Volume 5, Issue 6, June 2015 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

A Hybrid Approach of Similarity Based SVM and CRF for Named Entity Recognition

Sobhana N V
Associate Professor
Rajiv Gandhi Institute of Technology
Kottayam, Kerala, India

Abstract— Named entity recognition is the most important and fundamental task of text mining. Search/Information Access, Social media monitoring, E-Discovery, Records Management, National Security or Intelligence, Enterprise Business Intelligence/Data Mining, Competitive Intelligence etc are some of the major areas of text mining. Machine learning methods like CRF, MEMM and SVM have been widely used for learning to recognize such entities from an annotated corpus. In this paper, we propose a novel kernel function for SVM and a combined approach including SVM and Conditional Random Fields (CRF) for named entity recognition (NER). The proposed kernel is calculated based on a novel distance function between the string based features and different contextual information of the words along with the variety of features that are helpful in predicting the various named entity (NE) classes. The distance function makes use of certain statistics that are derived from the training data and K-Means clustering information. 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 81.99% Precision and 78.36 Recall.

Keywords: Named Entity Recognition, Geological Corpus,Classification, Clustering,Precision, Recall

I. INTRODUCTION

Named Entity Recognition (NER) is a principle part of information extraction system. NER incorporates identification of proper names in texts and their classification into a set of predefined categories of interest. Different categories can be person names, location names, organization names, date & time expressions etc. A wide range 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 use rules which are manually written by linguists. There are several rule based NER systems, which consists of 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 requires skilled linguistics in order to use hand crafted rules. The main disadvantage of these rule based method is that they need massive 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. The major 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 labelling task. Section 5 describes the details of Geological Corpus. Section 6 explains the experiments and Results. Conclusion comes in section 7.

II. GEOLOGICAL NAMED ENTITY RECOGNITION

Geology is the study of origin, history and structure of the earth. Text mining on geological documents is an important task 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. Identification and classification of named entities (NEs) are a 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. 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.

NER is a hard problem. Words have different applications and there is infinite number of proper names. The problem of identifying NEs is solved to some extent due to capitalization feature In English language. 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.

III. 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. Coordinates (longitude, latitude) is used to represent a geographic footprint.

A. 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. 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.

F= {W-2, W-1, Wi, 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 help to recognize miscellaneous NEs such as time expressions, date expressions, percentages, numerical numbers etc.

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

IV. THE PROPOSED KERNEL FOR SVM

Support Vector Machines uses a line or surface to separate the data. Thus, it is suitable for binary classification problems but not for multiple-class problems where there are more than two candidate objective classes. Named entity recognition is a multiple-class task. As a result, the initial binary SVM is not fit for most named entity recognition tasks. Two major types of approaches are used to solve multiple-class problems. One is to update an SVM kernel function that can merge the multiple classification surface problems into an optimization so as to solve multiple class classification in one pass. The alternative is to apply multiple binary classifiers until they finish the job.

The linear SVM computes the dot product between instances

$K(X, Y) = \phi(X, Y)$

If $x_1, x_2 \ldots.. xn$ are features of X, and $y_1, y_2 \ldots.. y_n$ are features of Y

$x_i . y_i = 1$ if x and y are same

$x_i . y_i = 0$ otherwise

However if we are dealing with word features and other string based features, then such dot product based similarity computation is not able to capture the NER task specific similarity between the strings. For example, the words 'Prof.' and 'Chairman' have some similarity in the context of the NER task as both occurs frequently at the preceding position of the person names; 'small' and 'large' are related words, both being adjectives used in similar contexts; 'town' and 'district' have similarity as both of these are common location terms and occur frequently at the surrounding positions of the location names. Such task specific similarity is important in word features as well as in other string features like suffix, prefix and n-grams.

We have attempted to capture this semantic similarity with the distance between the instances. It is based on K-Means clustering information. We have used these similarity functions as kernel in SVM. These individual functions are combined with a suitable weigh. The combined function is also used as a composite kernel.

V. K-MEANS CLUSTERING BASED KERNEL FOR NER

In the clustering based kernel we use cluster information of the feature values (e.g., words) as a measure of similarity between them. Cluster information has been used in different NLP tasks in the past. Several types of clustering techniques (e.g., Brown et al., 1992; Pereira et al., 1993; Ushioda, 1996; Biemann, 2006) Miller et al. (2004) have been proposed and used in various NLP tasks.

VI. CLUSTERING OF WORDS

Here we have used the K-Means clustering algorithm The input to the algorithm is a list of words to be clustered and a large raw corpus (we have used a raw corpus containing 2 lakh words). The output from the clustering algorithm provides the average distance from cluster members to the centre of each cluster. Kernel computation is used to obtain the similarity between the words. The similarity is obtained from the distance from the centre of the cluster.

