Exploiting Social Media for Natural Language Processing: Bridging the Gap between Language-centric and Real-world Applications

Simone Paolo Ponzetto
Research Group Data and Web Science
University of Mannheim
Mannheim, Germany
simone@informatik.uni-mannheim.de

Andrea Zielinski
Fraunhofer IOSB
Fraunhoferstraße 1
Karlsruhe, Germany
andrea.zielinski@iosb.fraunhofer.de

Introduction

Social media like Twitter and micro-blogs provide a goldmine of text, shallow markup annotations and network structure. These information sources can all be exploited together in order to automatically acquire vast amounts of up-to-date, wide-coverage structured knowledge. This knowledge, in turn, can be used to measure the pulse of a variety of social phenomena like political events, activism and stock prices, as well as to detect emerging events such as natural disasters (earthquakes, tsunami, etc.).

The main purpose of this tutorial is to introduce social media as a resource to the Natural Language Processing (NLP) community both from a scientific and an application-oriented perspective. To this end, we focus on micro-blogs such as Twitter, and show how it can be successfully mined to perform complex NLP tasks such as the identification of events, topics and trends. Furthermore, this information can be used to build high-end socially intelligent applications that tap the wisdom of the crowd on a large scale, thus successfully bridging the gap between computational text analysis and real-world, mission-critical applications such as financial forecasting and natural crisis management.

Tutorial Outline

1. Social media and the wisdom of the crowd. We review the resources which will be the focus of the tutorial, i.e. Twitter and micro-blogging in general, and present their most prominent and distinguishing aspects (Kwak et al., 2010; Gouws et al., 2011), namely: (i) instant short-text messaging, including its specific linguistic characteristics (e.g., non-standard spelling, shortenings, logograms, etc.) and other features – i.e., mentions (@), hashtags (#), shortened URLs, etc.; (ii) a dynamic network structure where users are highly inter-connected and author profile information is provided along with other metadata. We introduce these properties by highlighting the different trade-offs related to resources of this kind, as well as their comparison with alternative data publishing platforms – for instance, highly unstructured text vs. rich network structure, semi-structured metadata tagging (like hashtags) vs. fully-structured linked open data, etc.

2. Analyzing and extracting structured information from social media. We provide an in-depth overview of contributions aimed at tapping the wealth of information found within Twitter and other micro-blogs. We first show how social media can be used for many different NLP tasks, ranging from pre-processing tasks like PoS tagging (Gimpel et al., 2011) and Named Entity Recognition (Ritter et al., 2011) through high-end discourse (Ritter et al., 2010) and information extraction applications like event detection (Popescu et al., 2011; Ritter et al., 2012) and topic tracking (Lin et al., 2011). We then focus on novel tasks and challenges opened up by social media such as *geoparsing*, which aims to predict the location (including its geographic coordinates) of a message or user based on his posts (Gelernter and Mushegian, 2011; Han et al., 2012), and methods to automatically establish the credibility of user-generated content by making use of contextual and metadata features (Castillo et al., 2011).

3. Exploiting social media for real-world applications: trend detection, social sensing and crisis management. We present methods to detect emerging events and breaking news from social media (Mathioudakis et al., 2010; Petrović et al., 2010, *inter alia*). Thanks to their highly dynamic environment and continuously updated content, in fact, micro-blogs and social networks are capable of providing real-time information for a wide vari-

Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, pages 5–6,
Sofia, Bulgaria, August 4-9 2013.

ety of different social phenomena, including consumer confidence and presidential job approval polls (O'Connor et al., 2010), as well as stock market prices (Bollen et al., 2011; Ruiz et al., 2012). We focus in particular on applications that use social media for health surveillance in order to monitor, for instance, flu epidemics (Aramaki et al., 2011), as well as crisis management systems that leverage them for tracking natural disasters like earthquakes (Sakaki et al., 2010; Neubig et al., 2011) and tsunami (Zielinski and Bürgel, 2012; Zielinski et al., 2013).

References

Eiji Aramaki, Sachiko Maskawa, and Mizuki Morita. 2011. Twitter catches the flu: detecting influenza epidemics using Twitter. In *Proc. of EMNLP-11*, pages 1568–1576.

