Are fillers helpful to fulfill a flawless classification of Dialog Acts?

Gabriel KOUTCHINSKY[*] and **Mohamed BEN LASSOUAD**[†]

Abstract

The ability to classify Dialog Acts (DA) from a conversation can revolutionize our understanding of conversations and enable bots to respond appropriately. In language learning, it is essential to recognize the fillers and expressions used by speakers to buy time while they think about what to say next and how to articulate it in order to sound fluent. All languages, including sign languages, have their own unique fillers, and spotting them seems straightforward for a learning algorithm using deterministic rules. However, can we predict Dialog Acts by analyzing neighboring utterances of fillers? Our research suggests that the benefit gained from this approach is minimal and that utterances expressed after the fillers are more informative than those before the fillers.

1 Introduction

1.1 Dialog Acts (DA)

Dialog act classification is a crucial task in the field of Natural Language Processing (NLP), which involves recognizing the intended meaning or purpose of a user's statement in a conversation. With the increasing popularity of chatbots and conversational agents such as ChatGpt, YouChat, and ChatSonic, there has been a surge in research and development of algorithms and machine learning models for automated DA classification (Godfrey et al., 1992; Li et al., 2017; Leech and Weisser, 2003; Busso et al., 2008; Passonneau and Sachar., 2014; Thompson et al., 1993; Poria et al., 2018; Shriberg et al., 2004; Mckeown et al., 2013). This interest in DA classification is driven by the need to improve the conversational ability and responsiveness of chatbots (Colombo* et al., 2019; Jalalzai* et al., 2020) and virtual assistants, making them more effective in helping users achieve their goals. As such, DA classification is becoming a critical component of NLP systems that aim to provide more natural and efficient communication between humans and machines.

This task is essential for building smarter and more efficient dialogue systems that can understand and respond to human language more accurately. In this article, we focus on the role of fillers (e.g. "um" or "uh") in understanding conversations. Therefore, the possibility of having improvements in the modeling of spoken language. For this, we use a database of MRDAs labeled with their dialogue act.

A worthwhile refinement of input context-based classification is the modeling of inter-tag dependencies. This task is approached as sequence-based classification where output tags are considered as a DA sequence. In this perspective, we try to emphasize the role of fillers in improving language modeling speech and the classification of dialogue acts. As fillers contain useful information (i.e. the need to halt the flow of speaking) that can be exploited by deep contextualized integrations to better model spoken language, we try to use neural network models to predict speech acts.

[*] Gabriel KOUTCHINSKY
gabriel.koutchinsky@ensae.fr
[†] Mohamed BEN LASSOUAD
mohamed.benlassouad@ensae.fr

1.2 The role of fillers

Indeed, fillers (or *disfluencies*) are part of speech used to fill someone's speech while they are thinking about a proper way to put their following sentences. We can consider two types of fillers (Jonsson and Thyberg, 2016) : lexical fillers made of existing words such as "*you know*" or "*well*" and non-lexical ones, which mainly consist of onomatopeia like "*uhhh*" or "*ummm*". Let's notice that all onomatopeias cannot be considered to be fillers because a speaker exclaiming "*Oh!*" gives an information about his surprise, sadness or anger, whereas fillers are not supposed to carry any other meaning than the speaker searching for words. As those uninformative parts of speech seem pretty easy to spot and classify, we are wondering whether learning on utterances just before and just after filler-only utterance could help get a more accurate DA classifier.

2 Background

2.1 Existing research

Several approaches have been proposed to tackle the DA ranking problem ((Chapuis* et al., 2020), (Colombo* et al., 2020)). These methods can be divided into two different categories. The first works addressed the problem of sequence labeling as an independent classification of each statement. Among the techniques used for this type of classification: the Bayesian network (Simon Keizer, 2002), HMM ((Andreas Stolcke, 2000)) and SVM ((Dinoj Surendran, 2006)) and labels. These methods require large corpora to train models from scratch, such as: Switchboard Dialog Act (SwDA) ((J.J. Godfrey, 1992)), Meeting Recorder Dialog Act (MRDA) ((Elizabeth Shriberg, 2004)), Daily Dialog Act ((Yanran Li, 2017)), HCRC Map Corpus of Tasks (MT) ((Henry S. Thompson, 1993)). This makes it more difficult for them to be adopted by smaller datasets, such as: human-human dialogue corpus.

