Is one brick enough to break the wall of spoken dialogue state tracking?

Lucas Druart, Valentin Vielzeuf, Yannick Estève

IS ONE BRICK ENOUGH TO BREAK THE WALL OF SPOKEN DIALOGUE STATE TRACKING?

Lucas Druart [1,2] *Valentin Vielzeuf* [1] *Yannick Estève* [2]

[1] Orange Innovation, France [2] LIA - Avignon Université, France

[1]{lucas1.druart,valentin.vielzeuf}@orange.com,

[2]{first.last}@univ-avignon.fr

ABSTRACT

In Task-Oriented Dialogue (TOD) systems, correctly updating the system's understanding of the user's needs (*a.k.a* dialogue state tracking) is key to a smooth interaction. Traditionally, TOD systems perform this update in three steps: transcription of the user's utterance, semantic extraction of the key concepts, and contextualization with the previously identified concepts. Such cascade approaches suffer from cascading errors and separate optimization. End-to-End approaches have been proved helpful up to the semantic extraction step. This paper goes one step further paving the path towards completely neural spoken dialogue state tracking by comparing three approaches: (1) a state of the art cascade approach, (2) a locally E2E approach with rule-based contextualization and (3) a completely neural approach. Our study highlights that although they all outperform the recent DSTC11 best model, especially with a filtering post-processing step, (1) remains the most accurate approach. Indeed, both (2) and (3) have trouble propagating context as dialogues unfold showing that context propagation in completely neural approaches is an open challenge.

Index Terms— spoken dialogue systems, context adaptation, end-to-end, dialogue state tracking

1. INTRODUCTION

Digitization enables many tasks to be automated, nevertheless users sometimes require assistance to perform a specific task such as making a reservation at a restaurant or booking a hotel room. Task-Oriented Dialogue (TOD) systems are designed to assist such users. A common approach to implement them is to break the problem down to three iterative steps [1]: updating the system's understanding of the users' needs, reasoning over a database and domain knowledge to choose the next action and providing the user an answer. This paper focuses on the first step.

Traditionally the user's needs update consists of three components, respectively performing the transcription of the user's utterance, semantic extraction and contextualization of the extracted concepts [2]. Unfortunately, this method presents the inconvenience of propagating errors of a component on to the next one(s) (*i.e.* cascading errors) and of not optimizing all components on the final objective (*i.e.* separate optimization) [3]. End-to-End (E2E) approaches may address these issues by designing models in which the gradient (*i.e.* error signal) can back-propagate from the output all the way to the input [4].

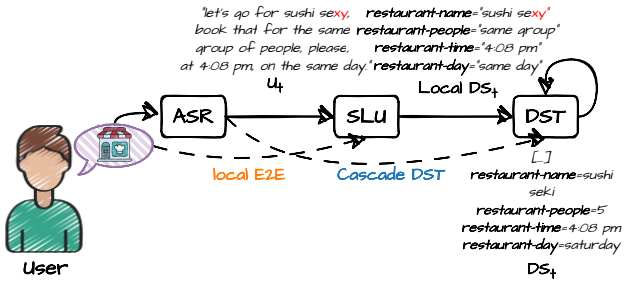

Fig. 1. Spoken Dialogue State Tracking alternatives. Red characters indicate potential cascading errors.

On the one hand, with the advent of deep-learning and textual embeddings, state of the art Dialogue State Tracking (DST) models work directly on automatic transcriptions [5]. However, such approaches require careful, dataset specific, mechanisms to catch and correct transcription errors together with data augmentation to increase the model's robustness to specific upstream errors.

On the other hand, Spoken Language Understanding (SLU) working directly on the speech signal, has successfully been applied to tasks which process utterances individually such as voice command slot filling [6] and dialogue act classification [7, 8]. Such systems often leverage transfer learning of previously trained models which is challenging because it requires a trade-off between learning new knowledge (*e.g.* domain's vocabulary, domain's ontology structure) and keeping previous capabilities (*e.g.* transcription of open vocabulary concepts) on a small amount of data [3].

