# Robust Cross-lingual Task-oriented Dialogue



LU XIANG, JUNNAN ZHU, and YANG ZHAO, National Laboratory of Pattern Recognition, Institute of Automation, CAS, School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing, China
YU ZHOU, National Laboratory of Pattern Recognition, Institute of Automation, CAS, School of Artificial Intelligence, University of Chinese Academy of Sciences, Fanyu AI Research, Beijing Fanyu Technology Co., Ltd, Beijing, China
CHENGQING ZONG, National Laboratory of Pattern Recognition, Institute of Automation, CAS, School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing, China

Cross-lingual dialogue systems are increasingly important in e-commerce and customer service due to the rapid progress of globalization. In real-world system deployment, machine translation (MT) services are often used before and after the dialogue system to bridge different languages. However, noises and errors introduced in the MT process will result in the dialogue system's low robustness, making the system's performance far from satisfactory. In this article, we propose a novel MT-oriented noise enhanced framework that exploits multi-granularity MT noises and injects such noises into the dialogue system to improve the dialogue system's robustness. Specifically, we first design a method to automatically construct multi-granularity MT-oriented noises and multi-granularity adversarial examples, which contain abundant noise knowledge oriented to MT. Then, we propose two strategies to incorporate the noise knowledge: (i) Utterance-level adversarial learning and (ii) Knowledge-level guided method. The former adopts adversarial learning to learn a perturbation-invariant encoder, guiding the dialogue system to learn noise-independent hidden representations. The latter explicitly incorporates the multi-granularity noises, which contain the noise tokens and their possible correct forms, into the training and inference process, thus improving the dialogue system's robustness. Experimental results on three dialogue models, two dialogue datasets, and two language pairs have shown that the proposed framework significantly improves the performance of the cross-lingual dialogue system.

CCS Concepts: • **Computing methodologies → Discourse, dialogue and pragmatics;**

Additional Key Words and Phrases: Cross-lingual, dialogue system, adversarial learning, knowledge, robustness

**ACM Reference format:**
Lu Xiang, Junnan Zhu, Yang Zhao, Yu Zhou, and Chengqing Zong. 2021. Robust Cross-lingual Task-oriented Dialogue. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.* 20, 6, Article 93 (July 2021), 24 pages.
https://doi.org/10.1145/3457571

The research work described in this article has been supported by the National Key Research and Development Program of China under Grant No. 2017YFB1002103.
Authors' addresses: L. Xiang, J. Zhu, Y. Zhao, Y. Zhou, and C. Zong, Intelligence Building, No. 95, Zhongguancun East Road, Haidian District, Beijing 100190, China; emails: lu.xiang@nlpr.ia.ac.cn, junnan.zhu@nlpr.ia.ac.cn, yang.zhao@nlpr.ia.ac.cn, yzhou@nlpr.ia.ac.cn, cqzong@nlpr.ia.ac.cn.

2375-4699/2021/07-ART93 $15.00
https://doi.org/10.1145/3457571

## 1 INTRODUCTION

Task-oriented dialogue systems aim to achieve specific user goals such as navigation inquiry or calendar scheduling within limited turns via natural language. Over the past several years, we have witnessed the explosions of research and real-world applications of dialogue systems from both academia and industry. Due to the simplicity and promising performance, end-to-end dialogue systems, which receive plain text as input and generate responses as output directly, currently attract great attention and have been applied to many virtual assistants and customer conversation services [5, 22, 23, 44, 45, 47].

With the rapidprogress of globalization and the popularity of cross-border e-commerce, it has become an urgent need for companies to develop and deploy cross-lingual dialogue systems. A natural way is to collect training data and train the dialogue system for each language. However, it is quite expensive and time-consuming to collect and perform human annotation of high-quality dialogue data. However, many internet companies, including Google, Baidu, Microsoft, and so on, have deployed their **machine translation (MT)** systems on-line, making the MT services easy to obtain. Therefore, in this article, we adopt MT systems as the language bridge before and after the dialogue system. The workflow is shown in Figure 1, and it can be divided into three steps. (1) *Translation step*: We adopt a machine translator to translate a user's utterance (language *e*) into the language that the system can deal with (language *f*). (2) *Dialogue step*: The dialogue system accepts the translated utterance and generates a response in language *f*. (3) *Back-translation step*: Another machine translator is adopted to translate the response back into language *e*.

Although the performance of MT systems has been greatly improved due to the rapid development of deep learning, general-purpose translators are usually developed based on the corpora in the news domain, which is different from the dialogue domain. Therefore, the translation step will inevitably introduce a variety of noises and errors to the dialogue module. The original dialogue system is usually trained on clean data, making it difficult to handle such noisy input properly. Many studies have demonstrated that even minimal changes in the input can fool state-of-the-art neural networks with high probability [16, 39]. For example, Belinkov and Bisk [4] has shown that **neural machine translation (NMT)** models can be easily brittle to small perturbations applied to inputs. End-to-end dialogue system is also facing the same problem. Considering the example in Figure 2, the end-to-end dialogue system will generate two entirely different responses for the translations of one sentence. Moreover, this problem is even crucial in dialogue systems since most conversations consist of multiple turns, and a small mistake in an early turn could cascade into a big misunderstanding later. In this article, instead of improving the NMT system's performance, we focus on making the dialogue system tolerate the perturbations introduced by machine translation to enhance the robustness of the whole system.

This problem has been studied in text classification [1, 12, 30, 51] and machine translation [4, 18] tasks. However, to the best of our knowledge, there has been little research focusing on the dialogue system's robustness. Furthermore, none of the existing work has focused on the robustness of the cross-lingual dialogue system. Therefore, this article studies a robust end-to-end dialogue system in the cross-lingual scenario that can overcome noises or errors introduced by MT. To this end, we propose a novel *MT-oriented noise enhanced framework* that exploits multi-granularity MT noises and injects such noises into the dialogue system to improve the dialogue system's robustness. Specifically, to capture the noises introduced in the MT process, we first design a method to construct multi-granularity MT-oriented noises automatically. Here, we adopt the word alignment and back-translation technique to extract multi-granularity MT-oriented noises, which are then used to generate multi-granularity adversarial examples at word-level, phrase-level, and sentence-level. These multi-granularity MT-oriented noises and adversarial examples contain abundant noise knowledge.

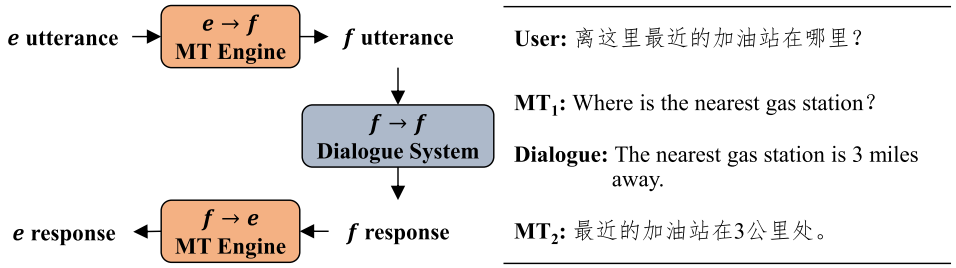


Fig. 1. The workflow of utilizing machine translators to deploy a cross-lingual dialogue system. Language $e$ is the user's language, and language $f$ is the dialogue system's language. $MT_1$ and $MT_2$ are $e \rightarrow f$ MT engine and $f \rightarrow e$ MT engine, respectively.

| User | 请告诉我最近的杂货店怎么走。<br>(qing gao su wo zui jin de za huo dian zen me zou) |
|---|---|
| Input_1 | give me directions to the closest grocery store. |
| Response_1 | The nearest grocery store is whole foods, it's 2 miles away. |
| Input_2 | tell me the direction of the latest grocery store. |
| Response_2 | the nearest grocery store is sigona farmers market, it's 4 miles away. |
| Gold Response | there are whole foods 2 miles away and sigona farmers market 4 miles away. where do we go? |

Fig. 2. The non-robust problem of the end-to-end dialogue system. The Input_1 and Input_2 are the translation results of the user's utterance by using two different machine translators. However, the responses generated by the dialogue system are quite different, shown as Response_1 and Response_2.

Furthermore, we propose two strategies to incorporate noise knowledge: Utterance-level adversarial learning and Knowledge-level guided learning. In the utterance-level adversarial learning, we propose a multi-granularity adversarial training method to enhance the dialogue system's robustness. This method adopts the adversarial training over the multi-granularity adversarial examples and the clean training data to learn a perturbation-invariant encoder, guiding the dialogue system to learn noise-independent hidden representations. In the knowledge-level guided learning, we enhance the robustness of the dialogue system from the perspective of knowledge. The multi-granularity noises contain the noise tokens and their possible correct forms. We incorporate the multi-granularity noises into both the training and the inference process to explicitly restrain the user's utterance, thus improving the dialogue system's robustness. One advantage of our proposed methods is that it does not require any data in the user's language, which means that a cross-lingual dialogue system can be deployed without any human effort once a MT system and an original end-to-end dialogue system are available.

We employ the proposed MT-oriented noise enhanced framework on several state-of-the-art end-to-end dialogue models. We conduct cross-lingual experiments on Chinese-to-English and German-to-English. Experimental results have shown that our proposed framework significantly improves the dialogue systems' performance on cross-lingual data.

Our main contributions are summarized as follows:

- We present a novel MT-oriented noise enhanced framework to improve the robustness of the end-to-end dialogue system in the cross-lingual scenario. To our best knowledge, this is the first work toward building a cross-lingual task-oriented dialogue from such an aspect.

- We propose a method to construct multi-granularity MT-oriented noises and multi-granularity adversarial examples. Furthermore, we propose two strategies to incorporate these two noise knowledge: (i) Utterance-level adversarial learning and (ii) Knowledge-level guided learning.
- The experimental results have shown that our proposed framework significantly improves the dialogue system's performance on cross-lingual data. Moreover, extensive experiments indicate that the combination of these two strategies can further improve the performance.

## 2 RELATED WORK

This article focuses on the robustness of the cross-lingual dialogue system, and the related work can be divided into the following three categories.

**Task-oriented dialogue systems.** Task-oriented dialogue systems are designed to assist users in achieving specific goals. They can be divided into two categories by the implementation method: *modularized* and *end-to-end*. For the modularized systems [45, 46], a set of modules including **spoken language understanding (SLU)** [3, 9, 29, 48], **dialog state tracking (DST)** [19, 52], dialogue policy learning [38, 41–43], and natural language generation [36] are used. These modules are designed separately, resulting in high costs and error propagation. However, end-to-end approaches [5, 13, 14, 20, 27, 45, 47, 49] have shown promising results. Wu et al. [47] proposed the **global-to-local memory pointer (GLMP)** network, which is composed of a global memory encoder, a local memory decoder, and a shared external knowledge to incorporate the external knowledge into the learning framework effectively. Unlike GLMP, Lei et al. [20] and Zhang et al. [49] explicitly model the process of belief tracking and proposed TSCP and LABES-S2S, respectively, both of which are a two-stage copy-augmented Seq2Seq model. Though the above methods have some differences, our proposed framework can be applied to any of the end-to-end dialogue systems. In this article, we use the GLMP [47], TSCP [20], and LABES-S2S [49] as the monolingual dialogue systems in the cross-lingual dialogue workflow shown in Figure 1.

