Event Detection and Tracking

Supervisor: João Paulo Cordeiro
Co-Supervisor: Sebastião Pais

July 16, 2018

Abstract

Social media have emerged as powerful means of communication for people looking to share and exchange information on a wide variety of real-world events. Short messages posted on Twitter can typically reflect these events as they happen. For this reason, the content of such social media sites is particularly useful for real-time identification of real-world events and their associated user-contributed messages. As such, a crowd can be viewed as a community sharing a common focus about some specific event. Event detection has intensively been studied in the last decade mainly due to the advent of social media [1]. Two different approaches have been proposed: document-pivot and feature-pivot. All the studied techniques are interesting but do not cover the overall picture. Once an event is detected, it is crucial to track it, i.e. to follow its evolution. Within this scope, topic models have shown successful results for text clustering tasks. However, they only rely on a term-document matrix to compute similarity, which may be insufficient for social media texts that are short and lack in contextual information. This is confirmed by the recent work of [2], who show that the use of locality-sensitive hashing combined with named entity recognition achieves better performance for detecting news than using the topic modeling approach. Moreover, topics are represented as sets of words, that may not all be bursty, and thus may include more general topics than specific ones.

Objectives

This work will focus on new strategy for event tracking, based on the most recent findings, as the work of [3] which proposed the Dual C-means clustering algorithm allowing to mix document-pivot and feature-pivot techniques into a unique model. Dual C-means showed to perform likewise topic models for word sense induction [4]. The advantage of the Dual C-means algorithm is that different similarity measures can be implemented, including knowledge-based metrics, that may lead to improved results in the line of [2], as tweets are linked to YAGO by their entities. Moreover, the clustering process can be driven by bursty keywords or named entities, and thus better adapt to the dynamic environment of Twitter, instead of relying on

all possible terms present in a time window. It will also allow the integration of richer sources of knowledge in clustering. One possibility is to explore entailment dependencies [5].

Task Description

The work is planed to be accomplished through the following tasks:

Task 1: Explore the state-of-the-art in event detection and Tracking, as well as textual clustering and entailment methods.

Task 2: Data gathering and setup of a dataset for evaluating event detection in social media.

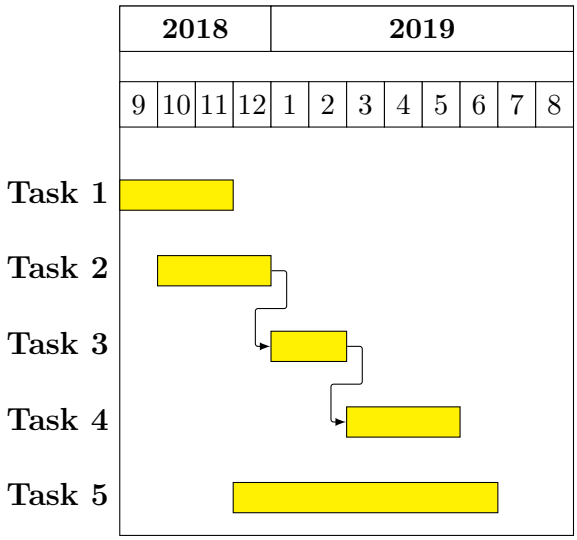
Task 3: Exploitation of a combination o methods for detecting and tracking events.

Task 4: Find out and propose a new approach for this problem.

Task 5: Write the dissertation and a scientific article.

Schedule

The schedule for the planed tasks are shown in the following Gantt chart:



Contacts

João Paulo Cordeiro (Office 4.3)
Sebastião Pais (Office 4.1)

UBI, Department of Informatics
Rua Marquês d'Ávila e Bolama
6201-001 Covilhã

References

[1] A. Farzindar and D. Inkpen, "Natural language processing for social media," *Synthesis Lectures on Human Language Technologies*, vol. 8, no. 2, pp. 1–166, 2015.

[2] H. M. Wold and L. C. Vikre, "Online news detection on twitter," Master's thesis, NTNU, 2015.

[3] J. G. Moreno, G. Dias, and G. Cleuziou, "Query log driven web search results clustering," in *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval*, pp. 777–786, ACM, 2014.

[4] S. Acharya, A. Ekbal, S. Saha, P. Santhanam, J. G. Moreno, and G. Dias, "Multi-objective word sense induction using content and interlink connections," in *International Conference on Applications of Natural Language to Information Systems*, pp. 366–375, Springer, 2016.

[5] S. Pais, G. Dias, R. Moraliyski, and J. Cordeiro, "Unsupervised and language-independent method to recognize textual entailment by generality," in *Proceedings of CLIB 2014*, pp. 82–90, Institute for Bulgarian Language, 2014.