Highlighting Psychological Features for Predicting Child Interjections During Story Telling

Gaël Lejeune, François Rioult, Bruno Crémilleux

INTERSPEECH 2016
September 8–12, 2016, San Francisco, USA

Highlighting Psychological Features for Predicting Child Interjections During Story Telling

Gaël Lejeune, François Rioult, Bruno Crémilleux

Normandie Univ, UNICAEN, ENSICAEN, CNRS, GREYC, 14000 Caen, France
firstname.lastname@unicaen.fr

Abstract

Conversational agents are more and more investigated by the community but their ability to keep the user committed in the interaction is limited. Predicting the behavior of children in a human-machine interaction setting is a key issue for the success of narrative conversational agents. In this paper, we investigate solutions to evaluate the child's commitment in the story and to detect when the child is likely to react during the story. We show that the conversational agent cannot solely count on questions and requests for attention to stimulate the child. We assess how (1) psychological features allow to improve the prediction of children interjections and how (2) exploiting these features with Pattern Mining techniques offers better results. Experiments show that psychological features improves the predictions and furthermore help to produce robust dialog models.

Index Terms: Conversational Agents, Event Prediction, Knowledge Discovery, Emotion Modelling

1. Introduction

As evidenced in [1], there is a strong interest in the Human Computer Interaction (HCI) community for improving the interaction between users and agents. In particular, this is useful to favor specific conversational settings such as online assistants [2], companionship for elderly care [3] or conversational toys [4]. This involves a better understanding of the user mental state in order to lessen the cognitive cost for the user and simultaneously improve the user satisfaction. A conversational agent should be able to maintain the user's commitment in the interaction and to detect when it is necessary to rekindle the conversation or, to the contrary, when the user wants to remain passive. The key issue is to avoid a rigid turn-taking scheme [5] so that the user has an impression of naturalness.

This article studies the specific case of interaction between a Conversational Agent (CA) and a child in a narrative environment. The objective is to keep the children committed in the story by improving the way the information is conveyed and how the user behavior is taken into account. This requires an appropriate model of dialog. Such a model offers the CA various options to measure and to influence the commitment of the child in the dialog. The CA can guide the dialog in order to favor or avoid child interjections, for instance to respect assigned objectives such as bringing the story to its end or put the listener in an active position. Furthermore, it allows the CA to measure how much it should wait for an answer to a direct question or a check for attention. If the model tells the CA that the child is very likely to speak, it is worthy to wait a few seconds. On the contrary, such a pause would be awkward if it is not followed by an interjection of the child. This can worsen the quality of the interaction.

We propose to tackle this task as a classification problem where the objective is, given an history of utterances (or breath groups), to predict who the next speaker will be. The proposed framework involves Pattern Mining and Machine Learning. It aims to accurately predict children interjections during the story using psychological features. Two phenomena show that predicting child interjection is an interesting task and that its resolution involves much more than obvious strategies. Firstly, children do not systematically react to solicitations. Secondly, they often interrupt the story teller without being solicited. Therefore, there is a need for more accurate strategies for modelling interaction. We show that pattern mining improves the dialog models and favors a better interaction between the child and the adult.

Related works on the subject are elaborated in Section 2, the dataset and the proposed methods are presented in Section 3. Finally, results are presented in Section 4 and we propose a discussion about our results and future works in Section 5.

2. Related work

Conversational agents (CA) are used for a wide range of applications. The main issue for these systems is to keep the user involved in the interaction [6]. Two different strategies can be identified to improve the quality of the interaction. The first one is to work on the context in order to obtain a setting comparable to a dialog between humans [7]. This objective may be achieved by improving the prosody of the system [8] or by embodying the CA [4]. The other solution is to improve the content: what is said and how it is said. That is, relying on the conveyed information to keep the listener involved in the interaction. Relying on a basic dialog model can lessen the attention of the listeners quickly by giving them the impression that they have to fill the gaps in order to keep the dialog in a very linear scheme.

The most obvious attempt to create a dialog model comes from keyword-based systems such as SEMAINE [9]. The linearity of such systems is a strong disadvantage and this kind of strategy fails to represent the various dimensions involved by each dialog utterance [10]. Dialog models have been created with rule-based approaches as developed by [11] or with statistical approaches [12]. Recent works in the field have focused on stochastic or semi stochastic models [13] and machine learning approaches including deep learning [14] and reinforcement learning [15]. Most models are based on the *Markov Decision Process* [16] but they require an explicit representation of the mental states which make them prone to rigid turn-taking results.

