©2010 International Journal of Computer Applications (0975 – 8887)
Volume 1 – No. 3

Conditional Random Field Based Named Entity Recognition in Geological text

Sobhana N.V
Indian Institute of Technology,
Kharagpur, West Bengal, India.

Pabitra Mitra
Indian Institute of Technology,
Kharagpur, West Bengal, India.

S.K. Ghosh
Indian Institute of Technology,
Kharagpur, West Bengal, India.

ABSTRACT
The paper describes about the development of a Named Entity Recognition (NER) system for Geological text using Conditional Random Fields (CRFs). The system makes use of the different contextual information of the words along with the variety of features that are helpful in predicting the various named entity (NE) classes. The NE tagged geological corpus was developed from the collection of scientific reports and articles on the geology of the Indian subcontinent has been used to build up the system. The training set consists of more than 2 lakh words and has been manually annotated with a NE tag set of seventeen tags. The system is able to recognize 17 classes of NEs with 75.8% F-measure.

Categories and Subject Descriptors
H. Information Systems : Information Storage and Retrieval

General Terms
Documentation

Keywords
Geological Corpus, Named Entity Recognition, Precision, Recall, F-measure, Geographic references

1. INTRODUCTION
Named Entity Recognition (NER) is a key part of information extraction system. NER involves identification of proper names in texts and their classification into a set of predefined categories of interest. Different categories are usually person names, location names, organization names, date & time expressions etc. A variety of techniques has been used for NER. The different approaches to NER include a. linguistic approaches b. machine Learning (ML) based approaches c. hybrid systems. The linguistic methods usually use rules manually written by linguists. There are several rule based NER systems, containing mostly lexicalized grammar, gazetteer lists, and list of trigger words, which are capable of providing up to 92% F-measure accuracy for English [1]. Linguistic approach uses hand crafted rules which require skilled linguistics. The main disadvantage of these rule based method is that they need vast experience and grammatical knowledge of the particular language or domain and these systems are not easily adaptable to other domains or languages [2]. Machine learning approaches are trainable and are thus much cheaper than that of rule-based ones. Some of the machine learning techniques used for the NER tasks are hidden markov model [3], Maximum Entropy Markov Model (MEMM) [4], Conditional Random Fields [5],[6]. Hybrid systems have been generally more effective for NER. Combination of MaxEnt, hidden markov model (HMM) and handcrafted rules for make creating NER is explored in [7].

Section 2 gives Characteristics of geological text. Section 3 discusses features used for Geological NER. Section 4 gives brief introduction to Conditional Random Fields, a machine learning approach to sequence labeling task. Section 5 describes the details of Geological Corpus. Section 6 explain the experiments and Results. The paper is concluded in section 7.

2. GEOLOGICAL NAMED ENTITY RECOGNITION
Geology is the study of origin, history and structure of the earth. Text mining on geological documents is an important area in scientific data mining. These documents contain spatial references and geo references in the form of spatial coordinates stored in database. They contain geospatial and temporal information. This spatial and temporal information is very important but normal text mining algorithms will fail to extract such information.

Named Entity Recognition (NER) is an important tool in almost all Natural Language Processing (NLP) application

areas. Proper identification and classification of named entities (NEs) are very big challenge to the NLP researchers. Geological NER has applications in several domains including information extraction, information retrieval, question answering [8], automatic summarization, machine translation [9] etc from Geological text.

Named entity recognition (NER) (also known as entity identification and entity extraction) is a subtask of information extraction that seeks to locate and classify atomic elements in text into predefined categories such as the names of persons, organizations, locations, expressions of times, quantities, monetary values, percentages, etc. The different categories of Geological Named Entities considered here are Country, State, City, Region, Mountain, Island, Water bodies, River, Village, Mineral, Year, Organization, Measures, Person, Time, Fault and Rock. We have used corpus based machine learning technique to recognize, classify, and identify these geological entities.

In general NER is a hard problem. Words have different applications and there is an infinite number of proper names. In English, the problem of identifying NEs is solved to some extent due to capitalization feature. Most of the named entities begin with a capital letter which is a discriminating feature for classifying a token as named entity. Geological documents contain spatial references and geo references in text. Major task in handling geographical references in text is Name Resolving. Many names do not have existence also. The same place name can be used for several locations (referent ambiguity) and same location can have multiple names (reference ambiguity). So if we have to uniquely identify place names in Geological document, the task becomes more complex and Geological named entity recognition, disambiguation are issues which need proper attention.