**Cross-lingual dialogue systems.** Adapting dialogue systems to different languages is a challenging task and has not yet been explored thoroughly enough. The previous work mainly focuses on the cross-lingual transfer of some modules in the modularized system. Calvo et al. [6, 7] proposed different strategies to convert the SLU training data into the target language through MT systems so as to train the corresponding SLU model. Bai et al. [2] proposed to use reinforcement learning to improve the translation for SLU language transferring. Liu et al. [24, 26] proposed to refine the cross-lingual word embeddings and introduced a latent variable model to improve the performance of zero-shot cross-lingual SLU. Chen et al. [8] studied the problem of cross-lingual DST and proposed a teacher-student framework for building cross-lingual DST. Schuster et al. [35] presented a multilingual intent and slot filling dataset and explored different cross-lingual transfer learning methods to improve intent and slot detection models for other languages. Liu et al. [25] leveraged a mixed language training framework for cross-lingual transfer of SLU and DST. Qin et al. [34] introduced a data augmentation framework to generate multi-lingual code-switching data for zero-shot cross-lingual tasks including SLU and DST. Different from prior work, our work directly uses a machine translator to translate user utterance into target language to interact with the system. Moreover, we mainly concentrate on building a robust cross-lingual dialogue system toward the machine translator. This is a totally different aspect from prior work.

**Adversarial learning in Natural Language Processing.** Using adversarial learning to improve the system's robustness has been applied to various natural language processing tasks, including text classification [12, 30], machine translation [4, 10, 11], dialogue generation [21], and so on. The basic idea is to attack the well-trained network by constructing adversarial samples so

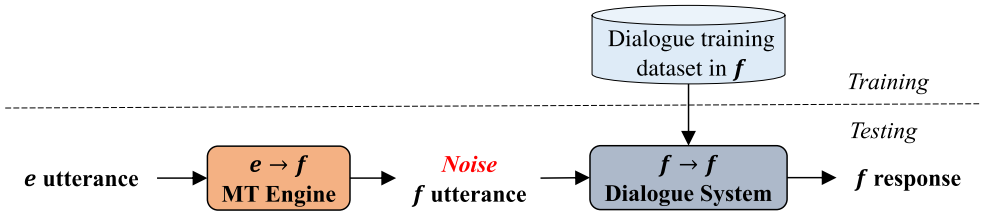


Fig. 3. Task description of cross-lingual dialogue system.

that the network parameters can be adjusted to improve the robustness and resist these attacks. Belinkov and Bisk [4] improved the robustness of character-based neural machine translation models by including adversarial examples in the training data. Niu and Bansal [31] revealed the over-sensibility and over-stability of the generative dialogue models, in which they generated adversarial examples using different strategies, including random swap, stopword dropout, data-level paraphrasing, generative-level paraphrasing and grammar errors. Unlike previous work, ours generates adversarial examples oriented to the MT so that the trained dialogue system can uttermost handle the noises or errors introduced by the machine translator.

## 3 TASK DESCRIPTION

Our cross-lingual dialogue system's objective is to allow the users to interact with the system using other languages different from the training language, avoiding the expensive cost of collecting training data for each language. Thus we employ machine translators to bridge the language gap between users and the dialogue agent. We use the following two resources to deploy the cross-lingual dialogue system:

(1) **Dialogue dataset in language *f*:** $\mathcal{D}_f = \{(X_f, B, Y_f)\}$, where $X_f$ denotes the dialogue context and $B$ denotes the **knowledge base (KB)** information. $Y_f$ denotes the response given the dialogue context and KB. This dialogue dataset is used to train the end-to-end dialogue agent in language $f$.

(2) **MT engines**, which can translate sentence from $e$ to $f$ (denoted as $MT_{e \Rightarrow f}$) and back-translate from $f$ to $e$ (denoted as $MT_{f \Rightarrow e}$). Hence, the cross-lingual dialogue system for language $e$ can be formalized as the following three steps:

$$X_f = g_1(X_e | MT_{e \Rightarrow f}), \tag{1}$$

$$Y_f = g_2(X_f | \mathcal{D}_f), \tag{2}$$

$$Y_e = g_3(Y_f | MT_{e \Rightarrow f}). \tag{3}$$

As illustrated in Figure 3, the first step utilizes the machine translator to convert the user's input $X_e$ into $X_f$. This will always introduce unexpected noises and errors, such as the translations of entities, the expression modes, and words chosen by the translator. However, the dialogue systems are usually trained on clean datasets, making them vulnerable to these noises and produce undesirable and unintended outputs. For example, in Figure 2, both input_1 and input_2 are the translation results of the same source sentence using two different machine translators. When feeding these two translations to a dialogue system separately, the system generates two entirely different responses. We do not aim to improve the machine translator's performance because there are many translation services on-line for hundreds of language pairs, and we can directly use such services to quickly deploy a cross-lingual dialogue system. Instead, our goal is to enhance the robustness of the dialogue system to handle the noisy input generated by machine translation systems. Our

goal can be formalized as follows:

$$Y_f = Robust(X_f|\mathcal{D}_f, MT_{e\Rightarrow f}, MT_{e\Rightarrow f}). \tag{4}$$

## 4 BACKGROUND

End-to-end task-oriented dialogue systems usually use the dialogue history $X = (x_1, \ldots, x_n)$ and the KB information $B$ as input and directly output system response $Y = (y_1, \ldots, y_m)$. The probability of a response is defined as follows:

$$p(Y|X, B) = \prod_{i=1}^{m} p(y_i|y_1, \ldots, y_{i-1}, X, B). \tag{5}$$

Unlike the other typical Seq2Seq text generation task, the success of a task-oriented dialogue system heavily depends on KB queries' accuracy. The current end-to-end models can be divided into two sub-categories. The first category is those that extend the Seq2Seq architecture, which does not explicitly perform belief tracking. Meanwhile, Eric et al. [13], Madotto et al. [27], and Wu et al. [47] adopt a copy mechanism that allows copying information retrieved from the KB to the generated response. The latter two work adopt Memory Networks to memorize the KB entities and words appearing in the dialogue history.

The second category lies in explicitly modeling the process of belief tracking, which is necessary to form KB quires. All of these works [17, 20, 37, 49, 50] are based on the copy-augmented Seq2Seq learning framework proposed by Lei et al. [20], which introduced the belief span to track the dialogue belief states and facilitate KB retrieval. The TCSP [20] is a two-stage decoding process. In the first stage, the Seq2Seq model decodes the belief span $B_t$ unconditionally. The $B_t$ is used to perform KB search, resulting in $k_t$. Then in the second stage, the Seq2Seq model continues to generate a machine response $R_t$ on the additional conditions of $B_t$ and $k_t$. The two-stage decoding can be formularized as follows:

$$B_t = seq2seq(B_{t-1}R_{t-1}U_t|0, 0), \tag{6}$$

$$R_t = seq2seq(B_{t-1}R_{t-1}U_t|B_t, k_t). \tag{7}$$

Without loss of generality, we use several end-to-end dialogue models from the above two categories, including GLMP model [47], TSCP model [20], and LABES-S2S model [49], to evaluate our proposed MT-oriented noise enhanced framework.

## 5 OUR METHOD

This article aims to learn a robust dialogue system that can overcome perturbations in the input sentences introduced by the machine translator under the cross-lingual scenario. The critical problem is how to bridge the semantic gap between the noise text and clean text. To achieve that, we present a novel MT-oriented noise enhanced framework to improve the dialogue model. Specifically, we first construct multi-granularity MT-oriented noises. Then, we make use of the multi-granularity MT-oriented noises to construct adversarial examples. Finally, we propose two strategies to make the dialogue system learn from the adversarial examples and MT-oriented noises. We will describe the above three steps in the following sections.

### 5.1 Multi-Granularity MT-Oriented Noises Construction

To overcome MT noises, we need to know what kind of words or phrases the MT prefers to output. We utilize the bi-directional MT systems and monolingual data in language $f$ to construct word-level and phrase-level MT noises. Given a sentence of $k$ words in language $f$ $S_f = \{w_1, w_2, \ldots, w_k\}$,

**ALGORITHM 1:** Multi-granularity MT-oriented noises construction.

**Input:**
A set of monolingual data in language $f$: $\mathcal{S}_f = \{S_f^n\}_{n=1}^N$; Two MT systems:$\{MT_{e\Rightarrow f}, MT_{f\Rightarrow e}\}$; predefined lexical translation threshold: $\mu_1$; predefined phrase translation threshold: $\mu_2$

**Output:**
multi-granularity MT-oriented noises

Sentence pair set $SP \leftarrow \emptyset$
**for** each $S_f^n \in \mathcal{S}_f$ **do**
  Translate $S_f^n$ into $T_e^n$ through $MT_{f\Rightarrow e}$; Back-translate $T_e^n$ into $G_f^n$ through $MT_{e\Rightarrow f}$;
  Add sentence pair $(S_f^n, G_f^n)$ to the set $SP$;
**end for**
Conduct bi-directional word alignment on $SP$ using GIZA++;
Obtain the word alignment information $A_{SP}$ by the intersection of the two alignments;
Extract lexical and phrase translation table $L = \{(w_i, w_i^{'})\}$ and $M = \{(m_i, m_i^{'})\}$ from $A_{SP}$;
Word-level noises $LEX_{noise} \leftarrow \emptyset$; Phrase-level noises $PHR_{noise} \leftarrow \emptyset$;
**for** each $(w_i, w_i^{'}) \in L$ **do**
  **if** $p(w_i|w_i^{'}) \geq \mu_1$ and $p(w_i^{'}|w_i) \geq \mu_1$ **then**
    Add $(w_i, w_i^{'})$ to the set $LEX_{noise}$;
  **end if**
**end for**
**for** each $(m_i, m_i^{'}) \in M$ **do**
  **if** $(len(m_i) \geq 2$ *and* $len(m_i^{'}) \geq 2)$ *and* $(p(m_i|m_i^{'}) \geq \mu_2$ *and* $p(m_i^{'}|m_i) \geq \mu_2)$ *and* (the boundary words are not function words and must be aligned) **then**
    Add $(m_i, m_i^{'})$ to the set $PHR_{noise}$;
  **end if**
**end for**
**return** $LEX_{noise} \cup PHR_{noise}$;

we first employ the $MT_{f\Rightarrow e}$ to translate the sentence from $f$ to $e$ and then use the $MT_{e\Rightarrow f}$ to back-translate the sentence into $f$ $G_f = \{w_1^{'}, w_2^{'}, \ldots, w_l^{'}\}$. Through this cyclic translation procedure, we create a group of pseudo data $\{(S_f, G_f)\}$, and such data contain the noises introduced by MT. In this article, borrowing the idea of statistical machine translation, two granularity noises are constructed utilizing the pseudo data $\{(S_f, G_f)\}$. The process for multi-granularity MT-oriented noises construction is shown in Algorithm 1.

(1) *Word-level Noises* $\{(w_i, w_i^{'})\}$, where $w_i$ is the correct word and $w_i^{'}$ is the corresponding noise word. In this article, word alignment [32] is used to construct lexical noises. First, GIZA++[1] is adopted to conduct bi-directional word alignment on $\{(S_f, G_f)\}$ to establish the relationship between words in $S_f$ and those commonly used in the MT system, such as "nearest" and "latest." Based on the word alignment result, we can extract lexical translation table $L = \{(w_i, w_i^{'})\}$. However, the word alignment always contains many noises, and directly using the alignment will result in unexpected errors. We adopt the following two strategies to improve the quality of word alignment: (a) bidirectional word alignment, which means that only the alignments existing in both directions are retained; (b) threshold filtering, that is, for a word pair $(w_i, w_i^{'})$, if its lexical translation probability is smaller than a preset threshold, it will be removed from the word alignment

[1] https://github.com/moses-smt/giza-pp.

result. The lexical translation table obtained by the above method is denoted as word-level noises $LEX_{noise}$.

