Fundamental Exploration of Evaluation Metrics for Persona Characteristics of Text Utterances

Chiaki Miyazaki **Saya Kanno** **Makoto Yoda** **Junya Ono** **Hiromi Wakaki**
Sony Group Corporation, Japan
{chiaki.miyazaki, saya.kanno, makoto.yoda,
junya.ono, hiromi.wakaki}@sony.com

Abstract

To maintain utterance quality of a persona-aware dialog system, inappropriate utterances for the persona should be thoroughly filtered. When evaluating the appropriateness of a large number of arbitrary utterances to be registered in the utterance database of a retrieval-based dialog system, evaluation metrics that require a reference (or a “correct” utterance) for each evaluation target cannot be used. In addition, practical utterance filtering requires the ability to select utterances based on the intensity of persona characteristics. Therefore, we are developing metrics that can be used to capture the intensity of persona characteristics and can be computed without references tailored to the evaluation targets. To this end, we explore existing metrics and propose two new metrics: persona speaker probability and persona term salience. Experimental results show that our proposed metrics show weak to moderate correlations between scores of persona characteristics based on human judgments and outperform other metrics overall in filtering inappropriate utterances for particular personas.

1 Introduction

Maintaining utterance quality is important for commercial dialog systems. To achieve better quality, methods of filtering inappropriate utterances have been proposed from the perspectives of offensive language (Xu et al., 2020), grammar, topics (Tsunomori et al., 2020), discourse relation (Otsuka et al., 2017), and so on. In addition to these perspectives, we need a filter for **personas** of dialog systems. Persona-aware dialog systems are important in that having a consistent persona makes a dialog system believable (Higashinaka et al., 2018) and entertaining (Miyazaki et al., 2016). Throughout this paper, we use the term *persona* to indicate individuals such as real-life people and fictional characters. In addition, we use the term *persona characteristics* to indicate the distinctive qualities of a persona.

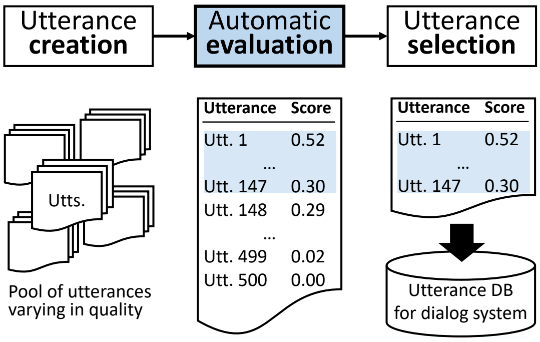


Figure 1: Process of selecting appropriate utterances for dialog system responses.

Figure 1 shows how we would like to automatically evaluate the appropriateness of a large number of arbitrary utterances and select utterances to be registered in the utterance database of a retrieval-based dialog system. Doing this is preferable for commercial use in terms of preventing unexpected utterances from being output. Evaluation metrics based on word overlap between an evaluation target and a reference (or a “correct” utterance) are often used to evaluate persona-aware utterance generation (e.g., *F1*, *BLEU*, and *ROUGE* in (Wolf et al., 2019; Madotto et al., 2019; Olabiyi et al., 2019)). However, these metrics are not applicable to utterance selection because preparing references for a large number of arbitrary utterances is extremely time-consuming. In other words, these metrics are not supposed to be used to evaluate utterances outside a predefined evaluation dataset. Therefore, metrics need to be computed without the references tailored to the evaluation targets. In addition, practical utterance selection requires the ability to select utterances based on the intensity of persona

Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue, pages 178–189
July 29–31, 2021. ©2021 Association for Computational Linguistics

characteristics.

Accordingly, we explore the metrics that can be used to capture the intensity of persona characteristics and can be computed without the references tailored to the evaluation targets. The contributions of this paper are as follows:

- We provide summaries of existing metrics used for evaluating persona-aware utterances.
- We propose two new metrics to evaluate persona characteristics without the references tailored to the evaluation targets.
- We investigate the effectiveness of the existing metrics and our proposed metrics in capturing the intensity of persona characteristics.

The rest of this paper is structured as follows. In Section 2, we introduce related work. In Section 3, we overview the existing evaluation metrics. In Section 4, we propose two new metrics. In Section 5, we investigate the correlation coefficient of the metrics between human judgments. In Section 6, we investigate filtering inappropriate utterances considering the practicality of the utterance selection.

2 Related Work

Since the release of the PERSONA-CHAT dataset (Zhang et al., 2018), many more studies have been conducted on persona-aware utterance generation (Song et al., 2019; Jiang et al., 2020; Liu et al., 2020), including studies by the 23 teams that participated in the ConvAI2 competition (Dinan et al., 2019). The PERSONA-CHAT dataset was created by crowdworkers who were asked to converse as the personas described in the given descriptions. Each description consisted of five sentences on average, such as “I am a vegetarian,” “I like swimming,” “My father used to work for Ford,” “My favorite band is Maroon5,” and “I got a new job last month, which is about advertising design.” In this manner, facts about the personas are described. However, the linguistic styles of the personas were not focused on.

Linguistic style is also an important aspect of persona-aware utterances. For example, Big Five personalities (Mairesse and Walker, 2007), gender, age, and area of residence (Miyazaki et al., 2015) can affect the linguistic styles of the utterances. In text style transfer, transfer success is often measured by transfer accuracy (Krishna et al., 2020). For example, when transferring negative sentences into positive ones, transfer success is measured by the fraction of sentences that are classified as positive (Fu et al., 2018).

Category		Metric
Persona-description-based	Trained	Persona accuracy
	Untrained	P-F1
		P-Cover
Sample-monologue-based	Trained	Personality classification accuracy
		uPPL
	Untrained	MaxBLEU

Table 1: List of existing metrics.

