Fusion of Long Distance Dependency Features for Chinese Named Entity Recognition Based on Markov Logic Networks

Zejian Wu[1], Zhengtao Yu[1,2,*], Jianyi Guo[1,2], Cunli Mao[1,2], and Youmin Zhang[1]

[1] The School of Information Engineering and Automation, Kunming University of Science and Technology, Kunming 650051, China
[2] The Institute of Intelligent Information Processing,Computer Technology Application Key Laboratory of Yunnan Province, Kunming 650051, China
{liipwujian,ztyu,gjade86}@hotmail.com,Maocunli@163.com,316935430@qq.com
http://www.liip.cn

Abstract. For the issue that existing methods for Chinese Named Entity Recognition(NER) fail to consider the long-distance dependencies, which is common in the document. This paper, Fusion of long distance dependency, proposes a method for Chinese Named Entity Recognition(NER) based on Markov Logic Networks(MLNs), which comprehensively utilizes local, short distance dependency and long distance dependency features by taking advantage of first order logic to represent knowledge, and then integrates all the features into Markov Network for Chinese named entity recognition with the help of MLNs. Validity of proposed method is verified both in open domain and restricted domain, experimental result shows that proposed method has better effect.

Keywords: Chinese Named Entity Recognition; Markov Logic Networks; Long distance dependency; Statistic Rational Learning; Natural Language Processing

1 Introduction

Named Entity Recognition (NER) is one of the key techniques in the fields of Information Extraction, Question Answering, Parsing, Metadata Tagging in Semantic Web, etc. NER task is to identify the three categories (entity, time and digital), seven subclasses (person, organization, place, time, date, currency and percentage) of named entities in the text [1-2]. Because of word flexibility of person, organization and place, their recognition is very difficult, while, composition of time, date, currency and percentage tends to have more obvious rules. Therefore, NER generally refers to the recognition of person, organization and place. As early as in 1995, MUC-6 established the special evaluation of NER, greatly promoted the development of English NER technology. MUC-6 and MUC-7 also

[*] Corresponding author.

M. Zhou et al. (Eds.): NLPCC 2012, CCIS 333, pp. 132–142, 2012.

established a multi-language entity recognition evaluation task MET (Multilingual Entity Task), including Japanese, Spanish, Chinese and other languages. BAKEOFF-3 and BAKEOFF-4, Held in 2006 and 2008, established a special evaluation task for Chinese NER. In 2003 and 2004, "Chinese information processing and intelligent human-machine interface technology evaluation" task, held by "863 Program", established Chinese Named Entity Recognition evaluation task. These evaluation tasks had played a very important role in promoting the development of Chinese Named Entity Recognition. Compared to English NER, Chinese NER is more difficult [2]. The main differences between Chinese NER and English NER lie in: (1) Unlike English, Chinese lacks the capitalization information which can play very important roles in identifying named entities. (2) There is no space between words in Chinese, so we have to segment the text before NER. Consequently, the errors in word segmentation will affect the result of NER. In this paper, we propose a Chinese NER model based on Markov Logic Networks with emphasizes on (1) Combining local feature, short distance dependency feature and long distance dependency feature into a unified statistical relational learning model; (2) Integrating probabilistic learning and relational learning for Chinese NER by MLNs. In order to deduce the complexity of the model and the searching space, we divide the recognition process into two steps: (1) word segmentation and POS tagging; (2) named entity recognition based on the first step. Proposed method is tested on the open domain and restricted domain corpus. The Precision, Recall and F1 in open domain and restricted domain are respectively (78.39%, 85.89%, 81.97%) and (85.39%, 88.89%, 87.10%). Experimental result shows that proposed model is better than Condition Random Fields when only use local and short distance dependency features, and long distance dependency features can effectively improve the recognition effect.