In TOD such models must also adapt to the dialogue's current context. While E2E SLU models have been efficiently

designed for independent single user utterance processing [6, 9, 10], contextually dependant E2E SLU remains unexplored to the best of our knowledge. Indeed, dialogue history integration to guide the current turn's prediction (*e.g.* better spelling of technical vocabulary) has been already implemented [7, 8, 11, 12]. Yet, for the tasks described in these studies, the context was not mandatory for solving the task (*e.g.* the task does not need to process cross-turn references resolution).

Producing contextual semantic annotations is expensive because of the cognitive load required to analyse the context and adapt the annotation. Such datasets exist for chat based dialogue understanding [13, 14] but lack for spoken dialogues, explaining the gap between E2E SLU and DST. To the best of our knowledge, the vocalized version of MultiWoz [15] is the only dataset providing speech, transcriptions and contextual semantic annotations (*i.e.* representing the user's needs from the beginning of the dialog up to the current turn).

This paper lies at the intersection of both directions. We focus on contextually dependant semantic extraction, such as DST, in which the previous dialogue turns are mandatory to correctly process the current one. In fact, the spoken DST models presented in this paper output a summary of the user's needs since the beginning of the dialogue. State of the art DST systems use cascade approaches which create a textual bottleneck both in terms of data and model inference. E2E approaches do not require ground-truth transcriptions and can be jointly optimized. We pave the path towards E2E spoken DST by comparing (1) a state of the art cascade approach, (2) a locally E2E approach with rule-based contextualization and (3) a completely neural approach.

2. METHOD

2.1. Task-Oriented Dialogues

In TOD users require assistance from an agent to complete a task such as making a reservation at a restaurant or booking a hotel room. More formally, let us define a TOD as a sequence of t dialogue turns $U_1, A_2, \ldots, A_{t-1}, U_t$ where A_{t-1} and U_t respectively correspond to textual agent's turn $t-1$ and spoken user's turn t. The goal of DST is to keep up to date a condensed representation of the user's needs. In this paper, users needs are represented as Dialogue States (DS) and correspond to a list of n slot-value pairs linearized as $\texttt{slot}_1\texttt{=value}_1\texttt{;...;slot}_n\texttt{=value}_n$. At a given turn t, a TOD system is thus inputted the previous context[1] and the current user turn from which it should output the updated user needs DS_t.

[1]For $t = 1$, the context is empty.

2.2. Context Propagation

As the dialogue unfolds, the user might refer to previously mentioned entities. In order to design a contextually dependant SLU model, we need to propagate the context of the previous turns $DS_{t-2} + A_{t-1}$ to inform the prediction $\hat{DS}_t$ of the current user turn U_t. This paper compares three alternatives of contextually dependant SLU models shown in Fig. 2. For each approach we train the model(s) over 10 epochs on a single 24Gb GPU and use the last checkpoint at inference[2].

2.2.1. Cascade DST Approach

The cascade approach consists of an Automatic Speech Recognition (ASR) model which transcribes the user's turn U_t and concatenates it to the previous' turns context $DS_{t-2} + A_{t-1}$ as the input of a Natural Language Understanding (NLU) model which then predicts the next Dialogue State DS_t.

Both components are trained separately: WavLM [16], with two additional linear layers outputting tokens' probabilities, to transcribe the user turns (fine-tuned with CTC loss) and T5 Encoder-Decoder [17] to output dialogue states. Note that the NLU model is trained with user turns transcriptions of the ASR model in order to be as close as possible to its inference regime.

2.2.2. Local E2E Approach

DS are updated through three operations: addition of a new slot-value pair, modification and suppression of a previously mentioned slot-value pair. In order to update DS_{t-2} and obtain DS_t with a rule based system we need to encode these operations into local DS. While additions and modifications of slots can remain the same in local DS, we mark suppressed slots by assigning them the value `<unk>`.

Finally, for references to previously mentioned slots, as shown in Fig. 2, we explicit the reference through the name of the referred slot's value present in DS_{t-2}.