The same idea can be used to evaluate persona-aware utterances. In fact, there is a study that uses a similar evaluation metric called *personality classification accuracy* (Su et al., 2019), which is the accuracy of the speaker classification for the evaluation target utterances. We utilize and modify this idea so that we can measure the persona characteristics of each utterance.

3 Existing Metrics

This section introduces the existing evaluation metrics for persona-aware utterances that can be computed without the references being tailored to the evaluation targets. Table 1 shows the list of the existing metrics. The metrics are roughly divided into those that are based on the persona descriptions as used in the PERSONA-CHAT dataset and those that are based on the sample monologues of the personas. In addition, they can be categorized by the involvement of machine learning, i.e., trained or untrained. Hereinafter, we use the term *monologue* to refer to a set of independent utterances that are not associated with the preceding or the following utterances in a dialog.

3.1 Metrics Based on Persona Descriptions

3.1.1 Persona Accuracy

Persona accuracy (Zheng et al., 2020) is the accuracy with which the binary classification distinguishes if a persona description is expressed in the evaluation target utterances.

3.1.2 Persona F1 (P-F1)

P-F1 is an untrained evaluation metric used by Jiang et al. (2020) that was adapted from a previous study (Dinan et al., 2018). P-F1 is the harmonic mean of *persona precision* and *persona re-*

call, which are computed based on the number of non-stop words shared between an evaluation target and a persona description.

3.1.3 Persona Coverage (P-Cover)

P-Cover is another untrained metric used by Jiang et al. (2020) that was adapted from a previous study (Song et al., 2019). Though this is also based on the non-stop words shared between an evaluation target and the persona description, it utilizes inverse term frequency[1] to place weight on words.

3.2 Metrics Based on Sample Monologues

3.2.1 Personality Classification Accuracy

Personality classification accuracy (Su et al., 2019) is the speaker classification accuracy for the evaluation targets. The speaker classification can be achieved by building a classifier to distinguish the speakers of the utterances in a monologue corpus of the target personas.

3.2.2 User Language Perplexity (uPPL)

uPPL (Wu et al., 2020) is a metric that evaluates whether an utterance satisfies the linguistic style of a given persona. It can be obtained by building a statistical language model for a persona using a sample monologue and computing the perplexity of an evaluation target given by the language model. Wu et al. (2020) employed users of the Chinese social networking service Douban as personas and used their postings to train the language models.

3.2.3 MaxBLEU

Su et al. (2019) used *MaxBLEU* (Xu et al., 2018) to measure similarities between the evaluation target and the monologue of a persona. The MaxBLEU of an evaluation target can be obtained by calculating the BLEU score for each utterance in the monologue and finding the largest score. MaxBLEU is the only untrained metric among the existing sample-monologue-based metrics presented in this paper.

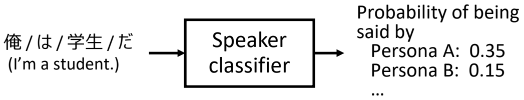
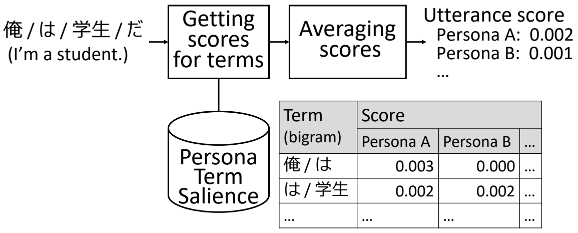


Figure 2: Process of obtaining an utterance score using PSProb.



Term (bigram)	Score		
	Persona A	Persona B	...
俺/は	0.003	0.000	...
は/学生	0.002	0.002	...
...	...	...	...

Figure 3: Process of obtaining an utterance score using PTSal.

4 Proposed Metrics

We propose a trained *persona speaker probability* (PSProb) metric and an untrained *persona term salience* (PTSal) metric.

4.1 Persona Speaker Probability (PSProb)

To measure the intensity of the persona characteristics of an utterance, we use the probability of the utterance being said by a persona. Figure 2 shows the process of obtaining an utterance score. First, we train a multinomial classifier to distinguish which persona is the speaker of each utterance in the training data. Then, we estimate the speaker to obtain the probability of an arbitrary utterance being said by a persona. This idea is quite similar to personality classification accuracy (Su et al., 2019). The sole difference is in their output: Persona classification accuracy is a metric that evaluates a set of utterances as a whole, while PSProb can be used to evaluate each utterance individually.

4.2 Persona Term Salience (PTSal)

We propose a metric that can be obtained without using machine-learning-based persona classification. We refrain from using such a classification to avoid complex conditions such as classification performance, machine learning algorithms, and training parameters. We assume evaluation metrics should be as simple as possible.

We define PTSal as the score that measures the importance of a term for a persona. Figure 3 shows the process of obtaining a score for an utterance.

[1] Though Jiang et al. (2020) and Song et al. (2019) used the term "inverse document frequency" for this, we chose the term used in the PERSONA-CHAT paper (Zhang et al., 2018) to avoid confusion with the inverse document frequency (IDF) used in the calculation of term frequency-inverse document frequency (TF-IDF), which will be mentioned in Section 4.2.

