Towards Mental Model-driven Conversations

Francesca Alloatti[a,b], Federica Cena[b], Luigi Di Caro[b], Roger Ferrod[b] and Giovanni Siragusa[b]

[a]CELI - Language Technology, Turin, Italy
[b]University of Turin, Torino, Italy

Abstract

In recent years conversation has become a key channel for human-computer interaction. Dialogue personalization could result in an important aspect, making sense of users' features when engaged in a conversation with a machine. A feature that has been properly taken into account is the user's mental model, a crucial aspect since it determines users' expectations and the way they interact with a chatbot. In this position paper, we propose a theoretical framework that combines existing meta-mental models (behaviour-based and lexical-based) in a computational model that can be used to automatically detect the users' mental model from the dialogues with a chatbot by exploiting Linguistic theory and Machine Learning techniques.

Keywords

mental model, conversational agents, neural networks

1. Introduction

Recent years have seen the rise of conversational agents (CAs) in everyday life: chatbots that use written communications above all, but also vocal digital assistants. It is safe to say that conversation is becoming a key mode of human-computer interaction. However, despite much recent success in natural language processing and dialogue research, the communication between a human and a machine is still in its infancy. In this context, dialogue personalization could be the solution to narrow part of the gap, making sense of users' features (e.g., preferences, expertise, communication style, emotions, personality) when engaged in a conversation with a machine. In this context, the users' mental model is yet to be properly explored, while it is a crucial aspect since it determines users' expectations and the way they interact with a chatbot.

The notion of mental model was originally postulated by the psychologist Kenneth Craik [1] as a small-scale model of how the world works. This model is used to anticipate events, to reason, to generate explanation. A mental model can be seen as a reasoning mechanism in a person's working memory or as a set of beliefs and understandings that help users' decisions and interaction with world [2].

mental model is often confused with the notion of *User Model*, since both produce reciprocal expectation. mental model are the expectations users have about a computer's behaviour [3], while the User Model is the representation the system has of a user's [4]. The fundamental distinction is that a mental model can be found in the mind of people, while User Model resides inside a computer.

mental model has been historically applied to Intelligent Tutoring System (ITS) field , where detecting mental model shifts during learning represents an important step in diagnos-

Joint Proceedings of the ACM IUI 2021 Workshops, April 13–17, 2021, College Station, USA
francesca.alloatti@celi.it (F. Alloatti)

ing ineffective learning and intervening by providing appropriate feedback. Another field where mental model has extensively studied is Human-computer Interaction (HCI), where it has been defined as "internal representations of a system" [5], which "are formed through knowledge (education), experience or a combination of the two" [6], and thus can differ from the institutionalized, legitimated conceptions held by experts.

It is extremely important to understand the mental model that people use in the interaction in order to design more effective interfaces for the users [5]. This is especially true for conversational interfaces. For instance, if a chatbot is aware of the user's mental model, it may change its answers depending on the person's beliefs about the chatbot in order to make them more understandable and helpful.

Since the mental model is not directly observable, the challenge is how to understand it. Usually, qualitative user studies have been performed to this aim [7, 8, 9]. Instead, we propose to combine existing meta-mental models (behaviourally-based [8] and lexical-based features [10]) in a computational fashion that can be used to automatically detect the users' mental model from the dialogues with a chatbot, exploiting Linguistic theory and Machine Learning techniques.

2. Background: Users' Mental Model Detection

Because mental models are not directly observable, it is necessary to use some evidence to infer their features (explanations, reaction time, questions..), usually exploiting qualitative techniques (interviews, focus groups, etc).

For example, [7] takes into account 14 semi-structured interviews conducted with existing users of CA systems (Siri, Cortana) and present four key areas where current systems fail to support effective user interaction. In the majority of cases users were unable to make accurate judgments about system capability. Users had poor mental models of how their agents worked and these were reinforced through a lack of meaningful feedback mechanisms.

" Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems. 2018. Another example is [8], who conducted an experiment where 20 novice users had to complete 5 tasks (factual, instrumental, controversial, predictive) while interacting with a Google Home device, and asked them to think aloud as they completed those tasks.

