Information Processing and Management 51 (2015) 32–49

Contents lists available at ScienceDirect

Information Processing and Management

journal homepage: www.elsevier.com/locate/infoproman

Analysis of named entity recognition and linking for tweets

Leon Derczynski [a,*], Diana Maynard [a], Giuseppe Rizzo [b,d], Marieke van Erp [c], Genevieve Gorrell [a], Raphaël Troncy [b], Johann Petrak [a], Kalina Bontcheva [a]

[a] University of Sheffield, Sheffield S1 4DP, UK
[b] EURECOM, 06904 Sophia Antipolis, France
[c] VU University Amsterdam, 1081 HV Amsterdam, The Netherlands
[d] Università di Torino, 10124 Turin, Italy

ARTICLE INFO

Article history:
Received 12 November 2013
Received in revised form 16 October 2014
Accepted 22 October 2014
Available online 19 November 2014

Keywords:
Information extraction
Named entity recognition
Entity disambiguation
Microblogs
Twitter

ABSTRACT

Applying natural language processing for mining and intelligent information access to tweets (a form of microblog) is a challenging, emerging research area. 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. Information extraction from tweets is typically performed in a pipeline, comprising consecutive stages of language identification, tokenisation, part-of-speech tagging, named entity recognition and entity disambiguation (e.g. with respect to DBpedia). In this work, we describe a new Twitter entity disambiguation dataset, and conduct an empirical analysis of named entity recognition and disambiguation, investigating how robust a number of state-of-the-art systems are on such noisy texts, what the main sources of error are, and which problems should be further investigated to improve the state of the art.

1. Introduction

Information Extraction (IE) (Cardie, 1997; Appelt, 1999) is a form of natural language analysis, which takes textual content as input and extracts fixed-type, unambiguous snippets as output. The extracted data may be used directly for display to users (e.g. a list of named entities mentioned in a document), for storing in a database for later analysis, or for improving information search and other information access tasks.

Named Entity Recognition (NER) is one of the key information extraction tasks, which is concerned with identifying names of entities such as people, locations, organisations and products. It is typically broken down into two main phases: *entity detection* and *entity typing* (also called classification) (Grishman & Sundheim, 1996). A follow-up step to NER is Named Entity Linking (NEL), which links entity mentions within the same document (also known as entity disambiguation) (Hirschmann & Chinchor, 1997), or in other resources (also known as entity resolution) (Rao, McNamee, & Dredze, 2013). Typically, state-of-the-art NER and NEL systems are developed and evaluated on news articles and other carefully written, longer content (Ratinov & Roth, 2009; Rao et al., 2013).

* Corresponding author.
E-mail address: leon@dcs.shef.ac.uk (L. Derczynski).

http://dx.doi.org/10.1016/j.ipm.2014.10.006
0306-4573/

In recent years, social media – and microblogging in particular – have established themselves as high-value, high-volume content, which organisations increasingly wish to analyse automatically. Currently, the leading microblogging platform is Twitter, which has around 288 million active users, posting over 500 million tweets a day,[1] and has the fastest growing network in terms of active usage.[2]

Reliable entity recognition and linking of user-generated content is an enabler for other information extraction tasks (e.g. relation extraction), as well as opinion mining (Maynard, Bontcheva, & Rout, 2012), and summarisation (Rout, Bontcheva, & Hepple, 2013). It is relevant in many application contexts (Derczynski, Yang, & Jensen, 2013), including knowledge management, competitor intelligence, customer relation management, eBusiness, eScience, eHealth, and eGovernment.

Information extraction over microblogs has only recently become an active research topic (Basave, Varga, Rowe, Stankovic, & Dadzie, 2013), following early experiments which showed this genre to be extremely challenging for state-of-the-art algorithms (Derczynski, Maynard, Aswani, & Bontcheva, 2013). For instance, named entity recognition methods typically have 85–90% accuracy on longer texts, but 30–50% on tweets (Ritter, Clark, Mausam, & Etzioni, 2011; Liu, Zhou, Wei, Fu, & Zhou, 2012). First, the shortness of microblogs (maximum 140 characters for tweets) makes them hard to interpret. Consequently, ambiguity is a major problem since semantic annotation methods cannot easily make use of coreference information. Unlike longer news articles, there is a low amount of discourse information per microblog document, and threaded structure is fragmented across multiple documents, flowing in multiple directions. Second, microtexts exhibit much more language variation, tend to be less grammatical than longer posts, contain unorthodox capitalisation, and make frequent use of emoticons, abbreviations and hashtags, which can form an important part of the meaning. To combat these problems, research has focused on microblog-specific information extraction algorithms (e.g. named entity recognition for Twitter using CRFs (Ritter et al., 2011) or hybrid methods (van Erp, Rizzo, & Troncy, 2013)). Particular attention is given to microtext normalisation (Han & Baldwin, 2011), as a way of removing some of the linguistic noise prior to part-of-speech tagging and entity recognition.

In light of the above, this paper aims to answer the following research questions:

RQ1 How robust are state-of-the-art named entity recognition and linking methods on short and noisy microblog texts?
RQ2 What problem areas are there in recognising named entities in microblog posts, and what are the major causes of false negatives and false positives?
RQ3 Which problems need to be solved in order to further the state-of-the-art in NER and NEL on this difficult text genre?

Our key contributions in this paper are as follows. We report on the construction of a new Twitter NEL dataset that remedies some inconsistencies in prior data. As well as evaluating and analysing modern general-purpose systems, we describe and evaluate two domain specific state-of-the-art NER and NEL systems against data from this genre (NERD-ML and YODIE). Also, we conduct an empirical analysis of named entity recognition and linking over this genre and present the results, to aid principled future investigations in this important area.

The paper is structured as follows.[3]

Section 2 evaluates the performance of state-of-the-art named entity recognition algorithms, comparing versions customised to the microblog genre to conventional, news-trained systems, and provides error analysis.

Section 3 introduces and evaluates named entity linking, comparing conventional and recent systems and techniques. Sections 2 and 3 answer **RQ1**.

Section 4 examines the performance and errors of recognition and linking systems, making overall observations about the nature of the task in this genre. This section addresses **RQ2**.

Section 5 investigates factors external to NER that may affect entity recognition performance. It introduces the microblog normalisation task, compares different methods, and measures the impact normalisation has on the accuracy of information extraction from tweets, and also examines the impact that various NLP pipeline configurations have on entity recognition.

In Section 6, we discuss the limitations of current approaches, and provide directions for future work. This section forms the answer to **RQ3**.

In this paper, we focus only on microblog posts in English, since few linguistic tools have currently been developed for tweets in other languages.

2. Named entity recognition

Named Entity Recognition (**NER**) is a critical IE task, as it identifies which snippets in a text are mentions of entities in the real world. It is a pre-requisite for many other IE tasks, including NEL, coreference resolution, and relation extraction. NER is difficult on user-generated content in general, and in the microblog genre specifically, because of the reduced amount of contextual information in short messages and a lack of curation of content by third parties (e.g. that done by editors for newswire). In this section, we examine some state-of-the-art NER methods, compare their performance on microblog data, and analyse the task of entity recognition in this genre.

[1] http://news.cnet.com/8301-1023_3-57541566-93/report-twitter-hits-half-a-billion-tweets-a-day.
[2] http://globalwebindex.net/thinking/social-platforms-gwi-8-update-decline-of-local-social-media-platforms.
[3] Some parts of Sections 3.2, 5.1, 5.2 and 5.3 appeared in an earlier form in Derczynski et al. (2013).