Lecture Notes in Electrical Engineering 714

Erik Marchi · Sabato Marco Siniscalchi · Sandro Cumani · Valerio Mario Salerno · Haizhou Li *Editors*

Increasing Naturalness and Flexibility in Spoken Dialogue Interaction

10th International Workshop on Spoken Dialogue Systems

Springer

Lecture Notes in Electrical Engineering

Volume 714

The book series *Lecture Notes in Electrical Engineering* (LNEE) publishes the latest developments in Electrical Engineering - quickly, informally and in high quality. While original research reported in proceedings and monographs has traditionally formed the core of LNEE, we also encourage authors to submit books devoted to supporting student education and professional training in the various fields and applications areas of electrical engineering. The series cover classical and emerging topics concerning:

- Communication Engineering, Information Theory and Networks
- Electronics Engineering and Microelectronics
- Signal, Image and Speech Processing
- Wireless and Mobile Communication
- Circuits and Systems
- Energy Systems, Power Electronics and Electrical Machines
- Electro-optical Engineering
- Instrumentation Engineering
- Avionics Engineering
- Control Systems
- Internet-of-Things and Cybersecurity
- Biomedical Devices, MEMS and NEMS

For general information about this book series, comments or suggestions, please contact leontina.dicecco@springer.com.

To submit a proposal or request further information, please contact the Publishing Editor in your country:

China

Jasmine Dou, Editor (jasmine.dou@springer.com)

India, Japan, Rest of Asia

Swati Meherishi, Editorial Director (Swati.Meherishi@springer.com)

Southeast Asia, Australia, New Zealand

Ramesh Nath Premnath, Editor (ramesh.premnath@springernature.com)

USA, Canada:

Michael Luby, Senior Editor (michael.luby@springer.com)

All other Countries:

Leontina Di Cecco, Senior Editor (leontina.dicecco@springer.com)

**** This series is indexed by EI Compendex and Scopus databases. ****

More information about this series at http://www.springer.com/series/7818

Erik Marchi · Sabato Marco Siniscalchi ·
Sandro Cumani · Valerio Mario Salerno ·
Haizhou Li
Editors

Increasing Naturalness and Flexibility in Spoken Dialogue Interaction

10th International Workshop on Spoken Dialogue Systems

Springer

Editors
Erik Marchi
Apple
Cupertino, CA, USA

Sandro Cumani
Polytechnic University of Turin
Torino, Italy

Haizhou Li
National University of Singapore
Singapore, Singapore

Sabato Marco Siniscalchi
Kore University of Enna
Enna, Italy

Valerio Mario Salerno
Kore University of Enna
Enna, Italy

ISSN 1876-1100 ISSN 1876-1119 (electronic)
Lecture Notes in Electrical Engineering
ISBN 978-981-15-9322-2 ISBN 978-981-15-9323-9 (eBook)
https://doi.org/10.1007/978-981-15-9323-9

This Springer imprint is published by the registered company Springer Nature Singapore Pte Ltd.
The registered company address is: 152 Beach Road, #21-01/04 Gateway East, Singapore 189721, Singapore

Preface

This book contains a selection of revised papers that were presented at the 10th edition of the International Workshop on Spoken Dialogue Systems (IWSDS) that took place in the beautiful town of Syracuse in Sicily (Italy), from 24 to 26 April 2019. IWSDS is usually held every year and provides a platform to present and discuss global research and application of spoken dialogue systems.

This 10th edition of IWSDS named "Increasing Naturalness and Flexibility in Spoken Dialogue Interaction" focused specifically on the following topics:

- Context Understanding and Dialogue Management
- Human–Robot Interaction
- Dialogue Evaluation and Analysis
- Chatbots and Conversational Agents
- Lifelong Learning
- Question Answering and other Dialogue Applications
- Dialogue Breakdown Detection

Cupertino, USA — Erik Marchi
Enna, Italy — Sabato Marco Siniscalchi
Torino, Italy — Sandro Cumani
Enna, Italy — Valerio Mario Salerno
Singapore, Singapore — Haizhou Li

Contents

Human-Robot Interaction

Dialogue Evaluation and Analysis

Chatbots and Conversational Agents

Lifelong Learning

Context Understanding and Dialogue Management

Skip Act Vectors: Integrating Dialogue Context into Sentence Embeddings

Jeremy Auguste, Frédéric Béchet, Géraldine Damnati, and Delphine Charlet

Abstract This paper compares several approaches for computing dialogue turn embeddings and evaluate their representation capacities in two dialogue act related tasks within a hierarchical Recurrent Neural Network architecture. These turn embeddings can be produced explicitely or implicitely by extracting the hidden layer of a model trained for a given task. We introduce *skip-act*, a new dialogue turn embeddings approach, which are extracted as the common representation layer from a multi-task model that predicts both the previous and the next dialogue act. The models used to learn turn embeddings are trained on a large dialogue corpus with light supervision, while the models used to predict dialog acts using turn embeddings are trained on a sub-corpus with gold dialogue act annotations. We compare their performances for predicting the current dialogue act as well as their ability to predict the next dialogue act, which is a more challenging task that can have several applicative impacts. With a better context representation, the *skip-act* turn embeddings are shown to outperform previous approaches both in terms of overall F-measure and in terms of macro-F1, showing regular improvements on the various dialogue acts.

J. Auguste (✉) · F. Béchet
Aix-Marseille Univ, Université de Toulon, CNRS, LIS, Marseille, France
e-mail: jeremy.auguste@lis-lab.fr

F. Béchet
e-mail: frederic.bechet@lis-lab.fr

G. Damnati · D. Charlet
Orange Labs, Lannion, France
e-mail: geraldine.damnati@orange.com

D. Charlet
e-mail: delphine.charlet@orange.com

E. Marchi et al. (eds.), *Increasing Naturalness and Flexibility in Spoken Dialogue Interaction*, Lecture Notes in Electrical Engineering 714,
https://doi.org/10.1007/978-981-15-9323-9_1

1 Introduction

Following the successful application of continuous representation of words into vector spaces, or *embeddings*, in a large number of Natural Language Processing tasks [14, 15], many studies have proposed the same approach for larger units than words such as sentences, paragraphs or even documents [10, 11]. In all cases the main idea is to capture the *context of occurrence* of a given unit as well as the unit itself.