This model relies on a Whisper [18] backbone model with a fine-tuned decoder on the user turns. At inference time, the context $DS_{t-2} + A_{t-1}$ is added before the decoder input tokens to condition the decoding. Note that erroneous references and suppression (*e.g.* non-existing slots) are discarded by post-processing the outputs.

2.2.3. Completely Neural Approach

The completely neural approach leverages both previous approaches by fusing audio and semantic encoder's outputs and feeding them to a semantic decoder. The goal of this approach is to enable joint optimization of all components. More formally, we have:

[2]Code will be made available upon publication.

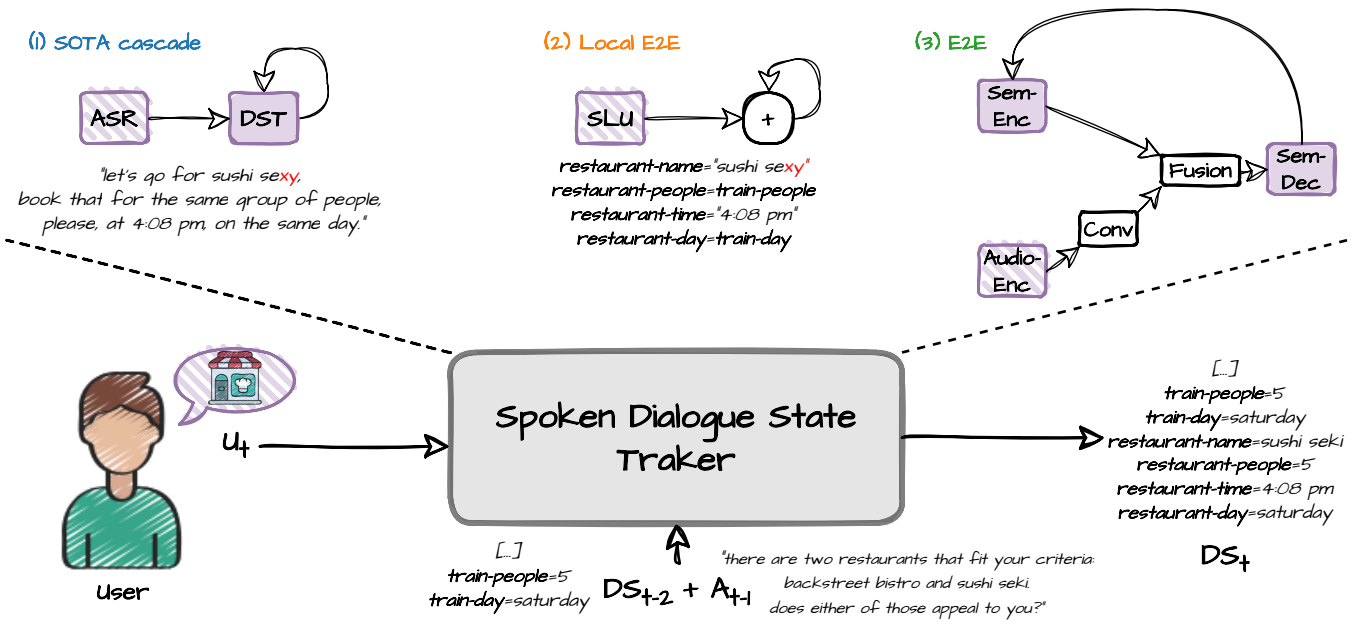

Fig. 2. Three approaches for context propagation in spoken DST: (1) SOTA cascade, (2) local E2E model with rule-based contextualization and (3) completely neural model. (1) and (2) present cascading errors in red characters. Hatched components are speech-related while solid ones are text-related.

$$
\begin{aligned}
h_1 &= \text{Sem-Enc}(DS_{t-2} + A_{t-1}) \\
h_2 &= \text{Audio-Enc}(U_t) \\
h &= \text{Fusion}(h_1 + \text{Conv}(h_2)) \\
\hat{DS}_t &= \text{Sem-Dec}(h)
\end{aligned}
$$

WavLM [16] and T5's encoder [17] respectively encode the current turn (audio) and the context (text). Given that both models do not have the same processing windows, two convolution layers (stride 3, kernel size 9) are added to downsample the audio encoder's outputs. The fusion layer is a MLP which enables the model to select and mix the information from both encoders. Finally, a T5 decoder outputs DS_t conditioned on the fusion of both encoders' outputs.

