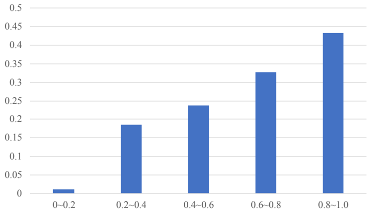


Figure 2: The distribution of ground-truth answers in the voted responses. The x-axis denotes the normalized vote number, and the y-axis represents the proportion of the ground truth answers.

4.1 Datasets

In view of the complex interactions involved in conversational search, the datasets used for metric evaluation should meet three criteria: 1) Each interactive dialogue should have specific search intents since the existence of search intents is one of the important features in conversational search [1]; 2) To refrain from collecting human annotations ourselves and reduce annotation bias, it is better that the open datasets contain human annotations that enable us to align metrics to the gold standard; 3) The dialogues in the datasets consist of multiple-round interactions so that the test environment can be closer to real conversational search scenario. After a comprehensive survey of existing datasets [6, 7, 11, 20, 25, 30, 32, 49, 50, 52], we chose three datasets that met the criteria: Topic-Chat [11], PersonaChat [52], and MSDialog [30]. **Topic-Chat (TC)** [11] is a large collection of knowledge-grounded human-to-human conversations that consists of 11,319 dialogues with 8 broad topics. Each conversation has a specific topic and each utterance in the conversation is rated on a 5-point scale of quality. **PersonaChat (PC)** [52] is a dataset of human-to-human persona-conditioned open-domain conversations that contain 10,907 dialogues with personal topics. Each partner is asked to act as a persona to converse with each other. **MSDialog** [30] is a large-scale dialogue corpus of question answering interactions between customers and a help desk from an online forum on Microsoft products. This dataset consists of more than 2,000 multi-round information-seeking conversations with 10,000 utterances.

Preprocessing of datasets. Given that we adopt predictive power (described in §4.3) to examine our metric, each conversation in above corpora needs to be converted to <question, reference, response1, response2> evaluation sets, which means one question can have one reference response (i.e., ground truth) and two candidate responses with different qualities, namely, a ‘good’ candidate response and a ‘bad’ candidate response. Note that these two candidate responses answer the same question. Given the dataset differences, we adopt different strategies to preprocess them.

In the original TC and PC corpora, a question only has one response, which is inadequate for our metric evaluation method. Therefore, we refer to the recent work by Mehri and Eskenazi [26] and use their publicly released collections[4], which are built on TC and PC corpora. Mehri and Eskenazi [26] conduct a human quality annotation of human-to-machine and human-to-human responses for both TC and PC to study the efficiency of their USR metric. In their collections, human annotation is carried out on sixty dialogue contexts. For each context, there are three or four system outputs (obtained from different generative models), one newly-written human response, and one ground-truth response. All the responses are labelled by three annotators with scores from six dimensions: Understandable (0-1), Natural (1-3), Maintains Context (1-3), Interesting (1-3), Uses Knowledge (0-1), and Overall Quality (1-5). In our study, we only consider the ‘Overall Quality’ scores because this score intuitively reflects human preference. The average value of ‘Overall Quality’ scores from three annotators is regarded as the final score for each response. Thus, all the system outputs and human response could be grouped into different <question, reference, response1, response2> sets in line with the final score for each context. If a response has a higher score than the other one, this response is deemed to be the ‘good’ response in this set. It is worth noting that we do not put the responses with the same final scores into one set, which means candidates in one set should have different annotation scores. Finally, we obtain 550 sets for TC and 328 sets for PC (shown in Table 5).

Table 5: The number of evaluation sets in each collection.

Corpus Name	Evaluation Sets
Topic-Chat(TC)	550
PersonaChat(PC)	328
MSDialog	3,000

Unlike TC and PC with explicit human annotation scores, MSDialog collects users’ judgements by using a variety of human feedback labels [5]. In this paper, we use two types of human labels, namely ‘vote’ and ‘is_answer’ to obtain evaluation sets. The tag ‘*vote*’ represents the number of ‘helpful’ votes for the answer from the community. If users agree with the response and think this answer may be helpful for this question, users can give one vote to this response. Note that users are not allowed to vote the same response more than once. The tag ‘*is_answer*’ is a binary tag, which indicates whether this answer is selected as the best answer in the dialogue session. Especially, this tag is often annotated by the user who posted the initial question and started the dialogue. Given that the ‘is_answer’ annotation can indicate that the response solves the issue from the questioners’ perspective, we deem the *is_answer* responses to be the reference responses (ground truth).

To shed light on the connection between the ‘*vote*’ and ‘*is_answer*’ tag, we further calculate the proportion of the ground truth against the voted responses. All the ‘*vote*’ values are normalized by the maximum ‘*vote*’ of the same dialogue in order to reduce the trendy questions’ bias. Fig 2 shows the distribution of ground-truth answers in the voted responses. It is observed that the proportion of ground-truth answers grows steadily with the increase of the normalized vote scores, which means a response is more likely to be annotated as a relevant answer if this response has more votes in the same conversation session. It is worth noting that the proportion is 43.26% when the normalized value is 1 (i.e., the response has the most votes in the dialogue). Therefore, we use the tag ‘*vote*’ as an indicator to simulate the different human judgements. If a response has a higher ‘*vote*’ value, it is regarded as the ‘good’ response in a set. Finally, we randomly select ~3,000 evaluation sets to test our metric.

[4] The collections (i.e., TC and PC) are available at http://shikib.com/usr

[5] MSDialog collection is available at https://ciir.cs.umass.edu/downloads/msdialog/

Table 6: The selected automatic evaluation metrics.