(2) *Phrase-level Noises* $\{(m_i, m_i^{'})\}$, where $m_i$ is the correct phrase and $m_i^{'}$ is the corresponding noise phrase, for example, "6 am" and "six in the morning." Phrase-level noises contain more contextual information than word-level noises. Similarly, we construct phrase-level noises from the bidirectional word alignment. To improve the quality of the phrase translation table, we only preserve the phrase pairs that meet the following conditions: (a) the length of the source phrase $m_i$ and the target phrase $m_i^{'}$ is no smaller than 2; (b) the boundary words of $m_i$ and $m_i^{'}$ are not function words; (c) phrase translation probability must be higher than a preset threshold. The phrase translation table obtained by the above method is denoted as phrase-level noises $PHR_{noise}$.

### 5.2 Multi-Granularity Adversarial Examples Generation

Under the cross-lingual scenario, the dialogue system receives the MT result of a user's utterance as input. However, due to the noises and errors introduced in the MT process, the dialogue system's performance will degrade dramatically. To alleviate this problem, we propose to transform the clean dialogue training data into noise data that contain MT noises so that the dialogue system can learn how to deal with the noises. After obtaining the above multi-granularity MT-oriented noises, we need to inject the noises into the dialogue training data and generate adversarial examples at the word and phrase level.

*(1) Word-Level Adversarial examples.*

We substitute words with their candidates from the word-level noises to produce diverse user utterances containing MT noises at the word level. The entire process is shown in Algorithm 2 and consists of the following two steps:

**Step 1: Important Words Selection (lines 4–8)** We observe that some keywords are more important than other words in user utterances. We divide words into two categories: intention-related and context-related. For example, in the utterance "*i want to find an* ***expensive*** *restaurant in the* ***east*** *part of town*," the words "expensive" and "east" are intention-related words. The other words like "want," "find," "restaurant," "part," "town" are context-related words. The intention-related words express the user's needs, and it will be much helpful if we can generate a variety of expressions about such words. Hence, we prioritize replacing intention-related words. On the other side, the context also affects the understanding of the user's utterance. We also randomly sample one word from the context-related words to replace. It is worth noting that we do not allow the replacement for qualifiers (such as the, a, an), modal verbs (such as can, cannot, could), and personal pronouns (such as she, hers, herself, it) since changing them could easily lead to inconsistent and even semantic changes.

**Step 2: Candidate Word Replacement (lines 11–16)** For a given word $w$ that needs to be replaced, we first lookup the word-level noises and obtain a set of candidates $\{w_1^{'}, w_2^{'}, \ldots, w_k^{'}\}$. Then we randomly select a word from the candidates to replace the given word $w$.

Overall, the algorithm first uses Step 1 to filter out words that cannot be replaced and select intention-related words and context words. Then, the algorithm repeats Step 2 to find replacements for each word and replace it with the candidate.

*(2) Phrase-Level Adversarial examples.*

Word-level adversarial examples are relatively straightforward. However, such kind of rewriting is limited to only a few words. In the real expression of human language, there exist various phrases with synonymous meanings. To further diversify the adversarial examples, we introduce the phrase-level adversarial examples through the phrase-level noises. Similarly to the word-level adversarial example generation, the process for generating phrase-level adversarial examples is also divided into two steps:

**ALGORITHM 2:** The Procedure of Word-Level Adversarial Examples Generation.

**Input:**
Dialogue training data in language $f$: $\mathcal{D}_f = \{(X_f, Y_f)\}$; Word-level noises: $LEX_{noise}$
**Output:**
Word-level adversarial examples: $\mathcal{W}_{adv} = \{(X_f^{'}, Y_f)\}$

Initialization: $\mathcal{W}_{adv} = \{\}$;
Selected words: $S = \{\}$;
**for** each $(X_f, Y_f) \in \mathcal{D}_f$ **do**
  Create a set $Lex$ of all words $w_i$ in $X_f$;
  Filter out qualifiers, modal verbs and personal pronouns in $Lex$;
  Divide $Lex$ into intention-related word set $Lex_1$ and context-related word set $Lex_2$;
  $S \leftarrow Lex_1$;
  $w_k = randomchoice(Lex_2)$; Add $w_k$ into set $S$;

  Initialization: $X_{adv} \leftarrow X_f$;
  **for** each word $w_i$ in $S$ **do**
    Initiate the set of candidates $Candidate$ by seeking $w_i$ in $LEX_{noise}$;
    $c_k$ = randomchoice($Candidate$)
    $X_{adv} \leftarrow$ Replace $w_i$ with $c_k$ in $X_{adv}$;
  **end for**
  Add $(X_{adv}, Y_f)$ into set $\mathcal{W}_{adv}$;
**end for**
**return** $\mathcal{W}_{adv}$;

Step 1: Important Phrase Selection. For a user utterance, we extract all the N-grams in the utterance and keep those N-grams with corresponding noise phrases in the phrase-level noises. Then we give priority to the N-grams, which contain the intention-related words. If none of the kept N-grams contain the intention-base words, we randomly select one to replace.

Step 2: Candidate Phrase Replacement. For the given phrase $m$ that needs to be replaced, we lookup the phrase-level noises to obtain a set of candidates $\{m_1^{'}, m_2^{'}, \ldots, m_l^{'}\}$. Then we randomly select a phrase from the candidates for a replacement to construct phrase-level adversarial examples.

*(3) Sentence-Level Adversarial examples.*

Besides the word-level and phrase-level transformation, we can also transform the utterance in sentence-level. In human conversations, there exist various synonymous expressions with different sentence structures. Here, we use the back-translation technique [53] to generate conversational data containing translation noise. We use the MT system $MT_{f \to e}$ to translate user utterances from language $f$ to $e$ and then use reverse MT system $MT_{e \to f}$ to translate the generated translations from $e$ back to the original language $f$. Through this way, we can generate adversarial examples with different expressions while conveying similar information. More importantly, the sentence-level adversarial examples directly contain the expression preferences of the MT system.

### 5.3 Adversarial Strategies

Given the original dialogue training data and the generated multi-granularity adversarial examples, to make the dialogue model robust to the noises introduced by MT in user utterances, a natural way is to expose it to the same pattern of noises during training. To achieve this, we present

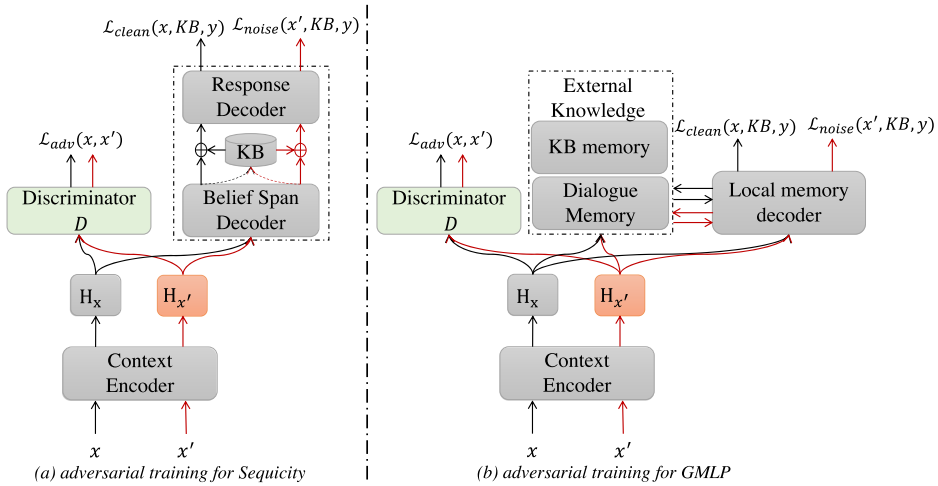


Fig. 4. The architecture of a dialogue system with utterance-level adversarial training. The dark solid arrow lines represent the forward-pass information flow for the clean input dialogue history $\mathbf{x}$, while the red dashed arrow lines for the noise dialogue history $\mathbf{x}'$.

two adversarial strategies: (a) Utterance-level adversarial learning and (b) Knowledge-level guided learning. The former strategy aims at learning a universal context encoder that makes the hidden representations of the clean dialogue context and adversarial examples as similar as possible so that the encoder can output robust hidden representations independent of noises. Compared to utterance-level adversarial learning, the latter strategy explicitly adds the multi-granularity noises into the KB, making the dialogue system learning the correlations between the noise tokens and their possible correct forms.

*5.3.1 Utterance-level adversarial learning.* To make the dialogue model robust to the utterance generated by the MT system, our basic idea is to maintain the consistency of representations through the context encoder of the dialogue model for the original input $\mathbf{x}$ and its perturbed input $\mathbf{x}'$. When fed with a perturbed input $\mathbf{x}'$, the sequence of representations will be disturbed. Thus, we aim at learning a perturbation-invariant encoder that can produce similar representations for $\mathbf{x}$ and $\mathbf{x}'$ so that the decoder can output correct response when the input is $\mathbf{x}'$.

Figure 4 illustrates the architecture of our approach. Given a clean dialogue history $\mathbf{x}$ and its perturbed input $\mathbf{x}'$, we hope that the encoded representation $\mathbf{H}_{\mathbf{x}'}$ and $\mathbf{H}_{\mathbf{x}}$ are very close so that the decoder is able to generate the robust response $\mathbf{y}$ given $\mathbf{x}'$. To this end, in this article, we introduce another two objectives besides the original objective to improve the robustness of the encoder:

- $\mathcal{L}_{\text{adv}}(\mathbf{x}, \mathbf{x}')$**:** This objective is used to encourage the encoder to output similar representations $\mathbf{H}_{\mathbf{x}}$ and $\mathbf{H}_{\mathbf{x}'}$ for clean dialogue history $\mathbf{x}$ and its perturbed dialogue history $\mathbf{x}'$ to achieve an invariant encoder. This will benefit the external knowledge selection and the decoder generation.
- $\mathcal{L}_{\text{noise}}(\mathbf{x}', \mathbf{KB}, \mathbf{y})$**:** This objective is used to guide the decoder to generate response $\mathbf{y}$ given the noise dialogue history $\mathbf{x}'$ and the knowledge base.
- $\mathcal{L}_{\text{clean}}(\mathbf{x}, \mathbf{KB}, \mathbf{y})$**:** This is the original training objective, which is used to guide the decoder to generate response $\mathbf{y}$ given a clean dialogue history $\mathbf{x}$ and knowledge base. It can guarantee the performance of the dialogue system while keeping the stability of the dialogue model.

Formally, given the clean dialogue training data and the multi-granularity adversarial examples, the adversarial training objective is given in Equation (8),

$$\mathcal{L} = \mathcal{L}_{\text{clean}}(\mathbf{x}, \mathbf{KB}, \mathbf{y}) + \mathcal{L}_{\text{noise}}(\mathbf{x}', \mathbf{KB}, \mathbf{y}) + \alpha * \mathcal{L}_{\text{adv}}(\mathbf{x}, \mathbf{x}'), \tag{8}$$

where $\alpha$ is an adversarial learning hyper-parameter. $\mathbf{x}$ and $\mathbf{x}'$ are clean dialogue history and corresponding adversarial dialogue history. $\mathbf{y}$ is the ground-truth response.