When processing dialog transcriptions, being able to model the *context of occurrence* of a given turn is of great practical use in applications such as automated dialog system for predicting the next action to perform, or analytics in order, for example, to pair questions and answers in a corpus of dialog logs. Therefore finding the best embedding representations for dialog turns in order to model dialog structure as well as the turns themselves is an active field of research.

In this paper, we evaluate different kinds of sentence-like (turns) embeddings on dialogue act classification tasks in order to measure how well they can capture dialog structures. In a first step, the dialogue turn embeddings are learned on large corpus of chat conversations, using a light supervision approach where dialogue act annotations are given by an automatic DA parser. Even if the annotations are noisy, this light supervision approach allows us to learn turn-level vector representations on a large amount of interactions. In a second step, the obtained turn-level vector representations are used to train dialogue act prediction models with a controlled supervised configuration.

After presenting the dialogue act parser architecture in Sect. 3, we will present the various dialogue turn embeddings approaches in Sect. 4. The corpus and the dialogue act annotation framework are presented in Sect. 5 while Sect. 6 describes the experimental results.

2 Related Work

In order to create and then evaluate the quality of embeddings, several different types of approaches have been proposed. For word embeddings, a lot of work has been done to try to evaluate how they are able to capture relatedness and similarity between two words by using manual annotation [4, 9, 12] and by using cognitive processes [2, 18]. However, on sentence embeddings, it is not easy to tell how similar or related two sentences are. Indeed, the context in which they appear is very important to truly understand the meaning of a sentence and how it interacts with other sentences.

Multiple papers propose different kinds of evaluation tasks in order to evaluate different kinds of sentence embeddings. In [8], the authors use the SICK [13] and STS 2014 [1] datasets to evaluate the similarity between sentences by using similarity ratings. They also use sentiment, opinion polarity and question type tasks to evaluate the embeddings. As these datasets are composed of sentence pairs with-

out context, the proposed sentence embeddings approaches are only based on the sentence itself. In [7], sentence embeddings are evaluated by looking at their ability to capture surface, syntactic and semantic information. Here again, this framework primarily focuses on the sentence itself and not on the context in which it is produced. In [5], a sentence embeddings evaluation framework is proposed that groups together most of the previous evaluation tasks in addition to inference, captioning and paraphrase detection tasks. In all of the above approaches, the focus is on the evaluation of sentence embeddings such as Skip-thoughts [10], ParagraphVectors [11] or InferSent [6] in order to find out the embeddings that have the best properties in general. However, none of these embeddings and evaluation tasks are built to take into account dialogues and more specifically, the structure and interactions in a dialogue. Some work has been done in order to take into account the dialogue context in [17]. In their work, the authors try to take into account this context by using a modified version of word2vec to learn sentence embeddings on dialogues. These embeddings are then evaluated by comparing clusters of sentence embeddings with manually assigned dialogue acts. This allows to see if the learned embeddings capture information about the dialogue context, however it does not use explicit dialogue structure information to learn the embeddings. In our work, we use a corpus with a noisy dialogue act annotation to learn specialized sentence embeddings that try to directly capture information about the context and interactions in the dialogue.

3 Dialogue Act Parser Architecture

In order to be able to create sentence embeddings that take into account the dialogue context, we will be using dialogue acts. They allow us to partially represent the structure and the interactions in a dialogue. We use two different kinds of models to parse these dialogue acts where one kind is used to create sentence embeddings, while the second kind is used to later evaluate the different embeddings.

The first architecture is a 2-level hierarchical LSTM network where the first level is used to represent the turns in a conversation, and the second level represents the conversation, as shown in Fig. 1. The input is the sequence of turns which are themselves sequences of words represented as word embeddings. The word embeddings are trained by the network from scratch. The dialogue acts are predicted using the output for each turn at the second level. Since we do not use a bidirectional LSTM, the model only makes use of the associated turn and the previous turns of a conversation in order to predict a given act. It has no information about the future, nor about the previous acts. This architecture allows us to use the hidden outputs of the first layer as the sentence embeddings of each turn.

The second architecture is a simple LSTM network which only has a single layer, as shown in Fig. 2. The input sequence that is given to the LSTM is the sequence of turns of a conversation where each turn is replaced by a pre-trained turn embedding.

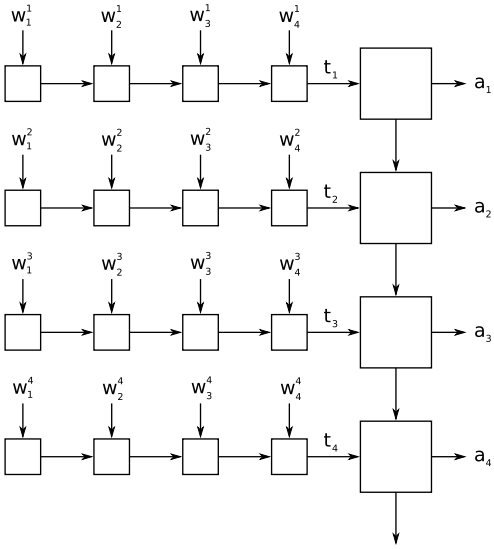

Fig. 1 Two level LSTM architecture used to create embeddings. w_i^j is the word i of turn j, t_j is the learned turn embedding and a_j is the predicted act

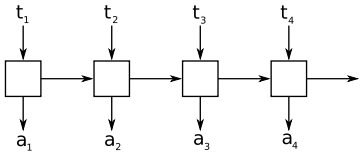

Fig. 2 LSTM architecture used for evaluation. t_i is a fixed pre-trained turn embedding and a_i is the predicted act

For each turn, the corresponding output in the LSTM is used to predict its dialogue act. This architecture is the one used to evaluate the different kinds of fixed pre-trained embeddings that are described in Sect. 4.