Many named entities (NEs) occur rarely in corpus. There are Ambiguity of NEs. There are many ways of mentioning the same NE. *Ex: Haradanahalli Doddegowda Deve Gowda* and *H. D. Deve Gowda* refer to the same person. *West Bengal, WB* both refer to the same location.

Our approach is the task of identifying and classifying tokens in a annotated Geological text corpus into predefined set of classes such as Country, State, City, Region, Mountain, Island, Water bodies, River, Village, Mineral, Year, Organization, Measures, Person, Time, Fault and Rock.

3. CHARACTERISTICS OF GEOLOGICAL TEXT

Geological documents contain textual description of geological phenomena, images and maps of geographic space in the form of spatial references, geo references and temporal information. Geographic references can be defined spatially using a point (ex. longitude and latitude) or a set of points. The information in the textual document such as place name and the corresponding linked geographic location is called geographic footprint. Geographic footprint is represented by coordinates (longitude, latitude)

2.1 Features used for Geological NER

Different features may be used for identifying NE's. The features aids in deciding to which class a named entity belongs. The main features for the NER task have been identified based on the different possible combination of available word and tag context. The features also include prefix and suffix for all words. The term prefix/suffix is a sequence of first or last few characters of a word, which may not be a linguistically meaningful prefix or suffix. We have considered the following feature set for NER task:

F= { W_{-2},W_{-1},W_i,W_{+1},W_{+2}, |prefix| <= 3, |suffix| <= 3, POS tag, Digit information, NE tag}

Context word feature: Previous and next words of a particular word can be as a feature.

Word prefix: A fixed length prefix of the current and/or the surrounding word(s) can be used as features.

Word suffix: Word suffix information assists in identifying NEs. This feature can be used in two different ways. The fixed length word suffix of the current and/or the surrounding word(s) can be used as a feature. For example, suffixes like -pur, -bad, etc are indicators of a name of a location.

Part of Speech (POS) Information: The POS of the current and/or the surrounding word(s) can be used as features.

Digit features: Several binary digit features have been considered depending upon the presence and/or the number of digits in a token (e.g., ContainsDigit [token contains digits], FourDigit [token consists of four digits], TwoDigit [token consists of two digits]), combination of digits and punctuation symbols (e.g., ContainsDigitAndComma [token consists of digits and comma], ContainsDigitAndPeriod [token consists of digits and periods]), combination of digits and symbols (e.g., ContainsDigitAndSlash [token consists of digit and slash], ContainsDigitAndHyphen [token consists of digits and hyphen], ContainsDigitAndPercentage [token consists of digits and percentages]). These binary valued features aids in recognizing miscellaneous NEs such as time expressions, date expressions, percentages, numerical numbers etc.

Named Entity Information: The NE tag of the current or previous word can be considered as the feature.

4. CONDITIONAL RANDOM FIELDS

Conditional Random Fields (CRFs) [10] are undirected graphical models used to calculate the conditional probability of values on designated output nodes given values assigned to other designated input nodes. A conditional random field (CRF) is a type of discriminative probabilistic model used for the labeling sequential data such as natural language text. Conditionally trained CRFs can easily include large number of arbitrary non independent features. The expressive power of models increased by adding new features that are conjunctions to the original features. When applying CRFs to the named entity recognition problem an observation sequence is the token sequence of a sentence or document of text and state sequence is its corresponding label sequence.

