Automatic analysis of online conversations as processes

Elena Epure*, Slavko Zitnik**, Dario Compagno***, Rebecca Deneckere*, Camille Salinesi*

*Centre de recherche en informatique, Université Paris 1 Panthéon-Sorbonne
Elena.Epure@malix.univ-paris1.fr, Rebecca.Deneckere, Camille.Salinesi@univ-paris1.fr
**Faculty of Computer and Information Science, University of Ljubljana
Slavko.Zitnik@fri.uni-lj.si
***Institut de la communication et des médias, Université Paris 3 Sorbonne Nouvelle
Dario.Compagno@univ-paris3.fr

The tremendous use of social media has changed the way society communicates and interacts nowadays, leading to a plethora of online conversations (Perrin et al., 2017). The increasing availability of these c*onversations as behavioral traces* has enabled automatic approaches for behavior discovery and analysis. These approaches, grounded in machine learning, data mining and language processing have become effective *predictive components* and intelligent *descriptive tools* for many domains. In *robotics*, online conversations have been used for training dialogue bots (Wu et al., 2002) ; in *politics* to analyze communication mechanisms between disseminators and public (Hemphill et Roback, 2014) ; in *security*, to enable modeling of narratives and the prediction of their influence on the crowd behavior (Houghton et al., 2013).

A widespread method to analyze automatically conversations emerges from pragmatics, specifically from speech act theory, which sustains that human communication is driven by intentions (Searle, 1969). Conversation analysis research (Searle et al., 1992) considers these intentions possible adequate concepts for representing conversations and inferring behavioral knowledge. This view has been also adopted by computer science community and subsequently exploited in automatic analyses. In general, such solutions rely on three steps : adopt an existing intention taxonomy or define a new one ; use or create a tagged corpus ; build the automatic technique either by defining relevant features for machine learning or by creating new algorithms based on text and language processing, and evaluate it on the tagged corpus.

Even though existing works brought significant contributions, there are several limitations and open issues to be tackled. *First*, the proposed intention taxonomies in linguistics are either *too general* (Searle, 1969) or *too detailed* (Vanderveken, 1990) to enable facile manual classification by non-experts. Further, the proposed intention taxonomies in computer science are often *specialized* for their target goals or corpora (Bhatia et al., 2016; Stolcke et al., 2000), making it challenging to *reproduce* on other types of online conversations. *Second*, conversation corpora created for enabling automatic intention identification are tagged per dialogue turn. However, turns of multiple sentences as often appear on social media has seldom a *unique intention* (Bhatia et al., 2016). *Third*, there is *scarce* computer science research on modeling conversations as processes though such view exists already in linguistics (Searle et al., 1992). The process mining community proposes automatic methods and techniques to discover processes and to analyze them interactively by relying on relevant and well formed logs of traced