Semantics-Enhanced Task-Oriented Dialogue Translation: A Case Study on Hotel Booking

Longyue Wang[†] **Jinhua Du**[†] **Liangyou Li**[§] **Zhaopeng Tu**[‡*] **Andy Way**[†] **Qun Liu**[†]
[†]ADAPT Centre, School of Computing, Dublin City University, Ireland
{longyue.wang, jinhua.du, andy.way, qun.liu}@adaptcentre.ie
[§]Noah's Ark Lab, Huawei Technologies, China
liliangyou@huawei.com
[‡]Tencent AI Lab, China
zptu@@tencent.com

Abstract

We showcase **TODAY**, a semantics-enhanced task-oriented dialogue translation system, whose novelties are: (i) task-oriented named entity (NE) definition and a hybrid strategy for NE recognition and translation; and (ii) a novel grounded semantic method for dialogue understanding and task-order management. TODAY is a case-study demo which can efficiently and accurately assist customers and agents in different languages to reach an agreement in a dialogue for the hotel booking.

1 Introduction

Applications of machine translation (MT) in some human–human communication scenarios still exist many challenging problems due to the characteristics of spoken languages and dialogues. For example, general-purpose MT systems cannot perform efficiently and effectively on specific tasks such as hotel booking because of the low accuracy of entity recognition and translation in dialogues between customers and hotel agents as shown below:

Source:	我想定个房间，{十二月二十五号} (星期二) [三点]入住。
Reference:	I would like to reserve a room on {December the 25th}, (Tuesday) and I will check in at [three o'clock].
Google:	I'd like to have a room, [three] on (Tuesday), {February 25}.
App1:	I want to book a room, (Two) or [Three] rooms at the {December 25} week.

In this example, *App1* is a commercialised translation system for the travel domain. We found that check-in date/time and week day were not translated correctly either by *Google* or *App1*. Wrong translations of these entities will impede communication between the customer and agent.

We showcase our task-oriented semantics-enhanced dialogue machine translation (DMT) system **TODAY**[1] which alleviates these problems for the hotel booking scenario.

2 System Description

In the hotel booking scenario, customers and agents speak different languages.[2] Customers access the hotel website to request a conversation, and the agent accepts the customer's request to start the conversion. Figure 1 shows the detailed workflow of TODAY. We first recognise entities by inferring their specific types based on information such as contexts, speakers etc. (cf. Section 2.1 and 2.2). Then, the recognised entities will be represented as logical expressions or semantic templates using the grounded semantics module (cf. Section 2.2). Finally, candidate translations of semantically represented entities will be marked up and fed into a unified bi-directional translation process.

2.1 Task-Oriented Named Entity Recognition and Translation

As standard types of entities (e.g. people, organizations, locations) cannot exactly match our task-oriented entity types, we define a series of task-oriented entity types in TODAY, including {*time, number, date, currency, room type, person name, hotel name, location, payment type* }. We combine rule-based and dictionary-based methods for our NE recognition and translation. For bilingual dictionary construction, we employ the ICE toolkit.[3] ICE can guide users through a series of linguis-

[*] Work was done when Zhaopeng Tu was working at Huawei Noah's Ark Lab.

[1] The demo system can be found at http://computing.dcu.ie/~lwang/demo.html.

[2] The rest of the paper will assume that customers speak English and agents speak Chinese.

[3] Available at http://nlp.cs.nyu.edu/ice.

The Companion Volume of the IJCNLP 2017 Proceedings: System Demonstrations, pages 33–36,
Taipei, Taiwan, November 27 – December 1, 2017.

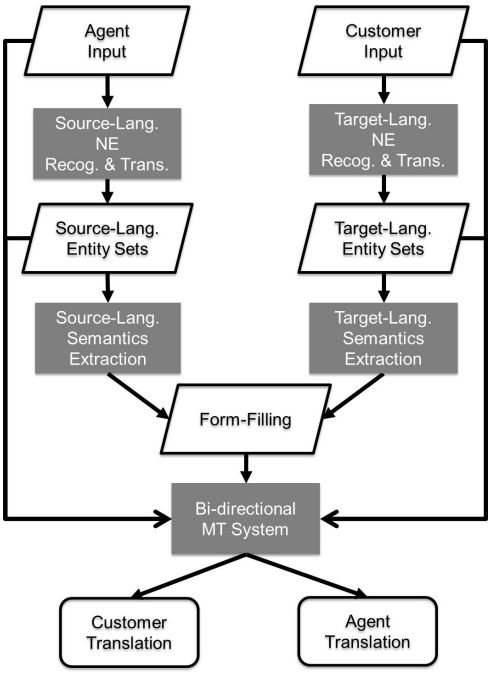

Figure 1: Workflow of TODAY.

tic processing steps, presents them with entities and dependency relations that are potential seeds, and helps them to expand the seeds by answering yes/no questions. Other types of NEs such as time, data etc. can be recognised and translated using rules because of their formulaic and common expressions.

