Tracking the Diffusion of Named Entities

TBD

Abstract

To do

Index Terms

To do

I. Introduction

The aim of this paper is to understand how named entities *emerge* and *spread* through social media based discourse. We are interested in exploring the following research questions:

1) **RQ1:** How can we accurately detect named entities in social media based discourse, given its myriad formats, often informal vernacular, and inherent noise (e.g. misspellings, abbreviations, etc.)?
2) **RQ2:** Under what conditions do entity mentions diffuse through discourse? And when are people *most likely* to be influenced into then discussing entities?
3) **RQ3:** How can we predict the discussion of certain named entities and who will begin talking about them?

II. Datasets

For this research we will use the following two datasets:

1) Twitter data - we have a large corpus of English tweets that we can use here.
2) Reddit data - download and access all of the data from the full dump.[1]

III. Research Stages

A. Stage 0: Data Preparation and NER

-To do:

–Annotate corpora with detected entities using basic typing of: person, location, organisation

–Run NER software over dataset and validate accuracy of this (using basic measures)

–Run NER over entire dataset to extract entities

[1]https://archive.org/details/2015_reddit_comments_corpus

B. Stage 1: Exploratory Analysis

-To do:

–Plot relative frequency distribution as a function of time for named entities, and characterise the *shape* of the entities

–Apply lifecycle model to profile users' NER citations over time and investigate how users' profiles are influenced by global, community, and prior behaviour dynamics

C. Stage 2: Diffusion Analysis

-To do:

–Model the spread of named entities through user profiles (could use multivariate diffusion models here)

D. Stage 3: Forecasting

-To do:

–Implement models to forecast if a user will mention an entity and who that will be (hard!)