We propose in Section 3 a rule-based and a pattern mining approach combined with machine learning to predict child interjection during the story telling. An hybridisation of the two methods is proposed as well.

http://dx.doi.org/10.21437/Interspeech.2016-527

3. Predicting child interjections

The task tackled in this article can be defined as follows: given the history of a dialog between a story teller (human or CA) and a listener (child), how likely will the next speaker be the child. The motivation is two folds. First of all, such a model offers the CA various options to measure and to influence the commitment of the child in the dialog. The CA can guide the dialog in order to favor or avoid child interjections, for instance to respect assigned objectives such as bring the story to its end or put the listener in an active position. Secondly, it allows the CA to measure how much it should wait for an answer to a direct question or a check for attention. If the model tells the CA that the child is very likely to speak, it is worthy to wait a few seconds. To the contrary, such a pause would be awkward if it is not followed by an interjection of the child. This can worsen the quality of the interaction [17].

In this section, we first present the dataset constructed for evaluation. Then, we describe a set of baselines and three original methods. Finally, we elaborate our learning framework.

3.1. Dataset

The dataset used for the experiments has been constituted from manual retranscriptions of a story told to 38 different children. The child is alone in the room and he follows the story using a computer screen and a microphone. Two narrators are involved: an adult and a wizard-of-oz. No instructions are given to the children so that they are free to interject whenever they want. In the introduction of the story, the adult warns the child that he will not be able to tell the story until its end and that he will have to be replaced. In the middle of the story, the first narrator leaves and the end of the story is told in a Wizard-of-Oz configuration. Each dialog has been manually segmented into breath groups. Statistics regarding the dataset are presented in Table 1.

	#BG	CBG	# W	#C	W/BG	C/BG
min	121.0	4%	1141.0	5733	5.84	28.6
max	249.0	26%	1528.0	7742	9.6	49.6
avg.	166.2	9%	1237.5	6261	7.5	38.1
stdev.	23.9	7	90.6	445.5	0.8	4.2

Table 1: Corpus Statistics, number of Breath Groups (#BG), proportion of BG from the child (CBG), length of the dialogs in words (#W) and characters (#C), average length of Breath Groups in words (W/BG) and characters (C/BG).

This corpus has been annotated by an expert (psychologist) with psychological features [18]. The annotation has been performed using the DIT++ framework (Dynamic Interpretation Theory release 5 [19]). Each breath group has been annotated with at least a Function-Dimension pair.Functions and Ddimensions are defined as follows:

(1) Function: informative (requiring or giving) or active (doing or requesting) intent of the speaker;

(2) Dimension: communicative contribution of the speaker in the conversation (task, feedback, turn management...).

Table 2 exhibits a sample of the corpus. Each line forms a breath group which is defined as a succession of words uttered without pause [20]. It corresponds to a part of a sentence, or one or more complete sentences. This definition offers an analysis unit more suitable than the sentence for a real-world dialog.

Speaker	Breath Group	Function (Dimension)
A	So, it's morning, chidren are coming to school	Inform (Task)
C	**Yes**	Contact (CM)
A	Look at this child	Suggestion (Task)
A	he does not seem happy to be there	Inform (Task)
C	**Yeah, I saw**	Confirm (Task)
A	And this boy, he has a ball...	Inform (Task)
A	Look,	Suggestion (Task)
A	it's Salim, he is calling his friends, [...]	Inform (Task)
C	**Uh Uh**	Stalling (Time)

Table 2: Three sequences ended by a child interjection : each line is a breath group, child (C) interjections are in bold.

-	-	A (Inform)	C (Contact)
-	A (Suggestion)	A (Inform)	C (Confirm)
A (Inform)	A (Suggestion)	A (Inform)	C (Stalling)

Table 3: Extraction from Table 2 of all Speaker (Function) sequences ended by a child interjection, each cell is a breath-group.

