INFORMATION RETRIEVAL AND EXTRACTION

PROJECT SCOPE DOCUMENT

Project #7 : Identifying Salient Named Entity of a Tweet

Mentor : Priya

Team #8

Shwetha G, 201405525

Surya Sashank Nistala, 201201083

Vibhav Prateek Srivastava, 201156030

Joshita Mishra, 201450831

Abstract

Unlike carefully authored news text and other longer content, tweets pose a number of new challenges, due to their short, noisy, context-dependent, and dynamic nature. We propose a solution to the problem of determining what a tweet is about through semantic linking: we add semantics to tweets by automatically identifying concepts that are semantically related to it and conduct an empirical analysis of named entity recognition and disambiguation. The identified concepts can subsequently be used for, e.g., social media mining, thereby reducing the need for manual inspection and selection.

Project Scope

The scope of project include following major tasks.

- Create around 5000+ tweet dataset with the format
  "<tweet><contained image URL><NE1,NE2..><salientNE>"
- Identifying **Named Entities** using either manual or automated tools

- Identifying **Salient Named Entity** representing the tweet.
- Evaluation of system for various approaches taken.

Related Systems

The basis for this work is two papers

1. E. Meij, W. Weerkamp, and M. de Rijke. Adding Semantics to Microblog Posts. In Proc. of the 5th ACM Intl. Conf. on Web Search and Data Mining (WSDM) , pages 563–572. ACM, 2012 - **Work on what a microblog post is about through semantic linking**

2. Yamada et al. “Evaluating the helpfulness of linked entities to readers” HT'14 Proposed System / Approach - **This paper propose a new method for evaluating the helpfulness of linking entities to users**

Project Approach:

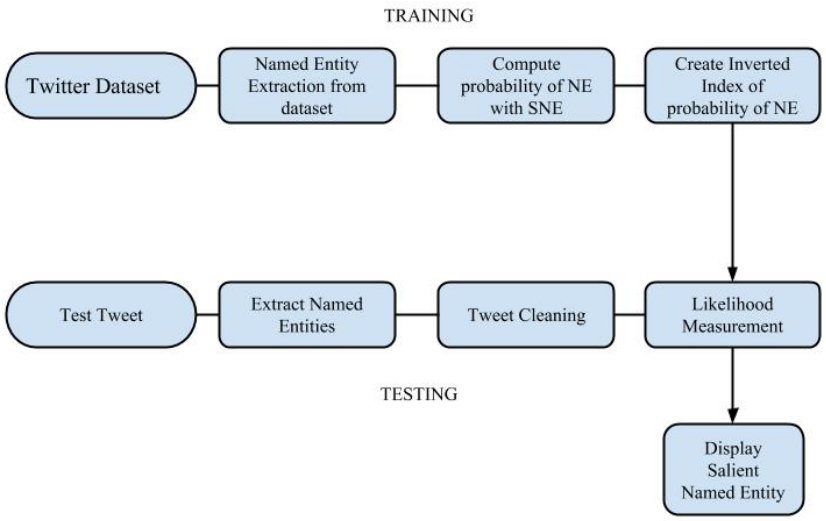

Figure 1: Overall working of the System

The following are the various steps considered in building the project.

- **Identify Named Entities:**
    - Use POS tagging approach and classify the named entities based on tags (eg: two nouns appearing together etc)
    - We use ANNIE from GATE, which uses gazetteer-based lookups and finite state machines to identify and type named entities in newswire text.
    - We use Stanford NER system, which uses a machine learning-based method to detect named entities, and is distributed with CRF models for English newswire text.
- Process the identified NEs to map corresponding **Salient Named Entity** or respective tweet from its set of NEs.
- **Evaluation of the system:**
    - Evaluate each approach taken against the manually classified results. Compute **Precision, Recall, F-Score** for each NE and salient NEs.
- Once system is stable, **provide a UI** to end users to enter a tweet and display the SNE representing the tweet as result.

Extra Contribution and Tentative Features

→ Try and implement spell correction for shorthand/slang /typos general using edit distance.

→ Query Expansion: Identify appropriate if SNE not present in set of NE.

→ Plot Graph of results of different algorithms

Datasets

Manually created dataset of 5000 tweet corpus of tweets with images, annotated with salient entities is used in the project. A row of this dataset will consist of the fields **<tweet> <contained ImageURL> <NE1,NE2..> <salientNE>**. All four fields are mandatory and will be used to train the application.

<tweet> contains tweet text

<image URL> URL of the image contained in the tweet

<NE> One or more named entities identified in the image

<salient NE> the NE the tweet is talking about.

Work Division:

- NE identification using POS tagging : Shwetha , Sashank
- NE identification using ANNE : Sashank
- NE identification using Stanford NER: Shwetha
- Dataset creation,Evaluation of System, regression testing : Joshita
- UI creation : Vibhav Srivastav
- Process the identified NEs to find Salient NE using ML algorithms:Vibhav, Joshita
- Documentation: Respective work documented by corresponding members.