International Journal on Recent and Innovation Trends in Computing and Communication ISSN: 2321-8169
Volume: 4 Issue: 11 41 – 46

Review Paper on Named Entity Recognition and Attribute Extraction using Machine Learning.

Hiba Momin
Student, Department of Information Technology, M.C.T Rajiv Gandhi Institute of Technology.
Mumbai, India.
mominhiba@gmail.com

Shubham Jain
Student, Department of Information Technology
M.C.T Rajiv Gandhi Institute of Technology.
Mumbai, India
shubhamj155@gmail.com

Hemil Doshi
Student, Department of Information Technology
M.C.T Rajiv Gandhi Institute of Technology.
Mumbai, India.
hemildoshi96@gmail.com

Prof. Ankush Hutke
Professor, Department of Information Technology
M.C.T Rajiv Gandhi Institute of Technology.
Mumbai, India.
ankush.hutke@mctrgit.ac.in

***Abstract*—**Named entity recognition (NER) is a subsidiary task under information extraction that aims at locating and classifying named entities in the text provided into pre-defined categories such as the names of people, locations, organizations, etc. In focused NER, once the entities are recognized we further aim at finding the most important named entities among all the others in a document, which we refer to as focused named entity recognition. We implement this using a classifier approach, i.e. Naïve Bayes classification, and we show that these focused named entities are useful for many natural language processing applications, such as document summarization, search result ranking, and entity detection and tracking. Attribute extraction on the other hand, involves automatic selection of attributes in your data (such as columns in tabular data) that are most relevant to the predictive problem you are working on. We try to implement an approach to extract the entities' attributes from unstructured text corpus. The proposed method is an unsupervised machine learning method that extracts the entity attributes utilizing deep belief network (DBN), we work on training data sets that we extract via web scraping tools, and test files for the same. Our goal can be twofold in this respect, firstly we can aim at simply organizing information so that it is useful to people, or put it in a semantically precise form to make further inferences.

Keywords: *Named Entity Recognition (NER), Naïve Bayes Classifications, Deep Belief Networks (DBN).*

*****

1. Introduction

The whole world has entered the era of big data. Dealing with this massive text effectively and efficiently has become an urgent problem in front of us. With the rapidly increasing growth of online electronic documents, a lot of technologies have been developed to deal with the enormous amount of information, such as: automatic text summarization, topic detection and tracking, and information retrieval. Information extraction involves finding and understanding very limited but relevant parts of documents. Based on this, structured representation of the relevant information is created.

An important task is to identify the main topics of a document; wherein subjects can be represented by words, sentences, concepts, and named entities. Entity refers to an independent existence of things. Each entity has its own characteristics, i.e., different entities have their specific attributes which are easily distinguished from all other entities. The name of Entity often represents the species, that have the same nature as other nouns. People specify that name for each entity, which is also known as Named Entity, NE. The entities with the same categories have similar attributes, but they are different in the values of the property. various types of entities have various properties. Any abstract object may be called an entity which is quite different from other information extraction applications. a user's interest can be also defined as an entity, such as people,products, etc. Moreover, objects that appear in the corpus can all be defined as an entity. Entities with various types have various attributes and information characteristics.

Our definition of focused named entities is concerned with What and Who. Therefore it is self-evident that the concept of focused named entity plays a key role for document understanding and automatic information extraction. Moreover, we shall illustrate that focused named entities may be used in other text processing tasks as well. For example, by giving more weight to focused named entities we can increase search results. We can define focused named entities as entities that are highly relevant to the main topic of a news article. we further investigated the machine learning approach to the problem, which is the focus. We discuss various issues encountered in this process of building a machine learning based system and show that our method may achieve near

human performance. Entity Attribute Extraction is yet another important technology in the field of natural language processing, and not only can the information obtained can be provided to the users directly but also used as the basis of building the intelligent query and data mining.

As a key aspect of information extraction, entity attribute may be used to define a new entity, handle entity mining and other applications. The sole purpose of the study on information extraction is to get the structured information from the natural language text. The task of entity attribute extraction is to let the computer fetch the attributes and their values by itself. While entities with the same category have the same attribute information structure, the value of each such attribute will be different. For eg there are general attributes of a people entity: full name, occupation, work units, mail, telephone, etc.; the attributes of organization or unit entity: the name of institution or unit, address, department, responsible person, the nature of services,etc. And the typical attributes of product entities: product name, manufacturer, product function, art, price, brand, characteristics, and so on. In recent years, with the rapid development of search engine technology, searching has become more and more intelligent. The search engine evolved from the "keywords search" to "SNS Search" and "Entity search". Entity Search is more complicated than the keywords Search. Although the traditional keyword search has developed well, the results provided by the search engine can help users find the information, but in fact, for the "Search Engine" system itself, it does not understand the meaning of the search. The primary focus on Entity search is not the "key words" but the object, such as people, institutions, organizations, etc. We hope that a conversion from keywords to an entity can help search engine understand and organize search results from a more subtle point of view