Johan Bollen, Huina Mao, and Xiaojun Zeng. 2011. Twitter mood predicts the stock market. *Journal of Computational Science*, 2(1):1–8.

Carlos Castillo, Marcelo Mendoza, and Barbara Poblete. 2011. Information credibility on Twitter. In *Proc of WWW-11*, pages 675–684.

Judith Gelernter and Nikolai Mushegian. 2011. Geoparsing messages from microtext. *Transactions in GIS*, 15(6):753–773.

Kevin Gimpel, Nathan Schneider, Brendan O'Connor, Dipanjan Das, Daniel Mills, Jacob Eisenstein, Michael Heilman, Dani Yogatama, Jeffrey Flanigan, and Noah A. Smith. 2011. Part-of-speech tagging for Twitter: Annotation, features, and experiments. In *Proc. of ACL-11*, pages 42–47.

Stephan Gouws, Donald Metzler, Congxing Cai, and Eduard Hovy. 2011. Contextual bearing on linguistic variation in social media. In *Proceedings of the Workshop on Language in Social Media (LSM 2011)*, pages 20–29.

Bo Han, Paul Cook, and Timothy Baldwin. 2012. Geolocation prediction in social media data by finding location indicative words. In *Proc. of COLING-12*, pages 1045–1062.

Haewoon Kwak, Changhyun Lee, Hosung Park, and Sue Moon. 2010. What is Twitter, a social network or a news media? In *Proc of WWW-10*, pages 591–600.

Jimmy Lin, Rion Snow, and William Morgan. 2011. Smoothing techniques for adaptive online language models: topic tracking in tweet streams. In *Proc. of KDD-11*, pages 422–429.

Michael Mathioudakis, Nick Koudas, and Peter Marbach. 2010. Early online identification of attention gathering items in social media. In *Proc. of WSDM-10*, pages 301–310.

Graham Neubig, Yuichiroh Matsubayashi, Masato Hagiwara, and Koji Murakami. 2011. Safety information mining – what can NLP do in a disaster –. In *Proceedings of IJCNLP-11*, pages 965–973.

Brendan O'Connor, Ramnath Balasubramanyan, Bryan R. Routledge, and Noah A. Smith. 2010. From tweets to polls: linking text sentiment to public opinion time series. In *Proc. of ICWSM-10*, pages 122–129.

Saša Petrović, Miles Osborne, and Victor Lavrenko. 2010. Streaming first story detection with application to Twitter. In *Proc. of NAACL-10*, pages 181–189.

Ana-Maria Popescu, Marco Pennacchiotti, and Deepa Paranjpe. 2011. Extracting events and event descriptions from Twitter. In *Comp. Vol. to Proc. of WWW-11*, pages 105–106.

Alan Ritter, Colin Cherry, and Bill Dolan. 2010. Unsupervised modeling of Twitter conversations. In *Proc. of NAACL-10*, pages 172–180.

Alan Ritter, Sam Clark, Mausam, and Oren Etzioni. 2011. Named entity recognition in tweets: an experimental study. In *Proc. of EMNLP-11*, pages 1524–1534.

Alan Ritter, Mausam, Oren Etzioni, and Sam Clark. 2012. Open domain event extraction from Twitter. In *Proc. of KDD-12*, pages 1104–1112.

Eduardo J. Ruiz, Vagelis Hristidis, Carlos Castillo, Aristides Gionis, and Alejandro Jaimes. 2012. Correlating financial time series with micro-blogging activity. In *Proc. of WSDM-12*, pages 513–522.

Takeshi Sakaki, Makoto Okazaki, and Yutaka Matsuo. 2010. Earthquake shakes Twitter users: real-time event detection by social sensors. In *Proc. of WWW-10*, pages 851–860.

Andrea Zielinski and Ulrich Bürgel. 2012. Multilingual analysis of Twitter news in support of mass emergency events. In *Proc. of ISCRAM-12*.

Andrea Zielinski, Stuart E. Middleton, Laurissa Tokarchuk, and Xinyue Wang. 2013. Social-media text mining and network analysis to support decision support for natural crisis management. In *Proc. of ISCRAM-13*.