4 Skip-Act Vectors

It is possible to construct sentence embeddings using several different means, each of them being able to capture different aspects of a sentence. In our case, we want to find out what kind of embeddings are the best at capturing information about the dialogical structure and the context in which appears a turn. Multiple different kind of embeddings are thus trained on the `DATCHA_RAW` corpus (the large unannotated corpus described in Sect. 5. The following self-supervised embeddings are trained:

Word Average This is simply the average of all the word embeddings in the turn. The word embeddings are learned with FastText [3] on the `DATCHA_RAW` corpus using a dimension of 2048 and a window size of 6. These can be considered as our baseline embeddings since they do not directly take into account the context in which the turns are produced.

Skip-thought These embeddings are learned using a skip-thought model [10]. This model tries to learn the sentence embeddings by trying to regenerate the adjacent sentences during the training. Thus, it tries to learn the context in which a sentence is produced.

In addition to these self-supervised embeddings, we also learned embeddings based on supervised tasks. To learn these embeddings, we use the 2-level LSTM architecture described in Sect. 3. The following supervised embeddings are trained:

RNN Curr Act These embeddings are learned by using a hierarchical neural network that is trained to predict the dialogue act of each turn. The embeddings are the hidden output from the turn layer of the network. Since the DATCHA_RAW corpus is not annotated with dialogue acts, we used a system developed during the DATCHA[1] project based on a CRF model developed in [16] (85.7% accuracy) to predict the dialogue acts of each turn of the corpus.

RNN Next Act These embeddings are created similarly to the RNN Curr Act embeddings but instead of predicting the current act for a given turn, the following act is instead predicted.

RNN Prev Act These embeddings are created similarly to the RNN Curr Act embeddings but instead of predicting the current act for a given turn, the previous act is instead predicted.

Skip-Act These embeddings combine the ideas of RNN Prev Act and RNN Next Act by using the same turn layer in the network for both tasks. This model shares the idea of the Skip-thought vectors by trying to learn the context in which the turns are produced. But instead of trying to regenerate the words in the adjacent turns, we try to predict the dialogue acts of the adjacent turns. This allows us to hope that the learned embeddings will focus on the dialogue context of turns. The architecture of this model is presented in Fig. 3.

5 Corpus

Chat conversations are extracted from Orange's customer services contact center logs, and are gathered within the DATCHA corpus, with various levels of manual annotations. The DATCHA corpus covers a wide variety of topics, ranging from technical issues (e.g.. solving a connection problem) to commercial inquiries (e.g.. purchasing a new offer). They can cover several applicative domains (mobile, internet, tv).

For our experiments, we use two different subsets of these chats:

- Chats from a full month that do not have any gold annotation (79000 dialogues, 3400000 turns) (DATCHA_RAW);

[1] http://datcha.lif.univ-mrs.fr.

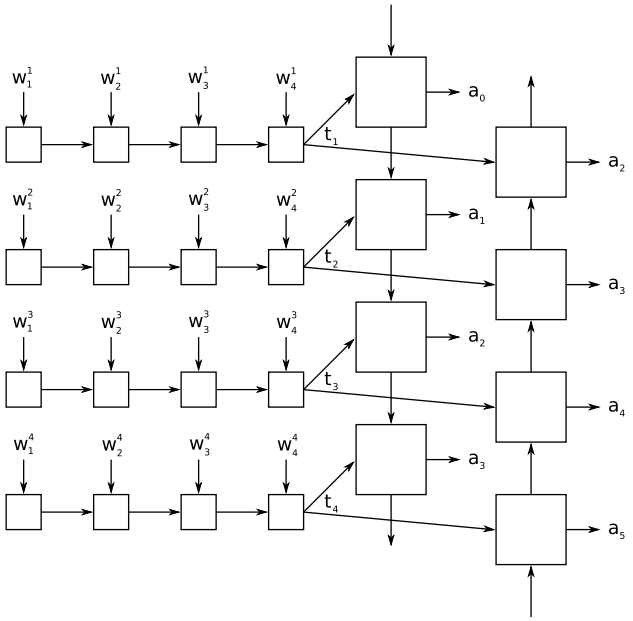

Fig. 3 Architecture used to create skip-act vectors. w_i^j is the word i of turn j, t_j is the learned turn embedding and a_j is the predicted act

- Chats annotated with gold dialogue act annotation (3000 dialogues, 94000 turns) (DATCHA_DA)

These subsets are partitioned into train, test and development parts. The label set used in the dialogue act annotation is as follows:

Label	Meaning	Description
OPE	Opening	Opening turns in the dialogue
PRO	Problem description	The client's description of his problem
INQ	Information question	Turn where a speaker asks for some information
CLQ	Clarification question	A speaker asks for clarification
STA	Statement	New information input
TMP	Temporisation	Starting a break of the dialogue
PPR	Plan proposal	Resolution proposal of the problem
ACK	Acknowledgement	A speaker acknowledges the other speaker's sayings
CLO	Closing	Closing turn
OTH	Other	For turns that don't match other described labels

This set has been designed to be as generic as possible, while taking into account some particular aspects of professional chat interactions (e.g.. *Problem description* or *Plan proposal*). The distribution of the different types of dialogue acts in the test split of the DATCHA_DA corpus can be found in Fig. 4. We also indicate the

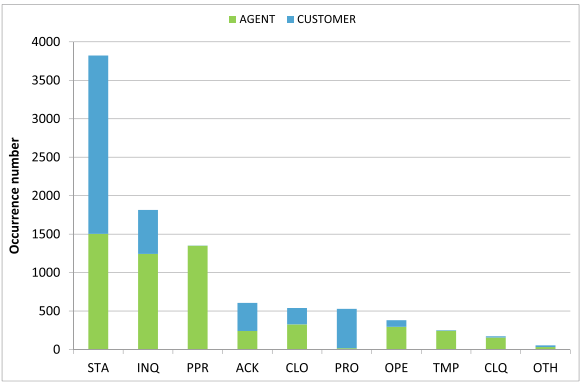

Fig. 4 Dialogue act distribution in the DATCHA_DA test corpus

distributions when considering only a single speaker since they use very different types of turns. For instance, *Plan proposals* are almost exclusively uttered by Agents while, conversely, *Problem descriptions* are mostly observed on Customers side.