3. RESULTS

3.1. Spoken Multi-Woz dataset

Multi-Woz is a human-human chat-based English Task-Oriented Dialogue (TOD) dataset commonly used for training and evaluating dialogue systems [14]. A spoken version of Multi-Woz with vocalized user turns was published in the context of the *Speech Aware Dialogue Systems* track of the 11th edition of the Dialogue System Technology Challenge[3] (DSTC11) [15]. The user utterances in the training set are available as synthetic speech, whereas the dev and test sets (**Dev**|**Test**) include both synthetic and human speech versions (**TTS**|**Human**). The dataset contains close to 10,000 dialogues with a 80/10/10 train-dev-test split and an average of 13.3 turns per dialogue. Among the pre-defined slots, we can distinguish 3 groups: categorical slots with a closed set of values (∼60%), non-categorical slots with an open set of values (∼30%) and time slots (∼10%). Note that, in order to reduce the value overlap across sets, non-categorical slots were replaced and time slots offset in the **Dev** and **Test** sets.

3.2. Evaluation

We evaluate all approaches with a turn-level exact match metric known as Joint-Goal Accuracy (JGA↑) [19]. This metric requires to post-process the coma separated slot-value output format to convert it into a valid dictionary which does not take into account the order of the slot-value pairs. Given that generative models, such as T5, are prone to hallucinations, a filtering step discards all slots which are not part of the predefined slots. We present here the results of all three approaches in two scenarios: with ground truth context DS_{t-2}, in Table 1 and with the previous prediction $\hat{DS}_{t-2}$ in Fig. 4.

We find that the cascade approach remains a tough competitor only 6 points behind the text oracle model. Although both the local and global E2E approaches achieve higher accuracies than the recent DSTC11 best model, especially with the filtering post-processing step, they are not competitive with a carefully designed cascade approach. It is notewor-

[3]https://dstc11.dstc.community/

	Dev		Test	
Text	81.2		80.3	
w/o filtering	*64.8*		*63.8*	
	TTS	**Human**	**TTS**	**Human**
DSTC11 baseline [15]	*n/a*		*n/a*	
w/o filtering	*38.4*	*31.8*	*n/a*	
DSTC11 best [5]	*n/a*		*n/a*	
w/o filtering	*47.2*	*43.2*	*44.0*	*39.5*
(1) Cascade	**75.2**	**71.9**	**75.4**	**71.8**
w/o filtering	*53.5*	*48.6*	*53.2*	*47.8*
(2) Local E2E	60.5	61.8	62.3	62.6
w/o filtering	*43.6*	*43.8*	*42.8*	*42.7*
(3) E2E	56.8	54.2	57.1	53.9
w/o filtering	*14.9*	*12.6*	*14.3*	*12.3*

Table 1. JGA of spoken DST with ground-truth previous state DS_{t-2}. Text line shows upper-bound performance. Note that [15] and [5] use dialogue history as input to the DST model.

thy that filtering out the undefined slots has a significant impact for all approaches. The completely E2E approach seems particularly prone to hallucinations which indicates that it has more trouble selecting the relevant information from the audio encoder's hidden states.

In order to get a more precise understanding of the differences between these approaches, we further evaluate each slot's F1-measure and present each slot groups average F1-measure on the **Test-Human** set in Fig. 3. Categorical slots present no difficulty while non-categorical and time slots are more challenging. Interestingly the backbone model seems to play a role in the output format given that the local E2E and completely E2E models' F1-measures are reversed for the Time slot group which requires careful formatting.

