Event summarization from news and tweets correlation

Cavallari Sandro
Giglio Marco
Morettin Paolo

ABSTRACT

This report we consider the task to extract a summary of the main event that cause a shift on the opinion of the Twitter users. In particular we refer to those shift as *sentiment shift* and we try to capture the event that cause those sentiment shift. In this work we present different language-independent and semantic-less meteorologies and we try to figure out the pros and cons of those methods.

Introduction

The main goal of this work is to found a summary of the events that cause some sentiment shift. While NLP technique for event extraction are mature, their performance on tweets inevitably degrades, due to the inherent sparsity in short texts. Since Tweet contains heterogeneous language structure and are at most long 140 character, extracting event form tweets text is quite tricky: instead linking tweet to news allow us to extract event form news text that have a well formed structure and contains more text respect to tweet.

Due to the fact that out dataset is composed by labelled tweet, we manage to extract news form the New York Times archive basing the research on tweet's label. Starting form a good Tweet-News classification, allowed us to correlate tweet and news that belong to the same macro-topics avoiding to correlate object of different arguments.

In particular document compression is one of the main topic in *NLP* research field: usually this task is achieved using abstractive methods and language dependent technique. Our work, rather, want to be as much as possible language independent, so we propose solutions for this problem using technique based on:

- a simple Bag of Word[BoW] combined with a Tf-idf approach
- a Latent Semantic Indexing[LSI] that uses a SVD mathematical technique [1]
- a N-Gram Graph[NGG] [2]

Related Work

Since our project have two main goal:

- compute the correlation between tweet and news
- create a summary of the main event that cause the sentiment shift

There are many work that are related with it. In particular, due to the incising and development of the Tweeter platform, in the last year, much work was done on the correlation and analysis of the tweeter data. Weiwei Guo et al. [3] have propose a framework for link tweet with news and extract form the resulting correlation some missing aspect of the tweet-event. Instead of using a LSI technique for compute a text-to-word representation, they propose a methodologies based on the Weighted Textual Matrix Factorization [4] [WTMF] model(the 2012 defacto standard). With the WTMF unsupervised model and the usage of cosine-similarity Weiwei Guo et al. were able to obtain good correlation performance and WTMF result to be a really strong tool for baseline creation.

In the literature there is a lot of work done on *recommendation system*; some of this systems are specifically developed for new recommendation: witch aims to recommend news articles based on some user features. This work seems to be complementary respect to our goal, but we focus on the correlation between tweet and news starting only with the tweet text content that is a much smaller and heterogeneous context respect to a user features.

Recently *Google* present a system for sentence compression, in particular this system try to achieve a *multi-sentence compression* [5] focusing on the importance of content selection and readable presentation. Our attention was caught by the fact that Google obtain good result summary using only redundancy information of the text, without the usage syntactic constraints, so potentially can be used as an other language-independent methodology. Similarly to NGG the Google MSC approach use a graph representation of the text, combining it with some part of speech information. The word-base graph

representation allow to manage the redundancy of the text, meanwhile the part of speech information is used to obtain a readable sentence compression in output.

The google methods seems to be quite similar to the NGG approach and obtain probably obtain better summarization results, but since we prefer a system that is as much as possible language-independent

Problem Definition
ProblemDefinition

Proposed Approach
As over mentioned, in order to solve our problem we try out different approaches, each one with his advantages and disadvantages. In this chapter we present all the different implementation of the solutions.

Bag of Word
LSI
N-Gram Graph
N-Gram Graph is a NPL tool initially proposed by

Experimental Evaluation
ExperimentalEvaluation

Conclusions
1. REFERENCES

[1] Indexing by Latent Semantic Analysis, Scott Deerwester , Susan T. Dumais*, George W. Furnas, Thomas K. Landauer and Richard Harshman

[2] Automatic Summarization from Multiple Documents : N-Gram Graph, George Giannakopoulos and Ncsr Demokritos, 2009

[3] Linking Tweets to News: A Framework to Enrich Short Text Data in Social Media, Weiwei Guo, Hao Li, Heng Ji and Mona Diab

[4] Modeling Sentences in the Latent Space, Weiwei Guo and Mona Diab , 2012

[5] Multi-Sentence Compression: Finding Shortest Paths in Word Graphs, Katja Filippova