6 Turn Embeddings Evaluation

6.1 Evaluation Protocol

We want to make sure that the generated embeddings are able to capture the different aspects of a dialogue. Dialogue acts are one way to partially represent the structure and interactions in a dialogue. Thus, we evaluate the different embeddings on two tasks. For the first task, we try to predict the dialogue act of a turn by only using the sequence of embeddings of the current and previous turns. For the second task, we do the same thing but instead of predicting the dialogue act of the current turn, we predict the act of the next turn (without giving the embedding of the next turn in the input). This second task allows us to tell if the learned embeddings manage to capture information about not only the turn but also about the context in which these turns are produced.

Some of the created embeddings are learned using tasks that involve dialogue acts, thus it is likely that these embeddings obtain the best results. But it is interesting to see if other embeddings are able to obtain similar or close results.

For both tasks, we use the architectures described in Sect. 3 with a hidden size of 512. For each turn, the corresponding output in the RNN is given to a decision layer which uses a softmax to output a probability distribution of the dialogue acts. We use cross-entropy as our loss function and Adam as the optimizer with a learning rate of 0.001. The PyTorch framework is used to build the different architectures.

In order to evaluate the quality of the different predictions, we primarily use 2 metrics:

- **accuracy**: the percentage of correct decisions;
- **macro F1**: the non-weighted average of the F1-measures of the 10 act labels. The F1-measure is the harmonic mean of precision P and recall R for a given label l such as $F1(l) = \frac{2 \times P(l) \times R(l)}{P(l)+R(l)}$;

6.2 Results and Analyses

Results of the prediction of the current and next acts are reported in Table 1. The first line corresponds to the first model described in Fig. 1 where no pre-trained embeddings are used and where the embeddings are learned jointly with the model's parameters on the DATCHA_DA corpus. The following lines correspond to the single turn-level architecture presented in Fig. 2 using several variants of fixed turn embeddings, pre-trained on the large DATCHA_RAW corpus. For each embedding type and task, we only report the results of the configuration that obtained the best results. We can first note a big difference in performances between the two tasks with the next act task being much harder than the current act task. It seems to be very difficult to predict the next act given the history of turns, particularly for some of them, as can be seen in Figs. 5 and 6 where some acts such as CLQ, INQ or PPR see a drop of 60 points in their F1-score while acts such as STA, CLO or OPE only have a drop of 20 points. This could be explained by the fact that closings and openings are easier to locate in the conversation, while statements are the most represented labels in conversations. On the other hand, it is not necessarily easy to know that the next turn is going to be a question or a plan proposal. We can also notice that the OTH act is not at all correctly predicted in the next act task, and even in the current act task it is the label with the worst F1-score. This is probably due to the fact that turns that are labeled OTH are usually filled with random symbols or words and are both very diverse and not frequent.

Table 1 Evaluation of the prediction of the current and next dialogue acts on all turns

		Current act		Next act	
LSTM architecture	Pre-trained embeddings	Accuracy	Macro-F1	Accuracy	Macro-F1
2-level hierarchical	None	83.69	78.15	46.21	26.45
Turn level	Word average	82.96	79.47	48.26	30.09
Turn level	Skip-thought	82.50	75.73	48.30	28.61
Turn level	RNN curr act	84.74	80.47	48.54	31.42
Turn level	RNN next act	84.40	81.42	49.97	34.47
Turn level	RNN prev act	83.02	80.44	48.77	31.96
Turn level	**Skip-act**	**85.24**	**82.16**	**49.96**	**35.33**

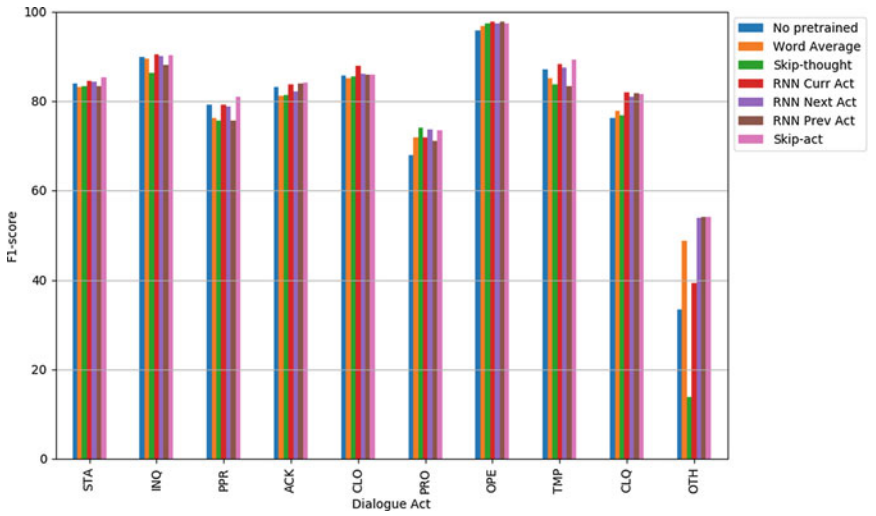

Fig. 5 F1-scores on the current act task on all turns

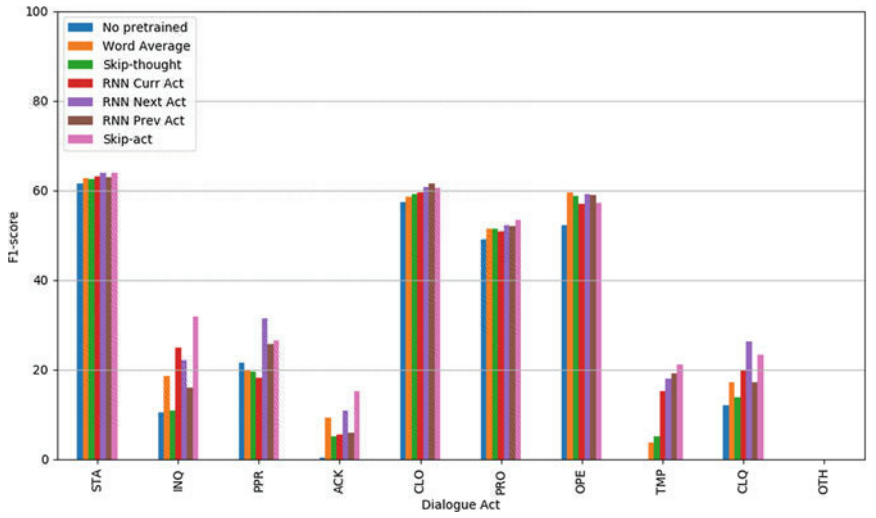

Fig. 6 F1-scores on the next act task on all turns

Unsurprisingly, for both tasks, the best results are obtained with embeddings learned using dialogue acts. However, the **Word Average** and **Skip-thought** vectors both achieve good results but they still are 2 points lower than the best results. It is interesting to note that the **Skip-thought** vectors do not achieve better results than **Word Average** vectors on the next act task. This can be surprising since they would have been expected to better capture information about the surrounding turns, however the generalization from word level prediction to turn level prediction is