Conv. ID	Topic	Character	Utterance (created by crowdworkers)
4	Movie	Asuna	気分転換に映画に行こうよ、何がいいかな？ (Let’s go see a movie for a change. What would you like to see?)
		Lizbeth	そおねぇ、なにか恋愛コメディがいいなぁ、何が上映中か、アスナ知ってる？ (I’d like to see a romantic comedy. Do you know what’s playing, Asuna?)
		Asuna	恋愛コメディかぁ、何があったかな？ちょっと映画館まで下見に行かない？ (A romantic comedy? I wonder what movies are playing now. Why don’t we go down to the movie theater and check it out?)
			...
18	Fashion	Kirito	参考までに聞くんだが…、シノンはどんなファッションが好きなんだ？ (Just for reference... What kind of fashion do you like?)
		Sinon	アンタも知っての通り、動きやすい服装、一本よ。 (As you know, I wear comfortable clothes. That’s all.)
		Kirito	はは、機能重視だもんな。実はちょっと雰囲気を変えたいと思ってさ。何かオススメがあったら教えてほしいな。 (Haha, you only care about function in fashion, right? Actually, I was thinking of changing my fashion a bit. If you have any suggestions, please let me know.)
			...

Table 2: Examples of crowdsourced conversations.

First, we prepare a table of the PTSal for each term observed in the sample monologues of the target personas. Then, we calculate the average score of the terms in an arbitrary utterance by consulting the prepared table.

To calculate the PTSal, we adapt and modify the calculation of *TF-IDF*, which is widely used to capture the importance of a term in a **document**. By adapting the metric, we can capture the importance of a term for a **persona**. PTSal can be calculated using the following formulae:

$$PTSal(t,p) = \mathit{UttFreq}(t,p) \cdot \mathit{SpkrRarity}(t)$$

$$\mathit{UttFreq}(t,p) = \frac{n(t,p)}{m(p)}$$

$$\mathit{SpkrRarity}(t) = \log \frac{|P|}{s(t)},$$

where $n(t,p)$ is the number of utterances with term t in the monologue of persona p and $m(p)$ is the total number of utterances in the monologue of persona p. $s(t)$ is the number of personas that used term t, and $|P|$ is the total number of personas. $UttFreq$ is used to capture how often a term is used by a persona, and $SpkrRarity$ is used to capture how few personas use a term. In short, $UttFreq$ is used instead of term frequency (TF), and $SpkrRarity$ is used instead of IDF.

5 Experiment 1: Correlation with Scores Based on Human Judgments

5.1 Purpose and Procedure

To examine whether the evaluation metrics can capture the intensity of persona characteristics, we calculated the correlation coefficient (Spearman’s rho) of the metrics between human judgments. We used ten characters from two popular anime series as personas: *Kirito*, *Asuna*, *Sinon*, *Leafa*, and *Lizbeth* from *Sword Art Online* (SAO) and *Ran*, *Sonoko*, *Shinichi*, *Heiji*, and *Kazuha* from *Case Closed* (CONAN), which is also known as *Detective Conan*. The characters are all Japanese high school students. Kirito, Shinichi, and Heiji are male, and the others are female.

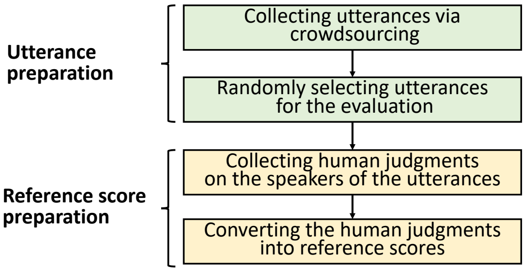

Utterance preparation
Collecting utterances via crowdsourcing
Randomly selecting utterances for the evaluation
Reference score preparation
Collecting human judgments on the speakers of the utterances
Converting the human judgments into reference scores

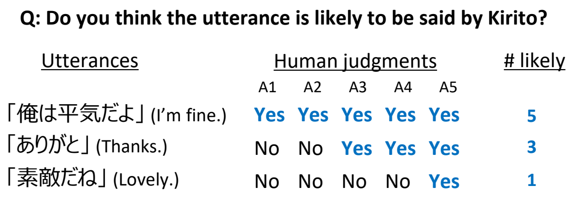
Figure 4: Process of preparing evaluation dataset.

5.2 Evaluation Dataset

We prepared the evaluation dataset by following the process shown in Figure 4. First, we collected utterances via crowdsourcing. To obtain the utterances that have characteristics of the target personas, we assigned a character to each crowdworker and asked the crowdworkers to converse as their characters. All the crowdworkers had watched the anime involved, with 92% of them having watched more than ten episodes. We included 26 topics (18 general topics and four topics specific to each anime) in the evaluation data and paired the crowdworkers to start conversations with an utterance regarding a given topic.

Anime	# utts.	# words	# uniq. words
SAO	498	12,779	1,797
CONAN	500	10,882	1,730

Table 3: Statistics of evaluation data.



Q: Do you think the utterance is likely to be said by Kirito?

Utterances	Human judgments					# likely
	A1	A2	A3	A4	A5	
「俺は平気だよ」(I'm fine.)	Yes	Yes	Yes	Yes	Yes	5
「ありがと」(Thanks.)	No	No	Yes	Yes	Yes	3
「素敵だね」(Lovely.)	No	No	No	No	Yes	1

Figure 5: Examples of human judgments with "likely" judgments being used as reference utterance scores.

The general topics consisted of self-introductions, movies, fashion, family, and so on. Table 2 shows examples of the crowdsourced conversations.

Through the data collection process, we obtained 2,070 utterances for each anime. For Experiment 1, we randomly extracted 100 utterances from each character and created a dataset that consisted of 500 utterances for each anime. Table 3 shows the statistics of the dataset. Note that the dataset for SAO consists of 498 utterances because there were misoperations for two utterances in the annotation process described in Section 5.3.

