Automatic analysis of online conversations as processes

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

Automatic analysis of online conversations as processes

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

*Centre de recherche en informatique, Universite Paris 1 Pantheon-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 medias, Universite 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

behavior (Aalst, 2011). However, process mining has been *rarely applied to text* and the existing methods are *unsuitable* for our goal (Osman et Zalhan, 2016).

Our aim is to create a *general approach for analyzing automatically online conversations* in order to reveal behavior as *processes* and to enable research in the affiliated domains. For that, we decided to apply standard process mining techniques on *logs* we design to capture *relevant conversation behavior*. The units of such logs are *intentions* and our next focus is to ensure these logs are *well-formed* and capable to reveal *reliable and correct processes*. So far, we aimed to improve the limitations of automatic intention discovery through : 1) a *general intention taxonomy* grounded in linguistic and empirical analysis, evaluated for *reproducibility* and *facile manual tagging* with non-experts ; 2) a Reddit corpus with dialogue turns of multiple utterances, *tagged per utterance* that complements existing corpora (2280 utterances) ; 3) *an evaluation of multiple classification algorithms and of features' importance for intention discovery* ; domain-independent, discourse features are proposed apart from classical content features ; weighted macro f-scores up to 78% are obtained in a 10-fold cross validation setup.

Références

Aalst, W. v. d. (2011). *Process mining* (1 ed.). Springer.

Bhatia, S., P. Biyani, et P. Mitra (2016). Identifying the role of individual user messages in an online discussion and its use in thread retrieval. *Journal of the Association for Information Science and Technology 67*(2), 276–288.

Hemphill, L. et A. J. Roback (2014). Tweet acts : How constituents lobby congress via twitter. In *17th ACM Conf. on Computer Supported Cooperative Work and Social Computing*, CSCW '14, pp. 1200–1210. New York, NY, USA : ACM.

Houghton, J., M. Siegel, et D. Goldsmith (2013). Modeling the influence of narratives on collective behavior case study : Using social media to predict the outbreak of violence in the 2011 london riots. In *International System Dynamics Conference*. System Dynamics Society.

Osman, C.-C. et P.-G. Zalhan (2016). From natural language text to visual models : A survey of issues and approaches. *Informatica Economica 20*(4/2016), 44–61.

Perrin, A., M. Duggan, et S. Greenwood (2017). Social media update 2016. Technical report, Pew Research Center : Internet, Science and Tech.

Searle, J. R. (1969). *Speech acts* (1 ed.). Cambridge University Press.

Searle, J. R., H. Parret, et J. Verschueren (1992). *(On) Searle on conversation* (1 ed.). John Benjamins.

Stolcke, A., N. Coccaro, R. Bates, P. Taylor, C. Van Ess-Dykema, K. Ries, E. Shriberg, D. Jurafsky, R. Martin, et M. Meteer (2000). Dialogue act modeling for automatic tagging and recognition of conversational speech. *Comput. Linguist. 26*(3), 339–373.

Vanderveken, DanielVanderveken, D. (1990). *Meaning and speech acts* (1 ed.). Cambridge University Press.

Wu, C.-H., G.-L. Yan, et C.-L. Lin (2002). Speech act modeling in a spoken dialog system using a fuzzy fragment-class markov model. *Speech Commun. 38*(1), 183–199.