3.2. Baselines

The intuition behind the baselines is that a child generally reacts to solicitations. Two types of solicitations are exploited:

Direct Question (DQ) The CA asks the child to give a piece of information, this can be a propositional question or a WH-question

Check for Attention (CfA) The CA requires the child to confirm that the interaction is still going normally (e.g. "Is it OK for you?")

This leads to three rule-based methods relying on the state of the current Breath-Group.

1. $DQ \rightarrow ChildInterjection$
2. $CfA \rightarrow ChildInterjection$
3. DQ OR $CfA \rightarrow ChildInterjection$

3.3. Direct Method

This method more deeply exploits the mental states annotations (psychological features) present in the corpus. The idea is quite simple: predicting the next speaker according to the annotations of the current breath-group. All the psychological features for a given breath group are used to predict the next speaker. The feature value stores the presence or absence of a given mental state in the current breath group. This method is expected to have better results than the baseline (more features are exploited) while keeping simplicity and interpretability.

In order to ease the reading, the given examples exploit only the functions of the DIT++ scheme. Table 3 presents sequences of psychological features extracted from Table 2. The direct method only exploits the attributes of the current breath group to predict who the next speaker will be. From these three sequences, the direct method extracts three candidate rules:

- $A(Inform)A$
- $A(Suggestion)A$
- $A(Inform)C$

The last rule is the only one that allows to detect child interjection. One can see that this particular case where the child interjects the adult could not be handled by the baselines.

3.4. Pattern Mining method

With the baselines and the direct method, only the annotations of the current breath-group are exploited to predict whether the child will be the next speaker. We propose to use pattern mining for extracting clues involving previous breath-groups. The objective is to build a strategy for the CA: what psychological features have to be activated in order to keep the child involved in the interaction and having a breeding ground for a future child interjection. We want to extract a reasonable amount of interesting patterns in order to favor similarities. With these objectives in mind, frequent closed patterns are computed using the algorithm proposed by Ukkonen [21]. These patterns have the following characteristics:

frequent : the pattern is supported by at least $minsup$ dialogs

closed : the pattern is not strictly included in a pattern supported by the same dialogs

The $minsup$ is set to 2 and no constraint is given on the length of the patterns. These patterns are used as features to train classifiers. If a pattern is closed (i.e. ends) in a given breath-group, the corresponding feature has a value of 1 (0 otherwise). From the examples in Table 3, this method extracts three patterns:

- $sup = 4$: $A(Inform)$
- $sup = 3$: $A(Inform)C$
- $sup = 2$: $A(Suggestion)A(Inform)C$

$A(Suggestion)A$ is not a closed pattern since it is always included in the pattern $A(Suggestion)A(Inform)C$.

3.5. Hybridisation

Closed patterns have the property of reducing redundancy but this can be a drawback for classification as shown by Brixtel [22]. The hybridisation aims to keep the good properties of the direct method and the pattern mining method. The features of the two methods are merged in order to train the classifiers. No weights are given to features but this setting gives a greater importance to short patterns since they can be detected with both methods. In our example, the pattern $A(Inform)C$ is extracted by both methods.

3.6. Learning framework

All the features used to build the classifiers are binary features: the pattern is present or absent in a given breath-group. A ten-fold cross validation has been performed to ensure results robustness. For all of our experiment, the WEKA implementation of all these classifiers[1] has been exploited. Since the interpretability of the produced model is a key issue, a focus has been given to decision trees. We also applied an ensemble learning method, random forests, in order to see if how much room for improvement we have. Other type of classifiers (SVM, bayesian networks, neural networks) as well as boosting method (Adaboost) have been tested. Only the interesting results will be reported in the next section.

[1] `http://www.cs.waikato.ac.nz/ml/weka/`

Method	Variant	P	R	F_1	$F_{0.5}$
Baselines	DQ	65.5	44.2	52.8	59.8
	CfA	83.1	15.5	26.2	44.4
	DQ OR CfA	69.3	59.7	64.2	67.2
Direct Method	NaiveBayes	67.5	64.6	66.0	66.9
	SMO	69.2	70.1	69.7	69.4
	C4.5 Tree	69.0	69.5	69.2	69.6
	Random Forest	69.2	67.6	68.4	68.0
Pattern Mining	NaiveBayes	66.7	65.1	65.9	66.4
	SMO	71.3	62.0	66.3	69.2
	C4.5 Tree	**76.1**	61.2	67.9	72.6
	Random Forest	69.5	60.4	64.6	67.5
Hybridisation	NaiveBayes	68.8	70.2	69.5	69.1
	SMO	72.6	70.1	71.3	72.1
	C4.5 Tree	71.1	70.7	70.9	71.0
	Random Forest	74.1	**71.5**	**72.8**	**73.6**

Table 4: Classification results for the three baselines, the direct method and the Pattern Mining method.