Our goal of the perturbation-invariant encoder is to make the representations produced by the encoder indistinguishable when fed with a clean dialogue history $\mathbf{x}$ and its perturbed noise dialogue history $\mathbf{x}'$. We adopt the adversarial learning framework [15] to solve the problem. The encoder serves as the generator $G$ that generates a sequence of hidden representations $\mathbf{H_x}$ given a dialogue history $\mathbf{x}$. We adopt another discriminator $D$ to distinguish $\mathbf{H_x}$ and $\mathbf{H_{x'}}$. The generator $G$, which refers to the encoder here, is to generate similar representations for $\mathbf{x}$ and $\mathbf{x}'$ that can fool the discriminator, while the discriminator $D$ tries to distinguish the two representations. $G$ and $D$ are optimized using a min-max loss function, and the loss function is formalized as follows:

$$\mathcal{L}_{\text{adv}}(\mathbf{x}, \mathbf{x}'; G, D) = E_{x\sim S}[-\log D(G(x))] + E_{x'\sim N(x)}[-\log(1 - D(G(x')))]. \tag{9}$$

We adopt the Multi-Layer Perception as the discriminator. The discriminator outputs a classification score given an input representation and tries to maximize $D(G(x))$ and minimize $D(G(x'))$. The objective encourages the context encoder to output similar representations for the clean dialogue history $x$ and noise dialogue history $x'$ so that the discriminator fails to distinguish them.

*5.3.2 Knowledge-level guided learning.* In the above utterance-level adversarial learning method, we improve the robustness by encouraging the context encoder to output similar representations for the clean data and noise data. Different from the above method, here we improve the robustness from the perspective of knowledge. We explicitly add the multi-granularity noises into the KB and incorporate the noise KB to improve the performance of the dialogue system.

In Section 5.1, we have constructed multi-granularity noises at the word and phrase level. We integrate the word-level and phrase-level noises to enhance the robustness of the cross-lingual dialogue system. The overall framework is depicted in Figure 5. This method's main idea is to use the multi-granularity noises to impact the encoder and decoder of the dialogue module. As shown in Figure 5, the integration process can be divided into the following two steps:

*Step 1: Noise KB Construction.* For each record in the KB, we randomly select one column and replace the value with the word or phrase in the multi-granularity noises. For example, in Figure 5, for the record "*Regent Street City Centre; Centre; Italian; Cheap; Pizza hut city centre*," the fourth column ("*PriceRange*") is selected, and we replace *Cheap* with *Economical*. Thus, we can inject the multi-granularity noises into the original clean KB and construct noise KB. We denote the original clean KB and the constructed noise KB together as the noise KB $KB^{'}$.

*Step 2: Noise KB Integration.* After obtaining the noise KB, we use the multi-granularity adversarial examples and the noise KB to retrain the dialogue system and adapt the dialogue system to be more robust to inputs containing MT noises. The optimization objective function is as follows:

$$\mathcal{L} = \mathcal{L}_{\text{clean}}(\mathbf{x}, \mathbf{KB}^{'}, \mathbf{y}) + \mathcal{L}_{\text{noise}}(\mathbf{x}', \mathbf{KB}^{'}, \mathbf{y}). \tag{10}$$

## 6 EXPERIMENTAL SETUP

### 6.1 Datasets

We conduct our experiments on two task-oriented dialogue datasets: CamRest676 [45] and SMD [13], which are collected through crowd-sourcing on the Amazon Mechanical Turk platform.

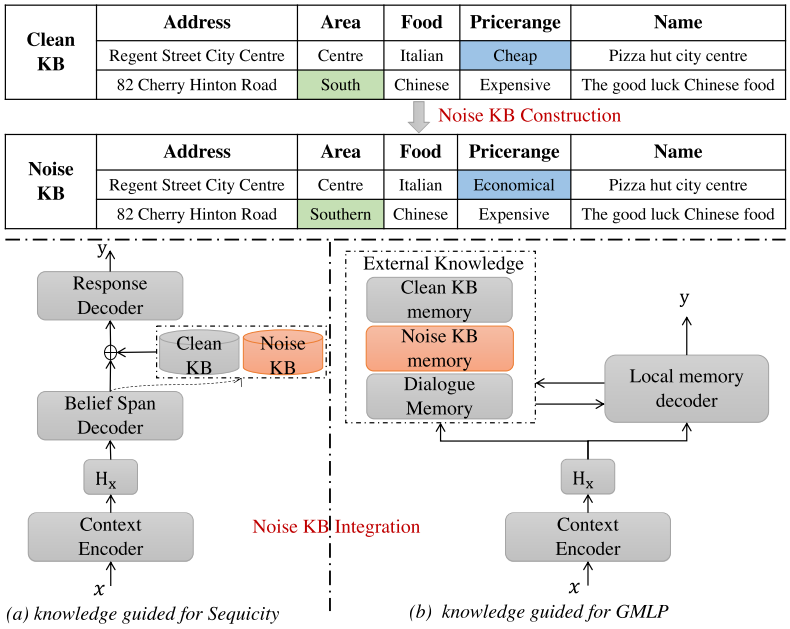


Fig. 5. The Framework of knowledge-level guided learning.

Table 1. Statistics of Datasets

| Dataset | Train | Valid | Test | Domain |
|---|---|---|---|---|
| CamRest676 | 408 | 136 | 136 | Restaurant reservations |
| SMD | 2425 | 302 | 302 | calendar scheduling<br>weather information retrieval<br>point-of-interest navigation |

CamRest676[2] is a single domain dialogue dataset about restaurant reservations. The dataset has 676 dialogs split into training, validation, and test set by the ratio of 3:1:1. SMD[3] is a multi-domain dialogue dataset with three distinct domains: calendar scheduling, weather information retrieval, and point-of-interest navigation. Data statistics are given in Table 1.

To conduct the cross-lingual experiment, we manually translate the two test sets into Chinese and German. For each language, one language expert is recruited to translate the test set. We also need machine translators corresponding to the above two language pairs to conduct experiments, including Chinese-to-English and German-to-English bi-directional MT systems.

We report results of two test scenarios: (1) Cross-lingual Test, which evaluates whether our proposed method makes the dialogue system more robust to the noise input translated by the MT system, and (2) Mono-lingual Test, which evaluates whether our proposed method makes the dialogue system perform comparable or better on the original clean inputs.

[2]The original CamRest676 dataset can be found in https://www.repository.cam.ac.uk/handle/1810/260970.
[3]https://nlp.stanford.edu/blog/a-new-multi-turn-multi-domain-task-oriented-dialogue-dataset/.

### 6.2 Dialogue Models and Evaluation Metrics

To ascertain the effectiveness and applicability of our method, we adopt the following representative task-oriented dialogue models as the monolingual dialogue systems shown in Figure 1 and implement the proposed MT-oriented noise enhanced framework on these models:

- GLMP [47]: The framework adopts the global-to-local pointer mechanism to query the knowledge base during decoding and achieves state-of-the-art performance.
- TSCP [20]: It is a two-stage CopyNet that consists of one encoder and two copy-mechanism augmented decoders. The first decoder decodes the belief state, which is used to facilitate KB retrieval. The second decoder generates a machine response based on the generated belief state and KB search result.
- LABES-S2S [49]: It is also a two-stage copy-augmented Seq2Seq model like Sequicity [20]. The difference is that in LABES-S2S, the belief states are represented as discrete latent variables and are jointly modeled with system responses given user utterances.

For adversarial examples generation, we compare our proposed multi-granularity adversarial examples generation with the other data augmentation methods [31]:

- Random Swap: Randomly swapping two adjacent words in a sentence, for example, changing "Where is the nearest gas station?" into "where is the gas nearest station?"
- Stopword Dropout: Randomly deleting stopwords in the user utterances, for example, changing "Where is the nearest gas station?" into "Where is nearest gas station?"
- Word Substitution: Replacing words with their synonyms from WordNet, for example, changing "Where is the nearest gas station?" into "where is the skinny gas station?"

For the GLMP model, we follow previous work [13, 27, 47] to evaluate our system on two automatic evaluation metrics, BLEU and entity $F_1$ score. BLEU [33] is used to measure the language quality of generated responses with golden responses guidance. For entity $F_1$, we micro-average over the entire set of dialogue system responses and compare the entities in plain text. This metric evaluates the model's ability to generate relevant entities from the underlying KBs and capture the dialogue's semantics.

For Sequicity and LABES-S2S model, which explicitly predict belief states, we use BLEU, **Entity Match rate (EMR)**, and Success $F_1$ (Succ. F1) following Lei et al. [20]. Entity match rate evaluates task completion, and it determines if a system can generate all correct constraints to search the indicated entities of the user. Success $F_1$ evaluates task completion, and it is the $F_1$ score of requested slots answered in the current dialogue. We also report a combined score (Comb) computed via (EMR + Succ. F1) $\times$ 0.5 + BLEU for overall quality measure inspired by Mehri et al. [28].

### 6.3 Experimental Settings

**Machine Translator.** We need machine translators to translate utterances from the user's language into the dialogue system's language to interact with the dialogue system. Besides, we also need the machine translators to construct the multi-granularity noises. We use a Chinese-to-English dataset containing 2.1M sentence pairs to train the Chinese-to-English and English-to-Chinese translation system. Another German-to-English dataset containing 2M sentence pairs is used to train the German-to-English and English-to-German translation system. We apply Byte-Pair encoding with 30K merge operations and maintain the source and target vocabularies to the most frequent 30K tokens. All of the systems are trained by the transformer [40] with the "base" version.

**Multi-granularity Noises.** We use an English monolingual dataset that contains about 800,000 spoken sentences to construct multi-granularity noises. We use our self-trained MT systems to

translate and back-translate the English dataset and then perform word alignment. Word pairs with bidirectional lexical translation probability greater than 0.01 are reserved as word-level noises. Phrase pairs meeting the conditions and the bidirectional phrase translation probability greater than 0.01 are kept as the phrase-level noises. The constructed multi-granularity noises are used to generate the multi-granularity adversarial examples.

**Implementation Settings.** For the base dialogue models, we use the model structures that follow the default settings in the open-source implementation of GLMP,[4] TSCP,[5] and LABES-S2S.[6] For *utterance-level adversarial learning*, the vocabulary size is 1,800 for CamRest676 and 2,800 for SMD. The adversarial learning hyper-parameter $\alpha$ is selected from [0.0, 0.01, 0.05, 0.1, 0.2, 0.3, 0.5, 0.8, 1, 1.5] for CamRest676 and SMD, respectively. For *knowledge-level guided learning*, the vocabulary size is 2000 for CamRest676 and 3000 for SMD. We use the Adam optimizer to train the models, with a learning rate of 0.003 and a decay parameter of 0.5.

## 7 RESULTS AND ANALYSIS

### 7.1 Cross-lingual Experimental Results

The experimental results using TSCP, LABES-S2S, and GLMP on SMD and CamRest676 are shown in Table 2, Table 3, and Table 4, respectively. Both Chinese to English (CH→EN) and German to English (DE→EN) cross-lingual experimental results are reported. The results are grouped into two columns according to the test set (Cross-lingual Test/Mon-lingual Test).