2 Related Work

Early approaches to Named Entity Recognition involve a lot of human effort, require researchers to write a series of complex regular expressions to match candidate entity, and need to develop a large dictionary of common entities. Moreover, these approaches could only be engineered to suit a specific domain [1]. These limitations motivate the development of machine learning systems in natural language processing. Bikel, et al first proposed a named entity recognition method based on Hidden Markov Models (HMM) [3]. While, because HMM is a generative model, theres only one feature variable with each state. However, we may want to incorporate more output features in our model to improve the accuracy of the tagger. For example, for a given token, we do not only want to consider the identity of the word. We would also want to take more output features into account such as the previous token, the next token, whether the token includes any digits or symbols etc. To maintain tractability of computation, HMM have to assume that observation features are independent of each other. In real life, most observations do have complex dependencies, and assuming independence between the features can severely impair the performance of the model. To handle the limitation of HMM, Liao, et al proposed many methods for NER based on linear-chain Conditional Random Fields (CRF) [4]. Even though linear chain CRF has many advantages over some of the more traditional models, it also has weaknesses. In a linear chain CRF, we assume that the only dependencies are between the labels of adjacent words. Thus, linear chain CRF is not able to use information from longer distance dependencies to assist label. While in real life, there are many long distance dependencies in a document. For example, if a word is tagged as a category, when the subsequent sections in the document again or repeatedly involved this word, it always appears in the same or similar form and its labeling is often the same. J Liu, et al, integrated long distance dependencies, proposed a method for biomedical named entities recognition based on skip-chain CRF [5]. Skip-chain CRF can handle relatively simple long distance dependencies, while, in real life, there are many complex long distance dependencies, skip-chain CRF is also unable to handle them.

As a result, researchers have turned to using a variety of statistical relational learning methods to increase the accuracy of English NER, while, related researches for Chinese NER is still rare. Statistical Relational Learning (SR-L) is a combination of probabilistic learning and relational learning [6]. The strength of probabilistic models is that they can handle uncertainty in learning and reasoning. Meanwhile, first order logic or relational databases can effectively represent a wide range of knowledge. SRL techniques attempt to combine the strength of the two approaches. This combined strength of probabilistic learning and relational learning gives SRLs more power in learning and inferences. Recently, there have been some studies in the application of SRL techniques to information extraction. Bunescu and Mooney have used Relational Markov Networks to identify protein names in biomedical text [7]. Domingos and Poon have applied Markov Logic Networks for the segmentation and entity resolution of bibliographic citations [8].

Here, we propose to use the power of Markov Logic Networks to model long distance dependencies for Chinese NER. To the best of our knowledge, this is the first research work that, integrated long distance dependencies, applies Markov logic Networks to Chinese NER. This paper is organized as follows. In section 3 we will review the related definitions of MLNs. In section 4 we will introduce our method for Chinese NER, followed by the experiment in section 5. The conclusions are given in section 6.

3 Markov Logic Review

Among many statistical relational learning methods, Markov Logic Networks (MLNs) is a powerful, direct approach. It is a first-order knowledge base with a weight attached to each formula which can be viewed as templates for features of Markov networks. It is defined as follows [9-10]:

A Markov Logic Network L is a set of pairs$\{(F_i, w_i)\}_{i=1}^{m}$, where F_i is a formula in first-order logic and w_i is a real number. Together with a finite set of constants

$C = \{C_1, C_2, C_3, ..., C_{|c|}\}$, it defines a Markov network $M_{L,C} =< X, E, \{\varphi_k\} >$ as follows:

1.$M_{L,C}$contains one binary node for each possible grounding of each predicate appearing in L. The value of the node is 1 if the ground atom is true and 0 otherwise.

2.$M_{L,C}$contains one feature for each possible grounding of each formulaF_i in L. The value of this feature is 1 if the ground formula is true and 0 otherwise. The weight of the feature is w_i associated with F_i in L.

For simplicity, a Markov logic network $M_{L,C}$ is a set of weighted first-order clauses. Together with a set of constants, it defines a Markov network with one node per ground atom and one feature per ground clause. The weight of a feature is the weight of the first-order clause that originated it. The probability of a state x in such a network is given by the log-linear model:

$$P(X = x) = \frac{1}{Z} \prod_{k=1}^{nc} \phi_k(x_k) = \frac{1}{Z} \prod_{i=1}^{m} (e^{w_i})^{n_i(x)} = \frac{1}{Z} \exp\{\sum_{i=1}^{m} [w_i \cdot n_i(x)]\} \quad (1)$$

Where Z is normalization constant, w_i is the weight of the i-th formula, and $n_i(x)$ is the number of satisfied groundings.

