Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 1249–1255
Marseille, 20-25 June 2022

Self-Contained Utterance Description Corpus for Japanese Dialog

Yuta Hayashibe
Megagon Labs, Tokyo, Japan, Recruit Co., Ltd.
7-3-5 Ginza Chuo-ku, Tokyo, 104-8227, Japan
hayashibe@megagon.ai

Abstract

Often both an utterance and its context must be read to understand its intent in a dialog. Herein we propose a task, Self-Contained Utterance Description (SCUD), to describe the intent of an utterance in a dialog with multiple simple natural sentences without the context. If a task can be performed concurrently with high accuracy as the conversation continues such as in an accommodation search dialog, the operator can easily suggest candidates to the customer by inputting SCUDs of the customer's utterances to the accommodation search system. SCUDs can also describe the transition of customer requests from the dialog log. We construct a Japanese corpus to train and evaluate automatic SCUD generation. The corpus consists of 210 dialogs containing 10,814 sentences. We conduct an experiment to verify that SCUDs can be automatically generated. Additionally, we investigate the influence of the amount of training data on the automatic generation performance using 8,200 additional examples.

Keywords: Dialog corpus, Natural language understanding, Utterance description, Text generation

1. Introduction

To develop a task-oriented dialog system that responds appropriately to input utterances, the intent of the utterances must be correctly understood. For example, if a customer says, "Looks good, but expensive?", the omitted phrases and the intent of the question must be recognized. Such an understanding of language has been formulated and studied using two major task frameworks.

One is dialog-act classification or slot filling (Liu and Lane, 2016; Gupta et al., 2019; Shi, 2020). For example, MultiWOZ (Budzianowski et al., 2018) defines act types and slots for task-oriented dialogs in seven domains such as hotels and restaurants. They require pre-defined labels or slots. This task framework is powerful if the dialog proceeds according to pre-defined scenarios. However, it lacks flexibility because it can only interpret utterances using pre-defined types and slots. Therefore, interpreting utterances to solve tasks in an exploratory way in consultative dialogs is difficult using this framework.

The other framework is the generation of summary text. This can handle a wide variety of utterances. Recently, studies have employed this framework in the medical domain (Joshi et al., 2020; Song et al., 2020; Krishna et al., 2021) or the call center domain (Favre et al., 2015). In these studies, the whole dialog is interpreted rather than the intentions of individual utterances. This approach can understand dialog after it is over, but is not suitable to comprehend dialog said in real time.

Therefore, we propose a task, Self-Contained Utterance Description (SCUD), to simultaneously describe the intent of an utterance in a dialog with multiple simple natural sentences. By automatically generating such sentences, a dialog system can be connected to an information retrieval system that uses sentences as inputs. For example, in a dialog about finding a place to stay, it is possible to suggest suitable accommodation candidates by accurately understanding the customer's intentions in the sentences. As an example of one such task, this study conducts an experiment involving the automatic generation of SCUDs using the collected dialogs. Additionally, we evaluate the influence of the training data amount on the generation performance.

2. SCUD

SCUDs allow the intent of an utterance to be understood by simply reading them. SCUDs are sentences that align to each sentence in an utterance. Consider the following example:

(1) a. [Operator] There are inns where you can enjoy dinner while looking at the night view.
夜景を見ながらディナーを楽しめる宿もあります。

b. [Customer] Sounds good, but expensive?
いいと思うんですけど，高いですか?

(2) a. I want an inn where I can enjoy dinner while looking at the night view.
夜景を見ながらディナーを楽しめる宿が良い。

b. I would like to know whether the inn where I can enjoy dinner while looking at the night view is expensive.
夜景を見ながらディナーを楽しめる宿が高いかどうか知りたい。

For example, the sentence Utterance (1b) is incomprehensible without Utterance (1a). SCUDs for Utterance (1b) are (2a) and (2b). They are expressions that specify the implied meaning and supplement the omitted phrases. By reading the SCUDs, the intent of the utterance can be understood.