We observe some common conclusions supported by the experimental results on the three dialogue models. First, the noises introduced in the MT process greatly influence the performance of the dialogue systems. Comparing the results of Cross-lingual Test and Mono-lingual Test on original dialogue models, the dialogue systems' performance drops down sharply when given the utterances translated by the MT systems. In the CH→EN cross-lingual experiment, TSCP and LABES-S2S drops from 1.0020 to 0.7385, 1.0422 to 0.6864 on SMD datasets, respectively (the first line and the fourth line in Table 2). In the DE→EN cross-lingual experiment, the performance of the three dialogue models also falls heavily. This demonstrates that the noises in the MT process do seriously affect the dialogue systems, and the robustness of the dialogue systems is poor.

Second, our proposed two strategies significantly improve the system's performance in Cross-lingual Test. Compared with the original dialogue model, whether in CH→EN cross-lingual test or DE→EN cross-lingual test, both the utterance-level adversarial learning and knowledge-level guided learning can significantly improve the performance. Our proposed methods bring substantial improvements for all the three dialogue architectures regarding all the evaluation metrics. Such improvements are consistent across both dialogue datasets and two language pairs, affirming the superiority and general applicability of our proposed methods. We can also find that utterance-level adversarial learning outperforms the knowledge-level guided learning in most cases. This may be because we incorporate the multi-granularity adversarial examples by adversarial learning to learn a robust context encoder in the utterance-level adversarial learning. In contrast, in the knowledge-level guided learning, we mix the multi-granularity adversarial examples directly into the clean training data. The improvement of utterance-level adversarial learning over knowledge-level guided learning suggests that the adversarial learning method provides a more robust way of utilizing the additional information in the multi-granularity adversarial examples. The proposed knowledge-level guided learning outperforms the original dialogue model since the noise KB provides strong clues for the MT translated noise utterance during the decoding process.

[4] https://github.com/jasonwu0731/GLMP.
[5] https://github.com/WING-NUS/sequicity.
[6] https://github.com/thu-spmi/LABES.

Table 2. Cross-Lingual Experimental Results on SMD Dataset

| Lang | Model | Cross-lingual Test | | | | Mono-lingual Test | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | **BLEU** | **EMR** | **Succ. F1** | **Comb** | **BLEU** | **EMR** | **Succ. F1** | **Comb** |
| CH ↓ EN | TSCP | 0.1744 | 0.4218 | 0.7063 | 0.7385 | 0.2081 | **0.7927** | 0.7951 | 1.0020 |
| | w/ Utt-adv | **0.1951**$^{*}$ | **0.4800**$^{\dagger}$ | **0.7986**$^{\ddagger}$ | **0.8344** | 0.2061 | 0.7855 | 0.7977 | 0.9976 |
| | w/ Kno-gui | 0.1913$^{*}$ | 0.4657$^{*}$ | 0.7951$^{\ddagger}$ | 0.8217 | **0.2097** | 0.7545 | **0.8304** | **1.0022** |
| | LABES-S2S | 0.1735 | 0.3835 | 0.6423 | 0.6864 | **0.2328** | **0.8393** | **0.7794** | **1.0422** |
| | w/ Utt-adv | **0.2029**$^{\dagger}$ | **0.4844**$^{\dagger}$ | **0.7234**$^{\ddagger}$ | **0.8068** | 0.2250 | 0.7747 | 0.7791 | 1.0019 |
| | w/ Kno-gui | 0.1856 | 0.4664$^{\dagger}$ | 0.6942$^{\ddagger}$ | 0.7659 | 0.2069 | 0.7582 | 0.7632 | 0.9676 |
| DE ↓ EN | TSCP | 0.1612 | 0.3964 | 0.7200 | 0.7194 | **0.2081** | 0.7927 | 0.7951 | 1.0020 |
| | w/ Utt-adv | 0.1729 | **0.5636**$^{\ddagger}$ | 0.7970$^{\ddagger}$ | 0.8532 | 0.1887 | **0.8109** | 0.8092 | 0.9987 |
| | w/ Kno-gui | **0.1862** | 0.5616$^{\ddagger}$ | **0.8005**$^{\ddagger}$ | **0.8672** | 0.2078 | 0.7899 | **0.8134** | **1.0094** |
| | LABES-S2S | 0.1958 | 0.3761 | 0.6649 | 0.7163 | 0.2328 | **0.8393** | 0.7794 | **1.0422** |
| | w/ Utt-adv | 0.2239$^{\ddagger}$ | **0.5021**$^{\ddagger}$ | **0.7635**$^{\ddagger}$ | **0.8567** | 0.2269 | 0.7802 | **0.7847** | 1.0093 |
| | w/ Kno-gui | **0.2257**$^{\ddagger}$ | 0.4774$^{\ddagger}$ | 0.7521$^{\ddagger}$ | 0.8405 | **0.2418** | 0.7527 | 0.7722 | 1.0043 |

*w/ Utt-adv* denotes that the model is trained using our proposed *utterance-level adversarial learning* method with the multi-granularity adversarial examples. *Kno-gui* denotes that the model is trained using our proposed *knowledge-level guided learning* method. The metrics Entity Match rate, Success $F_1$, and Combined Score are abbreviated as EMR, Succ. F1, and Comb. The best results in each group are highlighted in bold. The significant test is conducted under Cross-lingual Test. $*$ ($\dagger$, $\ddagger$) indicates that the improvement over the corresponding baseline is statistically significant where $p < 0.05$ (0.01, 0.001).

Table 3. Cross-Lingual Experimental Results on the CamRest676 Dataset

| Lang | Model | Cross-lingual Test | | | | Mono-lingual Test | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | **BLEU** | **EMR** | **Succ. F1** | **Comb** | **BLEU** | **EMR** | **Succ. F1** | **Comb** |
| CH ↓ EN | TSCP | 0.1731 | 0.4776 | 0.6485 | 0.7362 | 0.2001 | **0.9328** | 0.8204 | **1.0767** |
| | w/ Utt-adv | **0.2182** | **0.4776** | **0.7858** | **0.8499** | 0.2223 | 0.8731 | 0.7920 | 1.0549 |
| | w/ Kno-gui | 0.1974 | 0.4254 | 0.8092 | 0.8147 | **0.2085** | 0.8433 | **0.8398** | 1.0501 |
| | LABES-S2S | 0.1764 | 0.7273 | 0.7154 | 0.8978 | 0.2132 | **0.9727** | 0.8205 | 1.1098 |
| | w/ Utt-adv | **0.2340** | **0.7963** | **0.7943** | **1.0293** | **0.2561** | 0.9074 | **0.8438** | **1.1317** |
| | w/ Kno-gui | 0.2227 | 0.7182 | 0.7912 | 0.9774 | 0.2299 | 0.9000 | 0.8246 | 1.0922 |
| DE ↓ EN | TSCP | 0.1695 | **0.5299** | 0.6777 | 0.7732 | 0.2001 | **0.9328** | **0.8204** | **1.0767** |
| | w/ Utt-adv | **0.2210** | 0.4552 | 0.7714 | **0.8344** | 0.2217 | 0.8955 | 0.7748 | 1.0569 |
| | w/ Kno-gui | 0.2196 | 0.4478 | **0.7754** | 0.8312 | **0.2293** | 0.8731 | 0.7938 | 1.0627 |
| | LABES-S2S | 0.1941 | 0.7545 | 0.7631 | 0.9529 | 0.2132 | **0.9727** | **0.8205** | **1.1098** |
| | w/ Utt-adv | **0.2446** | **0.7963** | **0.8524** | **1.0690** | 0.2356 | 0.9074 | 0.8049 | 1.0918 |
| | w/ Kno-gui | 0.2376 | 0.6818 | 0.8005 | 0.9788 | **0.2600** | 0.8727 | 0.8139 | 1.1033 |

The best results in each group are highlighted in bold.

Third, our proposed methods can not only improve the performance of the dialogue system under the cross-lingual scenario but also can maintain a comparable performance for the source language. In some cases, our methods can improve the dialogue system's performance in dealing with the source language. For example, On SMD, the performance of the dialogue system can be improved from 1.0020 to 1.0094 (Mono-lingual Test results reported in line seven and nine of Table 2). On CamRest676, the Mono-lingual Test performance can be improved from 1.1098 to 1.1317 (line four and five in Table 3).

Table 4. Cross-lingual Experimental Results on SMD Dataset Using GLMP Model

| | CH → EN | | | | DE → EN | | | |
|---|---|---|---|---|---|---|---|---|
| | Cross-lingual Test | | Mono-lingual Test | | Cross-lingual Test | | Mono-lingual Test | |
| Model | **BLEU** | **Ent. F1** | **BLEU** | **Ent. F1** | **BLEU** | **Ent. F1** | **BLEU** | **Ent. F1** |
| GLMP | 0.0831 | 0.3998 | **0.1339** | **0.5408** | 0.0921 | 0.3982 | **0.1339** | **0.5408** |
| w/ Utt-adv | **0.1009** | **0.4452** | 0.1222 | 0.5357 | 0.1234 | **0.4583** | 0.1216 | 0.5318 |
| w/ Kno-gui | 0.0915 | 0.4258 | 0.1343 | 0.5255 | **0.1248** | 0.4246 | 0.1197 | 0.5000 |

The metric Ent. F1 denotes entity $F_1$ score. The best results in each group are highlighted in bold.

Table 5. Chinese to English Cross-Lingual Experimental Results Using Different Adversarial Examples

| Dataset | Adversarial Examples | Cross-lingual Test | | | | Mono-lingual Test | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | **BLEU** | **EMR** | **Succ. F1** | **Comb** | **BLEU** | **EMR** | **Succ. F1** | **Comb** |
| Cam | TSCP | 0.1731 | 0.4776 | 0.6485 | 0.7362 | 0.2001 | **0.9328** | **0.8204** | 1.0767 |
| | Swap | 0.1759 | 0.4851 | 0.6599 | 0.7484 | 0.2159 | 0.9104 | 0.7639 | 1.0530 |
| | Stopword | 0.1692 | 0.5000 | 0.6347 | 0.7365 | **0.2300** | 0.9179 | 0.7803 | **1.0791** |
| | Word Sub | 0.1805 | 0.4403 | 0.7051 | 0.7532 | 0.2159 | 0.9030 | 0.7824 | 1.0586 |
| | Word-Level | **0.1987** | 0.4104 | **0.7805** | 0.7942 | 0.2046 | 0.8507 | 0.8108 | 1.0353 |
| | Phrase-Level | 0.1869 | 0.4925 | 0.7231 | 0.7947 | 0.2020 | 0.8955 | 0.7672 | 1.0334 |
| | Sent-Level | 0.1595 | **0.5821** | 0.7179 | **0.8094** | 0.1632 | 0.8582 | 0.7116 | 0.9481 |
| SMD | TSCP | 0.1744 | 0.4218 | 0.7063 | 0.7385 | 0.2081 | 0.7927 | 0.7951 | 1.0020 |
| | Swap | 0.1751 | 0.4436 | 0.7122 | 0.7531 | 0.2056 | **0.8400** | 0.8033 | **1.0273** |
| | Stopword | 0.1676 | 0.4327 | 0.7183 | 0.7431 | 0.1961 | 0.8109 | 0.8016 | 1.0023 |
| | Word Sub | 0.1680 | 0.4145 | 0.7234 | 0.7370 | 0.1944 | 0.8109 | 0.7898 | 0.9947 |
| | Word-Level | 0.1851 | **0.4655** | 0.7455 | 0.7906 | **0.2092** | 0.8000 | 0.8197 | 1.0191 |
| | Phrase-Level | **0.1909** | 0.4545 | **0.7764** | **0.8063** | 0.1973 | 0.8109 | **0.8263** | 1.0159 |
| | Sent-Level | 0.1849 | 0.4291 | 0.7598 | 0.7793 | 0.2054 | 0.7964 | 0.7875 | 0.9973 |

*Cam* denotes the dataset CamRest676. *TSCP* means using cleaning training data to train the dialogue model without adversarial examples. *Swap*, *Stopword*, and *Word Sub* denote random swap, stopword dropout, and synonym substitution. The *Word-Level*, *Phrase-Level*, and *Sent-Level* are the adversarial samples of different granularity proposed in this article. The best results in each group are highlighted in bold.

