Intuitive Approaches for Named Entity Recognition and Classification: A survey

Mukta Takalikar
Research Scholar
Yeshwantrao Chavan College of Engineering, Nagpur

Manali Kshirsagar, Ph.D
Professor
Yeshwantrao Chavan College of Engineering, Nagpur

Gauri Dhopavkar
Research Scholar
G.H.Raisoni College of Engineering, Nagpur

ABSTRACT

The survey of research in the field of Named Entity Recognition and Classification (NERC) features, techniques and evaluation methods, is presented, though it is not extensive and may not cover all the languages. It gives the depth of previous work in the field. Automatic named entity recognition and classification in the text surely improves the quality of results while searching the web. Highly accurate Named Entity Recognition (NER) is a challenge even today. The output of NER system is used for question answering, document clustering, document summarization [9].

General Terms

Machine Learning, Text Mining

Keywords

Named Entity Recognition, Named Entity Extraction, Natural Language Processing

1. INTRODUCTION

Named entity recognition and classification is information extraction subtask where structured information referred as named entities are extracted from the unstructured text. The named entities are units carrying a well defined semantics. Names of persons, locations, organizations, phone numbers and dates are generic named entities whereas names of genes, proteins, enzymes in the biological domain are domain specific named entities [8].

2. CHALLENGES IN NER

It is a technical challenge to build NER systems that performs exactly as that of human because of complex interrelations among various parts of sentence and the variety of languages (e.g. Hindi does not have capitalization clues). The challenges are:
Open nature of vocabulary
Clues such as capitalization
Overlap between NE Types
Indirect occurrences of NE
Different ways of referring to same entity

The effective NER systems use large amount of common-sense knowledge.

2.1 Characteristics of system for NER

The NER system has to be robust while facing noise in the form of spelling and grammatical errors, highly accurate in output, portable or language independent, largely domain independent and extensible to extend the rules and gazzetters.

2.2 NER features

Descriptors or characteristic attributes of words designed for algorithmic consumption are referred as features. The features are :

- Word form and POS tags (if available)
- Orthographic features: Like capitalization, decimal, digits
- Word type patterns: Conjunction of types like capitalized, quote, functional etc.
- Bag of words: Word forms, irrespective of position
- Trigger words: Like New York **City**
- Affixes Like Hydera**bad**, Ram**pur**, Mehdi**patnam**, Lingam**pally**
- Gazetteer features: class in the gazetteer
- Left and right context
- Token length: Number of letters in a word
- Previous history: Classes of preceding Named Entities

3. APPROACHES

The various approaches are used for named entity recognition and classification

3.1 Supervised learning approaches

NER is treated as a classification problem with labelled training dataset as input used by the classification algorithm for the discovery of set of rules .Various approaches to Classification Algorithm uses machine learning, pattern recognition and statistical literature. The models used are Hidden Markov Model (Bikel *et al* 1999), (Seymore *et al* 1999), (Collier et al 2000), (Miller et al 1998), (Klein *et al* 2003). HMM approach has also been used for NER in languages other than English. HMM approach has also been used for Domain Specific Named Entity regognition; e.g., biomedical domain (Shen et al 2003), (Zhang et al 2002), (Zhao 2004); Liu *et al* 2005 who used HMM for identifying NE such as product names., The other models used are Maximum entropy model, Support Vector Machines, decision trees and conditional random fields[7]

3.2 Un-Supervised learning approaches

They are also referred as bootstrapping or weakly supervised approaches.NER uses seed list along-with large set of unlabelled examples.CRF is used to create gazetteers. The steps used in unsupervised learning are : Use Seed examples, train the classifier, add new examples and retrain. The task of fine-grained NER based on ontology uses unsupervised approaches.

In the unsupervised work (Watanabe et al 2003) uses CRF to create gazetteers from Wikipedia. (Jimeno *et al* 2008) compares various NER methods for automatically creating a gazetteer as well as an annotated NER corpus for disease names in medicine. Given a seed list of NE type examples, (Talukdar *et al* 2006) learns a pattern (as an automaton) from their contexts (*k* words before and after). The contexts are pruned using the IDF measure and then an automaton is induced from the context using a grammatical induction algorithm.