4. Results

Table 4 exhibits the results for the classification task of the baseline and the proposed methods. The positive class is the child interjection which is also the minority class. True Positives (TPs) are correctly predicted child interjections. False Positives (FPs) occur when a child interjection is wrongly predicted. False Negatives (FN) occur when there is an unpredicted child interjection. With these measures we compute Precision (P), Recall (R) and F_β-measure as follows:

- Precision: $P = TP/(TP + FP)$
- Recall: $R = TP/(TP + FN)$
- F-measure: $F_\beta = (1 + \beta^2) * \frac{P*R}{(\beta^2*P)+R}$

The F-measure is computed with the classical setting $\beta = 1$ and also with $\beta = 0.5$ in order to favour precision.

First of all, the baseline does not give as much precision as expected because children do not always react to a DQ. A CfA is more reliable but since it is more seldom used by humans, one can expect that its effect could drop if it is overused. Indeed, it can lead to a poor quality interaction. Even when the two rules are combined, the recall is still bad. As evidenced by [23], for rare events recall is more important than precision. Thus, this is precisely where our methods should give an added value.

The direct method improves the results in terms of recall, it shows that not only questions and checks for attention lead to child interjection. The pattern mining method improves precision, showing that in many cases what psychological features are activated before a direct question is of great importance. It shows that there is room for richer strategies from the CA where not only the current breath group is considered. The combination of the two methods gives slightly better results. The main reason is that patterns of length 1 count twice since they are features in both methods. Thus, they are twice more likely to be used by the random forest algorithm. Furthermore, it shows that the psychological feature of the current Breath-Group is of great importance.

5. Conclusion

In this paper, we tackled the problem of modeling dialog in a storytelling setting. The objective of the method is to accurately predict when the child is likely to take a turn of speech. The proposed method is a combination of data mining and machine learning exploiting information on dialog acts. We showed that an hybridisation between pattern mining and direct exploitation of psychological features showed the best results. Initial results are interesting for the community since they offer new opportunities for modeling dialog to improve the interaction between a child and a narrative conversational agent. Despite the small amount of training data, we achieved good results using closed patterns of psychological features showing that these features have good generalisation properties. Future work will have a two-fold focus: (1) discovering rules to enrich the text with psychological features and (2) computing a model suitable for real world applications with noise inherited from a speech-to-text component. We are developing a method to automatically annotate breath groups with psychological features. We hope that this method will be robust enough to be used with a noisy input. We are also investigating how to combine other dialog acts frameworks in the same pipeline.

6. Acknowlegements

This work is supported by the ANR (French Research National Agency) funded projects NARECA ANR-13-CORD-0015 and HYBRIDE ANR-11-BS002-002.

7. References

[1] W. Boisseleau, O. Serban, and A. Pauchet, "Building a narrative conversational agent using a component-based architecture," in *Proc. of the 2014 International Conference on Autonomous Agents and Multi-agent Systems*, ser. AAMAS '14. Richland, SC: International Foundation for Autonomous Agents and Multiagent Systems, 2014, pp. 1653–1654.

[2] P. B. de Byl, "An online assistant for remote, distributed critiquing of electronically submitted assessment," *Educational Technology & Society*, vol. 7, pp. 29–41, 2004.

[3] J. Broekens, M. Heerink, and H. Rosendal, "Assistive social robots in elderly care: a review," *Gerontechnology*, vol. 8, no. 2, 2009.

[4] J. Cassell, "Embodied conversational interface agents," *Commun. ACM*, vol. 43, no. 4, pp. 70–78, Apr. 2000.

[5] N. Ward and D. Devault, "Ten challenges in highly-interactive dialog system," *AAAI Spring Symposium Series*, 2015.