Our hybrid strategy can generates multiple translation candidates. For example, the date have various formats in both English and Chinese. Then the recognised entities and their candidate translations are formalise by XML markup in the source-side sentence. Finally the sentence is fed into the decoder to compete with the translation model.

2.2 Grounded Semantic Representation and Form Filling

We propose a specific semantic module for our task-oriented dialogues: *grounded semantics* module which is in the form of *Feature-Value* (FV) pairs.

In TODAY, we define two types of features (Table 1): customer features and room features. While the customer features include information about the customer of the current order, room features describe the details of the order, including check-in and check-out date, room type, room price, and payment information. All these features and their values can be recognised by the NE recogniser (cf. Section 2.1) but extended by adding extra patterns. These new patterns are used to label an entity with

Type	Feature	Example value
Customer	Name	John
	Tel. NO.	1234567
Room	Check-In Date	1st April 2017
	Check-Out Date	2nd April 2017
	Room Type	Single
	Price Per. Night	300 Dollars
	Payment Type	VISA
	Card NO.	987654321

Table 1: An example of a booking order which represents the grounded semantic of a dialogue.

more detailed categories, such as determining a date to be one of the two features: *check-in date* and *check-out date*. To achieve this, patterns take contexts into consideration. For example, according to the phrase *from 1st Jan to 2nd Jan*, *1st Jan* is a check-in date while *2nd Jan* is a check-out date.

After recognising all features and their values, we need to solve conflicts when a feature appears multiple times with different values.

Customer:	I'd like to have a **single** room.
Hotel:	Sorry. I only have **double** rooms available.
Customer:	OK. A **double** room would also be fine.

In this example, the feature *room type* appears three times with values *single* and *double*. To determine which value should be chosen, we propose to score each candidate value of a feature by comparing its contexts with predefined attributes of the feature. The candidate with the highest score is then taken as the final feature value.

We define four attributes on each feature:

- *Speaker*: Its value is either *hotel* or *customer*. For example, because the room type is usually chosen by a customer, we define the speaker attribute of this feature to be "customer".
- *Position*: It defines a range of positions in dialogues that a feature should be within. For example, we define the position attribute of a customer name as [1–3], because a dialogue usually starts from self-introduction and greetings, but in other features we set the position to infinity.
- $Pattern_{cur}$: It consists of a set of patterns that usually appear in the current utterance. For example, the customer name usually follows *I am* or *my name is*.
- $Pattern_{pre}$: It consists of a set of patterns that usually appear in the previous utterance.

Given these predefined attributes on a feature, we calculate a score for each candidate value of the

feature, according to Equation (1):

$$Score(v) = S \cdot P \cdot (\lambda_c + M_c) \cdot (\lambda_p + M_p) \quad (1)$$

where $S \in \{0, 1\}$ indicates whether the current speaker equals to the speaker defined in the feature or not, $P \in \{0, 1\}$ denotes whether the position of the feature is within the range given by its position attributes or not, and M_c and M_p are the number of matched patterns in the current utterance and previous utterance, respectively. We use $\lambda_c = 1$ and $\lambda_p = 1$ as smooth factors.

3 Experiments and Analysis

3.1 Setup

From the IWSLT DIALOG corpus,we select 1,023 and 1053 hotel booking sentences (34/36 dialogues) as development set and test set, respectively. We combine our home-made travel domain corpora as in-domain training data (180K). We also use domain adaptation techniques to select in-domain data from movie subtitles (Wang et al., 2016b).

We carry out our experiments using the phrase-based SMT model in Moses (Koehn et al., 2007) on Chinese (ZH)–English (EN). Furthermore, we train a 5-gram language model using the SRI Language Toolkit (Stolcke, 2002). We run GIZA++ (Och and Ney, 2003) for alignment and use MERT (Och, 2003) to optimize the feature weights. We develop TODAY on the basis of an open-source live support application Mibew[4] by integrating our semantics-enhanced SMT system and the semantic form filling.

3.2 Evaluation of Dialogue Translation

We first evaluate the domain adaptation and NE approaches on DMT, respectively. Then, we combine these best sub-models to further improve the translation quality.

The baseline systems are trained on the in-domain corpus and the results show that an MT system trained on small-scale data can only obtain 24.20 and 17.90 BLEU points on English–Chinese and Chinese–English, respectively. Combining the models trained on the selected pseudo in-domain data can improve the performance by at most +1.09 and +1.24 on EN-ZH (top-50K) and ZH-EN (top-50K), respectively. However, bring more pseudo in-domain data ($> top - 250K$), the performance drops sharply.

[4]Available at https://mibew.org.

System	EN-ZH	△	ZH-EN	△
In-domain	24.20	-	17.90	-
1-best Entity	30.70	+6.5	20.30	+2.4
N-best Entity	31.10	+6.9	20.20	+2.3

Table 2: Performance with task-oriented NE recognition.

Task	SYS	BLEU (%)
ZH-EN	Google	10.3
	App1	10.4
	TODAY	21.5
EN-ZH	Google	16.9
	App1	15.5
	TODAY	32.7

Table 3: Overall performance.