A Kernel computation

The distance between two string values (e.g., words) is also can be computed from the distance. Finally the individual distances are combined in a weighted fashion to obtain the kernel value of a feature group.

$$K(X,Y)=\text{lambda } i\ \phi(X,Y)$$

VII. CONDITIONAL RANDOM FIELDS

CRFs are often used for the labelling or parsing of sequential data, such as natural language text or biological sequences. CRFs work well in named entity recognition tasks. Many features can be used in CRFs. For example, term appearance (e.g., capitalization, affixes, etc.) and orthographic features (e.g., alphanumeric characters, dashes, Roman numeral characters, etc.) are used frequently.

Conditional Random Fields (CRFs) 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 labelling 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 can be 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

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

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

$$P(l/o) = \sum_{s:l(s)=l} P(s/o)$$

where l(s) is the sequence of labels corresponding to the labels of the states in sequences. *Zo* is a normalization factor over all state sequences. To make all conditional probabilities sum up to 1, we must calculate the normalisation factor

$$Z_o = \sum_{s}\exp\left(\sum_{t=1}^{T}\sum_{k}\lambda_k f_k(s_{l-1}, s_l, o, t)\right).$$

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.

However, CRFs have many drawbacks. First, CRFs use a limited size of context rather than the whole text because of computational limitation, thereby limiting the contextual information. Second, splitting the context of the whole text into small pieces of context will generally separate inherent relationships among them, and simply combining these pieces of context again cannot reproduce the original context due to the loss of relationships during splitting. For example, a CRF geological term identifier uses a two-word context. The whole text could be split into many pieces of two-word contexts. As a result, the same term in the different places of the text could be tagged with different results due to the variation in the context. However, SVM deals with the whole text so it does not have such restrictions. Third, CRFs are affected by the data distribution. If we want to achieve better results, the data should have an exponential distribution.

VIII. SVM-CRFS COMBINED GEOLOGICAL NAME ENTITY RECOGNITION

A new research area in machine learning is combining useful algorithms together to provide better performance or for achieving smooth and stable performance. SVM and CRFs are two conventional algorithms that can deal with named entity recognition tasks well. The feature context used by SVM is global and it does not have the same constraints as CRFs. SVM is initially the best fit for binary-class tasks and it does not perform well on multiple-class tasks. CRFs generally require more computational time and space than SVMs. Even though CRFs have many drawbacks, they are very good at sequential data tagging tasks, which is a typical problem in name entity recognition. Thus, we combined Similarity Kernel based SVM and CRFs because they can complement and facilitate each other.

In our approach, Geological named entity recognition was regarded as a two-step task. The first step was to determine whether a candidate term was a Geological one. If it was a Geological, we determine its class of entity. The first step was a binary classification task where the result was either yes or no. We then used CRFs to infer the type of Geological term. Finally, we merged the results returned by SVM and CRFs, before performing an amendment process.

IX. EXPERIMENTAL RESULTS

A NE tagged Geological corpus 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 labelled 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.

TABLE I EXPERIMENTAL RESULTS OF NER USING CRF

	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

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.

TABLE II EXPERIMENTAL RESULTS OF HYBRID APPROACH

	Class	Precision	Recall
1	Country	99	91
2	State	96.3	95
3	City	96	82
4	Region	98	93
5	Mountain	96	92.4
6	Water Bodies	86	86
7	Island	85	85
8	River	90.11	90.11
9	Village	85	85
10	Mineral	94.3	77
11	Organisation	47.5	47.5
12	Measures	93.2	93.2
13	Year	97	97
14	Person	69.11	69.11

15	Fault	34.53	20.1
16	Rock	90.2	92.1
17	Time	36.7	36.7
		81.99706	78.36588

We have got Precision of 81.99%, Recall of 78.36% for the combined approach. NE's such as Country, State and Year have high F-measure values because of their higher appearance in the corpus.

X. CONCLUSIONS

In this paper, we have developed a NER system using SVM and CRF with the help of a NE tagged Geological Corpus. We also presented a new named entity tag set 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. This proposed method has obtained better accuracy than CRF based method.

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] N. V Sobhana, S.K Ghosh, and Pabitra Mitra "Conditional random field based Named Entity Recognition in Geological text", International Journal of Computer Applications, ISSN 1076-5204, 2011

[12] Zhu, Fei, and Bairong Shen. "Combined SVM-CRFs for Biological Named Entity Recognition with Maximal Bidirectional Squeezing", PLoS ONE, 2012.

[13] Saha, S.K.. "A composite kernel for named entity recognition", Pattern Recognition Letters, 20100901

[14] Sobhana, N. V, S.K Ghosh, and Pabitra Mitra. "Entity Relation Extraction from geological text using Conditional Random Fields and subsequence kernels", 2012 Annual IEEE India Conference (INDICON), 2012.

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