5.3 Preparation of Reference Scores

To obtain reference scores of persona characteristics, we asked crowdworkers for annotations. We gave each crowdworker a list of utterances[2] and a character, and we asked them to answer if the character was likely to say each utterance on the list. Note that judgments about one persona are independent of judgments about other personas; therefore, an utterance can be labeled as "likely" for multiple personas. Five crowdworkers were assigned to judge each combination of an utterance and a character, so the number of crowdworkers who chose "likely" for each combination ranged from 0 to 5. Figure 5 shows examples of the annotation results. It should be noted that all the annotation crowdworkers had experience watching the anime involved, and 80% of them had watched more than ten episodes.

Hereinafter, we refer to the number of "likely" judgments as *NoL* for convenience. We used the NoL instead of a Likert scale because we wanted to make the annotation easier for crowdworkers. We considered binary judgment would be easier than judgment on a scale. Figure 6 shows the distribution of the NoL. Since the evaluation data is a mixture of the utterances of five characters, there are many utterances whose NoL is 0 for each character. For example, Kirito is the only male among the five characters chosen from SAO; therefore, many utterances from female characters did not suit Kirito and scored 0. Similarly, many utterances were scored 0 for Heiji of CONAN, who speaks with a strong Kansai dialect, which is spoken in the western region of Japan.

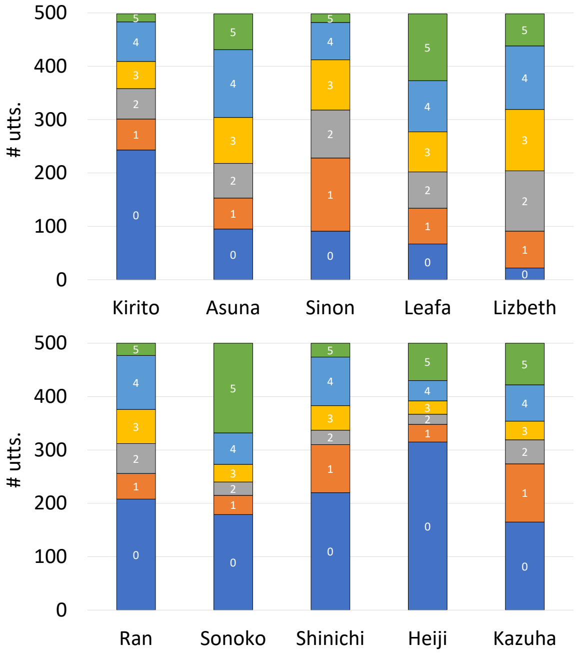


Figure 6: Numbers of utterances with NoL scores for each character (upper figure for SAO; lower figure for CONAN).

5.4 Metric Implementation Details

In this section, we describe the implementation details of the evaluation metrics used in this experiment, namely PSProb, PTSal, uPPL, and MaxBLEU. Of the metrics described in Section 3, persona accuracy and personality classification accuracy were not used because they are not applicable for scoring each utterance. Because P-F1 and P-Cover (based on persona descriptions) were proposed for evaluating utterances generated using persona descriptions, we assume they could be unsuitable for evaluating utterances created independently of the persona descriptions. Therefore,

[2] We split 500 utterances into ten lists consisting of 50 utterances per list and assigned five workers to each list, so we needed 50 crowdworkers for each character. Since we used ten characters, we used 500 crowdworkers in total for the annotation.

we evaluate these metrics as supplementary information in Section A of the appendix.

Unless otherwise noted, we tokenized utterances by using MeCab (Kudo et al., 2004) with the UniDic dictionary (Den et al., 2008). We chose that dictionary because it contains many colloquial expressions we consider suitable for tokenizing utterances.

5.4.1 Proposed 1: PSProb

As previously discussed, this metric is the probability of an utterance being said by a persona. We trained a multinomial classifier using logistic regression for SAO and CONAN. We used TF-IDF-weighted word unigrams as features. To train the models, we used monologue corpora consisting of lines extracted from SAO screenplays and subtitles from CONAN episodes. For SAO, we used screenplays for around 100 episodes; for CONAN, we used TV subtitles from 12 episodes. The lines in the subtitles are separated into short fragments, so we concatenated the consecutive lines of the same character. The numbers of lines, words, and unique words of the corpora are shown in Table 4. To adjust the imbalance of the data size among the characters, we randomly extracted the same number of lines for each character based on the smallest number. As a result, we used 1,955 lines for SAO (391 lines from each character) and 310 lines for CONAN (62 lines from each character). For each anime, we used 90% for training and used the remaining 10% for evaluating the classification performance. The performance of the speaker classifiers that we used to compute PSProb will be provided in Table B.2 of the appendix as supplementary information.

5.4.2 Proposed 2: PTSal

As previously stated, this is a metric to measure the importance of a term for a persona. We used all lines in the corpora shown in Table 4 as the sample monologues to calculate the PTSal. We used bigrams as terms of the words included in the lines. Table 5 shows example scores for the utterances. The first utterance, "俺" (first-person pronoun for male), strongly affected the score for Kirito. The second utterance, "キリトくん" ("Kirito-kun," a nickname for Kirito), strongly affected the score for Asuna because other characters rarely use the nickname to refer to or address Kirito. The third utterance, "お兄ちゃん" ("older brother"), strongly affected the score for Leafa because she mentions her brother frequently.

Anime	Character	# lines	# words	# uniq. words
SAO	Kirito	4,356	60,666	5,067
	Asuna	1,826	26,499	2,887
	Sinon	936	14,574	2,075
	Leafa	885	11,265	1,639
	Lizbeth	391	5,933	1,292
CONAN	Ran	241	2,765	603
	Sonoko	147	1,572	440
	Shinichi	103	1,844	559
	Heiji	94	1,684	482
	Kazuha	62	625	213

Table 4: Statistics of corpora used to compute PSProb, PTSal, uPPL, and MaxBLEU.