3.3 Rule based approaches

Set of handcrafted syntactic and semantic rules are used for identifying named instances. The robustness and portability is missing in these approaches.

Examples of well known rule based NER systems are Univ. of Sheffield’s LaSIEII (Humphreys *et al* 1998), ISOQuest’s NetOwl (Krupka and Hausman 1998), Facile (Black *et al* 1998), SRA (Aone *et al* 1998) and Univ. of Edinburgh’s LTG system (Mikheev *et al* 1999) and FASTUS (Appelt 1998) for English NER

3.4 Named Entity Extraction (NEX)

The NEX task is quite similar to the unsupervised approaches, except that the task is not to learn rules for NER but to create a *gazette* (*list* or *gazetteer*) of examples of the NE. Also, NEX is often applied to learn from web pages rather than documents. The idea is that once a comprehensive list of NE examples is created, NER in a given document corresponds to simple look up in this list.

An excellent survey of NER literature is done in (Nadeau and Sekine 2007. Detailed guidelines, issues and examples for NER are discussed in (Chinchor 1998), (Sang et al 2003). (Ratinov and Roth 2009) discuss some interesting issues and challenges in NER - particularly, the choice of an inference mechanism and representation of text chunks.

4. ARCHITECTURE

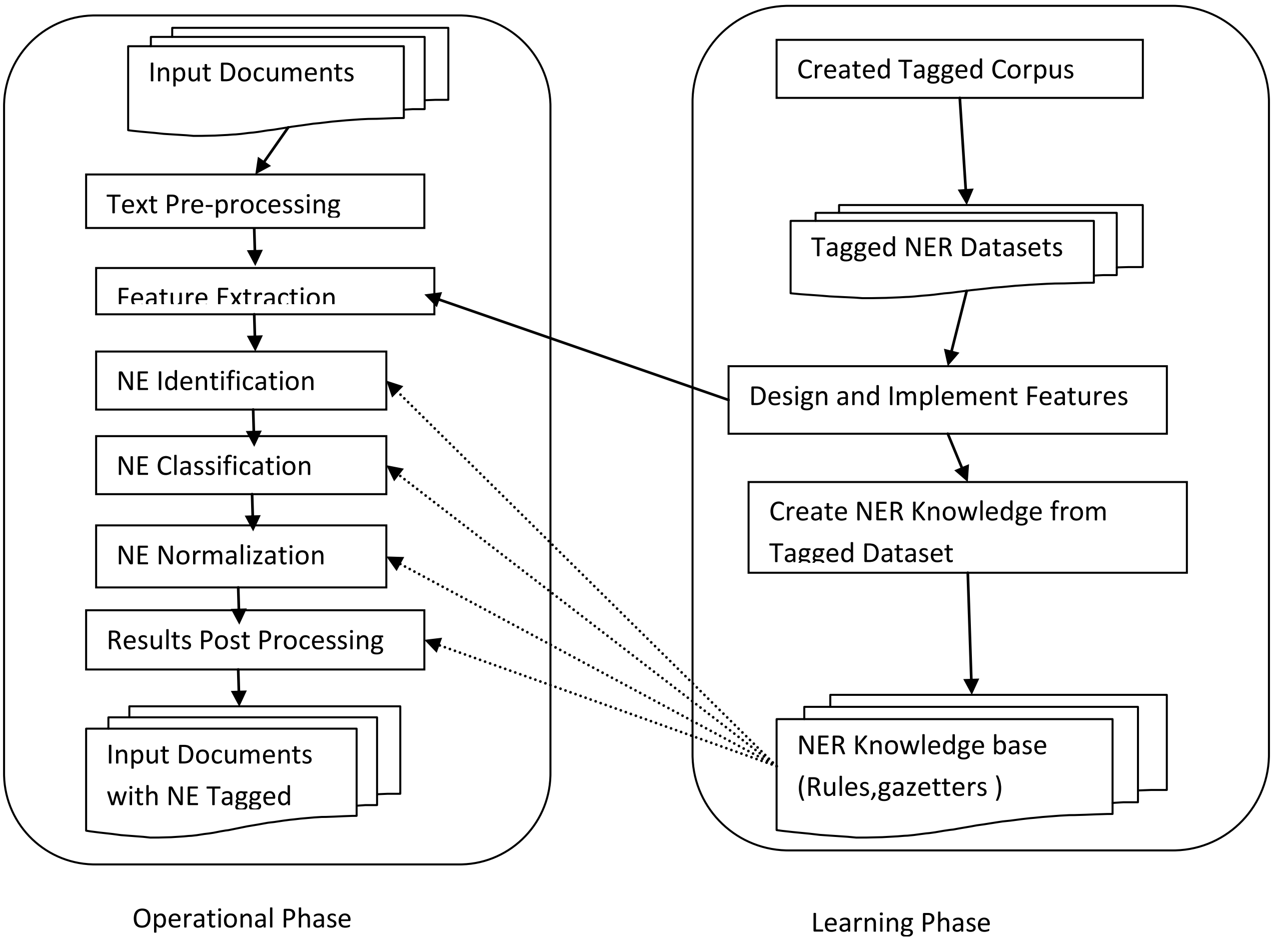

Fig. 1: Architecture of typical supervised learning system

5. ACCURACY OF ALGORITHM

Let $A = \{a1, a2, . . ., a_N\}$ denote (multi)set of N occurrences of the chosen type of NE in the test corpus.

Let $B = \{b1, b2, . . ., b_M\}$ denote the (multi)set of the M occurrences of the chosen type of NE identified by the algorithm in the test corpus.

An occurrence *bi € B* is classified as a *true positive* (*TP*) (as *false positive* (*FP*)) if *bi €A* (*bi ∉ A* respectively).