Metric Category	Metric
Word overlap-based	BLEU1-4, METEOR
Word embedding-based	Emebedding Average, BERT-Score
Learning-based	BERT-RUBER

4.2 Baseline Metrics

Since POS is a token-level label, using word-level baseline metrics may allow us to easily extend these metrics with POS labelling (shown in §3.3) and further find out the effect of part of speech in the evaluation process. Therefore, in this paper, three types of metrics were chosen (shown in Table 6):

Word overlap-based metrics. The basic idea of this type of metric is to count the number of words that co-occur in both the candidate responses and the ground truth. Here we choose BLEU [29] and METEOR [2][6] as the baselines, which are popular metrics used in the evaluation of open-domain dialogue.

Word embedding-based metrics. The basic idea of these metrics is to using embedding information to connect between words that are semantically similar and evaluate the similarity between candidate sentences based on these word embedding vectors. Embedding Average (EA) [8, 19, 27, 43], Embedding Extrema [9] and Embedding Greedy [34] are common representatives of this type of metrics. Since the performance of these three metrics are very close [22], we choose EA[7] as the baseline of embedding-base metrics. Additionally, BERT-Score [53] is one latest proposed metrics, which has been demonstrated to have a high correlation with human judgments. Therefore, BERT-Score[8] which is also selected for a baseline.

Learning-based metrics. This type of metric always consists of one or more training models, such as ADEM [24], RUBER [47], PONE [18], and BERT-RUBER [10]. Following previous work [10, 23], we select BERT-RUBER as the sole representative learning-based metric in this study given its superior performance. Since the performance of learning-based models could be influenced by the pre-prepared training dataset [23], we train and tune the model based on the specific dataset we use.

4.3 Evaluating Automatic Evaluation Metrics

Metric fidelity is the key concern in our metric design. Fidelity reflects the ability of a metric to measure what it intends to measure and agree with ultimate user preferences. Many recent studies [26, 45, 51] use *spearman's correlation coefficients* and *Pearson's correlation coefficients* to test the correlation between metric judgements and human annotations. However, given that there are many ties among human annotation scores, the correlation score could be very close and it may be difficult to distinguish the performance of candidate metrics. Therefore, in our experiment, we adopt predictive power [42] to capture the extent of a given evaluation metric's ability to predict a user's preference. Predictive power measures the ability of metrics to describe the agreement between metrics and user preferences [42]. The basic idea of predictive power is that if an evaluation metric agrees with the user's preference between two outputs, then that is a correct prediction [38]. The higher the predictive power score is, the more similar to the human judgements the metric is. We use predictive power to examine the similarity between metrics and human judgements in conversational search (as shown in Algorithm 1).

Algorithm 1: Computing predictive power.

```
Total=0;Correct=0;
for $d = 1 to N$ // for each dialogue
do
    foreach pair of responses($r_1$, $r_2$) in dialogue $d$ do
        Total++;
        $\delta X = X(r_1) - X(r_2)$;
        // $X$ is the metric score of response $r_j$
        $\delta X^* = X^*(r_1) - X^*(r_2)$;
        // $X^*$ is the judgement score of response $r_j$, such as votes
        if $((\delta X \times \delta X^*) > 0)$ // $X$ and $X^*$ positively agree
        then
            Correct++;
        end
    end
end
PredictivePower = Correct/Total;
```

Here, we summarize the overall process of metric evaluation as follows: 1) We first collect all the <question, reference, response1, response2> sets from the three datasets. 2) Candidate metrics are adopted to calculate the scores for both response1 and response2. 3) After that, we use predictive power to examine the coherence between metric judgements and human preference. 4) Finally, different candidate metrics are compared on their predictive power.

5 EXPERIMENTAL RESULTS

This section presents the experimental results for both POS-aware Adapted measures (§5.1) and POSSCORE metrics (§5.2). We systematically compare the performance of different POS combinations. We also discuss the effect of response length on POSSCORE (§5.3), the correlation between metrics (§5.4), and a case study (§5.5).

5.1 POS-aware Adapted Metrics

Table 7 shows the predictive power scores of PWE methods. The baseline is the results that are calculated with the original sentences. We use a two-sided T-test [46] to examine the difference between proposed methods and baselines.[9] The scores in the table indicate the agreements between metrics and user preferences. For example, the baseline BLEU1 metric has 60.4% (0.604) 'correct' prediction, which agrees with users' preference (i.e., selecting 'good' responses) within all the sessions in the TC collection. We can observe that:

- All PWE methods fail to defeat the original metrics based on word overlap (BLEU1-4 and METEOR), which means that only using the overlapping of POS words between candidates and references is inadequate for predicting users' preference. This is not surprising since the word reduction makes it more difficult for exact word matching.
- Although PWE with EA can perform better in PC, all the improvements are not significantly different to the baselines. This may be because of the small size of the PC collection. Meanwhile,

[6] Here we use the default settings of NLTK module, where $\alpha = 0.9$, $\beta = 3$, and $\gamma = 0.5$. The module link is https://www.nltk.org/_modules/nltk/translate/meteor_score.html

[7] We choose Fasttext embedding as the default embedding for EA and POSSCORE.

[8] Here we use the default settings of BERT-Score. The details of the settings are shown in https://github.com/Tiiiger/bert_score

[9] Since we find the trends of the significant test with Bonferroni correction are similar to the original T-test, we present the original T-test results and set 0.05 and 0.01 as the significant thresholds in this paper.

Table 7: The predictive power results of PWE methods. Baselines are calculated on the original sentences. The two-sided t-test is performed to detect any significant difference between proposed methods and baselines. * and ** represent significant value $p < 0.05$ and $p < 0.01$. The block colour represents the power scores' change of direction compared with the baseline in the same column (red denotes increment and blue represents decrements; the brightness of the colour indicates the change magnitude).