Date: June 6th, Reservation: 4 nights starting on July 5th, Area: Kyoto Prefecture, Number of people: 2 adults, 2 children.	
O	Thank you very much for visiting our service. Please let us know your preferences regarding accommodations. この度はご利用いただきまして、ありがとうございます。ご宿泊先につきまして、お客様のご希望をお聞かせいただけますでしょうか。
C	I would like to visit Kyoto with my husband and two young children. 夫と幼稚園児の子供 2 名と京都に行きたいとおもっています。
O	Is the destination in Kyoto city? Or is it in another area such as the Tango region? さようでございますか。京都の行先は京都市内でしょうか?それとも、丹後地方など別の地域でしょうか?
	...
C	We would like to have a buffet breakfast.$_{(A)}$ We are planning to eat dinner out in town. 朝食はバイキング希望です。$_{(A)}$ 夕食は外で食べる予定です。
O	All right. I will look into plans that include breakfast only. Do you have any other requests? 承知しました。それでは朝食のみ付いたプランをお調べします。ほかにご希望はございますでしょうか?
C	It would be nice to have a convenience store near the inn. コンビニが宿の近くにあるといいですね。
O	Okay. I will check for accommodations that have a convenience store nearby. What are the dates of your stay and how many nights will you be staying? かしこまりました。それでは、コンビニが近くにあるお宿をお調べします。ご宿泊のお日付と宿泊日数はいかがなさいますか?
C	We will be staying for 4 nights starting on July 5th. 7 月 5 日から 4 泊です。
O	I am sure you will have time for more than just sightseeing. Young children will be happy to play in the water in July. How about staying along the Kamo River? かしこまりました。4 泊となりますと、観光以外にもお時間があるかと思います。幼稚園児のお子様ですと、7 月は水遊びができると、お喜びいただけるかと思いますが、鴨川沿いのお宿はいかがでしょうか?
C	That is very good.$_{(B)}$ We can play in the Kamo River, can't we? Didn't know that.$_{(C)}$ それはとてもいいですね。$_{(B)}$ 鴨川で水遊びできるんですね。 知らなかったです。$_{(C)}$
	...

Table 1: Example of a collected dialog, where O represents the operator and C represents the customer. Underlined utterances are used as examples in Table 2.

3. Dialog Collection

For SCUD annotation, we collected Japanese dialogs between two people acting as a customer and an operator in a fictitious accommodation consultation service by using Slack[1], an online dialog platform. In a dialog, the customer informed the operator of their situation and needs. Then based on the information, the operator conducted a search to meet the customer's request. The dialog was finished once the operator judged that the requirements were specific enough to narrow appropriate accommodations.

3.1. Participants

All participants were native Japanese speakers with Slack experience. We asked 35 participants to play the role of the customer and two participants to play the role of the operator. One operator had experience in the tourism industry. The other did not. Each customer and operator pair engaged in six dialogs. Finally, we collected 210 ($= 35 \times 6$) dialogs[2].

3.2. Instructions for Participants

Each dialog had a random set of customer's situation settings consisting of the following elements and constraints: **date of the dialog** (month and day); **date of the reservation** (within three months of the date of the dialog and specified as either an exact date or span such as early, mid, or late in the month); **number of nights** (between one and four days); **areas** (one of the 51 areas in 47 prefectures, Tohoku, Kansai, Shikoku, and Kyushu regions); **number of people** (one or more for adults and zero or more for children[3]. The total number of adults and children is between one and four).

The constraints were presented to both the customer and the operator at the beginning of the dialog. We instructed the customer to ad-lib his or her requests based on the constraints. We also instructed the operator to finish the dialog when he or she judged that the requirements were specific enough to narrow down the accommodations.

[1] https://slack.com/

[2] Out of 210 dialogs, 126 dialogs were conducted by the operator who had experience in the tourism industry.

[3] 0 to 12 years old

Source	(A) We$_1$ would like to have$_2$ a buffet breakfast$_3$. 朝食はバイキング$_3$ 希望です$_2$。
SCUD	We$_1$ want to eat$_2$ buffet breakfast$_3$. 朝食はバイキング$_3$ 希望だ$_2$。
Source	(B) That$_1$ is$_2$ very good$_3$. それは$_1$ とてもいい$_3$ ですね$_2$。
Context	Young children will be happy to play in the water$_1$ in July. How about staying along the Kamo River$_1$? 幼稚園児のお子様ですと、7 月は 水遊び$_1$ ができると、お喜びいただけるかと思いますが、鴨川沿い$_1$ のお宿はいかがでしょうか?
SCUD	Accommodations along the Kamo River where we can play in the water$_{1\star}$ are$_2$ very good$_3$. 水遊びができる$_1$ 鴨川沿いの宿$_1$ がよい$_3$。
Source	(C) Didn't know$_2$ that$_3$. 知らなかったです$_2$。
Context	Young children will be happy to play in the water$_3$ in July. How about staying along the Kamo River$_3$? 幼稚園児のお子様ですと、7 月は水遊び$_3$ ができると、お喜びいただけるかと思いますが、鴨川沿い$_3$ のお宿はいかがでしょうか?
SCUD	I$_{1\star}$ didn't know$_2$ that we can play in the Kamo River$_{3\star}$. 鴨川で水遊びできるのを$_3$ お客様が$_1$ 知らなかった。$_3$

