Part-of-Speech is (almost) enough: SAP Research & Innovation at the #Microposts2014 NEEL Challenge

Daniel Dahlmeier
SAP Research and Innovation
#14 CREATE, 1 Create Way
Singapore
d.dahlmeier@sap.com

Naveen Nandan
SAP Research and Innovation
#14 CREATE, 1 Create Way
Singapore
naveen.nandan@sap.com

Wang Ting
SAP Research and Innovation
#14 CREATE, 1 Create Way
Singapore
dean.wang@sap.com

ABSTRACT

This paper describes the submission of the SAP Research & Innovation team at the #Microposts2014 NEEL Challenge. We use a two-stage approach for named entity extraction and linking, based on conditional random fields and an ensemble of search APIs and rules, respectively. A surprising result of our work is that part-of-speech tags alone are almost sufficient for entity extraction. Our results for the combined extraction and linking task on a development and test split of the training set are 34.6% and 37.2% F_1 score, respectively, and for the test set is 37%.

Keywords

Conditional Random Field, Entity Extraction, DBpedia Linking

1. INTRODUCTION

The rise of social media platforms and microblogging services has led to an explosion in the amount of informal, user-generated content on the web. The task of the #Microposts2014 workshop NEEL challenge is named entity extraction and linking (NEEL) for microblogging texts [1]. Named-entity extraction and linking is a challenging problem because tweets can contain almost any content, from serious news, to personal opinions, to sheer gibberish and both extraction and linking have to deal with the inherent ambiguity of natural language.

In this paper, we describe the submission of the SAP Research & Innovation team. Our system breaks the task into two separate steps for extraction and linking. We use a conditional random field (CRF) model for entity extraction and an ensemble of search APIs and rules for entity linking. We describe our method and present experimental results based on the released training data. One surprising finding of our experiments is that part-of-speech tags alone perform almost as well as the best feature combinations for entity extraction.

Published as part of the #Microposts2014 Workshop proceedings, available online as CEUR Vol-1141 (http://ceur-ws.org/Vol-1141)

#Microposts2014, April 7th, 2014, Seoul, Korea.

2. METHOD

2.1 Extraction

We use a sequence tagging approach for entity extraction. In particular, we use a conditional random field (CRF) which is a discriminative, probabilistic model for sequence data with state-of-the-art performance [3]. A linear-chain CRF tries to estimate the conditional probability of a label sequence $\mathbf{y}$ given the observed features $\mathbf{x}$, where each label y_t is conditioned on the previous label y_{t-1}. In our case, we use BIO CoNLL-style tags [5]. We do not differentiate between different entity classes for BIO tags (e.g, 'B' instead of 'B-PERSON').

The choice of appropriate features can have a significant impact on the model's performance. We have investigated a set of features that are commonly used for named entity extraction. Table 1 lists the features. The casing features

Feature	Example
words	Obamah
words lower	obamah
POS	^
title case	True
upper case	False
stripped words	obamah
is number	False
word cluster	-NONE-
dbpedia	dbpedia.org/resource/Barack_Obama

Table 1: Examples of features for entity extraction.

upper case and *lower case* and the *is number* feature are implemented using simple regular expressions. The *stripped words* feature is the lowercased word with initial hashtags and @ characters removed. The DBpedia feature is annotated automatically using the DBpedia Spotlight web API [1] and acts as a type of gazetteer feature. For a label y_t at position t, we consider features x extracted at the current position t and previous position $t-1$. We experimented with larger feature contexts but they did not improve the result on the development set.

2.2 Linking

For the linking step, we explore different search APIs, such as Wikipedia search[2], DBpedia Spotlight, and Google search to retrieve the DBpedia resource for a mention. We begin with using the extracted entities individually as query terms

[1] github.com/dbpedia-spotlight/dbpedia-spotlight
[2] github.com/goldsmith/Wikipedia

Feature	F_1 score
POS	0.622
+ is number	**0.629**
+ upper case	0.623

Table 2: Results for extraction feature selection.

to these search APIs. As ambiguity is a major concern for the linking task, for tweets where there are multiple entities extracted, we use the entities combined as an additional query term. For example, a tweet with annotated entities as *Sean Hoare* and *phone hacking*, *Sean Hoare* would map to a specific resource in DBpedia but *phone hacking* could refer to more than one resource. By using the query term *"phone hacking + Sean Hoare"*, we can help boost the rank for the resource *"News International phone hacking scandal"* to map to the entity *phone hacking* instead of a general article on *"Phone Hacking"*. In our system, we make use of the Web APIs for Wikipedia search and DBpedia Spotlight together with some hand-written rules to rank the resources returned. The result of the ranking step is then used to construct the DBpedia resource URL to which the entity is mapped.