Several research papers have been used to investigate the use of fillers, such as "um" and "uh" as classification features of dialogue acts. (Dinkar* et al., 2020) conduct experiments on two datasets of spoken language transcriptions and analyze the performance of various text representation methods with and without the inclusion of fillers. The results show that the incorporation of fillers can improve the results, both during language modeling and on a downstream task (Feeling Of Another's Knowing and position prediction).

2.2 The limitations of current models

Any DA classification model see fillers as a type of dialog act like any other. As it represents a break between what was being said before and what the speaker is going to tell, we had the idea of studying utterances preceding and following fillers in order to predict their Dialog Acts.

3 Problem statement

We start by formally defining the Sequence Labelling Problem using the notations from (Colombo, 2021). At the highest level, we have a set $D = (C_1, C_2, \ldots, C_{|D|})$ of conversations composed of utterances, i.e. $C_i = (u_1, u_2, \ldots, u_{|C_i|})$, with $Y_i = (y_1, y_2, \ldots, y_{|C_i|})$ being the corresponding set of Dialog Acts.

At a lower level each conversation is composed of utterances, i.e. $C_i = (u_1, u_2, \ldots, u_{|C_i|})$ with being the corresponding sequence of labels: each u_i is associated with a unique label $y_i \in \mathcal{Y}$, where $\mathcal{Y}$ is the set of all the possible dialogue acts. At the lowest level, each utterance can be seen as a sequence of words, i.e. $u_i = (\omega_1^i, \omega_2^i, \ldots, \omega_{|u_i|}^i)$.

Here we take, for each $C_i \in D$, a subset F_i corresponding to all fillers, i.e. $F_i = (u_{z_{i,1}}, u_{z_{i,2}}, \ldots, u_{z_{i,|F_i|}}$ with $\forall j \in [1, |F_i|], u_{z_{i,j}} = F$, F corresponding to the DA "Filler".

We try to predict the subsets Y_{be} and Y_{af}, with $Y_{be} = (y_{z_{1,1}-1}, y_{z_{1,2}-1}, \ldots, y_{z_{1,|F_i|}-1}, \ldots, y_{z_{|D|,1}-1}, y_{z_{|D|,2}-1}, \ldots, y_{z_{|D|,|F_i|}-1})$ and $Y_{af} = (y_{z_{1,1}+1}, y_{z_{1,2}+1}, \ldots, y_{z_{1,|F_i|}+1}, \ldots, y_{z_{|D|,1}+1}, y_{z_{|D|,2}+1}, \ldots, y_{z_{|D|,|F_i|}+1})$ (considering the absence of side values).

4 Experiments Protocol

4.1 Data Description

The dataset is formed from discussions of SwitchBoard corpus ((J.J. Godfrey, 1992)) and MRDA corpus ((Elizabeth Shriberg, 2004)). The first one consists of a pool of telephone conversations between American English speakers. This database has previously been annoted with dialog acts. The second dataset, the Meeting Recorder Dialogue Act Corpus, retranscripts and classifies 75 hours of conversation from meetings among 53 speakers. As the DA classes differ from one database to another, we decide to retreat those data in order to unify them. For instance, the SwitchBoard classes *Statement-non-opinion* and *Statement-opinion* were put together into one big *Statement* class.

Then, we create one dataframe containing utterances before a filler (including fillers when they preceeded another filler) and another dataframe with utterances after those fillers. Finally, in order to get a more accurate algorithm, we kept only the 6 main classes in terms of number of values in the datasets.

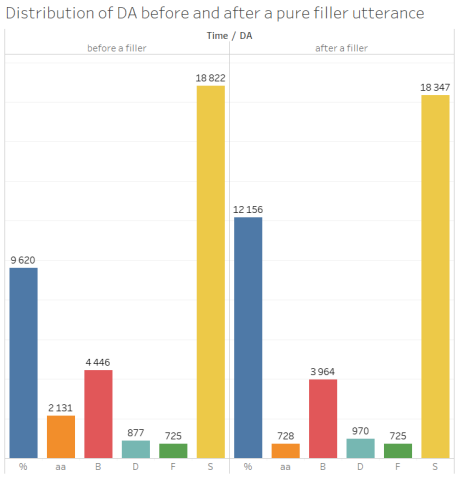

Figure 1: The distribution of the dialogue acts before and after the intervention of the filler.