### 7.2 Experimental Results of Different Adversarial Examples

In this part, we investigate the effectiveness of our multi-granularity adversarial examples generation with the other data augmentation methods, including Random Swap, Stopword Dropout, and Word Substitute Perturbation. For a fair comparison, we adopt the same dialogue model and the same way to utilize different kinds of adversarial examples. We directly add the adversarial examples to the clean training data and retrain the dialogue system. We conduct the CH→EN cross-lingual experiments using the TSCP model. The experimental results are given in Table 5.

Not surprisingly, as shown in Table 5, our multi-granularity adversarial examples outperform the baseline methods on most metrics in the Cross-lingual Test. This implies that our proposed method effectively generates multi-granularity adversarial examples that are more related to the noises introduced by MT systems. On CamRest676, utilizing the word-level adversarial examples can achieve the best BLEU and Succ. $F_1$, while the EMR reaches the best when using the sentence-level adversarial examples. On SMD, the best BLEU and Succ. $F_1$ are obtained by using phrase-level adversarial examples, while the best EMR is achieved by using word-level adversarial examples.

We observe that the EMR metric of the word-level (0.4104) on CamRest676 is much lower compared to other methods. This is potential because the size of the training data is relatively small

Table 6. All Adversarial Examples vs. Multi-Granularity Adversarial Examples

| Dataset | Adversarial Examples | Cross-lingual Test | | | | Mono-lingual Test | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | **BLEU** | **EMR** | **Succ. F1** | **Comb** | **BLEU** | **EMR** | **Succ. F1** | **Comb** |
| Cam | TSCP | 0.1731 | **0.4776** | 0.6485 | 0.7362 | 0.2001 | **0.9328** | **0.8204** | **1.0767** |
| | mix-all | 0.2058 | 0.4478 | 0.7636 | 0.8114 | **0.2256** | 0.8657 | 0.8033 | 1.0601 |
| | multi-granularity | **0.2124** | 0.4478 | **0.7655** | **0.8191** | 0.2255 | 0.8806 | 0.7916 | 1.0616 |
| SMD | TSCP | 0.1744 | 0.4218 | 0.7063 | 0.7385 | **0.2081** | **0.7927** | 0.7951 | 1.0020 |
| | mix-all | 0.1708 | 0.4691 | 0.7629 | 0.7868 | 0.1836 | **0.7927** | 0.8051 | 0.9825 |
| | multi-granularity | **0.1877** | **0.4727** | **0.7902** | **0.8192** | 0.2068 | 0.7891 | **0.8078** | **1.0052** |

*mix-all* denotes using all the adversarial examples, including random swap, stopword dropout, word substitute, word-level, phrase-level, and sentence-level. *multi-granularity* means only using the multi-granularity adversarial examples, including word-level, phrase-level, and sentence-level. The best results in each group are highlighted in bold.

with only 408 dialogues, making the learning of perturbations at the word level is much more complicated than the phrase level and sentence level. The word-level adversarial examples are generated by substituting words in user utterances with their candidates from the word-level noises, and we give priority to the substitution of the intention-related words. Therefore, the EMR score, which evaluates task completion and determines if a system can generate all correct constraints to search the indicated entities of the user, may drop despite the BLEU and Succ. $F_1$ increases.

We also observe some diverse results on Cross-lingual Test and Mono-lingual Test. Under the cross-lingual test setting, the improvements gained by our multi-granularity adversarial examples are higher than the baseline methods. For the monolingual test, Stopword Dropout achieves best on CamRest676, and Random Swap achieves best on SMD. Even then, the combined score of word-level (1.0191) and phrase-level (1.0159) is better than that of the original TSCP model (1.0020) on SMD. Hence, our proposed multi-granularity adversarial examples generation method is more helpful in enhancing the robustness of the dialogue system under the cross-lingual scenario.

We further experiment with all adversarial examples, including random swap, stopword dropout, word substitute, word-level, phrase-level, and sentence-level, to retrain the dialogue system. The experimental results are given in Table 6. Compared to the dialogue model trained using all adversarial examples, only using our proposed multi-granularity adversarial examples achieves better performance both under Cross-lingual Test and Mono-lingual Test.

### 7.3 Ablation Study

*7.3.1 Word-level vs. Phrase-level vs. Sentence-level.* In our framework, we implement three granularity adversarial examples at word-level, phrase-level, and sentence-level. Word-level enriches the original clean training data by substituting words with their candidates from the word-level noises. Phrase-level augments the original training data through phrase-level noises. And the sentence-level paraphrases the original training data through back-translation. We evaluate their performances on CamRest676 and SMD and report results in Table 7. Table 7 shows the Chinese to English cross-lingual experimental results during utterance-level adversarial learning. All three granularity adversarial examples improve the performance over the original dialogue model, including TSCP and LABES-S2S. We observe that in most cases, the sentence-level and phrase-level augmentation performs better than word-level on most evaluation metrics. One possible reason is that sentence-level and phrase-level capture more phenomenon in the MT process, while word-level is limited to only a few words.

*7.3.2 The effect of adversarial learning.* To investigate the impact of adversarial learning, the adversarial learning hyper-parameter $\alpha$ varies from [0.0, 0.01, 0.05, 0.1, 0.2, 0.3, 0.5, 0.8, 1, 1.5].

Table 7. Ablation Test for Word-Level, Phrase-Level, and Sentence-Level Adversarial Examples during Utterance-Level Adversarial Learning

| Dataset | Adversarial Examples | Cross-lingual Test | | | | Mono-lingual Test | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | **BLEU** | **EMR** | **Succ. F1** | **Comb** | **BLEU** | **EMR** | **Succ. F1** | **Comb** |
| Cam | TSCP | 0.1731 | 0.4776 | 0.6485 | 0.7362 | 0.2001 | **0.9328** | **0.8204** | **1.0767** |
| | w/ Word-Level | **0.2192** | 0.3955 | **0.7704** | 0.8022 | **0.2369** | 0.8433 | 0.8125 | 1.0648 |
| | w/ Phrase-Level | 0.1949 | 0.5075 | 0.7295 | 0.8134 | 0.2024 | 0.8657 | 0.7805 | 1.0254 |
| | w/ Sent-Level | 0.1972 | **0.5672** | 0.7439 | **0.8527** | 0.2099 | 0.8582 | 0.7841 | 1.0310 |
| | LABES-S2S | 0.1764 | 0.7273 | 0.7154 | 0.8978 | 0.2132 | **0.9727** | 0.8205 | 1.1098 |
| | w/ Word-Level | 0.2211 | 0.7636 | 0.7990 | 1.0024 | 0.2354 | 0.9455 | 0.8109 | 1.1136 |
| | w/ Phrase-Level | 0.2209 | 0.7818 | 0.7783 | 1.0010 | 0.2430 | 0.9455 | 0.8092 | 1.1203 |
| | w/ Sent-Level | **0.2303** | **0.7909** | **0.8010** | **1.0263** | **0.2570** | **0.9727** | **0.8210** | **1.1539** |
| SMD | TSCP | 0.1744 | 0.4218 | 0.7063 | 0.7385 | 0.2081 | 0.7927 | 0.7951 | 1.0020 |
| | w/ Word-Level | 0.1851 | 0.4655 | 0.7455 | 0.7906 | **0.2092** | 0.8000 | 0.8197 | 1.0191 |
| | w/ Phrase-Level | 0.1791 | **0.4945** | **0.7795** | 0.8161 | 0.1999 | 0.8364 | **0.8211** | **1.0287** |
| | w/ Sent-Level | **0.1927** | 0.4909 | 0.7582 | **0.8172** | 0.1931 | **0.8400** | 0.7767 | 1.0014 |
| | LABES-S2S | 0.1735 | 0.3835 | 0.7063 | 0.6423 | 0.2328 | **0.8393** | 0.7794 | **1.0422** |
| | w/ Word-Level | 0.1912 | 0.4756 | 0.7224 | 0.7902 | 0.2287 | 0.8132 | 0.7688 | 1.0197 |
| | w/ Phrase-Level | **0.2092** | **0.4889** | 0.7152 | **0.8113** | 0.2285 | 0.7857 | 0.7708 | 1.0068 |
| | w/ Sent-Level | 0.2050 | 0.4756 | **0.7242** | 0.8049 | **0.2409** | 0.7967 | **0.7804** | 1.0295 |

The best results in each group are highlighted in bold.

The results are shown in Figure 6, in which Figure 6(a) to (d) are the results using TSCP and Figure 6(e) to (h) is the result using LABES-S2S model. Each sub-figure in Figure 6 shows five lines representing the original dialogue model, adversarial learning with different adversarial learning weights using word-level, phrase-level, sentence-level, and multi-granularity adversarial examples, respectively.

We can find that: (1) Our proposed utterance-level adversarial learning outperforms the baseline. Especially for the LABES-S2S model, the adversarial learning outperforms the baseline by a big margin on all adversarial weights. (2) Training with multi-granularity adversarial examples can obtain the best performance in most cases except for the sub-figure (d). This implies that the adversarial examples in different granularities can reinforce each other and provide more diverse MT noises information. (3) The adversarial weight is more sensitive on the CamRest676. This is because the training size of the CamRest676 is relatively small, and the adversarial weight has a great influence on the experimental results.

We also observe that the curve trends are different between different language pairs. For example, in the first column of Figure 6, the trend of different granularity is different. For Chinese to English experimental results on CamRest676 (Figure 6(a) and (e)), the sentence-level outperforms the word-level and phrase-level in most cases, while the sentence-level underperforms the other two for German to English (Figure 6(b) and (f)). We guess that the sentence-level adversarial examples for German to English are generated by cycle translation using the German to English MT system, and this process introduces more noises than Chinese to English. Conversely, since we use the word alignment and some pre-defined rules, the word-level and phrase-level noises for German to English can filter out many noises. In any event, adversarial learning with multi-granularity adversarial examples can obtain relatively stable and excellent performance.