3. EXPERIMENTS AND RESULTS

In this section, we present experimental results of our method, based on the on the data released by the organizers.

3.1 Data sets

We split the provided data set into a training (first 60%), development (dev, next 20%), and test (dev-test, last 20%) set. We perform standard pre-processing steps. We perform tokenization and POS tagging using the Tweet NLP toolkit [4], lookup word cluster indicators for each token from the Brown clusters released by Turian *et al.* [6], and annotate the tweets with the DBpedia Spotlight web API.

3.2 Extraction

We train the CRF model on the training set of the data, perform feature selection based on the dev set, and test the resulting model on the dev-test set. We evaluate the resulting models using precision, recall, and F_1 score. In all experiments, we use the CRF++ implementation of conditional random fields[3] with default parameters. We found in initial experiments that the CRF parameters did not have a great effect on the final score. We employ a greedy feature selection method [2] to find the subset of the best features. Table 2 shows the results of the feature selection experiments on the development set. We can see that POS tags alone give a F_1 score of 62.2%. Adding the binary *is number* feature increases the score to 62.9%. Additional features, such as lexical features, word clusters, or the DBpedia Spotlight annotations, do not help and even decrease the score. Surprisingly the word token itself is *not* selected as one of the features. Thus, the CRF performs its task without even looking at the word itself! After feature selection, we retrain the CRF with the best performing feature set {*POS, is number*} and evaluate the model on the dev and dev-test set. The results are shown in Table 3.

[3] `code.google.com/p/crfpp/`

Data set	Precision	Recall	F_1 score
Dev	0.673	0.591	0.629
Dev-test	0.671	0.579	0.622

Table 3: Results for entity extraction.

3.3 Linking

To test our linking system, we follow two approaches. First, we measure the accuracy of the linking system using the gold standard where we observe an accuracy of 67.6%. As a second step, we combine the linking step with our entity extraction step and measure the F_1 score. Table 4 shows the results on the dev and dev-test split for the combined system.

Data set	Precision	Recall	F_1 score
Dev	0.436	0.287	0.346
Dev-test	0.477	0.304	0.372

Table 4: Results for entity extraction and linking.

4. CONCLUSION

We have described the submission of the SAP Research & Innovation team to the #Microposts2014 NEEL shared task. Our system is based on a CRF sequence tagging model for entity extraction and an ensemble of search APIs and rules for entity linking. Our experiments show that POS tags are a surprisingly effective feature for entity extraction in tweets.

5. ACKNOWLEDGEMENT

The research is partially funded by the Economic Development Board and the National Research Foundation of Singapore.

6. REFERENCES

[1] A. E. Cano Basave, G. Rizzo, A. Varga, M. Rowe, M. Stankovic, and A.-S. Dadzie. Making Sense of Microposts (#Microposts2014) Named Entity Extraction & Linking Challenge. In *Proc., #Microposts2014*, pages 54–60, 2014.

[2] A.L. Berger, V.J. Della Pietra, and S.A. Della Pietra. A maximum entropy approach to natural language processing. *Computational linguistics*, 22(1), 1996.

[3] J. Lafferty, A. McCallum, and F. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. 2001.

[4] O. Owoputi, B. O'Connor, C. Dyer, K. Gimpel, N. Schneider, and N. Smith. Improved part-of-speech tagging for online conversational text with word clusters. In *Proceedings of NAACL-HLT*, 2013.

[5] E.T.K. Sang and F. De Meulder. Introduction to the conll-2003 shared task: Language-independent named entity recognition. In *Proceedings of HLT-NAACL*, 2003.

[6] J. Turian, L. Ratinov, and Y. Bengio. Word representations: a simple and general method for semi-supervised learning. In *Proceedings of ACL*, 2010.