2. Literature Survey

The first paper of research NER was presented at the Seventh IEEE Conference on Artificial Intelligence Applications by Lisa F. Rau (1991). Rau's paper describe a system that ―extract and recognise [company] names, it relies on heuristics and handcrafted rules. From 1996, with the first major in task MUC-6, it never declined since then with steady research and numerous scientific events: HUB-4, MUC-7 and MET-2, IREX, CONLL, ACE and HAREM. The Language Resources and Evaluation Conference (LREC) has also been staging workshops and main conference tracks on the topic since 2000.

Named Entity Recognition Systems have been designed to use linguistic grammar-based techniques and statistical models. Handcrafted grammar-based systems are usually obtain better precision. Statistical NER systems typically require a large amount of manually commented training data. It normally finds the sequence of tags that maximizes the probability p(N|S), wherein S is the word sequence in a sentence, and N is the sequence of named-entity tags which are assigned to the words in S. English is the most popular langrage factor to research NER, but with the development of research in these areas, more and more kinds of languages have been researched. German is well studied in CONLL-2003 and in even earlier works. Similarly, Spanish & Dutch are strongly represented, boosted by a devoted conference: CONLL-2002. Japanese has also been studied in the MUC-6 conference, the IREX conference and in other works. Chinese is studied in an abundant literature, and so are French,Italian and Greek. And then many other languages have paid more attention to this area. Finally, Arabic has started to receive a lot of attention in large-scale projects like Global Autonomous Language Exploitation (GALE).

Parts of Speech Taggers also called word-category disambiguation or POS taggers. It reads a text in some language and assigns parts of speech to each word, such as noun, verb, adjective, etc. It based on both its definition, as well as its context, for example, relationship with adjacent and related words in a phrase, sentence, or paragraph. Labelling part of speech is harder than simply having a list of words and their parts of speech because some words can represent more than one part of speech at different times. For example, the word ―work, can be considered as noun or verb. Some of the NER systems are incorporated into Parts of speech taggers. Moreover, most of the NER systems are based on analyzing patterns of POS taggers.

Dean A. Pomerleau, in his research presented in the paper "Knowledge-based Training of Artificial Neural Networks for Autonomous Robot Driving," uses a neural network to train a robotic vehicle to drive on multiple types of roads (single lane, multi-lane, dirt, etc.). A large amount of his research is devoted to 1. extrapolating multiple training scenarios from a single training experience and 2. preserving past training diversity so that the system does not become over trained (if, for example, it is presented with a series of right turns – it should not always learn to turn right). These issues are very common in neural networks that must decide from amongst a variety of responses, but can be dealt with in a number of ways, for example by randomly shuffling the training examples, by using a numerical optimization algorithm that does not take too large steps when changing the network connections following an example, or by grouping examples in so-called mini-batches.

A. K. Dewdney, a computer scientist and a mathematician at University of Western Ontario and former Scientific American columnist, wrote, "Although neural nets do solve a few toy problems, their powers of computation are so limited that I am surprised anyone takes them seriously as a general problem-solving tool". There aren't any neural networks that have yet solved computationally difficult problems such as the n-Queens problem, the travelling salesman problem, or the problem of factoring large integers.

3. Motivation

In the current market scenario, big data is at the crowning point of the latest technology. Big data involves not just collection but manipulation in a way that we can develop prescriptive and predictive models from it, and extract patterns that prove of value to the customers. Machine learning involves the task of providing the computer with a decision making capability, basically a computer mimics the human mind and develops intelligent behavior. Within machine learning algorithms, we have the task of natural language processing. NLP has the subtask of Information extraction, and many other applications like named entity recognition, speech recognition, optical character recognition, word sense disambiguation, etc.

4. Methodology

A. Web Scraping

Web scraping is a software technique that facilitates information extraction from websites. This technique focuses on the transformation of unstructured data which is essentially in HTML format, on the web into structured data i.e. database or spreadsheets.

You can perform web scraping in various ways, including the use of Google Docs with almost any programming language. We would resort to Python because of its simplicity and rich eco-system. It has a library known as 'Beautiful Soup' which assists this task. Beautiful Soup is an incredible tool for pulling out information from a webpage. It can be used to extract tables, lists, paragraphs; filters can be used to extract information from web pages.