About NE component, we employ XML markup technique to insert bilingual entities into the translation. As the entity may have multiple translations, we also explore N-best entity lists. After inserting entities into the MT system, the performance improves by +6.5 (EN-ZH) and +2.4 (ZH-EN) BLEU points as shown in Table 2. When using the N-best entity method, it can further improve the performance by +0.4 BLEU on English–Chinese.

Based on the individual performance of each component, we design our DMT: 1) build translation models on selected top-50K data and combine it with baseline; 2) integrate N-best NE models to our MT our system. In Table 3, it shows that combination further improve the translation performance. Comparing with App1 and Google Translate, our system significantly outperforms these systems by +17.2 and +11.2 BLEU points on EN-ZH and ZH-EN, respectively.

3.3 Evaluation of Task-Oriented Named Entity and Translation

We manually annotated Chinese and English sentences in the test set to evaluate the proposed task-oriented NE recognition and translation in terms of accuracy, recall and F1. In Table 4, **Recog** indicates NE recognition on the source language, and **Trans** indicates translation task. All F1 scores are over 90% in terms of recognition and translation,

Lang	Task	Acc. (%)	Rec. (%)	F1 (%)
ZH-EN	Recog	98.21	99.76	98.99
	Trans	91.86	93.33	92.59
EN-ZH	Recog	97.78	96.04	96.90
	Trans	97.24	95.52	96.37

Table 4: Results of NE recognition and translation.

Task	SYS	Trans. Acc. (%)
ZH-EN	Google	72.08
	App1	85.28
	TODAY	97.24
EN-ZH	Google	58.11
	App1	66.42
	TODAY	95.52

Table 5: Comparison on entity translation with different systems.

Feature	Precision (%)
Customer Name	97.1
Customer Tel. NO.	91.2
Check-In Date	100.0
Check-Out Date	76.5
Room Type	73.5
Price Per. Night	79.4
Payment Type	100.0
Card NO.	97.1
Average	89.4

Table 6: Performance of form-filling (EN).

which shows that the proposed fine-grained NE definitions and hybrid strategy for NE recognition and translation is effective in TODAY.

We also compared TODAY with Google and App1 as shown in Table 5. Since we cannot obtain NE recognition information from both third-part applications, we manually inspected the top-300 sentences (in test set) and only calculate accuracy of translations of entities. If the translation of an entity matches the reference, we count it as correct; otherwise, it is regarded as incorrect. It shows that TODAY significantly outperforms both Google and App1 in terms of accuracy of entity translation.

3.4 Evaluation of Grounded Semantic Extraction

We tested our grounded semantic module on the test set in terms of feature recognition and form filling. Each dialogue is manually annotated with semantic features and has an associated order form as a reference. Since our feature recognizer is a simple extension of the NE recogniser, in this section we ignore the performance of the recognizer. Table 6 shows evaluation results of the form-filling given golden feature annotations.

Five features (*customer name*, *customer Tel. NO.*, *check-in date*, *payment type*, and *Card NO.*) achieve an accuracy over 90%. The reason of such high accuracy we analyse is that there are fewer conflicts for them in a single dialogue. By contrast, on the other 3 features, (*check-out date*, *room type*, and *price per. night*) the accuracy is between 70%–80%. By inspecting the dialogues, we found that (i) the *check-out date* is not always explicitly mentioned in the dialogues; (ii) the *room type* and *price per. night* have a relatively higher repetition. These observations suggest that it is harder to solve the conflicts on these three features.

4 Conclusion and Future Work

In this paper we described TODAY, a semantics-enhanced task-oriented dialogue translation system for hotel booking scenarios and evaluated its performance. In future work, we plan to integrate neural MT into our demo system based on our advanced approaches (Wang et al., 2016a, 2017).

Acknowledgments

This work is supported by the Science Foundation of Ireland (SFI) ADAPT project (Grant No.:13/RC/2106), and partly supported by the DCU-Huawei Joint Project (Grant No.:201504032-A (DCU), YB2015090061 (Huawei)).

References

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, and et al. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th ACL: Demo and Poster Sessions*, pages 177–180.

Franz Josef Och. 2003. Minimum error rate training in statistical machine translation. In *Proceedings of the 41st ACL*, pages 160–167.

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51.

Andreas Stolcke. 2002. SRILM – an extensible language modeling toolkit. In *Proceedings of the 7th ICSLP*, pages 901–904.

Longyue Wang, Zhaopeng Tu, Andy Way, and Qun Liu. 2017. Exploiting cross-sentence context for neural machine translation. In *Proceedings of the 2017 EMNLP*, pages 2816–2821.

Longyue Wang, Zhaopeng Tu, Xiaojun Zhang, Hang Li, Andy Way, and Qun Liu. 2016a. A novel approach for dropped pronoun translation. In *Proceedings of the 2016 NAACL*, pages 983–993.

Longyue Wang, Xiaojun Zhang, Zhaopeng Tu, Andy Way, and Qun Liu. 2016b. The automatic construction of discourse corpus for dialogue translation. In *Proceedings of the 10th LREC*, pages 2748–2754.