[6] W. Swartout, J. Gratch, R. W. Hill, E. Hovy, S. Marsella, J. Rickel, and D. Traum, "Toward virtual humans," *AI Mag.*, vol. 27, no. 2, pp. 96–108, Jul. 2006.

[7] C. Pelachaud, "Modelling multimodal expression of emotion in a virtual agent," *Philosophical Transactions of the Royal Society*, vol. 364, no. 1535, pp. 3539–3548, 2009.

[8] S. Kopp and I. Wachsmuth, "Synthesizing multimodal utterances for conversational agents: Research articles," *Comput. Animat. Virtual Worlds*, vol. 15, no. 1, pp. 39–52, Mar. 2004.

[9] M. Schröder, "The semaine api: Towards a standards-based framework for building emotion-oriented systems," *Adv. in Hum.-Comp. Int.*, vol. 2010, Jan. 2010.

[10] H. Bunt, "Multifunctionality in dialogue," *Comput. Speech Lang.*, vol. 25, no. 2, pp. 222–245, Apr. 2011.

[11] S. Varges, S. Quarteroni, G. Riccardi, A. V. Ivanov, and P. Roberti, "Leveraging pomdps trained with user simulations and rule-based dialogue management in a spoken dialogue system," in *Proc. of the 10th SIGDIAL Conference*. The Association for Computer Linguistics, 2009, pp. 156–159.

[12] D. Griol, F. Torres, L. F. Hurtado, E. Sanchis, and E. Segarra, "Different approaches to the dialogue management in the dihana project," in *10th Speech and Computer Conference(SPECOM)*, Amsterdam, The Netherlands, 2005, pp. 203–206.

[13] D. Griol, L. F. Hurtado, E. Segarra, and E. Sanchis, "A statistical approach to spoken dialog systems design and evaluation," *Speech Commun.*, vol. 50, no. 8-9, pp. 666–682, 2008.

[14] L. Deng, J. Li, J.-T. Huang, K. Yao, D. Yu, F. Seide, M. Seltzer, G. Zweig, X. He, J. Williams, Y. Gong, and A. Acero, "Recent advances in deep learning for speech research at microsoft," in *Acoustics, Speech and Signal Processing*, 2013, pp. 8604–8608.

[15] P. Su, D. Vandyke, M. Gasic, D. Kim, N. Mrksic, T. Wen, and S. J. Young, "Learning from real users: Rating dialogue success with neural networks for reinforcement learning in spoken dialogue systems," *CoRR*, 2015.

[16] T. Zhao, A. W. Black, and M. Eskenazi, "An incremental turn-taking model with active system barge-in for spoken dialog systems," in *Proc. of the 16th SIGDIAL Conference*, Prague, Czech Republic, 2015, pp. 42–50.

[17] D. Litman and K. Forbes-Riley, "Spoken tutorial dialogue and the feeling of another's knowing," in *Proc. of the SIGDIAL 2009 Conference: The 10th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, ser. SIGDIAL '09. Stroudsburg, PA, USA: Association for Computational Linguistics, 2009, pp. 286–289.

[18] A. Pauchet, F. Rioult, . Chanoni, Z. Ales, and O. Serban, "Interactive narration requires interaction and emotion." in *ICAART (2)*, J. Filipe and A. L. N. Fred, Eds. SciTePress, 2013, pp. 527–530.

[19] H. Bunt, "The dit++ taxonomy for functional dialogue markup," in *AAMAS 2009 Workshop, Towards a Standard Markup Language for Embodied Dialogue Acts*, 2009, pp. 13–24.

[20] J. Liscombe, J. Bell Hirschberg, and J. J. Venditti, "Detecting certainness in spoken tutorial dialogues," in *Proc. of Interspeech 2005*, 2005.

[21] E. Ukkonen, "Maximal and minimal representations of gapped and non-gapped motifs of a string," *Theoretical Computer Science*, vol. 410, no. 43, pp. 4341–4349, 2009.

[22] R. Brixtel, "Maximal repeats enhance substring-based authorship attribution," in *Recent Advances in Natural Language Processing, RANLP 2015, 7-9 September, 2015, Hissar, Bulgaria*, 2015, pp. 63–71.

[23] F. Salfner, M. Lenk, and M. Malek, "A survey of online failure prediction methods," *ACM Comput. Surv.*, vol. 42, no. 3, pp. 1–42, 2010.