A Keyword Sense Disambiguation Based Approach for Noise Filtering in Twitter

Sanjaya Wijeratne*
Kno.e.sis Center, Wright State University – USA
sanjaya@knoesis.org

Bahareh R. Heravi
Insight Centre for Data Analytics – NUIG
bahareh.heravi@insight-centre.org

Abstract

In this paper, we describe an approach to filter out noisy data generated by keywords-based tweet filtering methods by performing Word Sense Disambiguation on those keywords used to collect tweets. We present the noise filtering problem as a binary classification problem and discuss our evaluation strategy which is to be carried out in future.

1. Motivation

With growing popularity of streaming social media platforms such as Twitter for news reporting, locating timely and newsworthy information from them has become an essential step in Digital Journalism. Journalists use keywords-based tweet filtering to locate tweets created by eyewitnesses in order to create news stories. Keywords-based tweet filtering also brings a lot of irrelevant tweets as well. For example, a journalist who uses the keyword 'shoot' to find information about shooting incidents around the world via Twitter would get irrelevant tweets about photo/video shoots and football goal shoots because of the ambiguity of the term 'shoot'. The motivation of this work is to help journalists to find newsworthy content that interest them (tweets that are not noisy) from Twitter by filtering out noisy tweets collected by keywords-based tweet filtering.

2. Problem Statement

Let K be the set of all keywords used to collect T, which is the set of all tweets collected. Let S be the set of all senses (glosses) for all keywords in K. Let $S^+ \subset S$ be the set of all senses that could collect interesting tweets to the user for all $k_i \in K$. Let $P \subset K$ be the set of all tweet collecting keywords present in t. Given a tweet $t \in T$, K and S^+, how do we determine whether t is an interesting tweet to the user. In other words, can we determine t is not a noisy tweet given K and S^+? Hence this is a binary classification problem.

3. Related Work

Tweet classification in information filtering is a challenging problem because general text classification methods fail to address the problems with sparsity and non-standardized language used in tweets[1]. People have used supervised, semi-supervised and semantic relationships-based classification approaches to address the problem of short text classification[1]. But, according to the best of our knowledge, this is the first attempt of using word sense disambiguation on the set of tweet collecting keywords that are present in a tweet t to determine the intention of using them in the tweet's context C_t (C_t is defined in Section 6), which will then be used to classify t.

*This work is from author's ongoing internship at Insight Centre.

4. Research Question

We attempt to address how different senses of keywords in P can be used to determine whether t is noisy or not.

5. Hypothesis

Assume we have collected a tweet t, and know K and S^+, determining the senses of all keywords present in t which is $P \subset K$ can be used to determine whether t was intended to collect by the user using P or not.

6. Proposed Solution

For each tweet collecting keyword $k_i \in K$, we extract it's senses from BabelNet[1] to generate S. S_i is the set of all senses of $k_i \in K$. For each $k_i \in K$, the user will select a set of senses $S_i{}^+ \subset S_i$ that made the user to pick k_i as a keyword, which helps us to understand what senses of k_i would bring interesting tweets to the user. All senses of a keyword k_i that are not selected by the user $S_i{}^- \subset S_i$ are considered as senses that could bring noise for k_i. For each sense $s_i \in S_i$ of k_i, we generate a list of associated words (stopword removed and stemmed) using BabelNet synsets, glosses, entities and their types, which act as the context Cs_i for each sense $s_i \in S_i$ of k_i. Given a tweet t, we identify entities, their types, and remove stopwords and stem the remaining words to generate the context of the tweet C_t. For each keyword $p_i \in P$ in t, we disambiguate and assign the best sense to p_i using Simplified LESK algorithm[2] by calculating the overlap of each keyword's sense's context Cs_i with C_t. If the sense assigned to keyword p_i is from $S_i{}^-$, t will be classified as a noisy tweet for p_i and will be filtered.

7. Discussion on Evaluation

We plan to evaluate our approach using randomly selected tweet samples on each keyword that we used to collect tweets. We will manually remove any duplicate tweets in them. Accuracy will be measured on how precisely our approach identifies noisy tweets. In our initial evaluation for keyword 'shoot' with a sample set of 100 tweets (66 noisy), we achieved 89% accuracy in removing noisy tweets.

References

[1] G. Song, Y. Ye, X. Du, X. Huang, and S. Bie. Short text classification: A survey. *Journal of Multimedia*, 9(5):635–643, 2014.

[1] http://tinyurl.com/BabelNet
[2] http://tinyurl.com/SimpleLESK