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 datasets:

1) Reddit data – download and access all of the data from the full dump.[1]

2) CoNLL 2003 data – a corpus of newswire texts, annotated for named entity chunks and types. This describes where entity mentions are in the text, including locations, organisations, and person mentions.

3) Twitter data; unannotated – we have a large corpus of English tweets that we can use here.

4) Twitter data; annotated – there are two datasets annotated with named entities. These are from Ritter's 2011 EMNLP paper, and the W-NUT 2015 shared task.

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!)

IV. RELATED WORK

twitter info propagation

reddit compared to other OSNs: "Lifespan and propagation of information in On-line Social Networks: A case study based on Reddit" http://www.sciencedirect.com/science/article/pii/S1084804515001307

network structure of reddit: "Navigating the massive world of reddit: using backbone networks to map user interests in social media" https://peerj.com/articles/cs-4/

twitter nlp

V. DATA PREPARATION AND NER

To conduct our study, we need to convert 140GB of compressed Reddit posts into a set of interlinked and time-ordered conversations and the entities mentioned in each of them. This provides a number of sub-challenges: sampling of the Reddit data, creating a linked series of conversations, and picking out entity mentions in this text type. Reddit data is largely unexplored in the NLP community, despite the large volume of it and the especially rich metadata. This poses additional challenges: certainly, given the lack of work on Reddit text, there are no annotated datasets available yet, so supervised in-domain work is not directly possible yet. Additionally, the datasets are large, which makes it important to choose a good subset of data on which to do prototyping and development, in order to keep research cycles short. The result that we come to at the end of this stage is a rich dataset for tracking entity and concept diffusion within and across communities.

The Reddit dataset [1] is comprised of a sequence of comments, with one JSON record for each one. These are ordered temporally. Reddit itself is roughly similar to a forum, where top-level divisions are made by topic. Within each topic, or *subreddit*, there are posts, which begin with either a short piece of text or a link to an external resource – typically an image, video, or interesting article. Users then may publish comments for each post, and

Fig. 1. Example discussion around a Reddit post.

reply to each others' comments. This leads to a threaded discussion, centred on a particular topic, with a hierarchical comment structure (see Figure V).

A. Subreddit extraction

- subreddit extraction

top lists

describe top lists site

pick 100

extract these over whole sample

figures for raw dataset (# per year; total volume over time; volume-rank graph of subreddits; possible ref to appx)

B. Adapting NER for Reddit

– where is NER first mentioned and defined?

We model micro-topics in conversation as entity mentions. This allows tracking of topics at maximally fine granularity, looking at each user's interests at a low level, as opposed to monitoring broader topics such as "consumer electronics", "politics" and so on. In fact, these broader topics are already explicitly annotated by means of the subreddit topics.

Entity mentions are extracted through named entity recognition. Generally, this task aims to detect the boundaries of certain kinds of entities within a certain piece of text. In this instance, we tokenise text, splitting it into

sentences using the Punkt tokeniser [2], and subsequently word-sized chunks, using the `twokenize` tool with some adaptations [3]. This tool performs Penn Treebank-style tokenisation, a common standard, with some specific adaptations to enable it to handle the noise present in user-generated text. After this, we take a structured prediction approach to deciding which tokens in each sentence are part of an entity, and possibly the type of the entity. Finally, we concatenate entity tokens, and use these to build a list of entity mentions in any given input text. For example, given the input comment from the source JSON:

"body": "There are still some really good fighters on this card. Conor McGregor is on the card and so is Gunnar Nelson."

The following output entities should be collected:

"entity_texts": ["Conor McGregor", "Gunnar Nelson"]

Typically, many NER systems take a supervised approach; that is, they use data labelled by humans as training data, from which features are extracted to form training instances for a machine learning algorithm. However, NLP systems can be hard to transfer between text types; for example, NER systems for newswire might reach about 89% F1 on news articles, but only around 40% on tweets (a form of user-generated content), as found in [4]. One approach to overcoming this performance drop when changing text type is to train over a blend of text types. For example, Ritter [5] used both IRC[2] and newswire data when developing a part-of-speech tagger for tweets, as well as an unsupervised language model from the target text type. This led to strong performance improvements. We follow a similar approach, using a blend of NE-annotated corpora from both newswire and tweets. The newswire data is drawn from the CoNLL-2003 evaluation task set [6]; the twitter data is from Ritter's early experiments and also the W-NUT 2015 shared task [5], [7].

- blended text type approach as in ritter 2011
- CRF w/ classical features and all brown string depth features
- use entity-recognition toolkit [8]

C. Tuning entity recognition

Entity recognition needs to be tuned to fit Reddit data well. There are a number of parameters in our training data balance, feature extraction, and objective function that all reflect the nature of the data and the task at hand. We present our method for estimating of these factors, and intrinsic NER evaluation

- prefer precision; beta = 2 (as per ff-rpg paper)
- describe reddit annotation/tuning annotation approach (skim, find things it got wrong, annotate them, add to set)
- extract reddit tuning set
- estimate Brown c, w based on impact paper findings
- compare blended brown cluster src vs. reddit-only
- tune twitter / newswire balance

---

[2]Internet Relay Chat – informal internet conversation text

- one, three, four, ten entity classes?

References

[1] J. Baumgartner, "Complete public Reddit comments corpus," 2015.

[2] T. Kiss and J. Strunk, "Unsupervised multilingual sentence boundary detection," *Computational Linguistics*, vol. 32, no. 4, pp. 485–525, 2006.

[3] B. O'Connor, M. Krieger, and D. Ahn, "TweetMotif: Exploratory Search and Topic Summarization for Twitter," in *Proc. ICWSM*. AAAI, 2010.

[4] L. Derczynski, D. Maynard, G. Rizzo, M. van Erp, G. Gorrell, R. Troncy, J. Petrak, and K. Bontcheva, "Analysis of named entity recognition and linking for tweets," *Information Processing & Management*, vol. 51, no. 2, pp. 32–49, 2015.

[5] A. Ritter, S. Clark, O. Etzioni *et al.*, "Named entity recognition in tweets: an experimental study," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. ACL, 2011, pp. 1524–1534.

[6] E. F. Tjong Kim Sang and F. De Meulder, "Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition," in *Proceedings of the seventh Conference on Natural language learning at HLT-NAACL 2003-Volume 4*. ACL, 2003, pp. 142–147.

[7] T. Baldwin, Y.-B. Kim, M. C. de Marneffe, A. Ritter, B. Han, and W. Xu, "Shared tasks of the 2015 workshop on noisy user-generated text: Twitter lexical normalization and named entity recognition," *Proc. W-NUT (ACL-IJCNLP)*, pp. 126–135, 2015.

[8] L. Derczynski, I. Augenstein, and K. Bontcheva, "USFD: Twitter NER with Drift Compensation and Linked Data," *Proc. W-NUT (ACL-IJCNLP)*, pp. 48–53, 2015.