4 Chinese Name Entity Recognition Based on MLNs

4.1 Feature Selection and Their First Order Logic Representation

Local Features. Word itself, part of speech, word context and some specific dictionaries can be taken into consideration when select local features. However, for the reason that dictionary and some others external resources are the common entity resources by manual sorting, they have very small contribution to verity the effectiveness of proposed approach. Therefore, in order to better verify the validity of the method, we only select the inherent features of the document. Basic features we selected as follows:

1) Independent Feature: Represent information in the words of candidate entity. It includes word itself and its POS tag and it aims to inspect the internal information in candidate entity. For example, if a candidate entity includes word " 公司(Corporation)", the probability that it is labeled as organization will be increased. First-order logic formula that represents this feature as follows:

$$Word(+x) \wedge Tag(x, t) \Rightarrow Label(x, +l)$$

Denote that the labels of words in the text rely on the word itself and its POS tag information. "+" means separate weight is learned for different grounding formulas.

2) Local Context Feature: Represent information between current word and its adjacent words. For example, "机器学习领域著名学者周志华教授(Well-known scholar in the field of machine learning, "Zhou Zhihua" Professor)", suppose the word we want to tag is "周志华(Zhou Zhihua)". It is very difficult to directly label

the word, however, if we consider its previous adjacent word ”学者(scholar)” and its next adjacent word ”教授(professor)”, the probability that the word ” 周志华(Zhou Zhihua)” tagged as person will be increased. First order logic formula that represents this feature as follows:

$$Neighbour(x,y) \wedge Word(+x) \Rightarrow Label(y,+l)$$

$$Neighbour(x,y) \wedge Word(+y) \Rightarrow Label(x,+l)$$

$Neighbour(x,y)$ denotes that x is the previous adjacent word of y. The formulas denote that candidate entities'adjacent words have impact on its label.

Short Distance Dependency Features. Local features are the information that can be extracted directly from the corpus, while it does not take into account the dependencies between the labels of candidate entities. However, there is a wealth of relationship between the labels in a document. Such as, the label of an entity's previous word always is non-entity label, etc. For example, In the sentence ”来自北京大学、清华大学等200所院校现场接受考生及家长的咨询．(200 person in charge from Peking University, Tsinghua University and other institutions to the scene to accept the students and parents consulting.)”, ”北京大学(Peking University)” and ”清华大学(Tsinghua University)” are organizations, labels of their previous words (” 来自”(”from”) and ”，”) are all non-entity label. Therefore, dependencies of the labels of candidate entities should be used effectively. CRF is able to handle these short distance dependent features, and achieves fairly good result in the entity recognition task [12]. So we follow the idea of CRF, assume that the label of current word only relies on the label of its previous adjacent word and its next adjacent word. First order logic formula that represents this feature as follows:

$$Neighbour(x,y) \wedge label(x,+l) \Rightarrow Label(y,+label)$$

$$Neighbour(x,y) \wedge label(y,+l) \Rightarrow Label(x,+label)$$

The two formulas denote that the label of current word depend on the label of its previous adjacent word and its next adjacent word.

Long Distance Dependency Features. If a candidate entity is tagged as a category in a document, when the entity again or repeatedly involved in the subsequent sections of the document, it usually appears in the same or similar form. And some of the candidate words themselves has ambiguity, if there is no prior knowledge, it is very difficult to recognize. For instance:

1). ” 云南大学、昆明理工大学，云南师范大学等省内重点高校研招计划于昨日公布(Yunnan University, Kunming University of Science and Technology, Yunnan Normal University and other provincial key universities graduate enrollment plan announced yesterday)”

2). "云南大学、昆明理工大学仍是我省招生规模最大的两所大学(Yunnan University, Kunming University of Science and Technology are still the two universities whose enrollment is the largest in our province)"

3). "我省有包括云大、昆工、天文台等共17所招收研究生的高校及科研机构(In our province, there are 17 universities and research institutions, including , that including Yunnan University, Kunming University of Science and Technology and the Observatory, that recruit graduate students)"