Table 2: Examples of annotated SCUD and alignments. The correspondences are indicated by underlines with the same number. $\star$ indicates parts that require extrasentential information for the generation.

3.3. Statistics of Collected Dialogs

The minimum number of utterances[4] per dialog was 11, and the maximum was 35. The average was 19.0. The total number of utterances was 4,006. We also annotated the sentence boundaries and counted the number of sentences per dialog. The minimum number of sentences was 33, the maximum was 78, and the average was 51.0. The total number of sentences was 10,814. Table 1 shows an example of a dialog in which the operator proposed "accommodations along the Kamo River where you can play in the water" to the customer. This idea was one that the customer had not initially thought of, and it was appreciated

[4]In this paper, we refer to utterances as the chunk that a user enters into Slack at one time. The participants are allowed to keep writing multiple utterances.

Distance	0	1	2	3	4	5	6	7	8	9	More
Number	2,129	1,196	68	80	23	30	7	13	4	3	15

Table 3: Farthest distance between the utterance containing the source and the one containing the alignments.

4. Annotation of SCUDs

4.1. Annotation Methodology

For each customer's utterance, we annotated SCUDs. We simply call the sentence to be interpreted the "source." Table 2 shows examples. The annotation exploits the predicate-argument structure because it is a basic representation of a sentence (Fillmore, 1967). The predicate-argument structure indicates the relationship between a verbal expression and its case-labeled arguments such as subjects and objects. It can be regarded as a concise interpretation of the sentence, which is suitable to annotate SCUDs.

We created drafts of the SCUDs using the morphological analyzer JUMAN++[5] (Tolmachev et al., 2020) (Revision.1ee40d7), the dependency and case structure analyzer KNP[6] (Revision.165d699a) (Kawahara and Kurohashi, 2014), and simple rules which convert predicate-argument structures to natural sentences. We then performed the following manual modifications to use them as SCUDs. First, we fixed grammatical and semantic errors, which were due to analysis errors or conversion errors. Second, we complemented omissions of words or phrases. These included the complement of pro-verb and clauses not performed by KNP. All modifications were performed by one professional annotator.

4.2. Analysis of Annotated SCUDs

By annotation, we obtained 3,568 SCUDs for 2,848 sources in customer utterances. Sources with multiple predicates can have more than one SCUD. Out of the 2,848 sources, 2,213 sources (77.7%) had a single SCUD, 561 sources (19.7%) had two SCUDs, 64 sources (2.2%) had three SCUDs, and 10 sources (0.4%) had four or more SCUDs. We manually annotated phrase alignment between sources and SCUDs. The underlines in Table 2 show examples. While (A) can generate a SCUD without referring to anything other than the source, (B) and (C) must refer to other sentences marked with $\star$.

We counted the farthest distance between the utterance containing the source and the one containing the alignments. Table 3 shows that 93.2% SCUDs can be created by referring to the previous (the distance is one) utterance.

[5]https://github.com/ku-nlp/jumanpp
[6]https://github.com/ku-nlp/knp

	Dialog			Additional		
	R_1	R_2	R_L	R_1	R_2	R_L
T5	0.704	0.587	0.688	0.643	0.480	0.626
T5+	0.824	0.702	0.811	0.834	0.727	0.825

Table 4: Evaluation scores. R_1, R_2, and R_L indicate ROUGE-1, ROUGE-2, and ROUGE-L, respectively.

5. Benchmark

We benchmarked a pre-trained encoder-decoder model, T5, on our corpus to investigate how the state-of-the-art language generation model performs for SCUD generation.