not sufficiently efficient. It is also interesting to note that better results are achieved by **RNN Curr Act** embeddings (84.74%), which are learned on a corpus with a noisy annotation, compared to results achieved by the embeddings learned during the training on the DATCHA_DA corpus (83.69%) which has gold annotation. This results confirms our choice to train turn embeddings separately with light supervision on a significantly larger, even though noisy, training corpus.

Another interesting aspect of these results is the comparison of the different kinds of embeddings learned with dialogue act related tasks. Indeed, on the current act task, we can notice that **RNN Curr Act** embeddings obtain slightly lower results (−0.5 points) than **Skip-act** embeddings. This is surprising since **RNN Curr Act** are learned using the same task than the evaluation, while **Skip-act** are learned by trying to predict the next and previous acts only. These results could mean that **Skip-act** are more robust since they learn in what context the acts are produced. On the next act task, both the **RNN Next Act** and **Skip-act** achieve the same performances with 50% accuracy, while the **RNN Curr Act** embeddings obtain an accuracy of 48.5%.

We also reported in Tables 2 and 3 the results when considering only the turns from respectively the agent and the client for evaluation. It is important to note that the label distribution is very different depending on the speaker. Most of the questions (CLQ and INQ) and nearly all plan proposals (PPR) and temporisations (TMP) are from the agent while most of the problem descriptions (PRO) and the majority of statements (STA) are from the client. When evaluated on the agent side, **Skip-act** embeddings are again the best embeddings for both tasks, being 1 point higher than the **RNN Next Act** embeddings and 3.5 points higher than the **RNN Curr Act** embeddings. These results are interesting since the agent is the speaker with the most variety in the types of turns, including many turns with questions, plan proposals or temporisations. This seems to indicate that **Skip-acts** manage to capture more information about the dialogue context than the other embeddings.

Table 2 Evaluation of the prediction of the current and next dialogue acts on agent's turns

		Current act		Next act	
LSTM architecture	Pre-trained embeddings	Accuracy	Macro-F1	Accuracy	Macro-F1
2-level hierarchical	None	84.22	77.38	35.87	23.16
Turn level	Word average	82.48	77.31	37.78	27.02
Turn level	Skip-thought	80.36	74.75	37.07	25.39
Turn level	RNN curr act	84.70	79.01	38.90	29.00
Turn level	RNN next act	84.30	82.42	41.29	32.60
Turn level	RNN prev act	83.24	80.11	38.80	28.81
Turn level	**Skip-act**	**85.48**	**82.94**	**42.30**	**33.56**

Table 3 Evaluation of the prediction of the current and next dialogue acts on customer's turns

		Current act		Next act	
LSTM architecture	Pre-trained embeddings	Accuracy	Macro-F1	Accuracy	Macro-F1
2-level hierarchical	None	83.01	58.58	59.48	21.13
Turn level	Word average	83.59	60.97	61.71	21.80
Turn level	Skip-thought	85.31	59.13	62.70	20.49
Turn level	RNN curr act	84.78	64.16	60.89	21.74
Turn level	RNN next act	84.54	63.20	61.09	22.91
Turn level	RNN prev act	82.74	61.88	61.56	21.73
Turn level	**Skip-act**	**84.93**	**63.99**	**59.78**	**23.79**

We can also notice that this time, **Skip-thought** vectors obtain lower results than the simple **Word Average**. When evaluated on the customer side, **Skip-thought** vectors obtain the best scores on both tasks when looking at the accuracy (85.31% and 62.70%) but lower scores in terms of macro-F1. The scores on the next act task are higher but this is only due to the fact that the STA act represents 57.4% of the samples, whereas on all the turns and for the agent they respectively represent 40.2% and 27.8% of the samples.

7 Conclusion

We have proposed a new architecture to compute dialogue turn embeddings. Within the skip-act framework, a multitask model is trained in order to jointly predict the previous and the next dialogue acts. Trained in a lightly supervised way on a large corpus of chat conversations with an automatic dialogue act annotation, the output of the common hidden layer provides an efficient turn level vector representation that tends to capture the dialogic structure of the interactions. We have evaluated several dialogue turn embeddings configurations on two tasks, first predicting the associated dialogue act of the current turn, and then predicting the next dialogue act which is a more challenging task requiring a better representation of the dialogue structure. Skip-act embeddings achieve the best results on both tasks. In the future, it would be interesting to combine skip-thoughts and skip-acts in order to be able to capture the semantic and syntactic information in addition to the dialogue context of turns.

Acknowledgements This work has been partially funded by the Agence Nationale pour la Recherche (ANR) through the following programs: ANR-15-CE23-0003 (DATCHA), ANR-16-CONV-0002 (ILCB) and ANR-11-IDEX-0001-02 (A*MIDEX).