5.4.3 Existing 1: uPPL

To obtain the uPPL (Wu et al., 2020) of an utterance u, a statistical language model for the target persona LM_p should be trained first. Then, the uPPL can be calculated as the perplexity of u given by LM_p. Because the numbers of each persona's utterances are limited, Wu et al. (2020) trained a language model using all the training data and fine-tuned the model using each persona's utterances.

Because our monologue corpora are too small to construct a language model, we used a pre-trained Japanese BERT[3] as a language model, and we fine-tuned the model with our corpora shown in Table 4. We used 80% of the lines as training data, 10% as validation data, and 10% as evaluation data. We fine-tuned 100 epochs and chose the model whose validation loss was the lowest for each character. To calculate the perplexity of an utterance, first, we tokenized the utterance with the tokenizer for BERT, then we masked each word in the utterance, predicted the masked words using a language model, and obtained cross entropy loss for the probability distributions of predicted words. The perplexities of the evaluation data will be shown in Table B.3 of the appendix as supplementary information.

5.4.4 Existing 2: MaxBLEU

Based on a previous study (Su et al., 2019), we used MaxBLEU (Xu et al., 2018) as a metric that measures the similarities between an evaluation

[3] BERT-base_mecab-ipadic-bpe-32k_whole-word-mask obtained here: https://github.com/cl-tohoku/bert-japanese

Utterances (created by crowdworkers)	Kirito	Asuna	Sinon	Leafa	Lizbeth
こんにちは、どこから来たの？俺は桐ケ谷和人。埼玉県の川越市から来たんだ。 (Hello, where are you from? I'm Kazuto Kirigaya. I'm from Kawagoe City in Saitama Prefecture.)	0.0029	0.0001	0.0002	0.0000	0.0001
キリトくん、食べ物ばっかりだね …! (Kirito-kun, you keep talking about food...!)	0.0002	0.0042	0.0001	0.0001	0.0000
友達と一緒か、お兄ちゃんと一緒かなぁ〜。 (I'll be with my friends or with my brother.)	0.0000	0.0001	0.0001	0.0089	0.0011

Table 5: Examples of PTSal scores for utterances.

Character		r_s			
		PSProb	PTSal	uPPL	MaxBLEU
SAO	Kirito	**0.53** ***	**0.39** ***	-0.20 ***	0.17 **
	Asuna	0.28 ***	**0.33** ***	-0.06 n.s.	**0.32** ***
	Sinon	**0.21** ***	0.16 **	-0.03 n.s.	**0.37** ***
	Leafa	**0.35** ***	0.16 **	-0.02 n.s.	**0.27** ***
	Lizbeth	**0.32** ***	0.11 n.s.	-0.01 n.s.	0.03 n.s.
CONAN	Ran	**0.44** ***	**0.39** ***	-0.08 n.s.	0.07 n.s.
	Sonoko	**0.67** ***	**0.48** ***	-0.18 ***	0.02 n.s.
	Shinichi	**0.20** ***	**0.17** **	-0.11 n.s.	-0.01 n.s.
	Heiji	**0.52** ***	**0.48** ***	-0.45 ***	0.14 *
	Kazuha	**0.56** ***	**0.27** ***	-0.09 n.s.	0.10 n.s.

Table 6: Correlation coefficients (r_s) with NoL. "***," "**," and "*" indicate that r_s differs significantly from 0 at 0.1%, 1%, and 5%, respectively. "n.s." means r_s is not significantly different from 0. Significances are based on Holm-adjusted P-values.

target utterance and the sample monologue of a persona. We used the corpora shown in Table 4 as the sample monologues. We calculated the trigram BLEU score[4] between the evaluation target utterance and each utterance of the sample monologue, and we used the highest score as the evaluation target utterance score. To obtain the BLEU scores, we used `multi-bleu.perl` included in the Moses statistical machine translation system (Koehn et al., 2007) based on Xu et al. (2018).

5.5 Results

Table 6 shows the correlation coefficients (r_s) between the metrics and the NoL. In the table, the largest and the second-largest absolute values for each character are in bold. Note that the uPPL shows negative correlations because the smaller the perplexity is, the better the language model performs.

Our PSProb and PTSal metrics outperformed other metrics overall. The best and second-best performances were all PSProb or PTSal for CONAN in particular. The best performance of all was the case of PSProb for Sonoko, and the r_s was 0.67, which can be considered a strong correlation. Though PTSal could not perform as well as PSProb, PTSal did well without the assistance of machine learning. PTSal showed moderate to weak correlations for six out of ten characters, moderate correlations for Sonoko (0.48) and Heiji (0.48), and weak correlations for Kirito (0.39), Asuna (0.33), Ran (0.39), and Kazuha (0.27).

MaxBLEU was also computed without the assistance of machine learning; it did well for SAO, as we expected. However, it did not work well for CONAN, possibly because the size of the monologue corpus for CONAN was too small to find utterances sufficiently similar to the evaluation targets. In fact, while around 40% of the SAO utterances scored more than 20 in MaxBLEU, only around 9% of the CONAN utterances scored more than 20.

Although the uPPL did not work well overall, it performed well for Kirito and Heiji. The r_s of Kirito was -0.20, and the r_s of Heiji was -0.45, which can be considered weak to moderate correlations. As described in relation to Figure 6, their utterances have very different characteristics from other characters' utterances, assumedly a factor behind uPPL's good performance.