METHOD:

1. Import the library used to query a website
2. Specify the url.
3. Query the website and return the html to the variable.
4. Import the Beautiful soup functions to parse the data returned from the website
5. Parse the html in the variable, and store it in Beautiful Soup format.

B. Performing entity recognition

Named entity recognition (NER) is a subsidiary task under information extraction that aims at locating and classifying named entities in the text provided into pre-defined categories such as the names of people, locations, organizations, etc.

In the recent times the cost incurred on storing data has reduced whereas there has been a significant increase in computing power. Thus data scientists and data developers can build knowledge bases with copious amounts of entities and attributes assigned with those entities, that is, data and the metadata attached. These knowledge bases contribute in developing machine learning algorithms.

The classification is performed using naïve bayes techniques. **Classification** is the task of choosing the correct **label** for a given entry. In basic classification tasks, each input is considered isolated from all other inputs, and the set of labels is defined beforehand. Some examples of classification tasks are:

1. Deciding whether an email is spam or not.
2. Deciding what the topic of a news article is, from a fixed list of topic areas such as "sports," "technology," and "politics."
3. Deciding the context of a given occurrence of a word, for instance the word *bank* could be used to refer to a river bank, a financial institution, the act of tilting, or the act of depositing something in a financial institution.

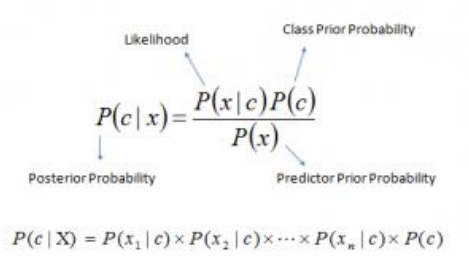
We need to train and test our algorithm, before which we need to split up the data into a training set and a testing set.

Training and testing should never be performed on the exact same data set as this would result in serious bias issues.

One technique to deal with this problem is shuffling the data set, we choose, say 80% of our data set values as the training set, that contains both the positive and negative labels, and then test against the remaining 20% to see the accuracy. This is termed as supervised machine learning, because what we are doing here is feeding data to the machine that segregates some data as positive and other as negative, after this 'training' is complete we show the machine some new data and ask the machine to classify it as positive or negative for us, based on what we taught it before, thus we are 'testing' the machines ability to make decisions.

A Naive Bayes classifier assumes that the presence of a particular featture in a class is unrelated to the presence of other features in it. For example, a fruit may be considered to be a mango if it is yellowish-orange, round at the bottom and caves upwards, and about 3 inches in diameter. Even if these features depend upon the existence of other features, all of these properties independently contribute to the probability of this fruit being a mango and that is why it is known as 'Naive'.

Bayes theorem provides a way of calculating posterior probability using bayes formula, P(c|x) from P(c), P(x) and P(x|c). Studying the equation below:



Above,
$P(c|x)$ is the posterior probability of *class* (c, *target*) given *predictor* (x, *attributes*).
$P(c)$ is the prior probability of *class*.
$P(x)$ is the prior probability of *predictors*.
$P(x|c)$ is the probability of *predictor* given *class*.

C. Attribute extraction

Deep Belief Network:

Deep Belief Networks (DBNs) are graphical models which learn to extract a deep hierarchical representation of the training data. By training the weights between neurons we can make the whole neural network generate the training data according to the maximum probability. The structure of our DBN network is a kind of deep neural network that is composed of several layers of Restricted Boltzmann Machines (RBM) and a layer of BP.

Feature Extraction:

Recognize named entities and then form the entity feature which is defined by the category of the recognized entity.

Analyze the text with semantic parser, get the Object Structure characteristics of the syntax tree.

Combine three features that have been discovered. Feed the integration of the feature set as parameters into DBN neural network, and then get the model of entity attributes extraction.

Table 4.1. Recognize Named Entities

Word	Tag
Confidence	B-NP
in	B-PP
The	B-NP
pound	I-NP
…	…

Feature vectors composition is as follows:
Etype: Entity category information feature.
Pos: Position features affects the relationship of words. In this paper, we concentrate on extracting nouns, verbs, quantifiers, prepositions and other features.
Tag: It is the result of the syntax parser.