References

1. Agirre E, Banea C, Cardie C, Cer D, Diab M, Gonzalez-Agirre A, Guo W, Mihalcea R, Rigau G, Wiebe J (2014) Semeval-2014 task 10: multilingual semantic textual similarity. In: Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014), pp 81–91
2. Auguste J, Rey A, Favre B (2017) Evaluation of word embeddings against cognitive processes: Primed reaction times in lexical decision and naming tasks. In: Proceedings of the 2nd workshop on evaluating vector space representations for NLP, pp 21–26. Copenhagen, Denmark
3. Bojanowski P, Grave E, Joulin A, Mikolov T (2017) Enriching word vectors with subword information. Trans Assoc Comput Linguist 5(1):135–146
4. Bruni E, Boleda G, Baroni M, Tran NK (2012) Distributional semantics in technicolor. In: Proceedings of the 50th annual meeting of the association for computational linguistics: long papers, vol 1, pp 136–145. Association for Computational Linguistics, Stroudsburg
5. Conneau A, Kiela D (2018) SentEval: an evaluation toolkit for universal sentence representations. In: Proceedings of the eleventh international conference on language resources and evaluation (LREC 2018). Miyazaki, Japan
6. Conneau A, Kiela D, Schwenk H, Barrault L, Bordes A (2017) Supervised learning of universal sentence representations from natural language inference data. In: Proceedings of the 2017 conference on empirical methods in natural language processing, pp 670–680 (2017)
7. Conneau A, Kruszewski G, Lample G, Barrault L, Baroni M (2018) What you can cram into a single \$&!#* vector: probing sentence embeddings for linguistic properties. In: Proceedings of the 56th annual meeting of the association for computational linguistics (volume 1: Long Papers), pp 2126–2136. Association for Computational Linguistics, Melbourne, Australia
8. Hill F, Cho K, Korhonen A (2016) Learning distributed representations of sentences from unlabelled data. In: Proceedings of NAACL-HLT, pp 1367–1377
9. Hill F, Reichart R, Korhonen A (2016) Simlex-999: Evaluating semantic models with (genuine) similarity estimation. Computational Linguistics 00173
10. Kiros R, Zhu Y, Salakhutdinov RR, Zemel R, Urtasun R, Torralba A, Fidler S (2015) Skip-thought vectors. In: Advances in neural information processing systems, pp 3294–3302
11. Le Q, Mikolov T (2014) Distributed representations of sentences and documents. In: International conference on machine learning, pp 1188–1196
12. Luong T, Socher R, Manning CD (2013) Better word representations with recursive neural networks for morphology. In: CoNLL, pp 104–113. 00192
13. Marelli M, Menini S, Baroni M, Bentivogli L, Bernardi R, Zamparelli R (2014) A SICK cure for the evaluation of compositional distributional semantic models. In: LREC, pp 216–223
14. Mikolov T, Chen K, Corrado G, Dean J (2013) Efficient estimation of word representations in vector space. In Proceedings of workshop at ICLR. 03267
15. Pennington J, Socher R, Manning CD (2014) Glove: global vectors for word representation. In: EMNLP, vol 14, pp 1532–1543. 01307
16. Perrotin R, Nasr A, Auguste J (2018) Dialog acts annotations for online chats. In: 25e Conférence Sur Le Traitement Automatique Des Langues Naturelles (TALN). Rennes, France
17. Pragst L, Rach N, Minker W, Ultes S (2018) On the vector representation of utterances in dialogue context. In: LREC
18. Søgaard A (2016) Evaluating word embeddings with fMRI and eye-tracking. ACL 2016, p 116. 00000

End-to-end Modeling for Selection of Utterance Constructional Units via System Internal States

Koki Tanaka, Koji Inoue, Shizuka Nakamura, Katsuya Takanashi, and Tatsuya Kawahara

Abstract In order to make conversational agents or robots conduct human-like behaviors, it is important to design a model of the system internal states. In this paper, we address a model of favorable impression to the dialogue partner. The favorable impression is modeled to change according to user's dialogue behaviors and also affect following dialogue behaviors of the system, specifically selection of utterance constructional units. For this modeling, we propose a hierarchical structure of logistic regression models. First, from the user's dialogue behaviors, the model estimates the level of user's favorable impression to the system and also the level of the user's interest in the current topic. Then, based on the above results, the model predicts the system's favorable impression to the user. Finally, the model determines selection of utterance constructional units in the next system turn. We train each of the logistic regression models individually with a small amount of annotated data of favorable impression. Afterward, the entire multi-layer network is fine-tuned with a larger amount of dialogue behavior data. An experimental result shows that the proposed method achieves higher accuracy on the selection of the utterance constructional units, compared with methods that do not take into account the system internal states.

K. Tanaka · K. Inoue (✉) · S. Nakamura · K. Takanashi · T. Kawahara
Graduate School of Informatics, Kyoto University, Kyoto, Japan
e-mail: inoue@sap.ist.i.kyoto-u.ac.jp

K. Tanaka
e-mail: tanaka@sap.ist.i.kyoto-u.ac.jp

S. Nakamura
e-mail: shizuka@sap.ist.i.kyoto-u.ac.jp

K. Takanashi
e-mail: takanashi@sap.ist.i.kyoto-u.ac.jp

T. Kawahara
e-mail: kawahara@sap.ist.i.kyoto-u.ac.jp

E. Marchi et al. (eds.), *Increasing Naturalness and Flexibility in Spoken Dialogue Interaction*, Lecture Notes in Electrical Engineering 714,
https://doi.org/10.1007/978-981-15-9323-9_2

1 Introduction

It is important for spoken dialogue systems to introduce internal states in order to realize human-like dialogue. By taking into account both input user utterances and system internal states, spoken dialogue systems are expected to generate more human-like natural utterances. Emotion has been considered as an internal state for spoken dialogue systems and virtual agents [2, 3, 13].

We address *favorable impression* to a user as an internal state of the system. We set up a speed-dating dialogue task where a male user talks with a female conversational robot about their profiles. In human-human speed-dating dialogue, their behaviors and attitudes sometimes reflect the degree of favorable impression to their interlocutors [9, 12]. In this study, to express the degree of favorable impression, we propose a dialogue system that selects utterance constructional units, inspired by a series of studies on the discourse analysis [17]. The utterance constructional units contain three parts: *response*, *episode*, and *question*. *Response* is a reaction to the user's utterance, such as feedbacks and answers to questions. *Episode* corresponds to information given by the system such as self-disclosure. *Question* is made by the system toward the user to elaborate the current topic or change the topic. Figure 1 illustrates the main idea of our proposed system. For example, when the degree of favorable impression to the user is high, the system tends to select multiple units such as the combination of *response* and *episode*, or another combination of *response* and *question*, to be more talkative. On the other hand, when the degree is low, the system would select only *response*.