5.1. Benchmark Settings

Text-to-Text Transfer Transformer (T5) (Raffel et al., 2020) performs strongly in various tasks. We used the implementation by HuggingFace[7] and the pre-trained Japanese T5 model[8].

As shown in Section 4.2, 93.2% SCUDs can be generated to refer to the source and the sentence just before the source. Therefore we concatenated the source and its preceding sentences as its context with special tokens >> and input to the model. Then we tagged a sentence in an utterance to generate SCUDs with special tokens `<target>` and `</target>`. The generated output was all the SCUDs for the source.

Of the 210 dialogs in our corpus, we used 125 for training, 41 for development, and 44 for testing. This generated 1,442 (training), 462 (development), and 581 (testing) examples. We performed the Unicode NFKC normalization for all inputs.

In the training, we set the number of tokens for sources to 128, that of SCUD to 64, the batch size to 40, and the training rate to 10^{-3}. The number of epochs to was 20. In the test, we did not limit the number of tokens for sources.

Below, we refer to the model as T5, and evaluate the results with the ROUGE measure[9] (Lin and Hovy, 2003).

5.2. Benchmark Results

When the source was regarded as the output without processing, the average ROUGE-1, ROUGE-2, and ROUGE-L were 0.565, 0.457, and 0.558, respectively as shown in Table 4. Such low scores indicate that SCUD generation requires considerable rewriting.

For the T5 generation, the average ROUGE-1, ROUGE-2, and ROUGE-L were 0.704, 0.587, and 0.688, respectively. Table 5 shows some examples of SCUD generation. The lines labeled "SCUD (T5)" are the predictions of the trained model. Example #1 is an example where the output is perfect.

We randomly sampled 30 of the cases with Rouge-L scores below 0.6 and analyzed the types of errors. The most common errors were incorrect extraction from the input and insufficient extraction. Each type had eight errors. Example #2 is an example of an error in which incorrect phrases were taken from the context to complement the customer's utterance. It is likely that the models are not sufficiently trained in how to complement from the context. Example #3 is an example of an error in which multiple SCUDs should be outputted but only one was generated. As described in Section 4.2, only about 20% of sources had multiple SCUDs. This may be because the training for such sources did not work well. Even the state-of-the-art model could not adequately handle these phenomena.

The next most common error was to produce a SCUD with a significantly different meaning from the correct answer. There were seven of this type of error. Example Example #4 is an example of this type. The correct answer is "breakfast is not necessary if the budget is exceeded", but T5 incorrectly generated "breakfast is necessary."

5.3. Additional Corpus

Based on the results, we created an additional corpus consisting of 8,200 examples. These contained errors identified by the error analysis such as those that require viewing the context to generate SCUDs from an utterance and those that generate multiple SCUDs from a single utterance. We use 6,499 examples for training, 811 for development, and 890 for testing.

We trained another SCUD generation model with the additional corpus. This model is referred to as T5+. The average ROUGE-1, ROUGE-2, and ROUGE-L were 0.824, 0.702, and 0.811 for the dialog corpus test examples, respectively. Table 4 shows the scores. Increasing the number of training cases significantly improved ROUGE-L from 0.688 to 0.811. Table 5 shows examples of SCUD generation. The lines labeled "SCUD (T5+)" are the predictions of the model.

We also investigated the performance of the two models on the additional corpus. The average ROUGE-1, ROUGE-2, and ROUGE-L with T5 were 0.64, 0.48, and 0.63, respectively. In contrast, the average ROUGE-1, ROUGE-2, and ROUGE-L with T5+ were 0.83, 0.73, and 0.83, respectively. The Rouge-L score of T5 was 0.626, which is lower than that of the Dialog corpus, suggesting that the corpus included many difficult cases. However, the performance of T5+ was similar to that of the Dialog corpus performance for the additional corpus.

Several methods may improve the SCUD generation performance. The most straightforward method is to increase the amount of training data as much as possible. Our experiments confirmed that this is a valid approach to enhance the performance.