Two mental models emerged from verbal strategies, which seemed to indicate differing approaches in the participants' information seeking behavior: i) *Push model.* People who employed the push model tried to "push" the system to give them more information by explaining to the machine their needs. This strategy is derived from the experience in a normal conversation with a human, which means that they believe articulating a specific personal information or a desire makes the miscommunication clear; ii) *Pull model.* the participants tried to "pull" information from the machine. They asked questions to see whether the system was able to draw out an answer for a particular type of information.

Only few work addresses the challenging problem of automatically detecting the mental model. Those studies usually pertain to the learning domain, where the mental model can be seen as the "students' level of understanding of a topic" [11]. They view the task of detecting the student mental models as a standard classification problem. The general approach is to combine textual features, which are automatically extracted, with supervised machine learning algorithms to automatically derive classifiers from expert-annotated data. For instance, [11] used content-based over-

lap methods, cohesion analysis of text, and word-weight based representations (like if-idf). They were combined with machine learning algorithms in order to automatically infer the underlying parameters. The authors evaluated their approach comparing the methods' predictions with human judgments.

3. Our proposed Metamodel of Mental Model

We propose to join different findings from mental model detection studies in a unique framework. The framework would be sustained by machine learning techniques; they would automatically compute the users' mental model by taking into account multiple features. Some of the features are behaviourally-based [8], while others take into account users' linguistic conduct during the human-machine dialogue [10].

In our perspective, users can be located along a spectrum: on one end, stands the comprehension of the system as it really is; it could be characterized as the mental model of those who developed the agent, and therefore we call these users *Developers*. On the opposite end, stand those who may be called *Primitives*. The differentiation between *Developers* and *Primitives* is exclusively linguistic: based on the lexicon they employ while interacting with the machine or the syntactic structure of their sentences, it is possible to compute a mental model "score" that situates a user along the spectrum.

However, in order to merge the linguistic mental model detection with the behavioural one, it's necessary to formalize the elements that lead to a Push or Pull model. [8] obtained the two models by observing the interactions and the verbal strategies. In our case, we propose a tagging system of the users' utterances that can lead to the identification of the appropriate model.

3.1. Automatic Detection of the Mental Model

The true challenge is to transpose the findings and theories about the mental model towards an automatic approach, i.e. an AI-driven framework that is able to compute the users' mental model autonomously, on the fly, from parameters that were clearly set. To achieve that, we identified some features of the dialogue that can help compute it:

- **Primitive or Developer mental model**. So-called Developers present compliance with the chatbot instructions; they avoid contextual references or out-of-context requests, since they are aware of the machine's capabilities. Primitives, on the other hand, will employ a more human-like form of conversation. They may use polite or formulaic expressions, or refer to world elements that are unknown to a machine.

- **Pull or Push mental model**. Instead of looking for specific linguistic features, in this case the whole utterance gets tagged with a single label. Labels could express the kind of request the user is making: for instance, is she trying to rephrase her request in order to pull the right information? Or is she trying to explain to the machine her needs in plain words? The distribution of "Push tags" or Pull ones in a conversation could then reveal the behavioural mental model of that user.

To capture the user' mental model, a single dataset of human-machine dialogues could be manipulated and analyzed: first, each utterance can get marked with a *Push* or *Pull* tag; then, their linguistic features can be captured

in order to situate a user along the *Developer* and *Primitives* spectrum. Values from both approach are taken into account to produce a customization of the system.

We plan to model the problem by building a modular architecture capable of extracting different features. For example, if it is necessary to tag the single words of a message, it is possible to refer to the sequence labeling problem and its best known techniques, such as Neural Networks and Conditional Random Fields [12]. Similarly, Recurrent and Convolutional Neural Network models can be used in the analysis of the context that surrounds words, thus capturing the most complex features related to the mental model. Psycholinguistic features such as greeting forms, chatbot anthropomorphization and out-of-context requests can be easily unveiled by attention mechanisms [13]. The latter ones, indeed, dynamically focus on words that are relevant for the task prediction. Finally, since it is necessary to aggregate all the intermediate results to generate a single score associated with the conversation and therefore with the user, we believe that an inter-attention technique [14, 15] is particularly suitable. The idea is to capture inter-related word features between utterances and select the relevant ones via a convolutional model to predict the score.

Once the various modules have been defined, it will be possible to integrate them into a single Neural Network model. The training can then continue separately or following the multi-tasking learning approach that allows to learn multiple tasks simultaneously.