We realize selection of utterance constructional units by a hierarchical structure of logistic regression models. The input is a set of features based on the user's dialogue behaviors. The output is a selection of the utterance constructional units of the next system turn. In the intermediate layer of the hierarchical structure, the degree of favorable impression is represented as an internal state. The proposed model predicts the favorable impression to the user and then the utterance constructional units step by step, where each step is realized with a logistic regression model. We train each logistic regression model with annotated labels of the favorable impression to the user. However, it is difficult to obtain a large number of training labels for the internal states. On the other hand, it is easier to get a large amount of data for the input and output behaviors because these are actual behaviors that can be objectively defined

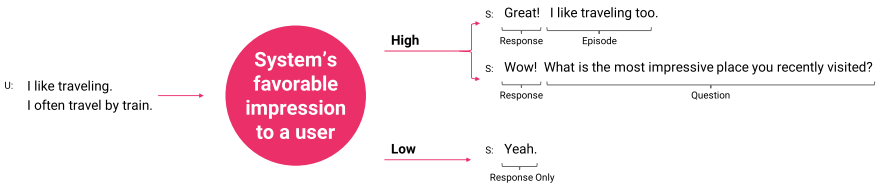

Fig. 1 Main idea of the proposed system that selects the next system utterance based on the system's favorable impression toward the user (U: user, S: system)

and observed in dialogue corpora. In this paper, we also propose an efficient model training to leverage the benefits of making use of internal states. At first, we pre-train each logistic regression model with a small number of training labels of the internal states. We then fine-tune the whole neural network with a larger amount of data of the input and output behaviors in an end-to-end manner. The pre-training captures the internal states, and the end-to-end fine-tuning scales up the amount of training data, which is vital for robust training. This study contributes to realizing dialogue systems that model internal states and also efficient model training where the amount of training data for the internal states is limited.

2 Speed-Dating Human-Robot Dialogue Corpus

In this section, we explain the dialogue data used in this study. We recorded a set of speed-dating dialogues where a male human subject talked with a female humanoid robot that was operated by another female subject. Right after the recording, we took a survey to obtain training labels of the internal states. We also manually annotated the utterance constructional units on the recorded dialogue data.

2.1 Dialogue Data Collection

We have collected a series of speed-dating dialogues between a male subject and a female humanoid robot named ERICA [7, 10]. ERICA was operated by another human subject, called an operator, who was in a remote room. When the operator spoke, the voice was directly played with a speaker placed on ERICA, and the lip and head motion of ERICA was automatically generated [8, 14]. The operator also controlled ERICA's behaviors such as eye-gaze, head nodding, and arm gestures. The snapshot of this data collection is shown in Fig. 2. We recorded 18 dialogue sessions which lasted 10 min and 55 s on average. The human subjects were 18 male university

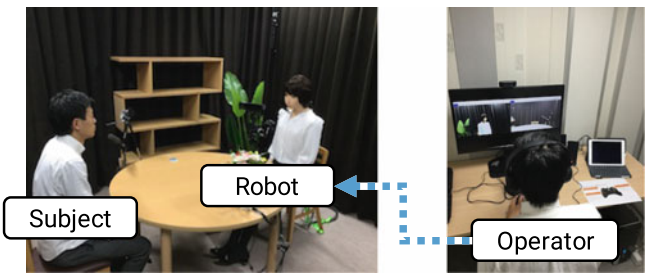

Fig. 2 Snapshot of data collection in WoZ setting

students (both undergraduate and graduated students). The ERICA's operators were 4 actresses whose ages ranged from 20s to 30s. Whereas each human subject participated in only one dialogue session, each ERICA's operator participated in several sessions. They are all native Japanese speakers. We used multimodal sensors that consisted of microphones, a microphone array, RGB cameras, and Kinect v2. We manually annotated utterances, backchannels, laughing, fillers, dialogue turns, and dialogue acts using recommended standards [5].

The dialogue scenarios and instructions are as follows. Since they met each other for the first time, they had to exchange their personal information to know well each other. In advance, we gave the participants a list of conversational topics that are likely to be talked about in first-encounter dialogues, such as hobbies, occupation, and hometown. We then instructed the participants to make a conversation based on the topic list. In the actual dialogue, participants often talked about the topics on the list such as favorite movies, sports, food, and recent trips. For the ERICA's operator, we instructed how to select the utterance constructional units together with the concept of the favorable impression. We asked the operator to select the utterance constructional units based on the degree of her favorable impression to the subject, but we also told that she did not necessarily need to follow this to keep the dialogue natural. We also told that the operator did not need to entertain the subject and the degree of her favorable impression to the subject could be not only positive but also negative.

After each dialogue session, we asked the operator to answer a survey. After the operator listed dialogue topics that they talked about, she rated the following items for each topic on the 7-point scale.

1. Operator's favorable impression to the subject
2. Subject's favorable impression to ERICA estimated by the operator
3. Operator's interest in the topic
4. Subject's interest in the topic estimated by the operator

The favorable impression is represented in one-dimension, positive and negative, as we regard it as a specific indicator in first-encounter dialogue. Although we conducted a similar survey to the male subjects, we used only the survey result from the operators. The reason is that the male subject was a different person on each dialogue session while the operators' survey should be consistent among sessions.

2.2 Analysis

First, we segmented all utterances by dialogue turns. In total, the number of turns of the operators was 899. Then, we manually annotated a set of utterance constructional units for each turn. This annotation was made by one annotator. The distribution of the patterns of utterance constructional units is reported in Table 1. As we see from the table, the majority of the patterns of utterance constructional units was response only (472 samples). Notably, the operators occasionally gave their episode and asked

Table 1 Distribution of the pattern of utterance constructional units

Utterance constructional units			Frequency
Response	*Episode*	*Question*	
✓	–	–	472
✓	✓	–	177
✓	–	✓	86
–	✓	–	69
–	–	✓	53
✓	✓	✓	8
Others			34
Total			899

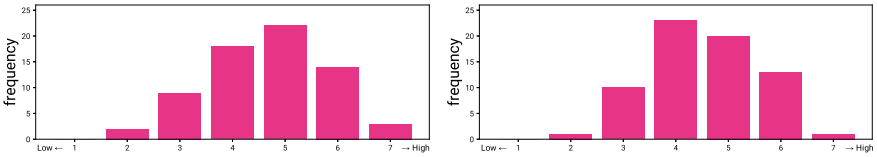

Fig. 3 Distribution of favorable impression reported by ERICA's operators

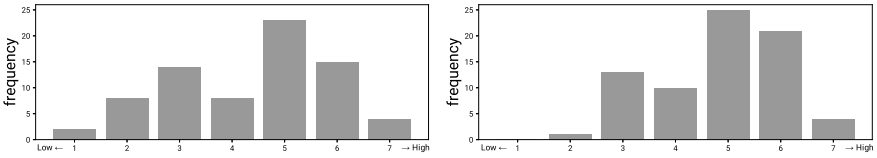

Fig. 4 Distribution of interest reported by ERICA's operators

back questions, but the cases having both an episode and a question was very rare (8 samples). We hypothesize that the operators reflected their favorable impression to the subjects on the utterance constructional units.