[7] https://github.com/huggingface/transformer

[8] https://huggingface.co/megagonlabs/t5-base-japanese-web-8k

[9] We used the SumEval implementation: https://github.com/chakki-works/sumeval

	Example #1	Example #2
Context	. . . If you're planning to take the bus, can I help you find a hotel that includes a ticket for the bus that goes around Kyoto? . . . バスをご利用でしたら、京都市内を回るバスのチケット付きの旅館をお探しましょうか?	Yes, sir. I'll check for inns with elevators. Would you like to have dinner? かしこまりました。それではエレベーター付きの旅館をお調べいたします。お食事はいかがなさいますか?
Source	I didn't know there was such a thing! I'm very happy to hear that. Please look for it. そういうものもあるんですね!すごく嬉しいです。お願いします。	I would like to have a common meal. I prefer to have a large amount. I would be nice if they would serve some local sake. 食事は普通で良いです。むしろ量が多い方が良いですね。あと、地元の銘酒とかを出して貰えるとありがたいのですが。
SCUD (Gold)	I need to find a hotel that includes a ticket for a bus that goes around Kyoto city. 京都市内を回るバスのチケット付きの旅館で探してほしい。	The more food, the better for me. 食事の量が多い方が良い。
SCUD (T5)	I need to find a hotel that includes a ticket for a bus that goes around Kyoto city. (R_1: 0.93, R_2: 0.86, R_L: 0.93) 京都市内を回るバスのチケット付きの旅館を探してほしい。	I want you to serve the best local sake. (R_1: 0.32, R_2: 0.00, R_L: 0.21) 地元の銘酒を出してほしい。
SCUD (T5+)	I need to find a hotel that includes a ticket for a bus that goes around Kyoto city. (R_1: 0.93, R_2: 0.86, R_L: 0.93) 京都市内を回るバスのチケット付きの旅館を探してほしい。	The more food, the better for me. (R_1: 1.00, R_2: 1.00, R_L: 1.00) 食事の量が多い方が良い。
	Example #3	**Example #4**
Context	We are four adults coming to Mie in mid-January for one night. 11 月中旬に三重に大人 4 名です。1 泊でお願します	Yes, sir. I'll also add hotels that offer discounts for consecutive nights to my search. How would you like meals? かしこまりました。連泊割引があるホテルも条件に加えておきますね。お食事の方はいかがなさいましょうか?
Source	I haven't decided on an area yet, but I'm thinking I'd like to stay in Ise-Shima. エリアはまだ決めてないんですがやっぱり伊勢志摩がいいかなと思ってます	If it is within my budget, I would like to have breakfast included. If it exceeds the budget, we can do without it. 予算内におさまるようでしたら、朝食サービスが欲しいです。予算を超えるようでしたら無しでもかまいません。
SCUD (Gold)	I haven't decided on an area yet. I'm thinking Ise-Shima would be good. エリアはまだ決めてない。伊勢志摩が良いかなと思っている。	We don't mind not having breakfast service if it exceeds our budget. 予算を超えるようなら朝食サービスは無しでもかまわない。
SCUD (T5)	I'm thinking Ise-Shima would be good. (R_1: 0.37 R_2: 0.24, R_L: 0.37) 伊勢志摩が良い。	If it is beyond our budget, we would like to have breakfast service. (R_1: 0.40 R_2: 0.17, R_L: 0.40) 予算を超えようでしたら、朝食サービスがほしい。
SCUD (T5+)	I haven't decided on an area yet. I'm thinking Ise-Shima would be good. (R_1: 0.76 R_2: 0.63, R_L: 0.76) エリアはまだ決めてない。やっぱり伊勢志摩が良い。	If breakfast is beyond our budget, we can do without it. (R_1: 0.58 R_2: 0.27, R_L: 0.50) 朝食が予算を超えれば無しでも構わない。

Table 5: Examples of the SCUD generation for underlined sentences. Pair of "Context" and "Source" is the input and "SCUD" is the output. T5 is the model trained only with dialogs and T5+ is the model trained with dialogs and additional data. R_1, R_2, and R_L indicate ROUGE-1, ROUGE-2, and ROUGE-L, respectively.

The second is to change the metric used to optimize the training. In this experiment, we directly optimized the evaluation index ROUGE. However, a classifier, which distinguishes between human-made sentences and system outputs, can be created to enhance fluency like Generative Adversarial Networks (Goodfellow et al., 2014). Then the predicted value of the classifier can be used for training. Another possibility is to build a classifier to determine whether the complement of ellipses is sufficient and use that classifier.