4. Conclusion

This work presents a theoretical framework to enhance task-oriented agents-human dialogues by taking the user mental model into account. The framework exploits linguistic theory disciplines, human-computer interactions findings [8] and machine learning techniques, such as advanced in neural networks. The objective is to personalize the dialogue taking the user model into account. For example, it could be possible to provide additional explanations about how the chatbot works and its goals to a user with a primitive mental model. At the same time, the chatbot could consider other user's features, such as her expertise, in order to adapt the conversation, for example providing explanations of concepts in relation to the level of knowledge of the user, changing the style and terminology according to user's expertise or helping the user in understanding her limits and potentialities (e.g. suggesting what the user may ask, or explaining what Angie is not able to). Moreover, this information can be combined with failure in conversation, in order to perform personalized recovery strategies that considers both user mental model and expertise. For example, for some users it may be viable to provide more information through links and external webpages, while for others it may be more practical to refer them to human agents. We are currently experimenting with the classification task. We obtained promising results with a proprietary dataset. Future work will include the extension of the task to open datasets for the sake of reproducibility. It is worth noting that although the whole system could be applied to all sort of task-oriented agents, user's profiling shall always depend on the specific context and requirements of the agent's domain, specifically for the technical expertise evaluation. The authors believe that this strategy could significantly improve the intelligence of the system, without the need to intervene on the neural dialogue model. The final goal would be to find new ways to build dialogue system that are closer and more compliant to the actual mechanisms of the human mind, by integrating findings from psycholinguistics and philosophy of the mind into computer systems.

References

[1] K. J. W. Craik, The nature of explanation, volume 445, CUP Archive, 1952.

[2] N. A. Jones, H. Ross, T. Lynam, P. Perez, A. Leitch, Mental models: an interdisciplinary synthesis of theory and methods, Ecology and Society 16 (2011).

[3] R. B. Allen, Mental models and user models, in: Handbook of human-computer interaction, Elsevier, 1997, pp. 49–63.

[4] P. Brusilovsky, E. Millán, User models for adaptive hypermedia and adaptive educational systems, in: The adaptive web, Springer, 2007, pp. 3–53.

[5] D. A. Norman, Some observations on mental models, Mental models 7 (1983) 7–14.

[6] N. Staggers, A. F. Norcio, Mental models: concepts for human-computer interaction research, International Journal of Man-machine studies 38 (1993) 587–605.

[7] E. Luger, A. Sellen, " like having a really bad pa" the gulf between user expectation and experience of conversational agents, in: Proceedings of the 2016 CHI conference on human factors in computing systems, 2016, pp. 5286–5297.

[8] J. Cho, Mental models and home virtual assistants (hvas), in: Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems, 2018, pp. 1–6.

[9] H. Candello, C. Pinhanez, F. Figueiredo, Typefaces and the perception of humanness in natural language chatbots, in: Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems, 2017, pp. 3476–3487.

[10] F. Alloatti, L. Di Caro, A. Bosca, Conversation analysis, repair sequences and human computer interaction - a theoretical framework and an empirical proposal of action, 2020. Accepted at The Fourth Workshop on Reasoning and Learning for Human-Machine Dialogues (AAAI 2021).

[11] R. Azevedo, A. Witherspoon, A. Chauncey, C. Burkett, A. Fike, Metatutor: A metacognitive tool for enhancing self-regulated learning, in: 2009 AAAI Fall symposium series, 2009.

[12] G. Lample, M. Ballesteros, S. Subramanian, K. Kawakami, C. Dyer, Neural architectures for named entity recognition, in: Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, San Diego, California, 2016, pp. 260–270. URL: https://www.aclweb.org/anthology/N16-1030. doi:10.18653/v1/N16-1030.

[13] D. Bahdanau, K. Cho, Y. Bengio, Neural machine translation by jointly learning to align and translate, arXiv preprint arXiv:1409.0473 (2014).

[14] Y. Tay, A. T. Luu, S. C. Hui, Compare, compress and propagate: Enhancing neural architectures with alignment factorization for natural language inference, in: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, 2018, pp. 1565–1575.

[15] B. Tian, Y. Zhang, J. Wang, C. Xing, Hierarchical inter-attention network for document classification with multi-task learning., in: IJCAI, 2019, pp. 3569–3575.