The three sentences above are extracted from one news report in accordance with the original order. In sentence 1), " 云南大学(Yunnan University)" and "昆明理工大学(Kunming University of Science and Technology)" can be judged as organization. The two entities are repeated in 2), we can obtain the correct category by the result of 1). "云大(YunDa)" and "昆工(KunGong)" are the abbreviation of "云南大学(Yunnan University)" and "昆明理工大学(Kunming University of Science and Technology)", and it is very difficult to label them only based on the local features as well as short distance features. While, once taking advantage of long distance features, they can be labeled straightforward by the correct category by the result of 1) and 2). Therefore, long distance features are very useful in the identification of candidate entities. This paper takes two types of long distance dependencies into consideration: 1) Homomorphism Repetition. Donate that if a candidate entity appears in different locations of the same document, the label of these entities should be labeled as the same. Regular expression is used to match homomorphism repetition to get the entity repetition information. 2) Abbreviation Repetition. Donate that if an entity appears in a document, and in the follow-up portion of the document, the abbreviation of the entity appears, the two candidate entities should be labeled as the same category. However, identification of the abbreviation of a name entity also is a difficult problem in NER task. In order to ensure the system's recognition accuracy, accuracy of extracting abbreviation repetition should be guaranteed, while recall rate of abbreviation repetition should be relatively relaxed. Abbreviation repetition can be identified by matching candidate word with the key words of entity's full name. In both cases, first order logic formulas can be uniformly represented as follows:

$$SameToken(x, y) \wedge label(x, +l) \Rightarrow Label(y, +l)$$

4.2 Weight Learning

Weight Learning in MLNs is to estimate the weights of formulas utilizing the training data [8-10]. We adopt Discriminative Weight Learning (DWL) to learn formulas'weights. The prerequisite of DWL is that it must be known priori that which predicates will be evidence and which ones will be queried. For the problem of Chinese NER, this is known. DWL divides grounding atoms in the domain into two sets: a set of evidence atoms X and a set of query atoms$Y$. In our approach,Y is all the grounding atoms of $Label(y, +l)$; others all belong to$X$.

The conditional likelihood (CLL) of Y given X is:

$$P(y|x) = \frac{1}{Z_x} \exp(\sum_{i \in F_y} w_i n_i(x,y)) = \frac{1}{Z_x} \exp(\sum_{j \in G_y} w_j g_j(x,y)) \tag{2}$$

Where Z_x is the partition function given X, $F_y$is the set of all MLNs clauses with at least one grounding involving a query atom, $n_i(x,y)$ is the number of true groundings of the $i-th$ clause involving query atoms,G_y is the set of ground clauses in $M_{L,C}$ involving query atoms, and $g_j(x,y) = 1$ if the $j-th$ ground clause is true in the data and 0 otherwise. By taking partial derivation of log-likelihood function of the formula above, we can obtain:

$$\frac{\partial}{\partial w_i} \log p_w(y|x) = n_i(x,y) - E_w[n_i(x,y)] \tag{3}$$

The time complexity of calculating $E_w[n_i(x,y)]$ accurately is enormous. Its approximate value can be calculated by calculating $n_i(x, y_w^*)$,y_w^* represents predicates in its Markov Blanket. Therefore, it translates into counting $n_i(x, y_w^*)$ in the maximum posteriori hypothesis state $y_w^*(x)$. Then, we can obtain the weight of the formulas.

4.3 Inference

Inference in Markov Logic Networks includes maximum likelihood inference, calculating marginal probabilities and calculating conditional probability [8-11]. This paper only needs the maximum possible explanation (MPE) which involves only the maximum likelihood inference. The following is a brief introduction to the method of maximum likelihood inference we used in Markov Logic Network. The maximum likelihood inference process can be stated as: given evidence set of X, seek the most probable state of the world Y. That is:

$$\max_y p(y|x) \tag{4}$$

According to Markov logic network's joint probability distribution, the equation above can be transformed into:

$$\max_y \sum_i w_i n_i(x,y) \tag{5}$$

Therefore, the MPE problem in Markov logic reduces to finding the truth assignment that maximizes the sum of weights of satisfied clauses. The most commonly used approximate solver is MaxWalkSAT, a weighted variant of the WalkSAT local-search satisfiability solver, which can solve hard problems with hundreds of thousands of variables in minutes. While, One problem with MaxWalkSAT is that they require propositionalizing the domain (i.e., grounding all atoms and clauses in all possible ways), which consumes memory exponential in the arity of the clauses. By taking advantage of the sparseness of relational domains, where most atoms are false and most clauses are trivially satisfied, MPE inference can be conducted by LazySAT algorithm which only ground atoms and clauses that is needed and can save memory exponentially. Therefore, we adopt LazeSAT for inference in proposed method.

