Learning to interrupt the user at the right time in incremental dialogue systems

Adam Chýlek, Jan Švec, and Luboš Šmídl *

NTIS - New Technologies for Information Society
Faculty of Applied Sciences, University of West Bohemia, Czech Republic
{chylek,honzas,smidl}@ntis.zcu.cz

Abstract. Continuous processing of input in incremental dialogue systems might result in the need of interrupting a user's utterance when clarification or rapport is needed. Being able to predict the right time when to interrupt the utterance can be another step to a more human-like dialogue. On the other hand, annotation of corpora with different types of possible interruptions requires additional human resources. In this paper, we discuss how to process a corpus that does not have interruptions specifically annotated. We also present initial experiments on two corpora and show that it is possible to model the desired behaviour from these corpora.

Keywords: incremental dialogue system, model of interruptions, corpora preparation

1 Introduction

Incremental dialogue systems are an important evolution of human-machine interaction [1]. While typical dialogue systems process user's input as a sequence of utterances that are separated by a silence of certain duration, the incremental systems have the ability to process shorter segments [2]. This can improve the user's experience and it can result in a behaviour that is closer to human conversation.

As a toy example and a motivation for this work, we can imagine a spoken dialogue system that has to take a phone number as an input from the user [3]. Let's assume that due to errors in automatic speech recognition (ASR), a noisy

* The final publication is available at Springer via https://doi.org/10.1007/978-3-030-00794-2_54. This work was supported by the European Regional Development Fund under the project Robotics for Industry 4.0 (reg.no.CZ.02.1.01/0.0/0.0/15_003/0000470) and by the grant of the University of West Bohemia, project No. SGS-2016-039. Access to computing and storage facilities owned by parties and projects contributing to the National Grid Infrastructure MetaCentrum provided under the programme "Projects of Large Research, Development, and Innovations Infrastructures" (CESNET LM2015042), is greatly appreciated.

environment or other sources of errors, the system recognizes a different number than what the user has said.

In a non-incremental dialogue system, the user has to wait for the end of the system's utterance. Let's say that the system reads the number back to the user and now she wants to correct the incorrectly recognized number. Depending on the design of a dialogue manager, the user may be able to correct only a part of the number (e.g. "The second digit was nine.") or in the worst case try to say the whole number again.

On the other hand, the incremental system should allow the user to barge in at any time during its turn. The system should remember what has been already said and change only the part of the number that the user is referring to (e.g. the user, after a second digit is read back, says "No, that was nine.").

In an incremental dialogue system, we can also easily imagine the following situation: the user dictates his number, the confidence score from the ASR for the last digit goes below a certain threshold and the system wants to ask the user to repeat that number. In this situation, it is the system that should plan whether and when to interrupt the user in order to get the number right. For example, the system could ask to clarify the second digit of a phone number when the user hesitates after first three digits (e.g. "Was that three one one?", the user could respond "No, three nine one"). Interrupting the user can be useful and it can also be another challenging part of the incremental dialogue systems research.

One could argue that being interrupted by a machine may be annoying, rude or hostile. The authors of [4] show that interruptions (from a linguistic and psychological point of view during human conversations) do not necessarily need to have a negative connotation (deemed as competitive). They can also be used to request clarification, convey rapport or help the other party finish their utterance. These can be classified as collaborative interruptions. A user can interrupt a dialogue system both in the competitive and collaborative fashion. It's up to a dialogue manager's strategy whether its interruptions will stay only collaborative. The definition of such strategy that would not be perceived as competitive is beyond the scope of this paper, as our method will only provide information whether it is the right time to interrupt.

As our research focuses mainly on unimodal spoken dialogue with a single user, we leveraged the availability of large corpora of speech data. We then processed them in a way that allowed us to obtain the training data for supervised learning methods using deep neural networks. Our addition to the existing research is also the description of the corpora preparation. The results of our experiments look promising, we can show that even though the used corpora were not created specifically for the task, it is possible to leverage the available data and predict the interruptions based on a short history of an audio signal.

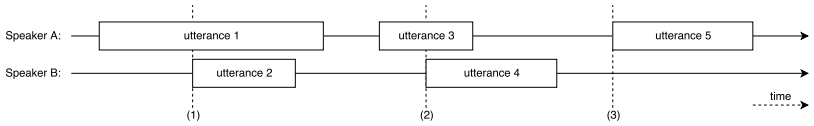

Fig. 1. Example of possible interruption types. The interruption (1) results in an internal overlap (INT), because utterance 2 ends before utterance 1. (2) marks beginning of an overlap that results into switching of speakers (OSW) and (3) marks the time of a clean switch without any overlap (CSW).