In the special case in which the output nodes of the graphical model are linked by edges in a linear chain, CRFs make first order Markov assumption and can be viewed as conditionally trained probabilistic finite automata (FSMs)

The conditional probability of a state sequence $s=<s_1,s_2\ldots.s_T>$ given an observation sequence $o=<o_1,o_2\ldots..o_T>$ is

$$P_{\wedge}(s/o)=\frac{1}{Z_o}\exp\sum_{t=1}^{T}\sum_{k}\lambda_k f_k(s_{l-1},s_l,o,t)$$

Where $f_k(s_{l-1},s_l,o,t)$ is a feature function whose weight λ_k is to be learned via learning. CRFs define the conditional probability of a label sequence based on total probability over the state sequences, $\mathbf{P(l/o)=\sum_{s:l(s)=l} P(s/o)}$ where l(s) is the sequence of labels corresponding to the labels of the states in sequences. Z_o is a normalization factor over all state sequences. To make all conditional probabilities sum up to 1, we must calculate the normalization factor $Z_{o=}\sum_{s}\exp\sum_{t=1}^{T}\sum_{k}\lambda_k f_k(s_{l-1},s_l,o,t)$. The feature functions could ask arbitrary questions about two consecutive states, any part of the observation sequence and the current position. For example a feature function may be defined to have a value 0 in most cases and have value 1 when s_{t-1}, s_t are certain states and the observation has certain properties.

5. DETAILS OF GEOLOGICAL CORPUS

The corpus was created out of a collection of scientific reports and articles on the geology of the Indian subcontinent. Many of these constituted reports submitted to the Earth Sciences Division of Department of Science and Technology, Government of India.

The corpus consists of about 200 documents with each document about 10,000 words long. They represented various aspects of geology. The lexicon consisted of several well known as well as rare geological terms. The titles of some of these documents are listed below.

Table I. Titles of documents in the corpus

Coastal Forms and Processes of the Godavari Delta
Retreat of Himalayan Glaciers: Indicator of Climate Change
Metamorphism of the Oddanchatram Anorthosite, Tamil Nadu, South India
K-T magmatism and basin tectonism in western Rajasthan, India, results from extensional tectonics

and not from Reunion plume activity
Crustal geotherm in southern Deccan Basalt Province, India: The Moho is as cold as adjoining cratons
Erosion and sedimentation in Kalpakkam (N Tamil Nadu, India) from the 26th December 2004 tsunami
Grain-size distribution, morphoscopy and elemental chemistry of suspended sediments of Pindari Glacier, Kumaon Himalaya, India

The NE tagged corpus help to identify named entities such as location, person, organization etc. Corpus has been manually annotated with the seventeen tags as described in Table II.

Table II. Named Entity Tagset

NE tag	Meaning	Example
COUNTRY	Country name	India
STATE	State name	West Bengal
WATERBODIES	Ocean name	Indian Ocean
MINERAL	Mineral name	Zinc
PERSON	Person name	Mahadevan
ORGANIZATION	Organization name	Geological Survey of India
COUNTRY_B COUNTRY_C COUNTRY_C	Beginning Internal or End of a multiword country name	India/COUNTRY_B
STATE_B STATE_C STATE_C	Beginning Internal or End of a multiword state name	West/STATE_B Bengal/STATE_C
WATERBODIES_B WATERBODIES_C WATERBODIES_C	Beginning Internal or End of a multiword ocean name	Indian/WATERBODIES_B Ocean/WATERBODIES_C
MINERAL_B MINERAL_C MINERAL_C	Beginning Internal or End of a multiword mineral name	Zinc/MINERAL_B
PERSON_B PERSON_C PERSON_C	Beginning Internal or End of a multiword person name	Jaya/ PERSON_B Surya/ PERSON_C
ORGANIZATION_B ORGANIZATION_C ORGANIZATION_C	Beginning Internal or End of a multiword organization name	Geological/ORGANIZATION_B Survey/ ORGANIZATION_C of/ ORGANIZATION_C India/ ORGANIZATION_C

6. EXPERIMENTAL RESULTS

A NE tagged Geological corpus (IITKGP-GEOCORP) has been used for NER experiment and it contains geology related information in India. This corpus is split into two sets. One forms the training data and the other forms the test data. They consist of 90% and 10% of the total data respectively. CRF is trained with training data and test data is tagged using CRF model.

More than 2 lakh words have been used as training set for the CRF based NER system. The size of the test file is 23K words and the data is labeled with 17 labels. We have used different standard measures such as Precision, Recall and F-measure for evaluation.

Recall is the ratio of number of NE words retrieved to the total number of NE words actually present in the file(gold standard).

Precision is the ratio of number of correctly retrieved NE words to the total number of NE words retrieved by the system.