*7.3.3 Utterance-level adversarial vs. Knowledge-level guide.* To integrate the multi-granularity adversarial examples and make the dialogue model robust to the noises introduced by the MT

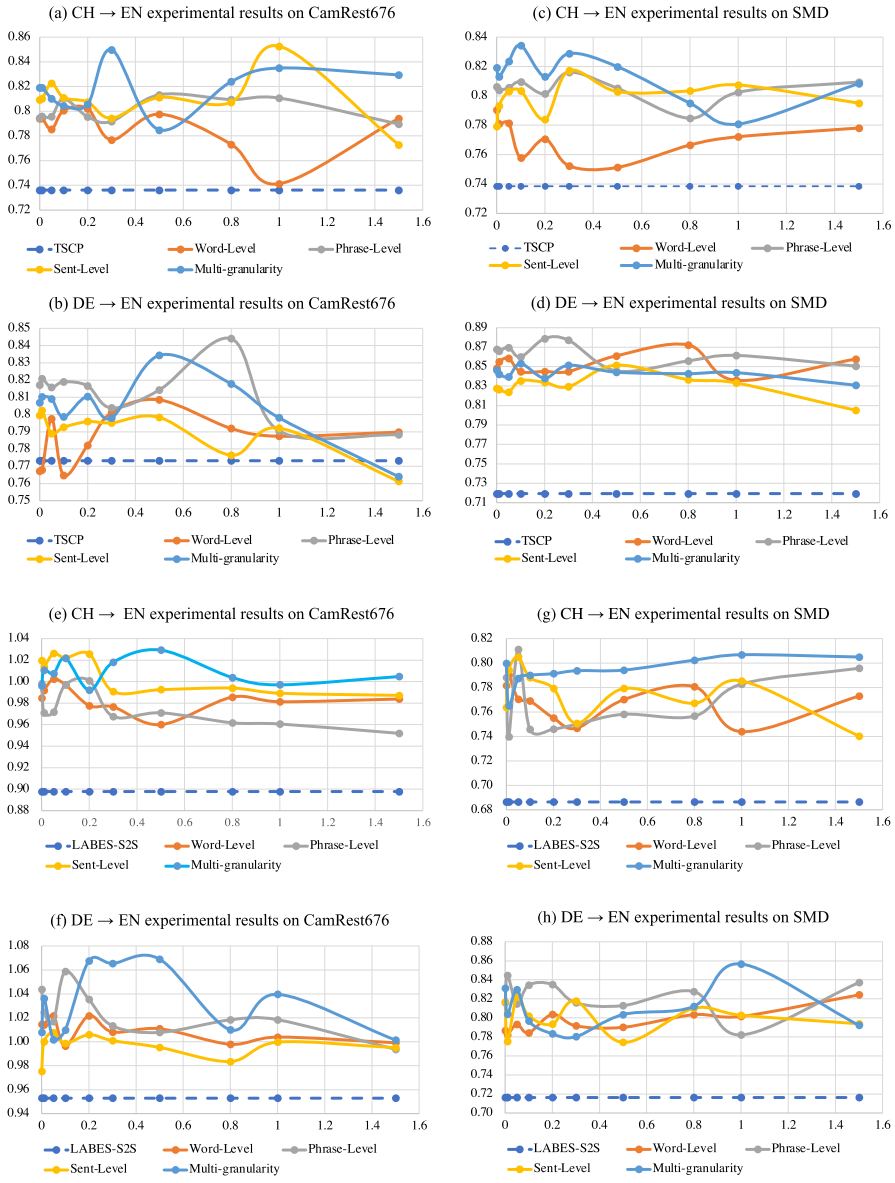


Fig. 6. Comparison of the training with different adversarial learning hyper-parameter $\alpha$. *Multi-granularity* denotes all the three granularity adversarial examples, including word-level, phrase-level, and sentence-level, are used during the training.

process, we propose two adversarial strategies. Utterance-level adversarial learning adopts adversarial learning to learn a perturbation-invariant encoder, which can output robust hidden representations independent of noises. In contrast, knowledge-level guided learning explicitly adds the multi-granularity noises into the KB, establishing connections between the noise tokens and their possible correct forms. The experimental results above demonstrate that both methods can improve system performance. These two methods improve system performance from different

Table 8. Experimental Results of Combining Utterance-Level Adversarial Learning and Knowledge-Level Guided Learning

| Dataset | Method | CH → EN | | | | DE → EN | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | **BLEU** | **EMR** | **Succ. F1** | **Comb** | **BLEU** | **EMR** | **Succ. F1** | **Comb** |
| SMD | LABES-S2S | 0.1735 | 0.3835 | 0.6423 | 0.6864 | 0.1958 | 0.3761 | 0.6649 | 0.7163 |
| | w/ Utt-adv | 0.2029 | 0.4844 | 0.7234 | 0.8068 | 0.2239 | 0.5021 | 0.7635 | 0.8567 |
| | w/ Kno-gui | 0.1856 | 0.4664 | 0.6942 | 0.7659 | 0.2257 | 0.4774 | 0.7521 | 0.8405 |
| | Combined | **0.2046** | **0.5112** | **0.7422** | **0.8313** | **0.2293** | **0.5144** | **0.7766** | **0.8748** |
| Cam | LABES-S2S | 0.1764 | 0.7273 | 0.7154 | 0.8978 | 0.1941 | 0.7545 | 0.7631 | 0.9529 |
| | w/ Utt-adv | 0.2340 | **0.7963** | 0.7943 | 1.0293 | **0.2446** | 0.7963 | **0.8524** | 1.0690 |
| | w/ Kno-gui | 0.2227 | 0.7182 | 0.7912 | 0.9774 | 0.2376 | 0.6818 | 0.8005 | 0.9788 |
| | Combined | **0.2437** | 0.7593 | **0.8281** | **1.0374** | 0.2279 | **0.8519** | 0.8441 | **1.0759** |

*Combined* denoted the results of combining the two methods.

perspectives. Hence, we want to explore whether combining the two methods can further improve the dialogue system's performance.

We conduct experiments using LABES-S2S on Chinese to English and German to English. The experimental results of Cross-lingual Test are shown in Table 8. As presented in Table 8, the combination of the two methods further boosts the dialogue system's performance, which affirms the effectiveness and robustness of the proposed approach.

### 7.4 Case Study

The above results show that our proposed methods can improve the performance of the task-oriented dialogue system when feeding the translated utterance. To better verify our methods, we conduct several case studies with the LABES-S2S model to illustrate the response generation quality.

Figure 7 compares the belief span (bspan) and system response generated by the original LABES-S2S model to those generated by the dialogue model, which is trained using the combination of our proposed utterance-level adversarial learning and knowledge-level guided learning. The first two are the examples selected from SMD, and the last one is the example from CamRest676. The red marks the difference between the original utterance and the MT result. From these three cases, we observe that when the clean utterance is given to the dialogue model, the system generates correct bspan, which is quite important since it determines whether the system can search the user's indicated entities. However, the original LABES-S2S fails to generate the right bspan after feeding the MT result. This also demonstrates that the end-to-end dialogue system is very sensitive to the perturbations in the input utterances. For example, in the first case, the MT system translates "最近的停车场 (the closest parking garage)" into "latest car park." Since this expression does not appear in the training data, the original LABES-S2S cannot handle it. Our method generates multi-granularity adversarial examples that contain different granularities of MT noises and integrates the noises into the dialogue model. Hence, our method can handle the utterance translated by the MT system and generate the right bspan.

What is more, the responses generated by our proposed method are even better than the original LABES-S2S to some extent. In the second example, the response generated by the original LABES-S2S is wrong. Although the response generated by our method is not perfect, it is much better than that of the original LABES-S2S. In the last example, our method generates an appropriate question "would you like a certain price range?" to ask user for further information about the recommended restaurant. These examples suggest that our proposed method can effectively enhance the robustness of the dialogue system toward the cross-lingual scenario.

| | | |
|---|---|---|
| **User:** 最近的停车场怎么走？<br>**Clean:** what are the direction to the closest parking garage ?<br>**MT:** how can the latest car park go ?<br>**Gold bspan:** navigate-poi_type : parking garage<br>**Gold Response:** the closest parking garage is [value_poi] , located [value_distance] away at [value_address] | **LABES-S2S (clean input)** | Generated bspan: *navigate-poi_type : parking garage*<br>Generated response: *the closest parking garage is [value_poi] at [value_address]* |
| | **LABES-S2S** | Generated bspan: *{}*<br>Generated response: *the [value_poi] is [value_distance] away .* |
| | **Combined** | Generated bspan: *navigate-poi_type : parking garage*<br>Generated response: *the nearest parking garage is [value_poi] at [value_address]* |
| **User:** 山景城星期五的最高气温是多少？<br>**Clean:** what will be the highest temperature on friday in mountain view ?<br>**MT:** what is the highest temperature in the mountain city on friday ?<br>**Gold bspan:** weather-date : friday , weather-location : mountain view , weather-weather_attribute : highest temperature<br>**Gold Response:** on friday , the highest temperature in mountain view is predicted to be [value_temperature] | **LABES-S2S (clean input)** | Generated bspan: *weather-date : friday , weather-location : mountain view , weather-weather_attribute : highest temperature*<br>Generated response: *it will not be [value_weather_attribute] in mountain view on friday .* |
| | **LABES-S2S** | Generated bspan: *weather-date : friday , weather-weather_attribute : highest temperature*<br>Generated response: *the highest temperature in mountain city is located at [value_address]* |
| | **Combined** | Generated bspan: *weather-date : friday , weather-location : mountain view , weather-weather_attribute : highest temperature*<br>Generated response: *the highest temperature on friday in mountain view is gonna be [value_weather_attribute] in mountain view .* |
| **User:** 嗨。我想在市区南部找一家好餐馆。<br>**Clean:** hi . i 'm trying to find a good restaurant in the south part of town .<br>**MT:** i 'd like to find a good restaurant in the southern part of the city .<br>**Gold bspan:** restaurant-area : south<br>**Gold Response:** there are several good restaurants in the south part of town . do you have a preference for the type of food or price range ? | **LABES-S2S (clean input)** | Generated bspan: *restaurant-area : south*<br>Generated response: *there are several restaurants in the south part of town . would you like more information ?* |
| | **LABES-S2S** | Generated bspan: *{}*<br>Generated response: *there are no restaurants serving [value_food] food . would you like to try something else ?* |
| | **Combined** | Generated bspan: *restaurant-area : south*<br>Generated response: *there are several restaurants in the south part of town . would you like a certain price range ?* |

Fig. 7. Comparison of response generation results before and after applying our proposed method. *User* denotes the user's utterance, and here the utterance is Chinese. *Clean* denotes the corresponding English utterance in the original clean test. *MT* denotes the translation result of the user's utterance using the trained Chinese-to-English MT system. *clean input* represents the input of the dialogue system is the clean utterance while the input of the other two is the MT result.

## 8 CONCLUSION

In the work, we present a novel MT-oriented noise enhanced framework to enhance the robustness of the end-to-end dialogue systems toward the cross-lingual scenario. We first design a method to automatically construct multi-granularity MT-oriented noises and multi-granularity adversarial examples containing abundant noise knowledge oriented to MT. Then, we propose two strategies to incorporate the noise knowledge: (i) Utterance-level adversarial learning and (ii) Knowledge-level guided method. These two strategies can also be used in combination. Our framework achieves significant improvements when applied to several end-to-end dialogue models on two datasets and two language pairs.

## REFERENCES

[1] Moustafa Alzantot, Yash Sharma, Ahmed Elgohary, Bo-Jhang Ho, Mani Srivastava, and Kai-Wei Chang. 2018. Generating natural language adversarial examples. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2890–2896.

[2] He Bai, Yu Zhou, Jiajun Zhang, Liang Zhao, Mei-Yuh Hwang, and Chengqing Zong. 2018. Source critical reinforcement learning for transferring spoken language understanding to a new language. In *Proceedings of the 27th International Conference on Computational Linguistics*. Association for Computational Linguistics, 3597–3607.

[3] He Bai, Yu Zhou, Jiajun Zhang, and Chengqing Zong. 2019. Memory consolidation for contextual spoken language understanding with dialogue logistic inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 5448–5453.

[4] Yonatan Belinkov and Yonatan Bisk. 2018. Synthetic and natural noise both break neural machine translation. In *Proceedings of the International Conference on Learning Representations (ICLR'18)*.

