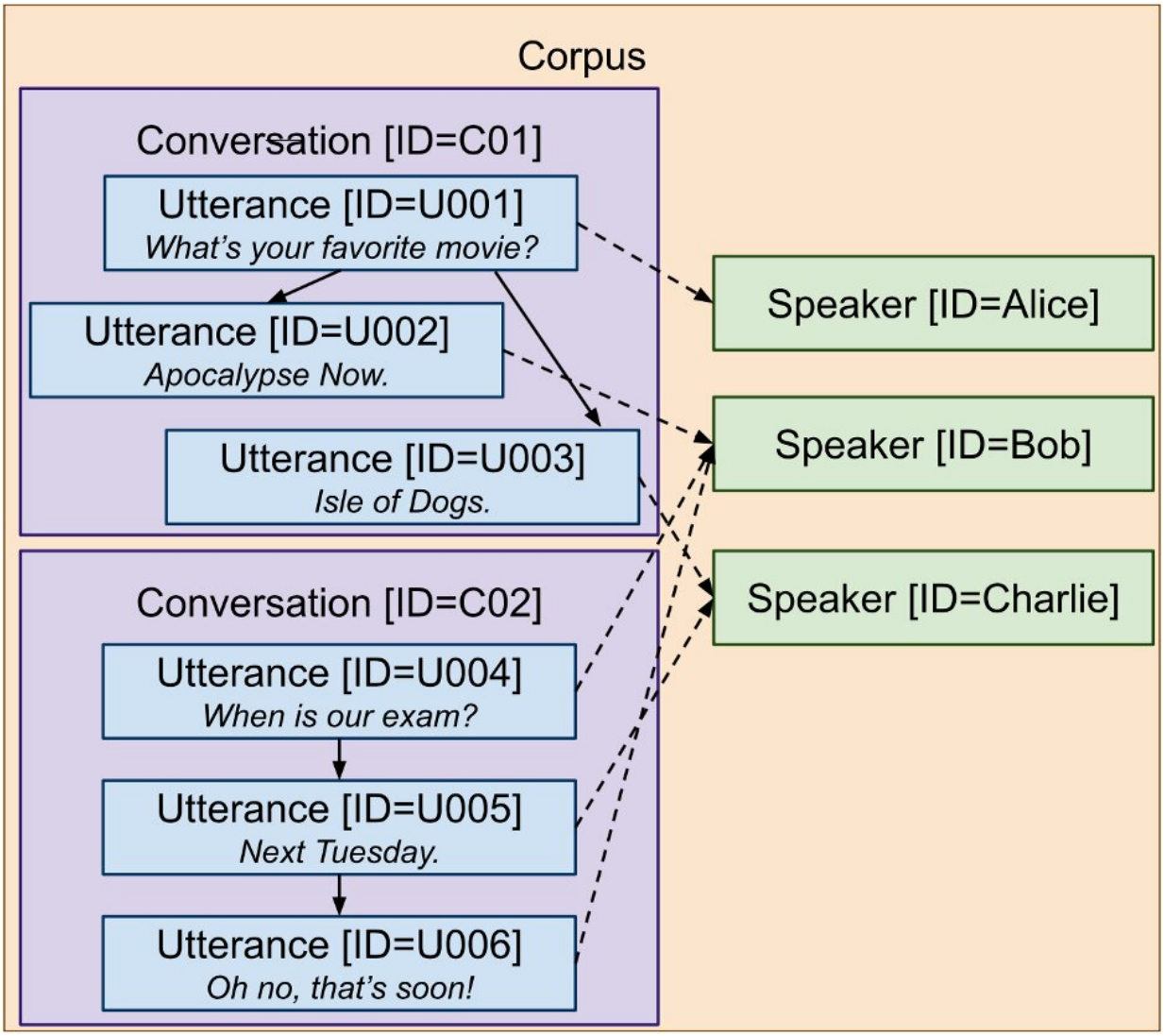
ConvoKit: An open-source toolkit to aid the analysis of conversations

June 1 2020, by Ingrid Fadelli



Credit: Chang et al.

In recent years, researchers have developed increasingly advanced natural language processing (NLP) techniques that can be trained to process, interpret and respond to sentences in human languages. In addition, some have developed toolkits that can guide researchers who are developing, training and evaluating NLP techniques.

Researchers at Cornell University have recently put together a new toolkit, dubbed ConvoKit, containing existing tools, methods and data that are ideal for developing and training NLP models designed to analyze human conversations and social interactions. This new toolkit, presented in a paper set to be presented at the SIGDIAL conference next month, makes a variety of cutting-edge techniques accessible to users with different levels of technical expertise.

"Through conversations, we discuss, collaborate, empathize and make our voices heard," Caleb Chiam, one of the researchers who developed the toolkit, told TechXplore. "Existing NLP toolkits, however, are not designed to work directly with conversational structures. ConvoKit fills that gap, as it is designed to make computational tools for conversational analysis accessible to users—no matter their technical background."

ConvoKit presents conversational data in a simple, user-friendly format. This basic format allows both expert and non-expert developers to explore and annotate the data, as well as running computations on it.

"Every conversation is about some set of individuals speaking to each other, saying certain things, in a specific order," Chiam explained. "We might typically record those conversations as transcripts—think, for example, of the transcripts we have of every "Friends' episode or every Supreme Court session (both of which are available in ConvoKit format, among many others). ConvoKit represents a set of such conversations as a 'corpus.'"

In ConvoKit, every corpus of conversations has three main elements or components, namely speakers (i.e., who is speaking), conversations (i.e., the overall exchange between two or more speakers) and utterances (i.e., what was said by a speaker at different points during a conversation). These three elements are considered 'first-class objects," which means that the toolkit enables their use as primary units of analysis.

A user could, for example, use ConvoKit to predict which speakers are more likely to mimic the linguistic style of other speakers, what conversations are more likely to become 'toxic' based on how they started off, or which utterances are polite and which ones are rude. This makes it ideal for conducting analyses that focus on specific aspects of conversations.

"ConvoKit's structure makes it easy to explore conversations," Chiam said. "For example, with these data structures, it is straightforward to pick any speaker in the dataset and go through the utterances made by that speaker and the conversations they have started. Similarly, you could choose any conversation in the dataset and iterate through the utterances that form the conversation or the speakers that were involved."

The new toolkit developed by Chiam and his colleagues also has a variety of transformers built into it, which enable more in-depth analyses. Transformers are modules that can be easily run on a conversational corpus, analyzing them using sophisticated machine learning and NLP methods.

"These computational methods can be adapted and applied to any given conversational corpus," Chiam said. "Moreover, ConvoKit users can design their own transformers for their own custom analysis. One can find examples of customized transformer features listed on convokit.cornell.edu. These include things like linguistic coordination,

politeness strategies, prompt types, and much more."

The new toolkit could prove extremely valuable for both developers and non-expert tech enthusiasts who are trying to create tools for the automatic analysis of conversations. ConvoKit is very easy to use and highly customizable, which makes it ideal for a variety of NLP applications.

"ConvoKit is in active development," Chiam said. "While much of the codebase is stable at this point, we have in the works many more methods and datasets that are currently being developed as part of our other active research. Also, since this is an open-source effort, we expect external contributions as well. Follow our GitHub page for the latest updates."

More information: ConvoKit: A Toolkit for the analysis of conversations. arXiv:2005.04246 [cs.CL]. arxiv.org/abs/2005.04246

github.com/CornellNLP/Cornell- … nal-Analysis-Toolkit