The third is to use auxiliary information such as alignment for training. As described in Section 4.2, we manually annotated phrase alignment between sources and SCUDs. If this annotation can be exploited, it may be possible to efficiently learn the information needed to complement the context.

6. Related Work

6.1. Label Classification and Slot Filling

Traditionally the task of capturing utterance intentions is designed as a label classification and slot filling task. This task has been annotated into corpora for training and evaluation, such as DSTC2 Corpus (Henderson et al., 2014), MultiWOZ (Budzianowski et al., 2018), the ICSI Meeting Recorder Dialog Act (MRDA) Corpus (Shriberg et al., 2004), and Action-Based Conversations Dataset (Chen et al., 2021).

In these corpora, utterances are understood by classifying them into pre-defined labels and filling in the slots. An advantage of this approach is that it is easy to handle due to the structured nature of the dialogs, which is sufficient when the topic is limited to a specific range. However, designing such a system is difficult for exploratory dialog.

6.2. Summarization

The method of understanding dialog by generating natural sentences is well studied in the field of dialog summarization. Summarization describes the main parts of the whole dialog while deleting the minor ones. Each sentence in the utterance is given a description, even if it has nothing to do with the conclusion of the dialog.

The corpus created by Fukunaga et al. (2018) is relevant to our study. Their corpus associated users' inexplicit intents to labels. Specifically, they focused on a scenario of a real estate search and associated utterances with labels. For example, the utterance, "I want to live alone." is associated with the "one-bedroom" label. In contrast, we aim to understand hidden intents in natural language in the form of SCUDs. The corpus created by (Yamamura and Shimada, 2018) is also relevant to our study. They annotated summarization for transcriptions of verbal dialogs per topic. Our annotation focuses on understanding users' intents rather than summarization.

6.3. Ellipsis Resolution

To generate SCUDs, both the target and other sentences must be referenced to generate the omitted expressions, which is ellipsis resolution. This task has been studied as semantic role labeling and predicate-argument structure analysis (PASA) (Gildea and Jurafsky, 2002; Kawahara and Kurohashi, 2004; Iida et al., 2005; Taira et al., 2008). From sentences, they extract relations such as "who did what to whom" that hold between a predicate and its arguments constituting a semantic unit of a sentence.

Although corpora annotated with texts such as newspaper articles (Baker et al., 1998; Palmer et al., 2005; Kawahara et al., 2002; Iida et al., 2007), blogs (Hashimoto et al., 2011), and web texts (Hangyo et al., 2012) are publicly available, most studies have focused on written texts. Imamura et al. (2014) constructed a corpus of 285 dialogs and performed PASA on the dialogs.

However, these tasks are designed to fill slots for predicates by extracting phrases. They are not designed to phrases based on existing texts. In addition, since almost all PASAs do not generate arguments but extract phrases, they cannot handle cases where complex expressions are omitted.

7. Conclusion

We have proposed the task to generate Self-Contained Utterance Description (SCUDs). Prior definitions of labels or slots are unnecessary to generate SCUDs in a human-readable format. We also constructed a dialog corpus, annotate SCUDs, and benchmark the proposed task against the recent state-of-the-art model T5 on automatic generation of SCUDs. The benchmark showed that increasing the amount of training data can improve the SCUD generation performance. In the future, we would like to improve the performance and create practical applications such as a question answering system.

Acknowledgments

We recognize Dr. Yuki Arase at Osaka University and Ms. Kayo Yamashita for the helpful discussions and insightful comments. Furthermore, we thank the anonymous reviewers for their careful reading and valuable comments.

8. Bibliographical References

Budzianowski, P., Wen, T.-H., et al. (2018). MultiWOZ - A Large-Scale Multi-Domain Wizard-of-Oz Dataset for Task-Oriented Dialogue Modelling. In *EMNLP*, pages 5016–5026.

Fillmore, C. J. (1967). The Case For Case. *Universals in Linguistic Theory*, pages 1–88.

Fukunaga, S.-y., Nishikawa, H., et al. (2018). Analysis of Implicit Conditions in Database Search Dialogues. In *LREC*, pages 2741–2745.

Gildea, D. and Jurafsky, D. (2002). Automatic labeling of semantic roles. *Computational Linguistics*, 28(3):245–288.

Goodfellow, I., Pouget-Abadie, J., et al. (2014). Generative Adversarial Nets. In *NeurIPS*, volume 27.