These two measures of performance combine to form one measure of performance, the F-measure, which is computed by the weighted harmonic mean of precision and recall.

$$F = \frac{(\beta^2 + 1)PR}{\beta^2 R + P}$$

where β^2 represents the relative weight of recall to precision (and normally has the value 1).We have used the C^{++} based OpenNLP CRF++ package [11], a simple, customizable, and open source implementation of Conditional Random Fields (CRFs) for segmenting or labeling sequential data.

TABLE III
PERFORMANCE OF THE FEATURE SET (prefix and suffix of length up to three of the current word, information about the surrounding words, POS information, digit features, and NE tag)

	Class	Precision	Recall	F-measure
1	Country	98.75	90.29	**94.33**
2	State	95.83	94.52	**95.17**
3	City	81.11	73	76.84
4	Region	82.54	71.23	76.47
5	Mountain	92.31	85.71	88.89
6	Water bodies	84	72.41	77.78
7	Island	78.71	78.57	78.64
8	River	88.89	88.89	88.89
9	Village	70.59	50	58.54
10	Mineral	94.26	80.42	86.79
11	Organization	46.79	91.82	61.99
12	Measures	91.53	92.05	91.79
13	Year	96	97.27	**96.63**
14	Person	68.38	80.69	74.03
15	Fault	33.33	14.29	**20.00**
16	Rock	71.15	75.51	73.27
17	Time	35.71	76.92	**48.78**
	Overall	77.05	77.27	75.81

NE's with highest F-measure and lowest F-measure values are highlighted in bold. NE's like Fault and Time have low F-measure. The reason is that they have fewer instances in training data and are more difficult to learn. We have got Precision of 77.05%, Recall of 77.27% and F-measure of 75.81% by the combination of features (prefix and suffix of length up to three of the current word, information about the surrounding words, POS information, digit features, and NE tag) for identifying named entities. NE's such as Country, State and Year have high F-measure values because of their higher appearance in the corpus.

7. CONCLUSION

In this paper, we have developed a NER system using CRF with the help of a NE tagged Geological Corpus (IITKGP-GEOCORP). We also presented a new named entity tagset that was developed for annotation of this corpus. We have considered features such as prefix and suffix of length up to three of the current word, POS information, digit features, information about the surrounding words and their tags. Analyzing the performance using other methods like MaxEnt and Support Vector Machines (SVMs) will be other interesting experiments.

REFERENCES

[1] Wakao, T., Gaizauskas, V. and Wilks, Y. 1996. Evaluation of an algorithm for the recognition and classification of proper names, In Proceedings of COLING-96.

[2] Singh, A. K. and Surana, H. 2007. Can Corpus Based Measures be Used for Comparative Study of Languages, In Proceedings of Ninth Meeting of the ACL Special Interest Group in Computational Morphology and Phonology, ACL 2007.

[3] Bikel, D. M. , Schwartz, R. L. and Weischedel R. M. 1999. An Algorithm that Learns What's in a Name. Machine Learning, pp. 211-231.

[4] Borthwick, 1999. Maximum Entropy Approach to Named Entity Recognition, Ph.D. thesis, New York University.

[5] Lafferty, J. D., McCallum, A. and Perera, F. C. N. 2001. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data, pp.282-289, ICML 2001.

[6] Li W and McCallum A. 2003. Rapid development of Hindi named entity recognition using conditional random fields and feature induction, ACM Transactions on Asian Language Information Processing (TALIP), pp.290–294.

[7] Srihari, R., Niu, C. and Li, W. 2000. A Hybrid Approach for Named Entity and Sub-Type Tagging, In Proceedings of the sixth conference on Applied natural language processing.

[8] Toral, A., Noguera, E. Llopis, F. and Munoz, R. 2005. Improving question answering using named entity recognition, In Proceedings of the 10th NLDB congress, Lecture notes in Computer Science.

[9] Spain, A., Babych B. and Hartley. 2003. A. Improving machine translation quality with automatic named entity recognition, Springer-Verlag 2003.

[10] Wallach, H. M. 2004. Conditional random fields: An introduction, Technical Report MS-CIS-04-21, University of Pennsylvania, Department of Computer and Information Science, University of Pennsylvania.

[11] Taku kudo. 2005. CRF++, an open source toolkit for CRF, http://crfpp.sourceforge.net.