POS tags	BLEU1			BLEU2			BLEU3			BLEU4			METEOR			EA		
	TC	PC	MSDialog	TC	PC	MSDialog	TC	PC	MSDialog	TC	PC	MSDialog	TC	PC	MSDialog	TC	PC	MSDialog
Baseline	0.604	0.500	0.554	0.616	0.488	0.552	0.620	0.466	0.546	0.607	0.439	0.545	0.651	0.543	0.547	0.642	0.662	0.556
PROPN + NOUN	0.458 **	0.252 **	0.461 **	0.452 **	0.248 **	0.457 **	0.445 **	0.248 **	0.453 **	0.426 **	0.252 **	0.453 **	0.463 **	0.282 **	0.489 **	0.636	0.636	0.58 **
ADJ	0.221 **	0.212 **	0.198 **	0.225 **	0.212 **	0.198 **	0.225 **	0.212 **	0.195 **	0.225 **	0.212 **	0.197 **	0.238 **	0.212 **	0.228 **	0.575	0.481	0.53
ADV	0.156 **	0.233 **	0.219 **	0.168 **	0.233 *	0.218 **	0.168 **	0.233	0.216 **	0.168 **	0.233	0.215 **	0.18 **	0.233 **	0.255 **	0.533 **	0.667	0.484 **
VERB	0.321 **	0.126 **	0.36 **	0.328 **	0.126 **	0.358 **	0.328 **	0.126 **	0.358 **	0.328 **	0.126 **	0.357 **	0.353 **	0.158 **	0.459 **	0.575 *	0.468 **	0.551
PRON	0.392 **	0.257 **	0.469 **	0.396 **	0.242 **	0.468 **	0.4 **	0.249 **	0.473 **	0.392 **	0.249 **	0.469 **	0.411 **	0.26 **	0.468 **	0.549 *	0.516	0.468 **
PROPN + NOUN + PRON	0.527 **	0.396 **	0.536 *	0.549 **	0.363 **	0.538	0.553 **	0.338 **	0.534	0.54 **	0.341 **	0.532	0.56 **	0.369 **	0.529 *	0.617	0.673	0.548
ADJ + PROPN + NOUN	0.48 **	0.284 **	0.473 **	0.47 **	0.28 **	0.468 **	0.466 **	0.28 **	0.47 **	0.456 **	0.284 **	0.468 **	0.47 **	0.301 **	0.505 **	0.638	0.635	0.583 **
ADV + VERB	0.35 **	0.129 **	0.389 **	0.354 **	0.134 **	0.387 **	0.354 **	0.134 **	0.387 **	0.356 **	0.134 **	0.387 **	0.374 **	0.17 **	0.452 **	0.587 *	0.573 *	0.543
VERB + PROPN + NOUN	0.51 **	0.262 **	0.501 **	0.501 **	0.266 **	0.488 **	0.492 **	0.266 **	0.488 **	0.495 **	0.262 **	0.487 **	0.523 **	0.318 **	0.524 **	0.628	0.651	0.574 **
ADJ + VERB + PROPN + NOUN	0.528 **	0.292 **	0.507 **	0.522 **	0.289 **	0.494 **	0.528 **	0.286 **	0.49 **	0.522 **	0.286 **	0.489 **	0.535 **	0.335 **	0.526 **	0.64	0.68	0.578 **
ADJ + PROPN + NOUN + PRON	0.567 *	0.421 **	0.537 *	0.556 **	0.412 **	0.541	0.545 **	0.399 *	0.537	0.538 **	0.378 *	0.538	0.556 **	0.415 **	0.54	0.626	0.689	0.552
ADV + VERB + PROPN + NOUN	0.533 **	0.283 **	0.502 **	0.52 **	0.28 **	0.49 **	0.52 **	0.283 **	0.491 **	0.522 **	0.283 **	0.494 **	0.55 **	0.344 **	0.521 **	0.627	0.685	0.577 **
ADV + ADJ + PROPN + NOUN	0.496 **	0.28 **	0.48 **	0.496 **	0.28 **	0.477 **	0.494 **	0.28 **	0.48 **	0.483 **	0.283 **	0.481 **	0.494 **	0.332 **	0.501 **	0.625	0.674	0.575 **
ADV + PROPN + NOUN + PRON	0.551 **	0.378 **	0.534 *	0.555 **	0.357 **	0.542	0.549 **	0.332 **	0.538	0.545 **	0.314 **	0.54	0.567 **	0.384 **	0.518 **	0.623	0.705	0.545
VERB + PROPN + NOUN + PRON	0.571 *	0.436 *	0.543	0.555 **	0.424 *	0.535 *	0.555 **	0.409 *	0.54	0.556 **	0.381 *	0.535	0.578 **	0.445 **	0.541	0.649	0.641	0.557
ADJ + ADV + VERB + PROPN + NOUN	0.542 **	0.305 **	0.508 **	0.537 **	0.299 **	0.495 **	0.535 **	0.299 **	0.497 **	0.537 **	0.296 **	0.495 **	0.554 **	0.378 **	0.524 **	0.631	0.692	0.578 **
ADJ + ADV + VERB + PROPN + NOUN + PRON	0.595	0.412 **	0.539 *	0.591	0.405 **	0.536 *	0.593	0.375 **	0.539	0.591	0.357 **	0.54	0.598 **	0.482 **	0.54	0.644	0.701	0.549

Table 8: The predictive power results of PTLC methods. Baselines are calculated on the original sentences without POS tags. The same annotation strategy of Table 7 is utilized for this table.