2 Related work

Although research that explores interruptions and turn-taking cues already exists, it was done either on a different kind of dialogue or with different intentions than our research (using corpora without explicit annotation of interruptions). For example, turn-taking phenomena were examined in multi-user multimodal systems [5], while our research focuses specifically on a spoken dialogue between two participants. An incremental dialogue system with interruptions presented in [6] focused on the dialogue as a whole and not so much on the mid-utterance interruptions. In [7] a simulated system with interruptions was also evaluated as a whole.

Many researchers also focus on an end-of-turn detection [8–10]. Their methods would allow the dialogue manager to know when the user stopped her thought (e.g. whether a pause in a speech meant an end of a sentence or a hesitation). This would, of course, be a meaningful time for the system to take the ground and start speaking. On the other hand, it may be possible to interrupt the user earlier and our approach should provide this information. In their work, we can also see the prominence of processing either raw audio signals or low-level features and the use of deep learning methods, which we also chose to adapt.

For a theory of interruptions and statistical corpus analysis, we'd like to mention [11] as a related work. Unfortunately, their interruption taxonomy was created with more detailed annotations in mind and their granularity could not be reached in our datasets, where we were missing such annotations. We had to resort to a simpler description of interruptions.

3 Choice of corpora

We have reviewed several speech corpora that we have obtained in the course of several years of ASR development and analyzed their suitability for our current task. Our main requirement was to have a separate channel for each speaker so we can clearly distinguish which speaker is speaking. Additional information about overlapping segments of speech will also allow us to distinguish several interruption types. These requirements resulted in experiments on two corpora: USC-SFI MALACH [12] and BH [13].

The BH corpus (Bezplatné hovory, standing for "free calls" in Czech) contains recordings of spontaneous telephone conversations between pairs of speakers in the Czech language. Each speaker was recorded on a separate channel. In contrast with similar datasets, there were no restrictions on the topics of the calls. The only restriction was the length of the call (10 minutes maximum). A large portion of the dataset has human-created transcriptions of speech, the annotations are time-aligned and assigned to the respective channels. This makes the data ideal for our task, as these conversations were rich in interruptions and overlapping speech. Although neither the interruptions nor the overlaps were specifically annotated, the ground-truth data could be automatically derived from the aligned transcription. The mixture of different topics and speakers could also help us create a more general model of interruptions.

The MALACH corpus contains interviews with holocaust survivors in Czech. The recordings have two channels - one for the interviewer and the other for the interviewee. Several hours of speech from the recordings were transcribed by human annotators. Timestamps of a start and of an end of overlapping segments were available.

We've also made an effort to find other corpora that were not only meeting our criteria, but that were freely available to other researchers. The closest match to our requirements was the CALLFriend corpus [14]. This corpus has timestamps of overlaps, but channel-to-speaker assignment is missing.

The approach for the automatic assignment that was used on MALACH archive could not be used here, because one channel could have more than one speaker assigned, breaking the already fragile automation. Therefore, our approach has not been tested on this corpora, but future work could allow us to test and compare our efforts with others on this dataset in the future.

4 Interruptions from overlaps

For the purpose of our studies, we derive interruption from overlapping speech segments. We defined three types of overlaps: internal overlap, overlap resulting in a new turn and a clean switch of turns, as illustrated in figure 1.

If speaker B starts speaking during speaker A's utterance, but speaker A continues her turn even after the end of the overlapping part, we mark the overlapping event as internal (INT). It is possible that some of the INT overlaps may be just a backchannel (e.g. "ehm"), but the annotated data did not have a consistent format for these events, so we consider them as interruptions. We can reason that even the dialogue system might want to provide a backchannel in its response and this would allow it to know when to do that.

If speaker A ends her utterance during the overlapping segment, we call that event an overlap resulting in a switch of turns (OSW). The last type of event is a clean switch (CSW) when there is no overlap and the other speaker starts the turn.

Only the OSW and INT events will be considered as interruptions. Even though these events may have a different meaning in a conversation [11], we

Table 1. Statistics of the datasets. For MALACH, the information on who continues the utterance was missing, so OSW and INT events could not be distinguished.

•	MALACH	BH
# of speakers	94	8150
# of overlaps (train/test)	665/312	91123/4658
# of OSW overlaps	-	56750/2910
# of INT overlaps	-	34373/1748

have decided we will also evaluate the system without differentiating between the INT and OCW events. We will refer to these simply as overlap (OVR) events. This had to be done because the MALACH corpus does not provide enough information to distinguish the type of an overlap.

From the point of a dialogue manager, we think that the OSW event may be more useful to the system than the INT. When the system decides it would like to interrupt the user, it would be more reasonable to use the moment when the user is more likely to end his turn after the interruption (the OSW type). As for the usefulness of INT, we can consider it as a good moment for a backchannel information.

