Multilingual Analysis of Twitter News in Support of Mass Emergency Events

Ulrich Bügel, Fraunhofer IOSB, Karlsruhe, Germany
Andrea Zielinski, Information Retrieval, Knowledge Management and Text Mining, Fraunhofer IOSB, Karlsruhe, Germany

ABSTRACT

Social media are increasingly becoming a source for event-based early warning systems in the sense that they can help to detect natural disasters and support crisis management during or after disasters. In this article the authors study the problems of analyzing multilingual twitter feeds for emergency events. Specifically, they consider tsunami and earthquakes as one possible originating cause of tsunami. Twitter messages provide testified information and help to obtain a better picture of the actual situation. Generally, local civil protection authorities and the population are likely to respond in their native language. Therefore, the present work focuses on English as “lingua franca” and on under-resourced Mediterranean languages in endangered zones, particularly Turkey, Greece, and Romania. The authors investigated ten earthquake events and defined four language-specific classifiers that can be used to detect earthquakes by filtering out irrelevant messages that do not relate to the event. The final goal is to extend this work to more Mediterranean languages and to classify and extract relevant information from tweets, translating the main keywords into English. Preliminary results indicate that such a filter has the potential to confirm forecast parameters of tsunami affecting coastal areas where no tide gauges exist and could be integrated into seismographic sensor networks.

Keywords: Cross-Lingual Information Extraction, Earthquakes, Emergency Events, Multilingual Event Detection, Social Sensor, Twitter

INTRODUCTION

Emergency information processing of social media can contribute effectively to identify regions affected by natural hazards such as earthquakes or tsunami, given that the feeds are real-time and often contain location information (ca. 1.2% with exact coordinates; ca. 50% city or state derived from the user profile). Due to the massive growth of Twitter data and its increasing number of users, it is however, a challenge to access and interpret the stream of data efficiently. Within the last years, there have been major achievements to make use of such “weak” human sensors as a complement to seismic sensors in some early warning systems (see Sakaki, 2010, Guy, 2010), focusing on English and Japanese. At present, there is no similar alerting system for the Mediterranean...
region. We try to fill this gap within the European TRIDEC project (www.tridec-online.eu) by adapting state of the art algorithms to the common Twitter languages in the endangered zones.

Social media often play a crucial role in disaster management during and after the crisis: citizens generally use Twitter postings or SMS messages to report emergencies. In this case, the information contained in them might be relevant for crisis management (relief and medical care for those affected, repair of broken infrastructure, etc.), so that there is a strong need to classify, cluster and extract such information effectively from large-scale noisy and unstructured data. As the messages are very short (max. 140 characters), NLP analysis is particularly difficult.

A number of text mining tools have been applied to recognize tactical, actionable information in tweets (Verma, 2011), to find messages that contain real-world or real-event information (Becker, 2011; Naaman, 2011), or to extract Named Entities (Neubig, 2011) or other news content (Sankaranarayanan, 2009) for one single language (mostly Japanese or English).

In some cases, though, it is crucial to cross language boundaries. For instance, when the epicenter is near the border of a country (e.g., Western Turkey and Greece), or when a twitter user reports an event in his/her native language (e.g., Romanian) that needs to be translated into a different language (e.g. English, German, or Spanish).

Therefore, our long-term goal within TRIDEC is to support the access to relevant information across languages, focusing on the translation of under-resourced Mediterranean languages like Turkish/Greek/Romanian into English.

The multilingual nature of the blogosphere has been a major hindrance during the Haitian earthquake, where reports ranged from Japanese, to English and Spanish. Caragea (2011)’s work is one of the few that deals with multilinguality, classifying either English or Spanish messages into one of 10 emergency classes.

For the same event, a Statistical Machine Translation (SMT) System has been built that supports crisis management for Haitian Creole (Lewis, 2010) and profits from crowdsourced human translations (Munro, 2010). From the author’s experience, a “codebook” is established that would enable and support the development of an SMT system for other less-resourced languages in short time and which could be integrated into an instant messaging system like twitter – provided, however, that a minimum of domain-specific bilingual parallel or comparable corpora and/or terminologies are available (Lewis et al., 2011). The need for rapid translation to aid the relief efforts has also come up during the tsunami event in Japan (Neubig et al., 2011).

EMERGENCY EVENTS

The most recent earthquakes with a magnitude > 4 in the Mediterranean shown in Figure 1 (See map offered by EMSC - European Mediterranean Seismological Centre) took place in regions for which few language resources exist. These are, in descending order, Greek, Hungarian, Bulgarian, Croatian, Bosnian, Albanian, Slovenian, Slovak, Turkish, Romanian, and Serbian (cf. http://matrix.statmt.org/resources/matrix?set=euro_all).

Most earthquakes were felt across national borders. While some events caused severe damage, e.g. the earthquake of Van, Turkey, for others no damage or injuries were reported (e.g. Romania).

Although the seismic system in the Mediterranean is quite dense and, in general, earthquakes can be accurately located and quantified, in the case of a tsunami, the network of tide gauges is quite sparse, and a human-sensor-
Related Content

Role of Accounting and Audit in the Recent Financial Crisis

A Descriptive Approach for Power System Stability and Security Assessment
[www.igi-global.com/chapter/a-descriptive-approach-for-power-system-stability-and-security-assessment/90792?camid=4v1a](www.igi-global.com/chapter/a-descriptive-approach-for-power-system-stability-and-security-assessment/90792?camid=4v1a)
Managing Crises in the Healthcare Service Chain: Lessons from the Past and Future Directions
[www.igi-global.com/article/managing-crises-healthcare-service-chain/52607?camid=4v1a](www.igi-global.com/article/managing-crises-healthcare-service-chain/52607?camid=4v1a)

Interaction Protocols Adaptation Based Coordination for Crisis Management Processes
[www.igi-global.com/article/interaction-protocols-adaptation-based-coordination-for-crisis-management-processes/156567?camid=4v1a](www.igi-global.com/article/interaction-protocols-adaptation-based-coordination-for-crisis-management-processes/156567?camid=4v1a)