POS words + POS tags	BLEU1			BLEU2			BLEU3			BLEU4			METEOR			EA		
	TC	PC	MSDialog	TC	PC	MSDialog	TC	PC	MSDialog	TC	PC	MSDialog	TC	PC	MSDialog	TC	PC	MSDialog
Baseline	0.604	0.500	0.554	0.616	0.488	0.552	0.620	0.466	0.546	0.607	0.439	0.545	0.651	0.543	0.547	0.642	0.662	0.556
PROPN + NOUN	0.563	0.462	0.532	0.557 *	0.441	0.536	0.572	0.441	0.537	0.563	0.437	0.533	0.58 *	0.487 *	0.537	0.611	0.555 **	0.55
ADJ	0.396 **	0.327 **	0.467 **	0.412 **	0.308 **	0.462 **	0.4 **	0.308 *	0.466 **	0.404 **	0.308 *	0.467 **	0.404 **	0.327 *	0.47 **	0.517 *	0.442	0.515
ADV	0.311 **	0.4	0.478 **	0.335 **	0.4	0.483 **	0.335 **	0.4	0.486 **	0.335 **	0.4	0.485 **	0.317 **	0.433	0.477 **	0.43 **	0.567	0.515 *
VERB	0.474 **	0.353 **	0.542	0.483 **	0.342 **	0.547	0.485 **	0.342 **	0.55	0.479 **	0.332 *	0.551	0.47 **	0.353 **	0.549	0.534 **	0.441 **	0.563
PRON	0.463 **	0.272 **	0.497 **	0.472 **	0.298 **	0.502 **	0.48 **	0.294 **	0.506 *	0.484 **	0.283 **	0.503 **	0.453 **	0.283 **	0.507 *	0.573	0.452	0.526
PROPN + NOUN + PRON	0.555 *	0.47	0.548	0.547 **	0.497	0.546	0.558 **	0.485	0.547	0.547 **	0.473	0.55	0.573 **	0.5	0.55	0.584 *	0.53 **	0.544
ADJ + PROPN + NOUN	0.58	0.518	0.549	0.586	0.489	0.556	0.602	0.486	0.557	0.594	0.475	0.558 *	0.584 **	0.528	0.549	0.617	0.606	0.562
ADV + VERB	0.495 **	0.478	0.54	0.523 **	0.442	0.544	0.54 **	0.46	0.549	0.526 **	0.42	0.553	0.479 **	0.491	0.539	0.502 **	0.509 **	0.558
VERB + PROPN + NOUN	0.599	0.561	0.557	0.575	0.57 *	0.556	0.577	0.554 *	0.554	0.586	0.534 *	0.556	0.625	0.567	0.556	0.628	0.581 *	0.551
ADJ + VERB + PROPN + NOUN	0.625	0.578 *	0.551	0.618	0.594 **	0.561	0.62	0.563 **	0.564 *	0.631	0.542 **	0.564 *	0.635	0.603	0.553	0.642	0.618	0.557
ADJ + PROPN + NOUN + PRON	0.575	0.485	0.56	0.591	0.503	0.557	0.573 *	0.509	0.56	0.562 *	0.512 *	0.557	0.593 **	0.515	0.559	0.602	0.587 *	0.549
ADV + VERB + PROPN + NOUN	0.6	0.564	0.558	0.581	0.567 *	0.56	0.601	0.545 *	0.554	0.609	0.503	0.556	0.64	0.557	0.556	0.618	0.592	0.557
ADV + ADJ + PROPN + NOUN	0.596	0.559	0.546	0.603	0.523	0.552	0.598	0.497	0.547	0.586	0.474	0.55	0.617	0.582	0.543	0.607	0.622	0.558
ADV + PROPN + NOUN + PRON	0.562	0.421 *	0.55	0.567 *	0.497	0.552	0.54 **	0.485	0.546	0.525 **	0.442	0.552	0.587 **	0.476	0.546	0.592 *	0.544 **	0.549
VERB + PROPN + NOUN + PRON	0.584	0.503	0.558	0.578	0.537	0.558	0.567 *	0.537 *	0.561	0.553 *	0.537 **	0.56	0.591 **	0.527	0.559	0.602	0.557 **	0.552
ADJ + ADV + VERB + PROPN + NOUN	0.618	0.579 *	0.552	0.618	0.582 **	0.562	0.618	0.534	0.562 *	0.631	0.509 *	0.572 **	0.635	0.598	0.552	0.637	0.631	0.562
ADJ + ADV + VERB + PROPN + NOUN + PRON	0.602	0.515	0.556	0.609	0.521	0.557	0.595	0.503	0.564 *	0.595	0.479	0.578 **	0.618	0.555	0.553	0.584 *	0.591 *	0.553

We notice that many cases of PWE with EA significantly outperform the baselines in MSDialog, which means that PWE methods can provide a notable improvement on original EA metrics.

- It is worth noting that PWE with EA has poor performance in the TC collection. Except for two cases ('VERB + PROPN + NOUN' and 'ADJ + ADV + VERB + PROPN + NOUN'), other cases perform worse than the original EA. In other words, PWE methods are sensitive to the variance of collections.

Table 8 presents the predictive power results of PTLC methods. The baseline follows the same settings in PWE. We can observe:

- In general, the incorporation of POS tags significantly improves the performance of word overlap-based metrics. Comparing with the results of PWE with BLEU1-4 and METEOR, it can be observed that a large number of POS tag combinations tend to outperform the baselines after incorporating the POS tag sequence. These improvements, to some extent, reveal the positive effect of POS distribution in the evaluation process of word overlap-based metrics.
- The effects of promotion within PTLC methods are different across collections. For example, there are a number of significant improvement cases in both PC and MSdialog, whereas PTLC methods do not reach significant differences in TC.
- PTLC methods are also influenced by the original framework of the metrics. We can find BLEU and METEOR achieve more significant improvements after considering the POS information.
- The selected POS tags have a considerable influence on the performance of PTLC. Particularly, it can be observed that both 'ADJ + VERB + PROPN + NOUN' and 'ADJ + ADV + VERB + PROPN + NOUN' are robust and create the most significant improvement cases across all three collections.
- Comparing with the PWE results, the predictive power of embedding-based metrics performs worse after incorporating POS tags - in most cases, they tend to perform poorer than the baselines. That means that rough incorporation of POS tags is inadequate for the improvement over the original embedding-based metrics.

