An Unsupervised Approach for Tweet Summarization

Jaya P A Srinath Srinivasa
Web Sciences Lab, IIIT Bangalore

Abstract

Text summarization is an important need for extracting latent semantics from short user-generated documents like social media activity, or operational notifications. The proposed research aims to (semi-) automate the process of summary generation from a given set of short documents to identify a set of candidate sentences that summarize one or more activity. To begin this study, we explore a completely unsupervised technique to identify candidate sentences from the document collection. These techniques do not require a domain ontology or training data, thus making them the first line of attack to address this problem. The entire document collection is represented as a collection of noun and verb phrases. Then, using Markov Random Field based factor computation technique, the potential relation between noun and verb phrases are identified in the form of $(subject, predicate, object)$ triples. Further high potential triples are used to identify candidate sentences for final summary. This is a generic unsupervised approach and the effectiveness is evaluated on a collection tweets related to various trending topics.

1 Summarization Approach

Among the various social networking sites compared to traditional media, twitter provides the most recent update about mainstream events. Twitter sent out more than 500 million tweets every day[1]. Among the million tweets, most of the tweets are conversational in nature and a few percentage of those tweets covers important information about the mainstream news. Here we are trying to summarize the tweets which are related to a recent mainstream event. The tweets related to an event are collected using twitteR and streamR packages.

Even though tweet conveys news in a very short and easily understandable form to human beings, the inherent nature of tweet as well as lack of context challenges the automatic tweet processing. The 140 character restriction imposes a lot of uncommon abbreviations, code words, emoticons to a tweet along with the twitter specific notations, such as hashtags, retweet tags, user mentions, hyperlinks, etc. Automatic processing of tweets are also challenged by the language variations as well as the idiosyncratic style of tweets.

In our summarization approach we exploit the energy associated with noun and verb phrases to identify the candidate tweet sentences. We mine the relation in-terms of verb phrases from a collection of noun phrases. This emerged triples

[1] http://www.internetlivestats.com/twitter-statistics/

information are used to identify the candidate tweet sentences. The summarization process involves i) Collation of information from tweets as noun and verb phrases. ii) Mine the relation information as triples and use that information to identify candidate tweet sentences.

The short and noisy nature of tweets challenges the use of standard Natural Language Processing (NLP) techniques. The NLP engines developed by the research community were designed to work on a grammatically correct English sentences.Cleaning and canonicalization has to be done on the tweet data-set to remove noise; by filtering hyperlinks, special symbols, repeated words etc, and also by replacing the abbreviations and spelling mistakes. This filtered and canonicalized tweet is then processed through the NLP pipeline stages such as, Tokenization, Sentence Boundary Detection, Parts Of Speech (POS) tagging, and Chunking. Further, POS tagged and chunked tweets are used to collect the noun and verb phrases.

Background Markov Random Filed(MRF): MRF is an undirected graphical model used to represent the correlation among nodes. It is widely used in image processing applications. In MRF the potential function is defined over the set of variables in a clique. In image processing the pairwise potential function is defined between neighboring pixels. The potential function takes a higher value if their colors are similar. The idea of energy computation using pairwise factors is motivated for finding the correlation between noun phrases.

The noun phrases represents the set of nodes in an undirected graph. The information of how important a particular verb phrase for a pair of noun phrases are used for mining the relation. The verb which maximizes the energy for a clique factor, becomes the potential verb for that clique which links the noun phrases. The information about the presence of noun phrases along with verb phrase and verb prior information are used for the factor computations. Noun-verb-noun (triples) which maximizes the factor value are used as potential triples. These triples are further used to identify the candidate tweet sentences. The sentences are sorted based on the scores of triples accumulated if multiple triples are mapped to same sentence.

Algorithm
Input: Collection of tweets related to a mainstream event or activity
Output: A set of tweet candidate sentences identified from tweets collection

1. T = The input tweets collection.

2. Perform POS tagging and chunking on the canonicalized tweets collection T.

3. Extract all Noun and Verb phrases

4. Identify the potential relation between noun and verb phrases based on MRF factor computation

5. Compute the verb prior ρ, as $\frac{T_{v_j}}{T_n}$, where T_{v_j} is the number of tweet sentences contains the verb v_j and T_n is the total number of tweet sentences.

6. for each noun phrase pair N_i and N_{i+1} from noun phrase collection do

7. for each verb feature v_j of N_i and N_{i+1} do

8. Compute factor $\phi(N_i, N_{i+1}) = \frac{max(c_1|c_2|c_3|c_4)}{c_1+c_2+c_3+c_4}$ where c_1, c_2, c_3, c_4 are the counts based on all four binary combinations of N_i and N_{i+1} with v_j
9. Compute $\theta_{v_j} = [[log[\phi(N_i, N_{i+1})] + log[\rho(v_j)]]]$
10. Identify v_j which maximizes (θ_{v_j})
11. Use v_j to generate candidate triples using N_i and N_{i+1}
12. Use the triples with highest score to identify the candidate tweet sentences. Add up scores of triples if multiple triples map to same sentence.
13. Sort sentences based on the score and use highest scored sentences as candidate sentences.

Our summarization approach is generic in nature and it can be used in any other domain where we can identify the key phrases as in terms of noun and verb phrases.