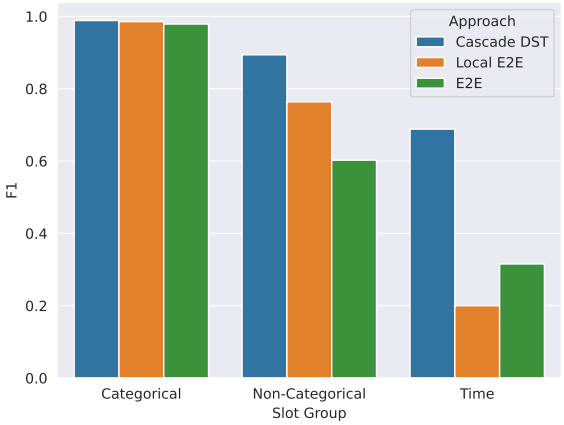

Fig. 3. Test-Human slot group average F1 measure.

In a more realistic scenario where we base our next prediction on the previous one[4] we observe that, as illustrated in Fig. 4, both the cascade and the completely E2E approaches perform as good with their predictions as with the ground-truth whereas the local E2E approach collapses. In fact it achieves a higher accuracy on the first turn and collapses on the next ones which highlights that using decoder prefixing to propagate the dialogue's context might not be the best method. Also note that the global E2E approach has more trouble handling long dialogues than its alternatives.

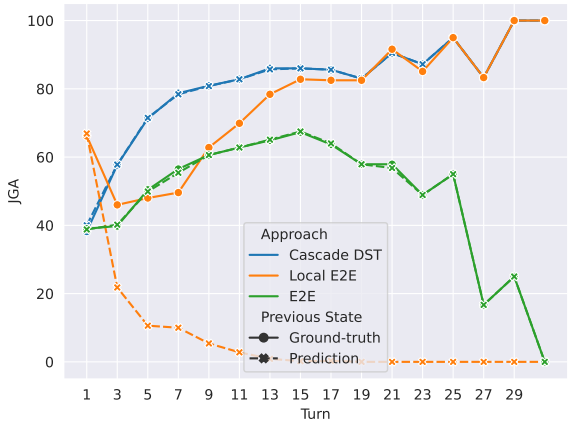

Fig. 4. Test-Human turn accuracy with and without ground-truth previous state for each approach. Note that there are fewer and fewer dialogues as the number of turns increases.

4. CONCLUSION

In order to pave the path towards E2E spoken DST, we compare a state of the art cascade approach with local and global E2E approaches. Our study highlights that although they all outperform the recent DSTC11 winner model, especially with a filtering post-processing step, the cascade approach remains the most accurate approach and context propagation in completely neural approaches an open challenge. Local E2E and global E2E approaches behave quite differently: the former is very accurate on the first dialogue turn but collapses when contextualizing its predictions and the latter hallucinates much more and has trouble with long dialogues.

Our results remain to be confirmed on other contextually dependant SLU datasets to come. Given the chat-based origin of Multi-Woz, the turns are assumed to be perfectly separable which is not the case in general. A more fine-grained evaluation to assess which errors are low-impact errors (*e.g.* rectified with the help of a database, with no impact on the dialogue trajectory) and improving the post processing with a more careful filtering is left as future work.

[4]Note that A_{t-1} and U_t remain unchanged which might lead to some incoherences.

5. REFERENCES

[1] Gokhan Tur and Renato De Mori, *SLU in Commercial and Research Spoken Dialogue Systems*, Wiley Telecom, 2011.

[2] Jason D. Williams, Antoine Raux, and Matthew Henderson, "The Dialog State Tracking Challenge Series: A Review," *Dialogue & Discourse*, 2016.

[3] Salima Mdhaffar, Valentin Pelloin, Antoine Caubrière, Gaëlle Laperrière, Sahar Ghannay, Bassam Jabaian, Nathalie Camelin, and Yannick Estève, "Impact Analysis of the Use of Speech and Language Models Pretrained by Self-Supersivion for Spoken Language Understanding," in *Conference on Language Resources and Evaluation (LREC)*, 2022.