To sum up, we find roughly using POS words (i.e., PWE and PTLC methods) may be inadequate for high-quality evaluation. However, the performance of PWE methods could be improved significantly when considering word embedding-based metrics. PTLC results reveal the positive effects of using POS tag distribution in the evaluation process of overlap-based metrics. However, PTLC methods have limited improvements on embedding-based metrics.

Table 9: The predictive power results of POSSCORE methods. Baselines are the original metrics’ results with the raw response sentences. The two-sided t-test is performed to detect any significant difference between proposed methods and best candidate baselines. sig_b means the significant test between POSSCORE and original metrics which achieve the best performance among all candidate baselines (e.g., METEOR in TC, and EA in PC and MSDialog), and sig_p mean the significant test between POSSCORE and the results of corresponding PWE methods. * and ** represent significant value $p < 0.05$ and $p < 0.01$. The block colour represents the power scores’ change of direction compared with the ‘best’ baseline in the same column (red denotes increment and blue represents decrement, and the brightness of the colour indicates the change magnitude).

	Metric Name or POS tag set	TC	sig_b	sig_p	PC	sig_b	sig_p	MSDialog	sig_b	sig_p
Baseline	BLEU1	0.604			0.500			0.554		
	BLEU2	0.616			0.488			0.552		
	BLEU3	0.620			0.466			0.546		
	BLUE4	0.607			0.439			0.545		
	METEOR	**0.651**			0.543			0.547		
	EA	0.642			**0.662**			**0.556**		
POSSCORE	PROPN + NOUN	0.689		**	0.61			0.571		
	ADJ	0.595	*	**	0.631			0.534	*	**
	ADV	0.585	**	**	0.655			0.516	**	
	VERB	0.585	*	**	0.588			0.567		
	PRON	0.487	**	*	0.671	*		0.481	**	
	PROPN + NOUN + PRON	0.596	*		0.628		*	0.556		
	ADJ + PROPN + NOUN	0.684		**	0.628			0.57		*
	ADV + VERB	0.58	*	**	0.628			0.528	**	
	VERB + PROPN + NOUN	0.693		**	0.61			0.587	**	
	ADJ + VERB + PROPN + NOUN	0.736	**	**	0.637			0.59	**	
	ADJ + PROPN + NOUN + PRON	0.598	*		0.646		*	0.566		
	ADV + VERB + PROPN + NOUN	0.707	*	**	0.668			0.573		
	ADV + ADJ + PROPN + NOUN	0.695		**	0.677			0.558		*
	ADV + PROPN + NOUN + PRON	0.615			0.64		**	0.54		
	VERB + PROPN + NOUN + PRON	0.635		*	0.57		*	0.581	*	*
	ADJ + ADV + VERB + PROPN + NOUN	0.74	**	**	0.689	**		0.569		
	ADJ + ADV + VERB + PROPN + NOUN + PRON	0.662		**	0.643		*	0.572		*

5.2 POSSCORE

Both PWE and PTLC methods are simple extensions of existing metrics. Although PTLC has achieved significant improvement over the original metrics, these extension methods are not adequate to capture the connection of POS distributions between references and candidates. This motivated our new development in POS-based metrics: POSSCORE. Table 9 presents the predictive power results for POSSCORE. In this experiment, we also adopt a two-sided T-test to examine significant differences. To comprehensively analyze our proposed metrics, we perform significance tests against both: (1) *The best original baseline metric (sig_b)*: for example, in TC, the original METEOR reaches the highest predictive power score and so METEOR is selected as the baseline of TC. In MSDialog, the original EA achieves the highest predictive power score and it is regarded as the baseline for MSDialog. (2) *The PWE methods with the ‘best’ original baseline metric (sig_p)*: this is because we find PWE methods can achieve better performance than the original baseline (as discussed in §5.1). By doing so, we can identify any improvement of POSSCORE over the gains from PWE methods.

With the above settings, we can observe in Table 9 that:

- Most configurations of POSSCORE can outperform the best performing baseline metrics. Especially, in TC, we can find that many POS tag sets can reach significant improvements against the original baseline. From this perspective, POSSCORE is better than PTLC methods since it can achieve significant improvements across all three collections.
- Although a corpus may have an influence on the extent of performance improvement, POSCORE can robustly outperform the original baseline across all the selected corpora. It can be observed that although the improvement in MSDialog is smaller than that in TC and PC, POSSCORE is still able to achieve higher predictive power than the baselines.
- Comparing with the results of PWE, we can see POSSCORE is significantly better than PWE across all three datasets.
- In terms of different POS combinations, it can be observed that ‘ADJ + ADV + VERB + PROPN + NOUN’ can robustly achieve high predictive scores across three collections, especially with significant improvements in both TC and PC. Therefore, we recommend to choose *‘ADJ + ADV + VERB + PROPN + NOUN’* for POSSCORE.

Table 10: The predictive power comparison with BERT-Score and BERT-RUBER. Two-sided t-test is performed to detect any significant difference between POSScore and two state-of-the-art metrics. * and ** represent $p < 0.05$ and $p < 0.01$.

	TC	PC	MSDialog
POSSCORE	**0.740**	**0.689**	0.569
BERT-Score	0.655^{**}	0.607^{*}	0.541^{*}
BERT-RUBER	0.509^{**}	0.561^{**}	$\mathbf{0.69}^{**}$

Table 10 shows the comparison between POSSCORE and other more recent state-of-the-art metrics: BERT-Score [53] and BERT-RUBER [10]. Here we use ‘ADJ + ADV + VERB + PROPN + NOUN’ to calculate POSSCORE as it generally performs the best (as shown in Table 9). Since BERT-RUBER needs to pre-train unreferenced models, we split each dataset into training datasets (80% of the whole datasets), develop datasets (10% of the whole datasets), and test datasets (10% of the whole datasets). Following previous work [10], we also use 2 layers of the bidirectional gated recurrent unit with the 128-dimensional hidden unit and apply three layers for MLP (Multilayer Perceptron Network) with 256, 512 and 128-dimensional hidden units. Learning rate decay is applied when no improvement was observed on validation data for five consecutive epochs.

