Computer Science Review 29 (2018) 21–43

Contents lists available at ScienceDirect

Computer Science Review

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

Review article

Recent Named Entity Recognition and Classification techniques: A systematic review

Archana Goyal [a], Vishal Gupta [b,*], Manish Kumar [c]

[a] PG Department of Information Technology, GGDSD College, Chandigarh, India
[b] University Institute of Engineering & Technology, Panjab University, Chandigarh, India
[c] Panjab University Regional Centre, Muktsar, Punjab, India

ARTICLE INFO

Article history:
Received 2 December 2017
Received in revised form 1 June 2018
Accepted 2 June 2018

ABSTRACT

Textual information is becoming available in abundance on the web, arising the requirement of techniques and tools to extract the meaningful information. One of such an important information extraction task is Named Entity Recognition and Classification. It is the problem of finding the members of various predetermined classes, such as person, organization, location, date/time, quantities, numbers etc. The concept of named entity extraction was first proposed in Sixth Message Understanding Conference in 1996. Since then, a number of techniques have been developed by many researchers for extracting diversity of entities from different languages and genres of text. Still, there is a growing interest among research community to develop more new approaches to extract diverse named entities which are helpful in various natural language applications. Here we present a survey of developments and progresses made in Named Entity Recognition and Classification research.

© 2018 Published by Elsevier Inc.

Contents

1. Introduction ... 22
   1.1. What is Named Entity? ... 22
   1.2. Named Entity Recognition and Classification (NERC) ... 22
   1.3. Motivation for conducting the survey ... 22
2. Factors affecting the performance of NERC task ... 23
   2.1. Language factor ... 23
   2.2. Textual genres or domain factor ... 23
   2.3. Entity type factor ... 23
3. Issues and challenges in NERC task ... 23
4. Applications of Named Entity Recognition ... 24
   4.1. Information Extraction ... 24
   4.2. Question-Answering ... 24
   4.3. Machine Translation ... 24
   4.4. Automatic Text Summarization ... 24
   4.5. Text Clustering ... 24
   4.6. Information Retrieval ... 24
   4.7. Knowledgebase or ontology population ... 24
   4.8. Opinion mining ... 24
   4.9. Semantic search ... 25
   4.10. Other applications ... 25
5. Techniques used in NERC ... 25
   5.1. Rule-based approaches ... 25
   5.2. Learning-based approaches ... 25
      5.2.1. Supervised learning ... 25
      5.2.2. Semi-supervised learning ... 26

* Corresponding author.
E-mail addresses: archana.goyal@ggdsd.ac.in (A. Goyal), vishal@pu.ac.in (V. Gupta), mkjindal@pu.ac.in (M. Kumar).

https://doi.org/10.1016/j.cosrev.2018.06.001
1574-0137/© 2018 Published by Elsevier Inc.

5.2.3. Unsupervised learning ... 26
5.3. Hybrid approaches ... 26
6. NERC systems ... 26
6.1. Rule-based NERC systems ... 26
6.2. Learning based NERC systems ... 27
6.2.1. Supervised NERC systems ... 27
6.2.2. Semi-supervised NERC systems ... 34
6.2.3. Unsupervised NERC systems ... 34
6.3. Hybrid systems ... 35
6.4. Analysis of results ... 36
7. Base classifiers for Named Entity Recognition ... 37
7.1. Naïve Bayes ... 37
7.2. Conditional Random Field (CRF) ... 38
7.3. Support Vector Machine (SVM) ... 38
7.4. Hidden Markov Model (HMM) ... 38
7.5. Maximum Entropy (MaxEnt) ... 38
8. Evaluation measures of NERC ... 38
8.1. Precision, recall and F-score ... 38
8.2. Matching predictions against Gold standard ... 38
8.3. Macro-and micro averaged F-score ... 39
8.4. Cross-validation ... 39
9. Future directions in NERC ... 39
10. Conclusion ... 39
References ... 40

1. Introduction

Today in the era of the internet, an abundance of information is available in digital form in many languages. The information stored in the structured or unstructured form needs to be processed and extracted in various natural language processing tasks. Extraction of meaningful information out of voluminous data is a big challenge which demands to develop new technologies to handle such a big data. Many areas of information extraction and natural language processing require certain pre-processing tools to analyze the lexical, morphological, phonetic, syntactic and semantic structure of the text. Named Entity Recognition is one of the text pre-processing tools which plays a vital role in different natural language applications such as Automatic Text Summarization [1], Machine Translation [2], Information Retrieval [3], Question Answering [4], etc.

1.1. What is Named Entity?

The term "Named Entity" was first considered important for information extraction task by the MUC-6 [5]. A named entity is a word form that recognizes the elements having similar properties from a collection of elements. It is called as a rigid designator or an atomic element or member of the semantic class which may vary depending upon the domain of interest. For example — in Biomedicine domain, entities of interest are gene and gene products; in general domain, person, location, organization, number, date, time, etc. are important entities; in the homeopathic domain, drug names and disease names are recognized as entities.

1.2. Named Entity Recognition and Classification (NERC)

Named Entity Recognition and Classification, an important subtask of Information Extraction [6], points to identify and classify members of rigid designators from data suited to different types of named entities such as organizations, persons, locations, etc. [7]. The concept of named entity came into existence with the introduction of MUC-6 [5]. To achieve the main goal of the conference, named entities played an important role by extracting ENAMEX (person, location, organization) and NUMEX (time, currency and percentage expressions) entities out of the structured information related to company activities as well as the unstructured text of military messages. After that various scientific events such as Information Retrieval and Extraction (IREX) Program, 2000 [8], Conference on Natural Language Learning 2002 (CONLL 2002) [9], Conference on Natural Language Learning 2003 (CONLL 2003) [10], Automatic Content Extraction (ACE) Program [11], HAREM [12], etc. gave major contribution in emergence of NER. Since then, Named Entity Recognition has become a fascinating field to be studied.

Till now, different entities have been recognized in different languages and in different domains using different approaches. Earlier systems were based on handcrafted rule-based algorithms which provided better results for restricted domains only but modern systems most often rely on machine learning based algorithms which overcome the drawbacks of rule-based systems. A number of factors are there which can make a difference to the performance of NERC like textual genres, types of entities, language, etc. NERC system built for one domain is quite challenging to port into another domain. To the extent of our knowledge, Named Entity Recognition techniques in different domains have not been found extensively which encouraged us to throw light on this field. So we present here a survey of developments and progresses made in NER research.

1.3. Motivation for conducting the survey

Today in the age of the internet, a vast amount of information is available online and is increasing every moment. The information stored can be in structured or unstructured form and needs to be extracted in a well-processed form suitable for the application using it. Robust information extraction techniques are needed to be explored to make the proper utilization of stored data on the web. Named Entity Recognition and Classification tool is the key component of Information Extraction. NERC recognizes and classifies the entity mentions of interest, useful in many natural language applications. The motivation of conducting this survey is to highlight the present status of NERC techniques developed by research community yet and to identify numerous issues and challenges as well as factors affecting the NERC performance which are to be considered carefully while designing these systems.

Download English Version:

https://daneshyari.com/en/article/6891621

Download Persian Version:

https://daneshyari.com/article/6891621

Daneshyari.com