We analyzed the survey results from the operators on the following items: (1) operator's favorable impression to the user, (2) subject's favorable impression to ERICA estimated by the operator, (3) operator's interest in each dialogue topic, and (4) subject's interest in each dialogue topic estimated by the operator. The distributions of the four items are plotted in Figs. 3 and 4. The number of dialogue topics was 74 in total. The distributions of interest tended to be more varied than those of favorable impression. This result suggests that the degree of interest more depends on the dialogue topics. On the other hand, this result also suggests that the favorable impression is more stable and gradually changes during the dialogue.

3 Problem Formulation

The task of this study is to select the utterance constructional units of the next system turn based on observed behavior features of the user. The problem formulation is illustrated in Fig. 5. The input feature vector is based on both the speaking and listening behaviors of the user. The speaking behavior feature is extracted during the preceding user turn, referred as $\boldsymbol{o}_s$. The listening behavior feature is computed during the last system turn, referred as $\boldsymbol{o}_l$. We concatenate the behavior feature vectors as:

$$\boldsymbol{o} := (\boldsymbol{o}_s \ , \ \boldsymbol{o}_l) \ . \tag{1}$$

The detail of the feature set is explained in Sect. 5. The output is the pattern of the utterance constructional units that consists of three elements: *response*, *episode*, and *question*. We refer the output as a system action $\boldsymbol{a}$. In this study, we take into account the internal states such as the system's favorable impression to the user. We define the internal states as a vector $\boldsymbol{s}$. In summary, the problem in this study is to predict the next system action $\boldsymbol{a}$ from the observation behaviors $\boldsymbol{o}$ by considering the internal states $\boldsymbol{s}$. This is a typical formulation in conventional studies on spoken dialogue systems where the internal states $\boldsymbol{s}$ correspond to dialogue states of slot filling. In the case of conventional studies such as task-oriented dialogues, the dialogue states were defined clearly and objectively, which makes it easy to collect a large number of training labels for statistical dialogue models such as Markov decision process (MDP) and partially observable Markov decision process (POMDP) [20]. In the current study on the first-encounter dialogue, however, the internal states correspond to states such as favorable impression. These states are ambiguous and subjective,

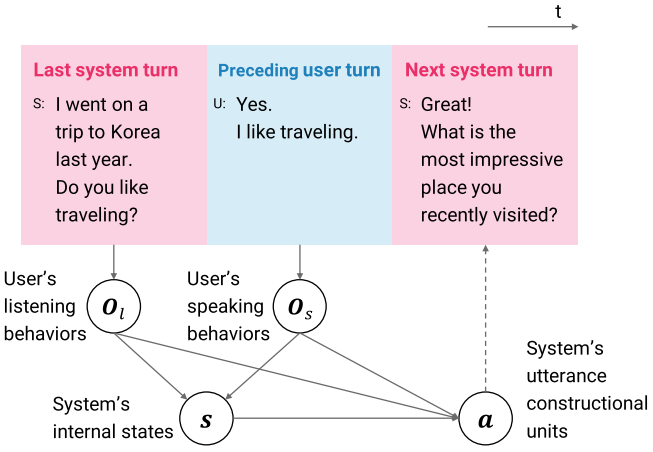

Fig. 5 Problem formulation for considering internal states to select the system next action

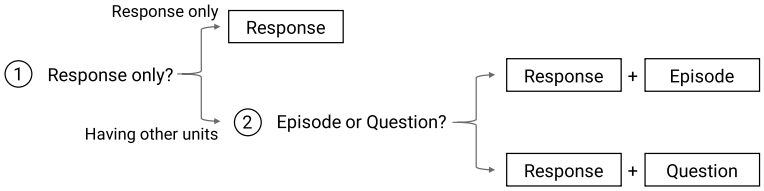

Fig. 6 Taxonomy for selection of the utterance constructional units. The numbers (1 and 2) in the figure correspond to classification tasks

which makes it difficult to prepare a sufficient number of training labels of them. Therefore, we propose efficient end-to-end training by facilitating a small number of labels of the internal states.

Since the distribution of the utterance constructional units is skewed as shown in Table 1, we do not directly select the utterance constructional units. Instead, we divide this problem into the following two sub-tasks. These sub-tasks can be defined as a taxonomy depicted in Fig. 6. The first task is to decide whether the system's turn consists of a response only or have other units (an episode and/or a question). If the decision is the latter case, the system triggers the second task which is to decide whether the system generates an episode or a question. Since we could observe only a few samples where all three utterance constructional units were used at the same time, we do not consider this rare case in the current formulation. In this study, we make the selection model for each task independently, but we combine them to decide the pattern of the utterance constructional units finally. The distribution and definition of labels of the utterance constructional units are summarized in Table 2. The first task corresponds to the selection between the majority pattern and the others. The second task focuses on the remainder steps.

4 End-to-end Modeling Using a Small Number of Labels of Internal States

We take into account the internal states such as favorable impression to the user in order to select the utterance constructional units of the next system turn. However, the number of training labels of the internal states is limited. Actually, in the current study, we could obtain the labels of favorable impression and interest only on each topic, whereas we have to generate the system's action for every turn. This is a universal problem in modeling internal states. On the other hand, we can easily obtain the labels of behaviors such as the observation $\boldsymbol{o}$ and the action $\boldsymbol{a}$ because these behaviors can be objectively observed.

We propose efficient end-to-end modeling for the selection of the utterance constructional units by using a small number of labels of the favorable impression and the interest. The proposed model is based on hierarchical neural networks where