It can be observed that POSSCORE consistently outperforms BERT-Score and BERT-RUBER in TC and PC. All the improvements in these two datasets are significant. It is worth noting that BERT-RUBER outperforms the other two metrics in MSDialog, while it performs the worst in TC and PC. We found that the learning-based metric BERT-RUBER is quite sensitive to the training collection. By learning BERT-RUBER on five different random samples of the MSDialog dataset, we found its predictive power ranges from 0.66 to 0.71. Considering the poor performance of BERT-RUBER in TC and PC, this demonstrated that BERT-RUBER is not robust. Comparatively speaking, the performance of POSSCORE is more stable across different collections and is much easier to interpret.

5.3 Effect of Response Length

Since POSSCORE performs counting on the number of POS and Non-POS words and tags, we would like to understand the influence of response length on its performance. To examine the effect of length bias, we created a new dataset based on each existing dataset, in which we simply modify the low-quality responses to make them repeat the content twice to increase their length. Thus, all the low

Table 11: The predictive power results of modified test collection with longer bad candidate responses where their contents are duplicated. Thus, all the bad responses have doubled their lengths compared to the original bad responses.

	TC	PC	MSDialog
BLEU2	0.718	0.652	0.535
BLEU4	0.715	0.698	0.541
METEOR	0.653	0.616	0.508
EA	0.642	0.662	0.556
POSSCORE	**0.756**	**0.71**	**0.587**

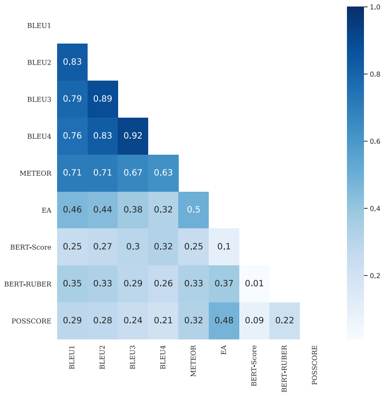


Figure 3: The Kendall's Tau correlation between different metrics on MSDialog collection (similar trends are observed in the other two collections).

quality (bad) responses are twice as long as the original responses. We compare POSSCORE against those best-performing metrics that were designed to be agnostic to the length (e.g., with length normalization), such as BLEU, METEOR and EA. The hypothesis is that the performance of those length-agnostic metrics should not be significantly affected. Table 11 presents the predictive power results of different metrics on the modified dataset. It can be observed that POSSCORE is not significantly affected by the length variation of the bad responses. All the results from POSSCORE are better than those selected baseline metrics.

5.4 Correlation Analysis

Following previous work [35, 36, 40], we compare the correlation between different metrics to analyze their relationships. Figure 3 presents Kendall's Tau correlation (r) between each pair of metrics given the original references and system responses. First of all, compared to correlations between traditional IR metrics based on relevance judgments of documents [40], we find conversational search metrics are generally more weakly correlated with each other. This is due to the nature of those metrics, which leverage different ways of modelling similarity between candidate responses and references, as a proxy of relevance. Secondly, not surprisingly, the metrics within the same category (Table 5) are more strongly correlated with each other, whereas metrics across categories are only moderately correlated. For example, BLEU1-4 and METEOR are strongly correlated ($r > 0.6$) while BERT-RUBER and BERT-Score are moderate ($r < 0.4$). It is worth noting that the correlation of our proposed POSSCORE with other metrics are moderate ($r < 0.5$), which means POSSCORE is measuring the responses substantially different from those metrics. Although POSSCORE is inspired by EA metrics, we can observe that the correlation between POSSCORE and EA is still less than 0.5. Therefore, this demonstrates that our proposed POSSCORE is a new metric that captures different aspects from those existing ones.

Table 12: An example of failure case of POSSCORE in the PersonaChat collection. The words with blue colour are the POS words that are recognized by POSSCORE.

Reference	I love blue too. I also enjoy mountain biking. Have you ever tried it?				
Candidate1	Do you have any hobbies? I enjoy mountain biking!				
Candidate2	I like pink . I think blue is too masculine color.				
Evaluation	Human	BLEU4	METEOR	BERT-Score	POSSCORE
Candidate1	**4**	**0.067**	**0.379**	**0.880**	1.437
Candidate2	3	0.031	0.147	0.859	**1.492**

5.5 Case Study

Despite that our proposed POSSCORE outperforms existing metrics, we present an example of its failure case in Table 12 to demonstrate its limitations for potential further improvements. We notice that in this example, all the automatic metrics except for POSSCORE are consistent with human annotations. This failure might be due to that POSSCORE is more sensitive to the POS distribution, especially when the POS word embeddings are very close (e.g., 'blue' vs. 'pink' in Table 12). Although the length of candidate1 is longer than that of candidate2, the number of POS words in candidate2 (n=7) is closer to the number in the reference (n=9). Since POSSCORE treats all the selected POS words equal and provides bonus scores (controlled by w) to the response if the distribution of POS words is similar to the reference, POSSCORE is more likely to overestimate the quality of this kind of response. In other words, the performance of POSSCORE might be constrained by the number of POS words in the candidates (i.e., the performance drops when the candidates have fewer POS words than the references). It can also be observed that the different mechanism between POSSCORE and other metrics, where POSSCORE considers both the word embedding similarity and the POS distribution similarity.

6 CONCLUSIONS

In this paper, we systematically analyzed whether and how POS words and POS tags could be used to benefit the automatic evaluation of conversational search systems. Based on the analysis, we propose two simple approaches to incorporate POS information into existing evaluation metrics, and present POSSCORE as a new simple metric for evaluating conversational search based on the distribution of POS words and tags. Extensive experiments on three publicly available datasets show that POSSCORE achieves significantly better alignments with human preferences than baseline metrics. Our work sheds light on the effectiveness of leveraging syntactic information for conversational search evaluation. In our future work, we plan to extend POSSCORE in two directions: (1) we currently treat all POS words within each syntactic group equally. However, we can observe in Table 7 that different POS groups play different roles. For example, the words with 'NOUN' tags have a stronger influence on the evaluation performance. (2) it has been shown that conversation context may contain useful explicit and implicit information for evaluation [21], which we did not yet consider. We would like to propose metrics that are both context- and syntactic-aware.