Thus the number of true positives identified by the algorithm is the number of occurrences which are in both B and A i.e., $\#TP = |A \cap B|$. The number of occurrences which are in B but not in A is the number of false positives: $\#FP = |B - A|$. An occurrence a_i in A is classified as a *false negative* (*FN*) if $a_i \notin B$. The number of occurrences which are in A but not in B is the number of false negatives: $\#FN = |A - B|$. Then the precision P, recall R and F-measure accuracy of the algorithm are:

$$P = \frac{\#TP}{|B|} = \frac{\#TP}{\#TP + \#FP}$$

$$R = \frac{\#TP}{|A|} = \frac{\#TP}{\#TP + \#FN}$$

$$F = \frac{2PR}{P + R}$$

6. STEPS FOR NER EVALUATION

a) Randomly select training and testing sets in 80%-20% proportion from the corpus

b) Training set is used to learn and tune the NER Knowledge base

c) The entities are extracted from the test set documents using learned and tuned NER knowledge base

d) Compute the precision, recall and overall accuracy measure i.e. F-Measure

6.1 English Language Tagged Datasets:

The table lists some tagged English corpora that has been used by researchers for different NER tasks

Table 1: Available tagged English corpora for NER

Corpus	Available at URL
ACE corpora	http://www.ldc.upenn.edu/Projects/ACE/
coNLL 2002 shared task corpora	http://cnts.uia.ac.be/conll2002/ner/
coNLL 2003 shared task corpora -	http://cnts.uia.ac.be/conll2003/ner/
GENIA	http://www-tsujii.is.s.u-tokyo.ac.jp/GENIA/home/wiki.cgi
MUC-7 corpus	http://www.itl.nist.gov/iaui/894.02/related projects/muc/proceedings/muc7proceedings

Tagging is same as NE annotation. (Fort *et al* 2009) has given guiding principles and a methodology for creating effective tagged NE datasets. Set of Guidelines are available at LDC. The Message Understanding conferences MUC-6 and MUC-7 had a special track for NER tasks.