6 Experiment 2: Filtering Inappropriate Utterances

6.1 Purpose and Procedure

Considering the practicality of the utterance selection, we conducted another experiment to examine whether inappropriate utterances for personas can be filtered using the evaluation metrics. We used the same metrics as those used in Experiment 1, namely PSProb, PTSal, uPPL, and MaxBLEU. The implementation details of the metrics are the

[4] We chose BLEU-3 because it performed the best among BLEU-1 to 4 on the evaluation of SAO. As for CONAN, MaxBLEU did not perform well overall in this experiment.

Anime	Charac-ter	AUPR			
		PSProb	PTSal	uPPL	MaxBLEU
SAO	Kirito	**0.83**	**0.72**	0.65	0.68
	Asuna	0.40	**0.42**	0.34	**0.43**
	Sinon	0.52	**0.53**	0.46	**0.63**
	Leafa	**0.45**	0.34	0.28	**0.38**
	Lizbeth	**0.33**	**0.29**	0.16	0.19
CONAN	Ran	**0.79**	**0.68**	0.53	0.65
	Sonoko	**0.87**	**0.66**	0.48	0.59
	Shinichi	**0.76**	0.69	0.61	**0.75**
	Heiji	**0.89**	**0.88**	0.86	0.82
	Kazuha	**0.78**	**0.68**	0.55	0.64

Table 7: AUPR for each metric.

same as those described in Section 5.4. We used the same data described in Section 5.2 and Section 5.3 as the evaluation dataset. In this experiment, we regarded the utterances whose NoL is 0 or 1 to be inappropriate and tried to extract them. For each PSProb, PTSal, and MaxBLEU, we extracted an utterance if the score for the metric was less than or equal to a threshold. As for uPPL, we extracted an utterance if the score for the metric was more than or equal to a threshold.

6.2 Results

Figure 7 shows precision-recall curves for extracting inappropriate utterances. The upper figure is for Kirito of SAO, and the lower figure is for Ran of CONAN. Table 7 shows the area under the precision-recall curve (AUPR) for all the characters. The larger the score is, the better the extraction performance. In the table, the largest and the second-largest scores for each character are in bold. As in Experiment 1, our PSProb and PTSal metrics outperformed other metrics overall. Except for the case of Shinichi, the best and second-best performances were all PSProb or PTSal for CONAN. MaxBLEU also performed well overall. It performed best for Asuna and Sinon and second best for Leafa and Shinichi. However, uPPL had the lowest performance for all the characters. The overall trend in the results of this experiment is consistent with Experiment 1.

7 Conclusion

We investigated the performances of existing metrics and new metrics (namely PSProb and PTSal) to find metrics that we can use to capture the intensity of persona characteristics and we can compute without the references tailored to the evaluation targets. Experimental results showed that our PSProb and PTSal metrics generally outperformed others in terms of correlation with scores based on human judgments and performance in filtering inappropriate utterances. We would like to clarify the strengths and weaknesses of the metrics by considering various practical cases of evaluating persona characteristics. In addition, we would like to investigate the effectiveness of the metrics on automatically generated utterances and utterances written in other languages.

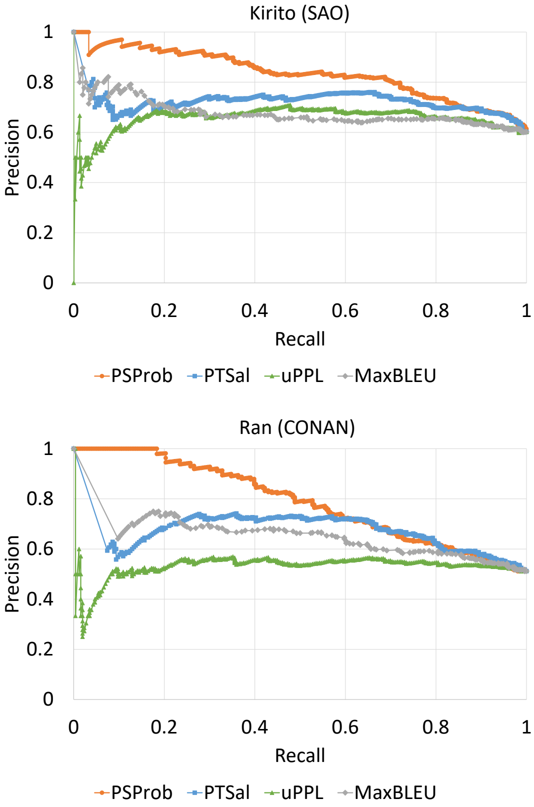


Figure 7: Precision-recall curves for utterance filtering (upper figure for Kirito of SAO; lower figure for Ran of CONAN).

Acknowledgments

The screenplay data for the Sword Art Online episodes was provided by Aniplex Inc.

References

Yasuharu Den, Junpei Nakamura, Toshinobu Ogiso, and Hideki Ogura. 2008. A proper approach to Japanese morphological analysis: Dictionary, model, and evaluation.

Emily Dinan, Varvara Logacheva, Valentin Malykh, Alexander Miller, Kurt Shuster, Jack Urbanek, Douwe Kiela, Arthur Szlam, Iulian Serban, Ryan

Lowe, et al. 2019. The second conversational intelligence challenge (ConvAI2). *arXiv preprint arXiv:1902.00098*.

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*.

Zhenxin Fu, Xiaoye Tan, Nanyun Peng, Dongyan Zhao, and Rui Yan. 2018. Style transfer in text: Exploration and evaluation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Ryuichiro Higashinaka, Masahiro Mizukami, Hidetoshi Kawabata, Emi Yamaguchi, Noritake Adachi, and Junji Tomita. 2018. Role play-based question-answering by real users for building chatbots with consistent personalities. In *Proceedings of the 19th Annual SIGdial Meeting on Discourse and Dialogue*, pages 264–272.