Gupta, A., Hewitt, J., et al. (2019). Simple, Fast, Accurate Intent Classification and Slot Labeling for Goal-Oriented Dialogue Systems. In *SIGdial*, pages 46–55.

Iida, R., Inui, K., et al. (2005). Anaphora Resolution by Antecedent Identification Followed by Anaphoricity Determination. *TALLIP*, 4(4):417–434.

Imamura, K., Higashinaka, R., et al. (2014). Predicate-argument structure analysis with zero-anaphora resolution for dialogue systems. In *COLING*, pages 806–815.

Kawahara, D. and Kurohashi, S. (2004). Zero Pronoun Resolution Based on Automatically Constructed Case Frames and Structural Preference of Antecedents. In *IJCNLP*, pages 334–341.

Kawahara, D. and Kurohashi, S. (2014). A Fully-Lexicalized Probabilistic Model for Japanese Syntactic and Case Structure Analysis. *Journal of Natural Language Processing*, 21(4):799–815.

Lin, C.-Y. and Hovy, E. (2003). Automatic evaluation of summaries using N-gram co-occurrence statistics. In *NAACL*, pages 71–78.

Liu, B. and Lane, I. (2016). Joint Online Spoken Language Understanding and Language Modeling With Recurrent Neural Networks. In *SIGdial*, pages 22–30.

Raffel, C., Shazeer, N., et al. (2020). Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Shi, H. (2020). A Sequence-to-sequence Approach for Numerical Slot-filling Dialog Systems. In *SIGdial*, pages 272–277.

Taira, H., Fujita, S., et al. (2008). A Japanese Predicate Argument Structure Analysis Using Decision Lists. In *EMNLP*, pages 523–532.

Tolmachev, A., Kawahara, D., et al. (2020). Design and Structure of The Juman++ Morphological Analyzer Toolkit. *Journal of Natural Language Processing*, 27(1):89–132.

9. Language Resource References

Baker, C. F., Fillmore, C. J., et al. (1998). The Berkeley FrameNet Project. In *ACL*, pages 86–90.

Chen, D., Chen, H., et al. (2021). Action-based conversations dataset: A corpus for building more in-depth task-oriented dialogue systems. In *NAACL*, pages 3002–3017.

Favre, B., Stepanov, E., et al. (2015). Call centre conversation summarization: A pilot task at multiling 2015. In *SIGdial*, pages 232–236.

Hangyo, M., Kawahara, D., et al. (2012). Building a Diverse Document Leads Corpus Annotated with Semantic Relations. In *PACLIC*, pages 535–544.

Hashimoto, C., Kurohashi, S., et al. (2011). Construction of a Blog Corpus with Syntactic, Anaphoric, and Sentiment Annotations (in Japanese). *Journal of natural language processing*, 18(2):175–201.

Henderson, M., Thomson, B., et al. (2014). The Second Dialog State Tracking Challenge. In *SIGdial*, pages 263–272.

Iida, R., Komachi, M., et al. (2007). Annotating a Japanese Text Corpus with Predicate-Argument and Coreference Relations. In *Proceedings of the Linguistic Annotation Workshop*, pages 132–139.

Joshi, A., Katariya, N., et al. (2020). Dr. Summarize: Global summarization of medical dialogue by exploiting local structures. In *EMNLP*, pages 3755–3763.

Kawahara, D., Kurohashi, S., et al. (2002). Construction of a Japanese Relevance-tagged Corpus. In *LREC*, pages 2008–2013.

Krishna, K., Khosla, S., et al. (2021). Generating SOAP notes from doctor-patient conversations using modular summarization techniques. In *ACL*, pages 4958–4972.

Palmer, M., Gildea, D., et al. (2005). The Proposition Bank: An Annotated Corpus of Semantic Roles. *Computational Linguistics*, 31(1):71–106.

Shriberg, E., Dhillon, R., et al. (2004). The ICSI Meeting Recorder Dialog Act (MRDA) Corpus. In *SIGdial*, pages 97–100.

Song, Y., Tian, Y., et al. (2020). Summarizing medical conversations via identifying important utterances. In *COLING*, pages 717–729.

Yamamura, T. and Shimada, K. (2018). Annotation and Analysis of Extractive Summaries for the Kyutech Corpus. In *LREC*, pages 3216–3220.