Table 4.2. Characteristics of Parameters and Threshold Range

Feature	Threshold range
Etype	Names,Palce,Organization
Pos	Noun,Verb,Adjective,Numeral…
Tag	SBV(subject-verb),VOB(verbobject),HED(head),IOB(indirect object),FOB(frontingobject),COO(coordinate) …

5. Proposed System

Step 1: The first step of collecting data is done by web scraping.Web scraping also called as web harvesting or web data extraction which is a software technique of extracting and retrieving information from websites. This is achieved by either directly implementing the Hypertext Transfer Protocol (HTTP) or embedding a web browser.

Step 2: The corpus is a huge chunk of plain text obtained from the Internet, or from some other source which provides any document readable by the computer, it contains noisy text. Therefore, we must firstly process the corpus to get pure unstructured text.

Step 3: Authority information associated with the words by comparing the words with the authority list that has the most common uses of the words is located in this step, this task enables us to extract named entities and based on historical data we also predict named topical entities.

Step 4:The compounds are processed through the neural networks to generate metadata guesses, we then implement a deep belief network which is described below.

DBN model training process is divided into two steps. Firstly, we train each layer of the RBM network respectively, ensuring that the feature vectors are being mapped to different feature space, and then the features are preserved.

Step 5: Finally we extract the attribute features through the deep belief network which has further applications.

Step 6: In this step the generated list of entities and attributes can be used for various NER applications..

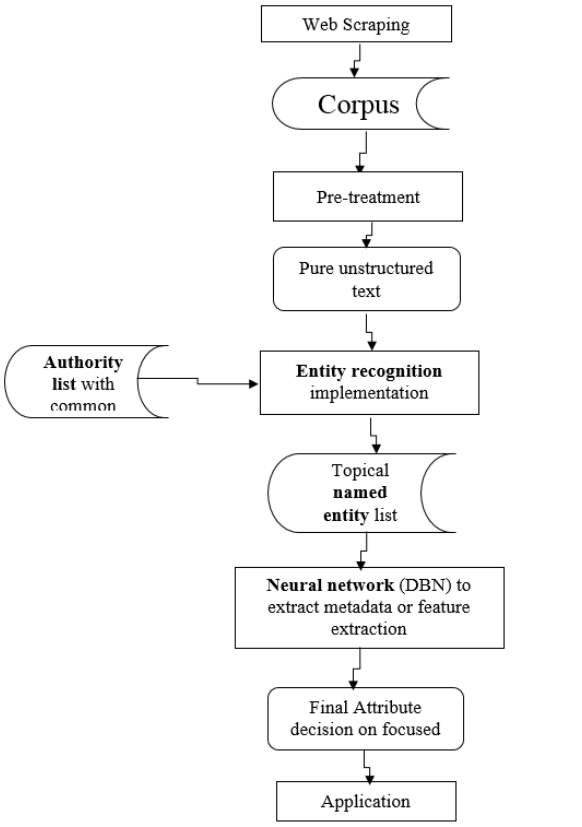


Fig: Proposed system

6. Feasibility Study

A. TECHNICAL FEASIBILITY:

Researches have shown that sometimes NER systems can be fragile as the system made for one domain cannot be used for any other domain. A lot of efforts are required to tune the NER systems so that they perform better in other domains.
In the nineties work in the NER domain was mainly focused at extraction from articles of journals. Later attention was turned to processing of reports and military dispatches. Further stages of content extraction included several types of informal text styles, like weblogs and text transcripts from telephone speech conversations. There has always been a huge amount of interest in entity identification in the domains of molecularbiology, bioinformatics, natural language processing. The name of the gene and gene products have been the most common entity of interest in these domains.
All the main efforts are aimed towards reducing the time consumed in annotation by employing semi-supervised learning, robust performance across domains and scaling up to fine-grained entity types. For supervised and semi-supervised machine learning approaches to NER many projects have turned towards crowdsourcing which is a promising solution for obtaining high quality human judgements.
Here we implement NER using neural networks. A common shortcoming of neural networks, is that they require a huge diversity in training for real-world operation. This isn't surprising, since any learning machine needs adequate representative examples in order to capture and understand the underlying structure that allows it to generalize and use it for new cases.

The major factors affecting the feasibility of the System are:

1. Length of the input text
2. Required application to be implemented.
3. Varying number and types of entities present in the huge chunk of text.
4. The capability of the system to handle large sizes of text.

From the factors above, the capability of the system to handle large sizes of text is a negligible point in itself, since the vast majority of systems nowadays can work with text seamlessly without any sort of issue. Since the scope of this project is limited to (at the current moment at least) simple documents having a maximum of 6-7 paragraphs of text, this should not be a problem since at the most, such a document will be of 100 kB.

