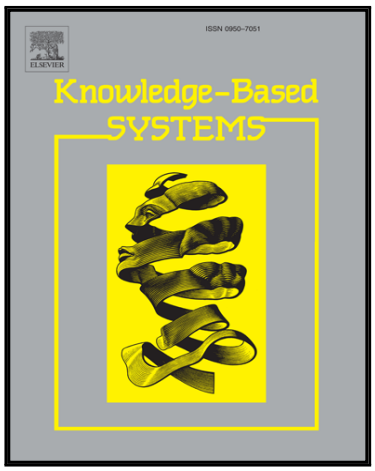


# A combination of active learning and self-learning for named entity recognition on Twitter using conditional random fields

Van Cuong Tran, Ngoc Thanh Nguyen, Hamido Fujita,
Dinh Tuyen Hoang, Dosam Hwang

PII: S0950-7051(17)30304-0
DOI: 10.1016/j.knosys.2017.06.023
Reference: KNOSYS 3951

To appear in: *Knowledge-Based Systems*

Received date: 22 March 2017
Revised date: 14 June 2017
Accepted date: 15 June 2017

Please cite this article as: Van Cuong Tran, Ngoc Thanh Nguyen, Hamido Fujita, Dinh Tuyen Hoang, Dosam Hwang, A combination of active learning and self-learning for named entity recognition on Twitter using conditional random fields, *Knowledge-Based Systems* (2017), doi: 10.1016/j.knosys.2017.06.023

# A combination of active learning and self-learning for named entity recognition on Twitter using conditional random fields

Van Cuong Tran[a], Ngoc Thanh Nguyen[b], Hamido Fujita[c], Dinh Tuyen Hoang[a], Dosam Hwang[a,*]

[a]*Department of Computer Engineering, Yeungnam University, Gyeongbuk, 38541, South Korea*
[b]*Faculty of Computer Science and Management, Wroclaw University of Science and Technology, 50-370 Wroclaw, Poland*
[c]*Faculty of Software and Information Science, Iwate Prefectural University, 020-0693, Iwate, Japan*

**Abstract**

In recent years, many applications in natural language processing (NLP) have been developed using the machine learning approach. Annotating data is an important task in applying machine learning to NLP applications. A common approach to improve the system performance is to train on a large and high-quality set of training data that is annotated by experts. Besides, active learning (AL) and self-learning can be utilized to reduce the annotation costs. The self-learning method discovers highly reliable instances based on a trained classifier, while AL queries the most informative instances based on active query algorithms. This paper proposes a method that combines AL and self-learning to reduce the labeling effort for the named entity recognition task from tweet streams by using both machine-labeled and manually-labeled data. We employ AL queries based on the diversity of the context and content of instances to select the most informative instances. The conditional random fields are also chosen as an underlying model to train a classifier for selecting highly reliable instances. The experiments using Twitter data show that the proposed method achieves good results in reducing the human labeling effort, and it can significantly improve the performance of the systems.

*Keywords:* Named entity recognition, active learning, self-learning, tweet streams

## 1. Introduction

Social networking services (SNSs) like Twitter, Facebook, and Google+ have attracted millions of users who publish and share the most up-to-date information, emergent social events, and their personal opinions, producing large volumes of data every day. For example, Twitter has more than 313 million monthly active users, and 500 million tweets are sent per day[1]. Information extraction over SNSs has recently become an interesting research topic. It attracts researchers in the fields of knowledge discovery and data mining. Normally, the data from SNSs are short, incomplete, noisy, but up-to-date [1],[2]. A big challenge for mining streamed data is the lack of labeled data due to rapid changes in distribution.

Named entity recognition (NER) is one of the most important subtasks in information extraction. It is defined as the identification of named entities within a text, and labels them with predefined categories, such as personal name, location, organization, etc. [3]. The traditional NER methods require prior labeling of the training data, and NER processes can be done offline. This is not suitable for SNSs, where the data are dynamically updated. To deal with the brief and up-to-date information in tweets, different NER methods have been proposed in recent years [4],[5],[6]. The majority of them have primarily focused on recognizing named entities in a pool of tweets [6],[7]. However, the emergent problem is how to effectively conduct the NER task in tweet streams.

The popular NER methods use supervised learning, requiring a large amount of training data with high accuracy to construct good classifiers. They achieve high performance if they are applied to well-formatted text and high-quality labeled data. However, the obtained results are often unsatisfying when applied to short and noisy messages like tweets [8]. On the other hand, annotating thousands of these data takes a great deal of human effort and even redundant. It is hard to annotate a corpus covering a broad domain, whereas unla-

*Corresponding author.
*Email address:* dosamhwang@gmail.com (Dosam Hwang)
[1]https://about.twitter.com/company