REFERENCES

[1] Avishek Anand, Lawrence Cavedon, Hideo Joho, Mark Sanderson, and Benno Stein. 2020. Conversational Search (Dagstuhl Seminar 19461). In *Dagstuhl Reports*, Vol. 9. Schloss Dagstuhl-Leibniz-Zentrum für Informatik.

[2] Satanjeev Banerjee and Alon Lavie. 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*. 65–72.

[3] Eric Brill. 1992. *A simple rule-based part of speech tagger.* Technical Report. PENNSYLVANIA UNIV PHILADELPHIA DEPT OF COMPUTER AND INFORMATION SCIENCE.

[4] Ye Chen, Ke Zhou, Yiqun Liu, Min Zhang, and Shaoping Ma. 2017. Meta-evaluation of online and offline web search evaluation metrics. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 15–24.

[5] Charles LA Clarke, Maheedhar Kolla, Gordon V Cormack, Olga Vechtomova, Azin Ashkan, Stefan Büttcher, and Ian MacKinnon. 2008. Novelty and diversity in information retrieval evaluation. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*. 659–666.

[6] Daniel Cohen, Liu Yang, and W Bruce Croft. 2018. Wikipassageqa: A benchmark collection for research on non-factoid answer passage retrieval. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*. 1165–1168.

[7] Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. 2018. Wizard of wikipedia: Knowledge-powered conversational agents. *arXiv preprint arXiv:1811.01241* (2018).

[8] Peter W Foltz, Walter Kintsch, and Thomas K Landauer. 1998. The measurement of textual coherence with latent semantic analysis. *Discourse processes* 25, 2-3 (1998), 285–307.

[9] Gabriel Forgues, Joelle Pineau, Jean-Marie Larchevêque, and Réal Tremblay. 2014. Bootstrapping dialog systems with word embeddings. In *Nips, modern machine learning and natural language processing workshop*, Vol. 2.

[10] Sarik Ghazarian, Johnny Tian-Zheng Wei, Aram Galstyan, and Nanyun Peng. 2019. Better automatic evaluation of open-domain dialogue systems with contextualized embeddings. *arXiv preprint arXiv:1904.10635* (2019).

[11] Karthik Gopalakrishnan, Behnam Hedayatnia, Qinglang Chen, Anna Gottardi, Sanjeev Kwatra, Anu Venkatesh, Raefer Gabriel, Dilek Hakkani-Tür, and Amazon Alexa AI. 2019. Topical-Chat: Towards Knowledge-Grounded Open-Domain Conversations.. In *INTERSPEECH*. 1891–1895.

[12] Seyyed Hadi Hashemi, Kyle Williams, Ahmed El Kholy, Imed Zitouni, and Paul A Crook. 2018. Measuring user satisfaction on smart speaker intelligent assistants using intent sensitive query embeddings. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*. 1183–1192.

[13] Seokhwan Kim, Michel Galley, Chulaka Gunasekara, Sungjin Lee, Adam Atkinson, Baolin Peng, Hannes Schulz, Jianfeng Gao, Jinchao Li, Mahmoud Adada, et al. 2019. The eighth dialog system technology challenge. *arXiv preprint arXiv:1911.06394* (2019).

[14] Julia Kiseleva, Kyle Williams, Ahmed Hassan Awadallah, Aidan C Crook, Imed Zitouni, and Tasos Anastasakos. 2016. Predicting user satisfaction with intelligent assistants. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*. 45–54.

[15] Julia Kiseleva, Kyle Williams, Jiepu Jiang, Ahmed Hassan Awadallah, Aidan C Crook, Imed Zitouni, and Tasos Anastasakos. 2016. Understanding user satisfaction with intelligent assistants. In *Proceedings of the 2016 ACM on Conference on Human Information Interaction and Retrieval*. 121–130.

[16] Julian Kupiec. 1992. Robust part-of-speech tagging using a hidden Markov model. *Computer speech & language* 6, 3 (1992), 225–242.

[17] Tian Lan, Xianling Mao, Heyan Huang, and Wei Wei. 2019. When to Talk: Chatbot Controls the Timing of Talking during Multi-turn Open-domain Dialogue Generation. *arXiv preprint arXiv:1912.09879* (2019).

[18] Tian Lan, Xian-Ling Mao, Wei Wei, Xiaoyan Gao, and Heyan Huang. 2020. PONE: A Novel Automatic Evaluation Metric for Open-Domain Generative Dialogue Systems. *arXiv preprint arXiv:2004.02399* (2020).

[19] Thomas K Landauer and Susan T Dumais. 1997. A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological review* 104, 2 (1997), 211.

[20] Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. Dailydialog: A manually labelled multi-turn dialogue dataset. *arXiv preprint arXiv:1710.03957* (2017).

[21] ALDO LIPANI, BEN CARTERETTE, and EMINE YILMAZ. 2021. How Am I Doing?: Evaluating Conversational Search Systems Offline. *ACM Transactions on Information Systems* (2021).

[22] Chia-Wei Liu, Ryan Lowe, Iulian V Serban, Michael Noseworthy, Laurent Charlin, and Joelle Pineau. 2016. How not to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. *arXiv preprint arXiv:1603.08023* (2016).

[23] Zeyang Liu, Ke Zhou, and Max L Wilson. 2021. Meta-evaluation of Conversational Search Evaluation Metrics. *arXiv preprint arXiv:2104.13453* (2021).