B. ECONOMIC FEASIBILITY

To implement large and effective software neural networks, a considerable amount of p storage and processing resources are required. Simulating even the most simplified Von Neumann architecture requires filling up of thousands and millions of rows of databases for its connections which require large amounts of hard disk space and computer memory. The designer of neural network systems will often need to simulate the transmission of signals through many of these connections and their associated neurons that requires incredible amounts of CPU processing power and time. The computing power continues to grow according to Moore's Law, that may provide sufficient resources to complete new tasks. Non von Neuman chips in the circuits are designed to implement neural nets and neuromorphic engineering caters to this difficulty directly.
Tensor Processing Unit(TPU) is a chip optimized for neural networks processing designed by Google.

References

[1] R. Barzilay and M. Elhadad. Using lexical chains for text summarization. In Proceedings of the ACL
[2] F. J. Damerau, T. Zhang, S. M. Weiss, and N. Indurkhya. Text categorization for a comprehensive time-dependent benchmark. Information Processing & Management, 2004.
[3] H. P. Edmundson. New methods in automatic abstracting. Journal of The Association for Computing Machinery, 16(2):264–285, 1969.
[4] J. Y. Ge, X. J. Huang, and L. Wu. Approaches to event-focused summarization based on named entities and query

words. In DUC 2003 Workshop on Text Summarization, 2003.

[5] E. Hovy and C.-Y. Lin. Automated text summarization in summarist. In I. Mani and M. Maybury, editors, Advances in Automated Text Summarization, pages 81–94. MIT Press, 1999.

[6] D. E. Johnson, F. J. Oles, T. Zhang, and T. Goetz. A decision-tree-based symbolic rule induction system for text categorization. IBM Systems Journal, 41:428–437, 2002.

[7] M.-Y. Kan and K. R. McKeown. Information extraction and summarization: domain independence through focus types. Columbia University Computer Science Technical Report CUCS-030-99.

[8] J. M. Kupiec, J. Pedersen, and F. Chen. A trainable document summarizer. In SIGIR '95, pages 68–73, 1995.

[9] D. Lawrie, W. B. Croft, and A. Rosenberg. Finding topic words for hierarchical summarization. In SIGIR '01, pages 349–357, 2001.

[10] F. Li and Y. Yang. A loss function analysis for classification methods in text categorization. In ICML 03, pages 472–479, 2003.

[11] C.-Y. Lin. Training a selection function for extraction. In CIKM '99, pages 1–8, 1999.

[12] C.-Y. Lin and E. Hovy. Identifying topics by position. In Proceedings of the Applied Natural Language Processing Conference (ANLP-97), pages 283–290, 1997.

[13] D. Marcu. From discourse structures to text summaries. In Proceedings of the ACL'97/EACL'97. Workshop on Intelligent Scalable Text Summarization, pages 82–88. ACL, 1997

[14] A. McCallum and K. Nigam. A comparison of event models for naive bayes text classification. In AAAI/ICML-98 Workshop on Learning for Text Categorization, pages 41–48, 1998.

[15] J. L. Neto, A. Santos, C. Kaestner, A. Freitas, and J. Nievola. A trainable algorithm for summarizing news stories. In Proceedings of PKDD'2000 Workshop on Machine Learning and Textual Information Access, September 2000.

[16] C. Nobata, S. Sekine, H. Isahara, and R. Grishman. Summarization system integrated with named entity tagging and ie pattern discovery. In Proceedings of Third International Conference on Language Resources and Evaluation (LREC 2002), 2002.

[17] C. D. Paice and P. A. Jones. The identification of important concepts in highly structured technical papers. In SIGIR '93, pages 69–78. ACM, 1993.

[18] J. R. Quinlan. C4.5: Programs for Machine Learning. Morgan Kaufmann, 1993.

[19] E. F. T. K. Sang and F. D. Meulder. Introduction to the conll-2003 shared task: Language-independent named entity recognition. In Proceedings of CoNLL-2003, pages 142–147, 2003.

[20] W.-M. Soon, H.-T. Ng, and C.-Y. Lim. A machine learning approach to coreference resolution of noun phrases. Computational Linguistics, 27(4):521–544, 2001

[21]] S. Teufel and M. Moens. Sentence extraction as a classification task. In ACL/EACL-97 Workshop on Intelligent and Scalable Text Summarization, 1997.

[22] T. Zhang. On the dual formulation of regularized linear systems. Machine Learning, 46:91–129, 2002.

[23] T. Zhang, F. Damerau, and D. E. Johnson. Text chunking based on a generalization of Winnow. Journal of Machine Learning Research, 2:615–637, 2002.