This figure 1 represents the distribution of the dialogue acts before and after the intervention of the fillers. The dialogue acts correspond to the following intentions:

- S = Statement,
- F = Filler,
- D = Disruption is mainly composed with utterances abandoned or interrupted before we could notice their types,
- B = Backchannel which corresponds to interjections made by speaker S2 to a main speech by speaker S1 in order to testify S2 gives full attention to S1. For example, a laugh or a "I see" answer are backchanneling,
- aa = Agree/Accept,
- % = Uninterpretable.

We can conjecture that before a filler, there are more parasitic elements (%) and less agree/accept than after a filler, maybe because before a typical use case of filler is, after agreeing a proposal, looking a way to qualify one's stance.

4.2 Metric used

As our distribution shows inequal repartition of data, accuracy shall not be appropriate to evaluate the performance of algorithms. Hence our choice to take the recall of the prediction, which corresponds to the proportion of true positives (i.e., correctly classified positive instances) out of all actual positive instances.

As we are in a multiclass problem, we have one recall by class and we decide to take the mere average of our recalls. As we have kept 6 classes, a random classifier would obtain 16.67% recall so let's see if we can beat that number!

4.3 Data Encoding

The first step in NLP is to turn the language into vectors. To do this, several models are used (Word2vec, BERT, etc.). Among the most used models, BERT (Bidirectional Encoder Representations from Transformers) which is a language model developed by Google in 2018. This method has significantly improved performance in automatic language processing. (Jacob Devlin, 2018)

5 Results

# of epochs	2	3	Random algo
Before	16.9	18.0	16.7
After	19.8	19.4	16.7

Table 1: The best results obtained with the same conditions for algorithm before after fillers.

Our results may seem disappointing but at least we have found something pretty interesting: our models always predict better DA of utterances after fillers than the ones before fillers. The results we get are only slightly better than the 16.67% corresponding to random classification. That means our algorithm makes it possible to learn from the data we give it.

The fact that post-filler data gets better results may come from the very distribution of labels inside the dataset. Indeed, as the part of uninterpretable data is larger in post-filler than in pre-filler data, our algorithm may have predicted well more of this class.

We were really surprise by the fact that, for post-filler data, our recall got better with 2 epochs than with 3 but that can result from our model overfitting the training dataset and getting thus weaker results of testing dataset.

6 Discussion/Conclusion

In conclusion, the impact of the incorporation of fillers in the classification of dialogue acts is quite negligible. However, the use of fillers is a promising approach to understand the intention of the interlocutor in a conversation and to improve natural language processing systems in dialog act classification tasks.

It would be interesting to conduct this kind of study after setting apart lexical fillers and non-lexical fillers because the first category looks even easier to spot in order to apprehend better the structure gravitating near fillers.

Fillers really look like the untapped depths of NLP and a better understanding of them could also help develop a conversational agent able to analyze non-verbal elements of speech as fillers represent the main bridge from verbal to non-verbal behaviours.

All that work may also become easier if the way to retranscript fillers was standardized using a rule about, for instance, how long a holding of filler could be reflected in the number of consonants in "*Uhhh*" or "*Ummm*", in order to be able to differentiate "*Uhh*" from "*Uhhhh*". In the future, it would be interested to extend this study on multilingual spoken datasets (Garcia* et al., 2019; Colombo et al., 2021).

References

John J. Godfrey, Edward C. Holliman, and Jane McDaniel. 1992. Switchboard: Telephone speech corpus for research and development. In *Proceedings of the 1992 IEEE International Conference on Acoustics, Speech and Signal Processing - Volume 1*, ICASSP'92, page 517–520, USA. IEEE Computer Society.

J. McDaniel J.J. Godfrey, E.C. Holliman. 1992. Switchboard: telephone speech corpus for research and development.

Henry Thompson, Anne Anderson, Ellen Bard, Gwyneth Doherty-Sneddon, Alison Newlands, and Cathy Sotillo. 1993. The hcrc map task corpus: natural dialogue for speech recognition.