[4] Dmitriy Serdyuk, Yongqiang Wang, Christian Fuegen, Anuj Kumar, Baiyang Liu, and Yoshua Bengio, "Towards end-to-end spoken language understanding," in *ICASSP*, 2018.

[5] Léo Jacqmin, Lucas Druart, Valentin Vielzeuf, Lina Maria Rojas-Barahona, Yannick Estève, and Benoît Favre, "OLISIA: a Cascade System for Spoken Dialogue State Tracking," in *Special Interest Group on Discourse and Dialogue (SIGDIAL)*, 2023.

[6] Siddhant Arora, Hayato Futami, Shih-Lun Wu, Jessica Huynh, Yifan Peng, Yosuke Kashiwagi, Emiru Tsunoo, Brian Yan, and Shinji Watanabe, "A study on the integration of pipeline and e2e slu systems for spoken semantic parsing toward stop quality challenge," in *ICASSP*, 2023.

[7] Vishal Sunder, Samuel Thomas, Hong-Kwang J. Kuo, Jatin Ganhotra, Brian Kingsbury, and Eric Fosler-Lussier, "Towards end-to-end integration of dialog history for improved spoken language understanding," in *ICASSP*, 2022.

[8] Jatin Ganhotra, Samuel Thomas, Hong-Kwang J. Kuo, Sachindra Joshi, George Saon, Zoltán Tüske, and Brian Kingsbury, "Integrating Dialog History into End-to-End Spoken Language Understanding Systems," in *Interspeech*, 2021.

[9] Guangzhi Sun, Chao Zhang, and Philip C. Woodland, "End-to-end spoken language understanding with tree-constrained pointer generator," in *ICASSP*, 2023.

[10] Valentin Pelloin, Nathalie Camelin, Antoine Laurent, Renato De Mori, Antoine Caubrière, Yannick Estève, and Sylvain Meignier, "End2end acoustic to semantic transduction," in *ICASSP*, 2021.

[11] Natalia Tomashenko, Christian Raymond, Antoine Caubrière, Renato De Mori, and Yannick Estève, "Dialogue history integration into end-to-end signal-to-concept spoken language understanding systems," in *ICASSP*, 2020.

[12] Vishal Sunder, Eric Fosler-Lussier, Samuel Thomas, Hong-Kwang J Kuo, and Brian Kingsbury, "ConvKT: Conversation-Level Knowledge Transfer for Context Aware End-to-End Spoken Language Understanding," in *Interspeech*, 2023.

[13] Xiangkun Hu, Junqi Dai, Hang Yan, Yi Zhang, Qipeng Guo, Xipeng Qiu, and Zheng Zhang, "Dialogue meaning representation for task-oriented dialogue systems," in *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2022.

[14] Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Iñigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gasic, "Multiwoz - a large-scale multi-domain wizard-of-oz dataset for task-oriented dialogue modelling," in *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2018.

[15] Hagen Soltau, Izhak Shafran, Mingqiu Wang, Abhinav Rastogi, Jeffrey Zhao, Ye Jia, Wei Han, Yuan Cao, and Aramys Miranda, "Speech aware dialog system technology challenge (dstc11)," in *Special Interest Group on Discourse and Dialogue (SIGDIAL)*, 2023.

[16] Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, Jian Wu, Long Zhou, Shuo Ren, Yanmin Qian, Yao Qian, Micheal Zeng, and Furu Wei, "Wavlm: Large-scale self-supervised pre-training for full stack speech processing," *IEEE Journal of Selected Topics in Signal Processing (JSTSP)*, 2021.

[17] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu, "Exploring the limits of transfer learning with a unified text-to-text transformer," *Journal Machine Learning Research (JMLR)*, 2020.

[18] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever, "Robust speech recognition via large-scale weak supervision," *ArXiv*, vol. abs/2212.04356, 2022.

[19] Victor Zhong, Caiming Xiong, and Richard Socher, "Global-locally self-attentive encoder for dialogue state tracking," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, 2018.