5 Corpora preparation

The time-aligned annotations of each channel in the BH corpus allow us to clearly detect any overlap of the speakers. We simply take the time of the beginning of one speaker's utterance and check whether it started during the other speaker's utterance. If it did start, we mark it as the beginning of an overlap. When either of the speakers stops speaking, we mark that time as the end of an overlap.

Although the MALACH had overlap annotations for each channel, they were assigned to a recording and not to its channels. This meant that we had to automatically assign who was responsible for the overlap (which channel initiated the interruption). Furthermore, complete textual transcriptions were not available, only an output of phoneme recognizer that did not prove to be useful for this task. The assignment of an overlap to a speaker has been done for each overlap automatically based on average energy levels before the beginning of an overlap. The channel with lower average energy was marked as the initiator of the overlap. We've tried to use the sequences of phonemes for the decision, but from the nature of the recording set-up, one speaker could often be heard on both channels and that introduced errors into the automatic assignment.

The statistics of the datasets created from the corpora are in table 1.

6 Experiment setup

In previous sections, we have shown how we infer the time of interruption from overlaps of different types. Now we can use speech data preceding these timestamps to predict whether it is the right time to interrupt the user.

We have focused our attention on the speech signal itself, as this offers us the chance to process the interruption detection in parallel with automatic speech recognition and it is also a common practice as mentioned in the related work. We always take t seconds of audio preceding each moment of interruption (for both the positive and the negative example) and extract several features that are described in the following paragraphs.

For a feature set that we will call MFCC, we have extracted 12 Mel-Frequency cepstral coefficients from the t seconds of an audio signal with a window length of 50 ms and also the frame's energy.

We have also extracted features using openSMILE and their Interspeech 2009 Emotion Challenge (IS09) feature set [15]. This contains not only the 12 Mel-Frequency cepstral coefficients but also root-mean-square signal frame energy, a frame-based zero-crossing rate of a time signal, a voicing probability from autocorrelation function and a fundamental frequency from the cepstrum. For these low-level descriptors, their moving averages with a window length of 3 were appended as well as first-order delta coefficients of the smoothed descriptors.

Although we had plenty of positive examples of interruptions, obtaining negative examples was a challenge, as these were not annotated.

It is clear that we can't mark everything that wasn't a positive example as a negative example. There could have been many opportunities when the speaker could have been interrupted, but the other party simply chose not to interrupt or did not have a reason to interrupt the speaker. The best way would be to let human annotators mark such examples, but that would defeat the purpose of using the already available datasets without any additional manual work.

To work around this issue, we have made an assumption that the current speaker was purposefully not interrupted during the t seconds preceding the actual interruption, making this our negative example. To reduce the space for the parameter search, we have also used t as the length of the audio segment preceding the interruption, from which the features were extracted. This means that when we see an interruption at time t_p, we generate features from time $t_p - t$ to t_p and mark them as our positive example and the features from $t_p - 2t$ to $t_p - t$ form the negative example.

Another assumption we've made was about the exact moment of an interruption. To compensate for possible annotation error in the range of milliseconds, we have added an offset parameter m. We augmented our data by offsetting the

Table 2. Results for both feature sets.

		IS09				MFCC			
Corpus	Type	accuracy	precision	recall	f-measure	accuracy	precision	recall	f-measure
BH	OVR	0.654	0.748	0.464	0.573	0.645	0.611	0.795	0.691
	OSW	0.693	0.737	0.601	0.662	0.692	0.697	0.682	0.689
	INT	0.639	0.633	0.662	0.647	0.629	0.625	0.649	0.636
MALACH	OVR	0.614	0.628	0.561	0.593	0.587	0.602	0.511	0.553

actual moment of interruption by 1 to m samples and the same principle has been used for the negative examples to keep the classes balanced.

We have conducted initial experiments on the development part of the BH dataset (2656 samples) using several ML methods, including support vector machines, decision trees, and neural networks. We have decided to use deep neural networks for their performance. Specifically, deep residual learning network (ResNet-152, [16]) was used, because it was shown to be performing well not only on image classification tasks but also in automatic speech recognition [17].

The input was normalized, the output was a softmax layer with 2 neurons. We have used categorical cross-entropy as a loss function and Adam as an optimizer. We have used a grid search to find parameters t and m using a development part of the BH dataset.

7 Results

The best performing setup used history of $t = 0.7$ seconds. The training data were augmented with an offset of up to $m = 3$ frames and the negative examples were generated 0.7 seconds before the actual interruption. The accuracy achieved on this setup and other monitored metrics can be seen in the table 2. Although this performance might not be suitable for production systems, it is significantly better than chance (binomial test, $p < 0.005$).