[5] Antoine Bordes and Jason Weston. 2017. Learning end-to-end goal-oriented dialog. In *Proceedings of the International Conference on Learning Representations (ICLR'17)*.
[6] Marcos Calvo, Fernando García, Lluís-F. Hurtado, Santiago Jiménez, and Emilio Sanchis. 2013. Exploiting multiple hypotheses for multilingual spoken language understanding. In *Proceedings of the 17th Conference on Computational Natural Language Learning*. Association for Computational Linguistics, 193–201.
[7] Marcos Calvo, Lluís-Felip Hurtado, Fernando Garcia, Emilio Sanchis, and Encarna Segarra. 2016. Multilingual spoken language understanding using graphs and multiple translations. *Computer Speech & Language* 38 (2016), 86–103. https://doi.org/10.1016/j.csl.2016.01.00
[8] Wenhu Chen, Jianshu Chen, Yu Su, Xin Wang, Dong Yu, Xifeng Yan, and William Yang Wang. 2018. XL-NBT: A cross-lingual neural belief tracking framework. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 414–424.
[9] Yun-Nung Chen, Dilek Hakkani-Tür, Gokhan Tur, Jianfeng Gao, and Li Deng. 2016. End-to-end memory networks with knowledge carryover for multi-turn spoken language understanding. In *Proceedings of the Conference of the International Speech Communication Association (Interspeech'16)*. 3245–3249.
[10] Yong Cheng, Lu Jiang, and Wolfgang Macherey. 2019. Robust neural machine translation with doubly adversarial inputs. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 4324–4333.
[11] Yong Cheng, Zhaopeng Tu, Fandong Meng, Junjie Zhai, and Yang Liu. 2018. Towards robust neural machine translation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, 1756–1766.
[12] Javid Ebrahimi, Anyi Rao, Daniel Lowd, and Dejing Dou. 2018. HotFlip: White-box adversarial examples for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Association for Computational Linguistics, 31–36.
[13] Mihail Eric, Lakshmi Krishnan, Francois Charette, and Christopher D. Manning. 2017. Key-value retrieval networks for task-oriented dialogue. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*. Association for Computational Linguistics, Saarbrücken,Germany, 37–49.
[14] Mihail Eric and Christopher Manning. 2017. A Copy-augmented sequence-to-sequence architecture gives good performance on task-oriented dialogue. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*. Association for Computational Linguistics, 468–473.
[15] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In *Advances in Neural Information Processing Systems 27*, Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger (Eds.). Curran Associates, Inc., 2672–2680.
[16] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. 2015. Explaining and harnessing adversarial examples. In *Proceedings of the 3rd International Conference on Learning Representations (ICLR'15)*.
[17] Xisen Jin, Wenqiang Lei, Zhaochun Ren, Hongshen Chen, Shangsong Liang, Yihong Zhao, and Dawei Yin. 2018. Explicit state tracking with semi-supervisionfor neural dialogue generation. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*. 1403–1412.
[18] Vladimir Karpukhin, Omer Levy, Jacob Eisenstein, and Marjan Ghazvininejad. 2019. Training on synthetic noise improves robustness to natural noise in machine translation. In *Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT'19)*. Association for Computational Linguistics, Hong Kong, China, 42–47.
[19] Sungjin Lee and Amanda Stent. 2016. Task Lineages: Dialog state tracking for flexible interaction. In *Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*. Association for Computational Linguistics, 11–21.
[20] Wenqiang Lei, Xisen Jin, Min-Yen Kan, Zhaochun Ren, Xiangnan He, and Dawei Yin. 2018. Sequicity: Simplifying task-oriented dialogue systems with single sequence-to-sequence architectures. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, 1437–1447.
[21] Jiwei Li, Will Monroe, Tianlin Shi, Sébastien Jean, Alan Ritter, and Dan Jurafsky. 2017. Adversarial learning for neural dialogue generation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2157–2169.
[22] Xiujun Li, Yun-Nung Chen, Lihong Li, Jianfeng Gao, and Asli Celikyilmaz. 2017. End-to-end task-completion neural dialogue systems. In *Proceedings of the 8th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Asian Federation of Natural Language Processing, Taipei, Taiwan, 733–743.
[23] Bing Liu and Ian Lane. 2018. End-to-end learning of task-oriented dialogs. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop*. Association for Computational Linguistics, 67–73.

[24] Zihan Liu, Jamin Shin, Yan Xu, Genta Indra Winata, Peng Xu, Andrea Madotto, and Pascale Fung. 2019. Zero-shot cross-lingual dialogue systems with transferable latent variables. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP'19)*. Association for Computational Linguistics, 1297–1303.

[25] Zihan Liu, Genta Indra Winata, Zhaojiang Lin, Peng Xu, and Pascale Fung. 2020. Attention-informed mixed-language training for zero-shot cross-lingual task-oriented dialogue systems. In *Proceedings of the 34th AAAI Conference on Artificial Intelligence (AAAI'20), the 32nd Innovative Applications of Artificial Intelligence Conference (IAAI'20), and the 10th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI'20)*. 8433–8440.

[26] Zihan Liu, Genta Indra Winata, Peng Xu, Zhaojiang Lin, and Pascale Fung. 2020. Cross-lingual Spoken language understanding with regularized representation alignment. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP'20)*. Association for Computational Linguistics, 7241–7251.

[27] Andrea Madotto, Chien-Sheng Wu, and Pascale Fung. 2018. Mem2Seq: Effectively incorporating knowledge bases into end-to-end task-oriented dialog systems. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, 1468–1478.

[28] Shikib Mehri, Tejas Srinivasan, and Maxine Eskenazi. 2019. Structured fusion networks for dialog. In *Proceedings of the 20th Annual SIGdial Meeting on Discourse and Dialogue*. Association for Computational Linguistics, 165–177.

[29] Gregoire Mesnil, Yann Dauphin, Kaisheng Yao, Yoshua Bengio, and Geoffrey Zweig. 2015. Using recurrent neural networks for slot filling in spoken language understanding. *IEEE/ACM Trans. Aud. Speech Lang. Process.* 23, 3 (2015), 530–539.

[30] Takeru Miyato, Andrew M. Dai, and Ian Goodfellow. 2017. Adversarial training methods for semi-supervised text classification. In *Proceedings of the International Conference on Learning Representations (ICLR'17)*.

[31] Tong Niu and Mohit Bansal. 2018. Adversarial Over-sensitivity and over-stability strategies for dialogue models. In *Proceedings of the 22nd Conference on Computational Natural Language Learning*. Association for Computational Linguistics, 486–496.

[32] Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Comput. Ling.* 29, 1 (2003), 19–51.

[33] 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 of the Association for Computational Linguistics*. Association for Computational Linguistics, 311–318.

[34] Libo Qin, Minheng Ni, Yue Zhang, and Wanxiang Che. 2020. CoSDA-ML: Multi-lingual code-switching data augmentation for zero-shot cross-lingual NLP. In *Proceedings of the 29th International Joint Conference on Artificial Intelligence (IJCAI'20)*. 3853–3860.

[35] Sebastian Schuster, Sonal Gupta, Rushin Shah, and Mike Lewis. 2019. Cross-lingual transfer learning for multilingual task oriented dialog. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, Minneapolis, Minnesota, 3795–3805.

[36] Shikhar Sharma, Jing He, Kaheer Suleman, Hannes Schulz, and Philip Bachman. 2017. Natural Language generation in dialogue using lexicalized and delexicalized data. arXiv:1606.03632. Retrieved from https://arxiv.org/abs/1606.03632.

[37] Lei Shu, Piero Molino, Mahdi Namazifar, Hu Xu, Bing Liu, Huaixiu Zheng, and Gokhan Tur. 2019. Flexibly-structured model for task-oriented dialogues. In *Proceedings of the 20th Annual SIGdial Meeting on Discourse and Dialogue*. Association for Computational Linguistics, 178–187.

[38] Pei-Hao Su, Milica Gašić, Nikola Mrkšić, Lina M. Rojas-Barahona, Stefan Ultes, David Vandyke, Tsung-Hsien Wen, and Steve Young. 2016. On-line active reward learning for policy optimisation in spoken dialogue systems. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, 2431–2441.

[39] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian J. Goodfellow, and Rob Fergus. 2014. Intriguing properties of neural networks. In *Proceedings of the International Conference on Learning Representations (ICLR'14)*.

[40] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.). Curran Associates, Inc., 5998–6008.

[41] Weikang Wang, Jiajun Zhang, Qian Li, Mei-Yuh Hwang, Chengqing Zong, and Zhifei Li. 2019. Incremental learning from scratch for task-oriented dialogue systems. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 3710–3720.

[42] Weikang Wang, Jiajun Zhang, Qian Li, Chengqing Zong, and Zhifei Li. 2019. Are you for real? detecting identity fraud via dialogue interactions. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP'19)*. Association for Computational Linguistics, 1762–1771.

[43] Weikang Wang, Jiajun Zhang, Han Zhang, Mei-Yuh Hwang, Chengqing Zong, and Zhifei Li. 2018. A teacher-student framework for maintainable dialog manager. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 3803–3812.

[44] Tsung-Hsien Wen, Milica Gašić, Nikola Mrkšić, Pei-Hao Su, David Vandyke, and Steve Young. 2015. Semantically conditioned lstm-based natural language generation for spoken dialogue systems. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 1711–1721.

[45] Tsung-Hsien Wen, David Vandyke, Nikola Mrkšić, Milica Gašić, Lina M. Rojas-Barahona, Pei-Hao Su, Stefan Ultes, and Steve Young. 2017. A network-based end-to-end trainable task-oriented dialogue system. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*. Association for Computational Linguistics, 438–449.

[46] Jason D. Williams and Steve Young. 2007. Partially observable Markov decision processes for spoken dialog systems. *Comput. Speech Lang.* 21, 2 (2007), 393–422.

[47] Chien-Sheng Wu, Richard Socher, and Caiming Xiong. 2019. Global-to-local memory pointer networks for task-oriented dialogue. In *Proceedings of the International Conference on Learning Representations (ICLR'19)*.

[48] Steve Young, Milica Gasic, Blaise Thomson, and Jason Williams. 2013. POMDP-based statistical spoken dialog systems: A review. *Proc. IEEE* 101 (05 2013), 1160–1179.

[49] Yichi Zhang, Zhijian Ou, Huixin Wang, and Junlan Feng. 2020. A probabilistic end-to-end task-oriented dialog model with latent belief states towards semi-supervised learning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics Online, 9207–9219.

[50] Yichi Zhang, Zhijian Ou, and Zhou Yu. 2020. Task-oriented dialog systems that consider multiple appropriate responses under the same context. In *Proceedings of the 34th AAAI Conference on Artificial Intelligence (AAAI'20), the 32nd Innovative Applications of Artificial Intelligence Conference (IAAI'20), and the 10th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI'20)*. AAAI Press, 9604–9611.

[51] Zhengli Zhao, Dheeru Dua, and Sameer Singh. 2018. Generating natural adversarial examples. In *Proceedings of the International Conference on Learning Representations (ICLR'18)*.

[52] Victor Zhong, Caiming Xiong, and Richard Socher. 2018. Global-locally self-attentive encoder for dialogue state tracking. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, 1458–1467.

[53] Junnan Zhu, Qian Wang, Yining Wang, Yu Zhou, Jiajun Zhang, Shaonan Wang, and Chengqing Zong. 2019. NCLS: Neural cross-lingual summarization. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP'19)*. Association for Computational Linguistics, 3054–3064.

Received November 2020; accepted March 2021