Bin Jiang, Wanyue Zhou, Jingxu Yang, Chao Yang, Shihan Wang, and Liang Pang. 2020. PEDNet: A persona enhanced dual alternating learning network for conversational response generation. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4089–4099.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th annual meeting of the association for computational linguistics companion volume proceedings of the demo and poster sessions*, pages 177–180.

Kalpesh Krishna, John Wieting, and Mohit Iyyer. 2020. Reformulating unsupervised style transfer as paraphrase generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 737–762.

Taku Kudo, Kaoru Yamamoto, and Yuji Matsumoto. 2004. Applying conditional random fields to Japanese morphological analysis. In *Proceedings of the 2004 conference on empirical methods in natural language processing*, pages 230–237.

Qian Liu, Yihong Chen, Bei Chen, Jian-Guang Lou, Zixuan Chen, Bin Zhou, and Dongmei Zhang. 2020. You impress me: Dialogue generation via mutual persona perception. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1417–1427.

Andrea Madotto, Zhaojiang Lin, Chien-Sheng Wu, and Pascale Fung. 2019. Personalizing dialogue agents via meta-learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5459.

François Mairesse and Marilyn Walker. 2007. PERSONAGE: Personality generation for dialogue. In *Proceedings of the 45th annual meeting of the association of computational linguistics*, pages 496–503.

Chiaki Miyazaki, Toru Hirano, Ryuichiro Higashinaka, Toshiro Makino, and Yoshihiro Matsuo. 2015. Automatic conversion of sentence-end expressions for utterance characterization of dialogue systems. In *Proceedings of the 29th Pacific Asia Conference on Language, Information and Computation*, pages 307–314.

Chiaki Miyazaki, Toru Hirano, Ryuichiro Higashinaka, and Yoshihiro Matsuo. 2016. Towards an entertaining natural language generation system: Linguistic peculiarities of Japanese fictional characters. In *Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 319–328.

Oluwatobi Olabiyi, Anish Khazane, Alan Salimov, and Erik Mueller. 2019. An adversarial learning framework for a persona-based multi-turn dialogue model. In *Proceedings of the Workshop on Methods for Optimizing and Evaluating Neural Language Generation*, pages 1–10.

Atsushi Otsuka, Toru Hirano, Chiaki Miyazaki, Ryuichiro Higashinaka, Toshiro Makino, and Yoshihiro Matsuo. 2017. Utterance selection using discourse relation filter for chat-oriented dialogue systems. In *Dialogues with Social Robots*, pages 355–365. Springer.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing*, pages 1532–1543.

Haoyu Song, Wei-Nan Zhang, Yiming Cui, Dong Wang, and Ting Liu. 2019. Exploiting persona information for diverse generation of conversational responses. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence*, pages 5190–5196.

Feng-Guang Su, Aliyah R Hsu, Yi-Lin Tuan, and Hung-Yi Lee. 2019. Personalized dialogue response generation learned from monologues. In *INTERSPEECH*, pages 4160–4164.

Yuiko Tsunomori, Ryuichiro Higashinaka, Takeshi Yoshimura, and Yoshinori Isoda. 2020. Improvements in the utterance database for enhancing system utterances in chat-oriented dialogue systems. *Journal of Natural Language Processing*, 27(1):65–88.

Thomas Wolf, Victor Sanh, Julien Chaumond, and Clement Delangue. 2019. TransferTransfo: A transfer learning approach for neural network based conversational agents. *arXiv preprint arXiv:1901.08149*.

Bowen Wu, Mengyuan Li, Zongsheng Wang, Yifu Chen, Derek Wong, Qihang Feng, Junhong Huang, and Baoxun Wang. 2020. Guiding variational response generator to exploit persona. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 53–65.

Jing Xu, Da Ju, Margaret Li, Y-Lan Boureau, Jason Weston, and Emily Dinan. 2020. Recipes for safety in open-domain chatbots. *arXiv preprint arXiv:2010.07079*.

Zhen Xu, Nan Jiang, Bingquan Liu, Wenge Rong, Bowen Wu, Baoxun Wang, Zhuoran Wang, and Xiaolong Wang. 2018. LSDSCC: a large scale domain-specific conversational corpus for response generation with diversity oriented evaluation metrics. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2070–2080.

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? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, pages 2204–2213.

Yinhe Zheng, Rongsheng Zhang, Minlie Huang, and Xiaoxi Mao. 2020. A pre-training based personalized dialogue generation model with persona-sparse data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 9693–9700.

A Evaluation of Metrics Based on Persona Descriptions

Regarding Experiment 1, we report evaluating the metrics based on persona descriptions, namely P-F1 and P-Cover. The evaluation dataset and the reference scores used for this evaluation are the same as those described in Section 5.

A.1 P-F1

P-F1 is a metric that evaluates how well a persona is expressed in an utterance (Jiang et al., 2020). It can be calculated using the following formulae:

$$\text{Persona F1} = \frac{2 \cdot \text{Recall} \cdot \text{Precision}}{\text{Recall} + \text{Precision}}$$
$$\text{Persona Recall} = \frac{\max_{i \in [1,L]} |W_{Y \cap d_i}|}{|W_{d_i}|}$$
$$\text{Persona Precision} = \frac{\max_{i \in [1,L]} |W_{Y \cap d_i}|}{|W_Y|},$$

where W_Y is a set of non-stop words in utterance Y and W_{d_i} is a set of non-stop words in the sentence d_i in the persona description.