[24] Ryan Lowe, Michael Noseworthy, Iulian V Serban, Nicolas Angelard-Gontier, Yoshua Bengio, and Joelle Pineau. 2017. Towards an automatic turing test: Learning to evaluate dialogue responses. *arXiv preprint arXiv:1708.07149* (2017).

[25] Ryan Lowe, Nissan Pow, Iulian Serban, and Joelle Pineau. 2015. The ubuntu dialogue corpus: A large dataset for research in unstructured multi-turn dialogue systems. *arXiv preprint arXiv:1506.08909* (2015).

[26] Shikib Mehri and Maxine Eskenazi. 2020. USR: An Unsupervised and Reference Free Evaluation Metric for Dialog Generation. *arXiv preprint arXiv:2005.00456* (2020).

[27] Jeff Mitchell and Mirella Lapata. 2008. Vector-based models of semantic composition. In *proceedings of ACL-08: HLT*. 236–244.

[28] Jekaterina Novikova, Ondřej Dušek, Amanda Cercas Curry, and Verena Rieser. 2017. Why we need new evaluation metrics for NLG. *arXiv preprint arXiv:1707.06875* (2017).

[29] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 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, 311–318.

[30] Chen Qu, Liu Yang, W Bruce Croft, Johanne R Trippas, Yongfeng Zhang, and Minghui Qiu. 2018. Analyzing and characterizing user intent in information-seeking conversations. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*. ACM, 989–992.

[31] Filip Radlinski and Nick Craswell. 2017. A theoretical framework for conversational search. In *Proceedings of the 2017 conference on conference human information interaction and retrieval*. ACM, 117–126.

[32] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250* (2016).

[33] Adwait Ratnaparkhi. 1996. A maximum entropy model for part-of-speech tagging. In *Conference on empirical methods in natural language processing*.

[34] Vasile Rus and Mihai Lintean. 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, 157–162.

[35] Tetsuya Sakai. 2006. Evaluating evaluation metrics based on the bootstrap. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 525–532.

[36] Tetsuya Sakai. 2007. On the reliability of information retrieval metrics based on graded relevance. *Information processing & management* 43, 2 (2007), 531–548.

[37] Tetsuya Sakai. 2012. Evaluation with informational and navigational intents. In *Proceedings of the 21st international conference on World Wide Web*. ACM, 499–508.

[38] Tetsuya Sakai. 2013. Metrics, statistics, tests. In *PROMISE winter school*. Springer, 116–163.

[39] Tetsuya Sakai et al. 2005. The Effect of Topic Sampling on Sensitivity Comparisons of Information Retrieval Metrics.. In *Proceedings of NTCIR-5*.

[40] Tetsuya Sakai and Noriko Kando. 2008. On information retrieval metrics designed for evaluation with incomplete relevance assessments. *Information Retrieval* 11, 5 (2008), 447–470.

[41] Tetsuya Sakai and Zhaohao Zeng. 2019. Which Diversity Evaluation Measures Are" Good"?. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 595–604.

[42] Mark Sanderson, Monica Lestari Paramita, Paul Clough, and Evangelos Kanoulas. 2010. Do user preferences and evaluation measures line up?. In *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*. ACM, 555–562.

[43] Iulian Vlad Serban, Alessandro Sordoni, Ryan Lowe, Laurent Charlin, Joelle Pineau, Aaron Courville, and Yoshua Bengio. 2017. A hierarchical latent variable encoder-decoder model for generating dialogues. In *Thirty-First AAAI Conference on Artificial Intelligence*.

[44] Grigori Sidorov, Alexander Gelbukh, Helena Gómez-Adorno, and David Pinto. 2014. Soft similarity and soft cosine measure: Similarity of features in vector space model. *Computación y Sistemas* 18, 3 (2014), 491–504.

[45] Koustuv Sinha, Prasanna Parthasarathi, Jasmine Wang, Ryan Lowe, William L Hamilton, and Joelle Pineau. 2020. Learning an Unreferenced Metric for Online Dialogue Evaluation. *arXiv preprint arXiv:2005.00583* (2020).

[46] Mark D Smucker, James Allan, and Ben Carterette. 2007. A comparison of statistical significance tests for information retrieval evaluation. In *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*. 623–632.

[47] Chongyang Tao, Lili Mou, Dongyan Zhao, and Rui Yan. 2018. Ruber: An unsupervised method for automatic evaluation of open-domain dialog systems. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

[48] Alexandra Vtyurina, Denis Savenkov, Eugene Agichtein, and Charles LA Clarke. 2017. Exploring conversational search with humans, assistants, and wizards. In *Proceedings of the 2017 chi conference extended abstracts on human factors in*

computing systems. 2187–2193.

[49] Yi Yang, Wen-tau Yih, and Christopher Meek. 2015. Wikiqa: A challenge dataset for open-domain question answering. In *Proceedings of the 2015 conference on empirical methods in natural language processing*. 2013–2018.

[50] Fanghua Ye, Jarana Manotumruksa, and Emine Yilmaz. 2021. MultiWOZ 2.4: A Multi-Domain Task-Oriented Dialogue Dataset with Essential Annotation Corrections to Improve State Tracking Evaluation. *arXiv preprint arXiv:2104.00773* (2021).

[51] Tsuta Yuma, Naoki Yoshinaga, and Masashi Toyoda. 2020. uBLEU: Uncertainty-Aware Automatic Evaluation Method for Open-Domain Dialogue Systems. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*. 199–206.

[52] Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. Personalizing dialogue agents: I have a dog, do you have pets too? *arXiv preprint arXiv:1801.07243* (2018).

[53] Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating Text Generation with BERT. In *International Conference on Learning Representations*. https://openreview.net/forum?id=SkeHuCVFDr

[54] Ke Zhou, Ronan Cummins, Mounia Lalmas, and Joemon M Jose. 2012. Evaluating aggregated search pages. In *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*. ACM, 115–124.