Ellen Gurman Bard Gwyneth Doherty-Sneddon Alison Newlands Cathy Sotillo Henry S. Thompson, Anne Anderson. 1993. The hcrc map task corpus: Natural dialogue for speech recognition.

Noah Coccaro Elizabeth Shriberg Rebecca Bates-Daniel Jurafsky Paul Taylor Rachel Martin Carol Van Ess-Dykema Marie Meteer Andreas Stolcke, Klaus Ries. 2000. Dialogue act modeling for automatic tagging and recognition of conversational speech.

Anton Nijholt Simon Keizer, Rieks op den Akker. 2002. Dialogue act recognition with Bayesian networks for Dutch dialogues.

Geoffrey Leech and Martin Weisser. 2003. Generic speech act annotation for task-oriented dialogues.

Elizabeth Shriberg, Raj Dhillon, Sonali Bhagat, Jeremy Ang, and Hannah Carvey. 2004. The ICSI meeting recorder dialog act (MRDA) corpus. In *Proceedings of the 5th SIGdial Workshop on Discourse and Dialogue at HLT-NAACL 2004*, pages 97–100, Cambridge, Massachusetts, USA. Association for Computational Linguistics.

Sonali Bhagat Jeremy Ang Hannah Carvey Elizabeth Shriberg, Raj Dhillon. 2004. The icsi meeting recorder dialog act (mrda) corpus.

Gina-Anne Levow Dinoj Surendran. 2006. Additional cues for mandarin tone recognition.

Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower Provost, Samuel Kim, Jeannette Chang, Sungbok Lee, and Shrikanth Narayanan. 2008. Iemocap: Interactive emotional dyadic motion capture database. *Language Resources and Evaluation*, 42:335–359.

Gary Mckeown, Michel Valstar, Roddy Cowie, Maja Pantic, and M. Schroder. 2013. The semaine database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent. *Affective Computing, IEEE Transactions on*, 3:5–17.

R. Passonneau and E. Sachar. 2014. Loqui human-human dialogue corpus (transcriptions and annotations).

Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. Dailydialog: A manually labelled multi-turn dialogue dataset.

Xiaoyu Shen Wenjie Li Ziqiang Cao-Shuzi Niu Yanran Li, Hui Su. 2017. Dailydialog: A manually labelled multi-turn dialogue dataset.

Kenton Lee Kristina Toutanova Jacob Devlin, Ming-Wei Chang. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. 2018. Meld: A multimodal multi-party dataset for emotion recognition in conversations.

Pierre Colombo*, Wojciech Witon*, Ashutosh Modi, James Kennedy, and Mubbasir Kapadia. 2019. Affect-driven dialog generation. *NAACL 2019*.

Alexandre Garcia*, Pierre Colombo*, Slim Essid, Florence d'Alché Buc, and Chloé Clavel. 2019. From the token to the review: A hierarchical multimodal approach to opinion mining. *EMNLP 2019*.

Pierre Colombo*, Emile Chapuis*, Matteo Manica, Emmanuel Vignon, Giovanna Varni, and Chloe Clavel. 2020. Guiding attention in sequence-to-sequence models for dialogue act prediction. *AAAI 2020*.

Hamid Jalalzai*, Pierre Colombo*, Chloé Clavel, Éric Gaussier, Giovanna Varni, Emmanuel Vignon, and Anne Sabourin. 2020. Heavy-tailed representations, text polarity classification & data augmentation. *NeurIPS 2020*.

Emile Chapuis*, Pierre Colombo*, Matteo Manica, Matthieu Labeau, and Chloe Clavel. 2020. Hierarchical pre-training for sequence labelling in spoken dialog. *Finding of EMNLP 2020*.

Tanvi Dinkar*, Pierre Colombo*, Matthieu Labeau, and Chloé Clavel. 2020. The importance of fillers for text representations of speech transcripts. *EMNLP 2020*.

Pierre Colombo, Emile Chapuis, Matthieu Labeau, and Chloe Clavel. 2021. Improving multimodal fusion via mutual dependency maximisation. *EMNLP 2021*.

Pierre Colombo. 2021. *Learning to represent and generate text using information measures*. Ph.D. thesis, (PhD thesis) Institut polytechnique de Paris.