The personas used by Jiang et al. (2020) are those in the PERSONA-CHAT dataset (Zhang et al., 2018), which means that each persona consisted of five sentences on average. In this experiment, we used persona descriptions that consisted of 20 sentences on average. We created the persona descriptions by extracting character descriptions from Wikipedia and removing sentences inappropriate for persona description (e.g., background of the anime series). The following is an excerpt of Kirito's persona description extracted from Wikipedia[5]:

> In the work, his birthday is October 7, 2008. He lives in Kawagoe City, Saitama Prefecture. He lost his parents in an accident shortly after his birth, and he was adopted by the Kirigaya family consisting of his mother's sister and her husband.

A.2 P-Cover

P-Cover is another metric that evaluates how well a persona is expressed in an utterance (Jiang et al., 2020). It can be calculated by the following formulae:

Character		r_s			
		P-F1		P-Cover	
SAO	Kirito	0.13	*	0.09	n.s.
	Asuna	0.00	n.s.	0.05	n.s.
	Sinon	-0.06	n.s.	-0.08	n.s.
	Leafa	0.00	n.s.	-0.04	n.s.
	Lizbeth	-0.05	n.s.	-0.01	n.s.
CONAN	Ran	0.04	n.s.	-0.02	n.s.
	Sonoko	0.08	n.s.	0.01	n.s.
	Shinichi	-0.10	n.s.	-0.11	n.s.
	Heiji	0.01	n.s.	0.00	n.s.
	Kazuha	-0.03	n.s.	-0.02	n.s.

Table A.1: Correlation coefficients (r_s) with NoL. "*" indicates that r_s differs significantly from 0 at 5%. "n.s." means r_s is not significantly different from 0. Significances are based on Holm-adjusted P-values.

$$\text{Persona Coverage} = \max_{i \in [1,L]} \frac{\sum_{w_j \in W_{Y \cap d_i}} \alpha_j}{|W_{Y \cap d_i}|}$$
$$\alpha_j = \frac{1}{1 + \log(1 + tf_j)}$$
$$tf_j = \frac{1e6}{idx_j^{1.07}},$$

where idx_j is the GloVe index and tf_j is computed via Zipf's law. The computation of tf_j was adapted from Zhang et al. (2018). We trained the GloVe (Pennington et al., 2014) using all the data shown in Table 4 and the persona descriptions. It should be noted that Jiang et al. (2020) seems to use the same GloVe model for both utterance generation and evaluation, but our evaluation target utterances were manually created independently of the GloVe model and the data used to train the model. The persona descriptions used for P-Cover are identical to those used for P-F1.

A.3 Results

Table A.1 shows the correlation coefficients (r_s) between the metrics and the NoL. The table indicates that neither P-F1 nor P-Cover showed significant correlation for most of the cases, primarily because the utterances did not have many exact words in common with the persona descriptions.

B Supplementary Information for Metric Implementation

B.1 PSProb

Table B.1 shows the breakdown of the data used for PSProb. As previously discussed, we used

[5] The original sentences are in Japanese.

Anime	Character	# lines		
		Total	Train	Eval.
SAO	Kirito	391	349	42
	Asuna	391	356	35
	Sinon	391	351	40
	Leafa	391	351	40
	Lizbeth	391	352	39
	All	1,955	1,759	196
CONAN	Ran	62	57	5
	Sonoko	62	56	6
	Shinichi	62	56	6
	Heiji	62	55	7
	Kazuha	62	55	7
	All	310	279	31

Table B.1: Breakdown of data used for PSProb.

Anime	Character	Precision	Recall	Chance rate
SAO	Kirito	0.47	0.64	0.21
	Asuna	0.51	0.51	0.18
	Sinon	0.55	0.53	0.20
	Leafa	0.56	0.45	0.20
	Lizbeth	0.42	0.36	0.20
CONAN	Ran	0.38	0.60	0.16
	Sonoko	0.50	0.50	0.19
	Shinichi	0.50	0.67	0.19
	Heiji	1.00	0.43	0.23
	Kazuha	0.83	0.71	0.23

Table B.2: Classification performance of models used to compute PSProb.

1,955 lines for SAO and 310 lines for CONAN, and we separated the lines into training data (90%) and evaluation data (10%).

Table B.2 shows the performance of the speaker classifiers that we used to compute PSProb. Though the scores do not seem to be that high, the precisions and recalls were all higher than the chance rates. All the precisions and recalls for SAO were significantly different from the chance rates ($p<0.05$; two-sided binomial test). The sample sizes for CONAN were too small to test for significance.

B.2 uPPL

Table B.3 shows the perplexities of the language models that we used to compute uPPL. Except for Lizbeth and Sonoko, the perplexity being at its lowest when characters of a model and evaluation data were identical meant the models were appropriately fine-tuned in general.

Model	Evaluation data				
	Kirito	Asuna	Sinon	Leafa	Lizbeth
Kirito	**24.1**	47.1	41.3	56.8	107.1
Asuna	80.2	**28.8**	55.8	66.8	96.3
Sinon	123.9	83.9	**40.4**	102.8	172.5
Leafa	179.5	100.7	121.8	**69.9**	188.3
Lizbeth	219.6	**163.5**	165.1	181.8	166.4

Model	Evaluation data				
	Ran	Sonoko	Shinichi	Heiji	Kazuha
Ran	**254.3**	1,576.0	604.2	1,258.8	457.8
Sonoko	**386.1**	773.3	771.0	2,497.0	1,304.3
Shinichi	1,177.5	4,211.4	**612.6**	3,262.7	2,271.5
Heiji	1,348.2	1,538.7	1,072.1	**263.7**	465.8
Kazuha	3,444.4	3,592.7	2,529.8	1,824.8	**392.6**

Table B.3: Perplexities for language models fine-tuned on each character (upper table for SAO; lower table for CONAN). Scores in bold are lowest perplexity for each model.