Moreover, we've reasoned in previous sections that we perceive the task of predicting OSW interruption as more valuable to the system. This type was the best performing in terms of accuracy and f-measure. It may be reasonable to pursue only this type of interruptions in further research.

Detecting only the general OVR type of interruption proved to be less useful than anticipated. This means that using datasets similar to MALACH (where we weren't able to automatically distinguish between the more specific OSW and INT interruptions) may not be possible without additional manual work.

The IS09 feature set did not achieve significantly better results (in accuracy). This means that we can use computationally much less expensive MFCC features without risking significantly worse performance.

8 Conclusion

We can conclude that it is possible to predict the right time to interrupt the user even when the source of the data was not intended for this task and segments where speakers overlap were used instead of specific interruption annotations.

The results of initial experiments indicate that making an effort to differentiate between the OSW and INT types of overlaps might result in an improved performance. This differentiation could also be used by different strategies in a dialogue management.

One of the directions of future research to improve the performance is the incorporation of features from an output of a phonetic recognizer or a spoken

language understanding system. Another direction is to train the classifier on one dataset and test it on a different one to see how well can be the interruption prediction generalized.

In a more distant future, analysis of more possible metrics and most importantly their relation to a factual improvement of the dialogue from the user's perspective may be needed.

References

1. Ward, N.G., Devault, D.: Ten Challenges in Highly-Interactive Dialog Systems. AAAI Spring Symposium on Turn-Taking and Coordination in Human-Machine Interaction (2015) 104–107
2. Schlangen, D., Skantze, G.: A general, abstract model of incremental dialogue processing. In: Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics (2009) 710–718
3. Walker, M., Langkilde, I., Wright, J., Gorin, A., Litman, D.: Learning to Predict Problematic Situations in a Spoken Dialogue System: Experiments with How May I Help You? Proceedings of the 1st North American chapter of the Association for Computational Linguistics conference (2000) 210–217
4. Li, H.Z.: Cooperative and Intrusive Interruptions in Inter- and Intracultural Dyadic Discourse. Journal of Language and Social Psychology **20**(3) (2001) 259–284
5. Skantze, G., Johansson, M., Beskow, J.: Exploring Turn-taking Cues in Multi-party Human-robot Discussions about Objects. Proceedings of the 2015 ACM on International Conference on Multimodal Interaction (2014) 67–74
6. Zhao, T., Black, A.W., Eskenazi, M.: An Incremental Turn-Taking Model with Active System Barge-in for Spoken Dialog Systems. 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue (September) (2015) 42–50
7. Khouzaimi, H., Laroche, R., Evre, F.: Turn-taking phenomena in incremental dialogue systems. In: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. Number September (2015) 1890–1895
8. Heeman, P.A., Lunsford, R.: Turn-Taking Offsets and Dialogue Context. Interspeech2017 (2017) 1671–1675
9. Masumura, R., Asami, T., Masataki, H., Ishii, R., Higashinaka, R.: Online End-of-Turn Detection from Speech based on Stacked Time-Asynchronous Sequential Networks. Interspeech2017 (2017) 1661–1665
10. Maier, A., Hough, J., Schlangen, D.: Towards Deep End-of-Turn Prediction for Situated Spoken Dialogue Systems. Interspeech2017 (2017) 1676–1680
11. Gravano, A., Hirschberg, J.: A Corpus-Based Study of Interruptions in Spoken Dialogue. Interspeech-2012 (2012)
12. Psutka, J., Radová, V., Ircing, P., Matoušek, J., Müller, L.: USC-SFI MALACH Interviews and Transcripts Czech LDC2014S04 (2014)
13. Valenta, T., Šmídl, L., Švec, J., Soutner, D.: Inter-annotator agreement on spontaneous Czech language: Limits of automatic speech recognition accuracy. In: Proceedings of 17th International Conference TSD 2014, Brno, Czech Republic. Volume 8655 LNAI. (2014) 390–397
14. Canavan, A., Zipperlen, G.: Callfriend american english-non-southern dialect. Linguistic Data Consortium, Philadelphia **10** (1996) 1
15. Schuller, B., Steidl, S., Batliner, A.: The INTERSPEECH 2009 emotion challenge. Interspeech (2009) 312–315

16. He, K., Zhang, X., Ren, S., Sun, J.: Deep Residual Learning for Image Recognition. Multimedia Tools and Applications (dec 2015) 1–17
17. Xiong, W., Droppo, J., Huang, X., Seide, F., Seltzer, M., Stolcke, A., Yu, D., Zweig, G.: The microsoft 2016 conversational speech recognition system. ICASSP 2017 5255–5259