7. RELATED WORK

NER is often accompanied by some post-processing to correct classification errors that may have occurred. (Lin *et al* 2004) propose a simple method to correct classification errors. A method for correcting NE boundary errors when only part of the NE has been detected correctly (e.g., rules for extending the detected NE to right or left) is also proposed by them. Combining the outputs of several NER systems, in the spirit of classifier ensembles, also has not received as much attention as it should have. Such classifier ensemble methods have shown promise in that the overall accuracy is better than that of the constituent classifiers, in standard statistical classification tasks (not necessarily NER). (Florian *et al* 2003) uses a class-error based voting scheme to combine the outputs of NER classifiers based on ME, HMM, Robust risk minimization and transformation-based learning. (Thao *et al* 2007) compares 3 voting mechanisms (majority, total accuracy, class-wise accuracy) to combine CRF, SVM, Naive Bayes and decision tree based NER classifiers for Vietnamese (see also (Tsai *et al* 2006)). (Wang and Patrick 2009) reports a combination scheme to combine SVM, ME and CRF classifiers and its application to perform NER from clinical notes. (Ekbal and Bandyopadhyay 2010) use a majority voting approach to combine NER classifiers for Bengali based on ME, CRF and SVM and demonstrate an increase of about 11% over the best performing SVM classifier for this task. Systematic comparison of various NER techniques, particularly for different languages, over different domains and across varied and unseen corpora, is an important issue. (Krishnarao *et al* 2009) compare CRF and SVM based NER systems for Hindi. (Petasis et al 2004) compares the performance of different NER systems on English, French, Greek and Italian web-pages. (Sekine and Eriguchi 2000) compare various techniques (ME, Decision Tree, HMM as well as hand-crafted

In ACE04 conference (Doddington *et al* 2004).Kernel-based approaches are being explored for Named Entity Relation Regognition(NERR) in particular and relation extraction in general; refer, for example, (Zhao and Grishman 2005) and (Culotta and Sorensen 2004) [9]

Indic Language/Asian/ European NER

Lack of capitalization, different word order, richer morphology, gender sensitive word forms in other languages make NER harder task in comparison with English language NER The techniques are developed for other languages based on Linguistic knowledge and characteristics of a particular language called as Language specific NER techniques and it is feasible to apply language independent NER technique to the class of related languages [9]

8. REFERENCES

[1] Barthowick, A. 1999, A maximum entropy approach to named entity recognition. Unpublished doctoral dissertation, New York University.

[2] Bikel, D. M., Schwartz, R., Weischedel, R. M. 1999. An algorithm that learns what's in a name. Machine Learning, 34, 211–231. doi:10.1023/A: 1007558221122

[3] Collins, M. Singer, Y. 1999. Unsupervised models for named entity classification. Proc. EMNLP,pp. 100-110.

[4] Etzioni, O., Cafarella, M., Downey, D., Popescu, A.-M., Shaked, T., Soderland, S., Weld, D.S. and Yates, A. 2005. Unsupervised named-entity extraction from the Web: An experimental study. Artificial Intelligence, 165, pp. 91–134.

[5] Ekbal, A., & Bandyopadhyay, S. 2010. Improving the performance of a NER system by post-processing and voting.In Structural, Syntactic and Statistical Pattern Recognition, LNCS 5342 (pp. 831−841), Springer.

[6] Krishnarao, A., Gahlot, H., Srinet, A., & Kushwaha, D. 2009 A comparison of performance of sequential learning algorithms on the task of named entity recognition for Indian languages. In proceedings of the International Conference on Computational Science (ICCS 2009), LNCS 5544, (pp. 123–132), Springer.

[7] Nadeau, D., Turney, P. and Matwin, S. 2006. Unsupervised named-entity recognition: generating gazetteers and resolving ambiguity. Proc. 19th Canadian Conf. Artificial Intelligence.

[8] Nadeau, D., & Sekine, S. 2007.A survey of named entity recognition and classification. Lingvisticae Investigationes, 30, 3–26. doi:10.1075/li.30.1.03nad

[9] Palshikar, G.K., 2011. Techniques for named entity recognition: a survey. TRDDC Technical Report, pp.191-217.

[10] Thelen, M. and Riloff E. 2002. A bootstrapping method for learning semantic lexicons using extraction pattern contexts. Conference on Empirical Methods in natural Language Processing (EMNLP 2002).

[11] Talukdar, P., Brants, T., Liberman, M., & Pereira, F. 2006.A context pattern induction method for named entity extraction .In proceedings of the 10th Conference on Computational Natural Language Learning (CoNLL-2006), (pp. 141–148).