Submitted by: Raytheon BBN Technologies
Presenter: Sean Colbath
Topic: Language and Translation Challenges in Social Media

The explosive growth of social media has led to a wide range of new challenges for machine translation and language processing. The language used in social media occupies a new space between structured and unstructured media, formal and informal language, and dialect and standard usage. Yet these new platforms have given a digital voice to millions of user on the Internet, giving them the opportunity to communicate on the first truly global stage – the Internet.

Social media covers a broad category of communications formats, ranging from threaded conversations on Facebook, to microblog and short message content on platforms like Twitter and Weibo – but it also includes user-generated comments on YouTube, as well as the contents of the video itself, and even includes 'traditional' blogs and forums. The common thread linking all of these is that the media is generated by, and is targeted at individuals.

This talk will survey some of the most popular social media platforms, and identify key challenges in translating the content found in them – including dialect, code switching, mixed encodings, the use of "internet speak", and platform-specific language phenomena, as well as volume and genre. In addition, we will talk about some of the challenges in analyzing social media from an operational point of view, and how language and translation issues influence higher-level analytic processes such as entity extraction, topic classification and clustering, geo-spatial analysis and other technologies that enable comprehension of social media. These latter capabilities are being adapted for social media analytics for US Government analysts under the support of the Technical Support Working Group at the US DoD, enabling translingual comprehension of this style of content in an operational environment.