5 Experiment

To objectively evaluate the effect of proposed method, we organize two set of experiments in open and restricted domain. Open domain experiment is based on People's Daily's open corpus in January 1998. For the reason that there are more repetitions of attractions, places and other entities in the fields of tourism which could reflect the effect of proposed method better, restricted domain experiments are based on the corpus in tourism field of Yunnan by manual collection.

5.1 Data

In open domain experiment, effectiveness of proposed method is tested on People's Daily's open corpus in January 1998, in which the average repetition of entities in each document is about three times. we select three types of entities (person, place and organization), and then process the corpus in specific way: First, each word in the corpus is divided into a separate row, then, tag the label behind the corresponding word and its POS tag, which each entity tag is labeled in the form of beginning, intermediate and end label, each non-entity is labeled by non-entity label. Example of corpus after pre-processing is as follows:
Original corpus:

[黔南州(Qian Nan Zhou) /ns民族(nationality) /n 干部(cadre) /n 学校(school) /n]nt

Processed corpus：

黔南州(Qian Nan Zhou) /ns ntb
民族(nationality) /n ntm
干部(cadre) /n ntm
学校(school) /n nte

Experiment in restricted domain is based on artificially collected 2000 documents in the field of Yunnan tourism. Firstly, pre-process the corpus utilizing word segmentation and POS tagging tools. Then, manually tag the corpus into eight categories: Attraction (jd), Number (Numbers in Chinese, e.g. "五十三(Fifty-three)")(m), Person (pn), Snack(xc), Place(dd), Specialty(tc), Festival(jr),Time(Time in Chinese, e.g. " 二十一世纪(Twenty-first century)")(t). Example of the corpus after pre-processing is as follows:

也许(Maybe) /d o
你(You) /r o
逛(visit) /v o

了/u o
束河白(Shu He Bai) /nr jdB
沙(Sha)/nr jdE
雪嵩(Xue Song) /nr jdB
石鼓(Shi Gu) /n jdE

The first column is the segmentation result of original text, the second column is corresponding POS tag, and the third column is corresponding entity label tagged by manual where "o" indicates a non-entity label. The average repetition of entities in each document is about fifteen times. Detailed statistics of data collections are shown in Table 1.

Table 1. Table1 statistic of corpus

Number of Documents	Train Corpus	Open Test Corpus	Closed Test Corpus
2000	800	1200	400
Attractions(jd)	Number(m)	Person name(pn)	Snack(xc)
76	130	306	51
Place name(dd)	Specialty(tc)	Festival(jr)	Time(t)
128	31	79	61

5.2 Experimental Analysis

Three comparative experiments are organized for each of the two experiments: the first experiment is based on Conditional Random Fields; the second experiment is based on Markov Logic Networks which only use local features and short distance features; the third experiment is also based on Markov Logic Networks with comprehensive utilizing local, short distance dependency and long distance dependency features. Closed and open comparative experiments both are organized for each of the three types of experiments. Evaluation of the two sets of experiments is based on the following three indicators:

$$Precision = \frac{NumberCorrect}{TotalTagged}$$

$$Recall = \frac{NumberCorrect}{ExpectedLabels}$$

$$F1 = \frac{2 * (Precision * Recall)}{Precision + Recall}$$

Table 2 gives detailed statistics of the comparative experimental result.

The experimental result shows that the Precision, Recall and F1 of proposed method in open domain and restricted domain are respectively (78.39%, 85.89%, 81.97%) and (85.39%, 88.89%, 87.10%). The reason why the accuracy of the experimental result is not very prominent is that, in order to verify the

Table 2. Table 2: Detailed comparative experimental result

	Open /Closed	Open Domain			Restricted Domain		
		Precision/%	Recall/%	F1/%	Precision/%	Recall/%	F1/%
CRF	Closed	70.30%	82.22%	75.79%	84.30%	87.22%	85.74%
	Open	66.33%	79.75%	72.42%	81.33%	78.75%	80.02%
MLNs(Local + Short)	Closed	85.19%	87.47%	86.31%	86.19%	87.47%	86.83%
	Open	73.28%	81.62%	77.23%	81.28%	82.62%	81.94%
MLNs(Local+ Short+ Long)	Closed	88.27%	93.46%	90.79%	90.27%	92.46%	91.35%
	Open	78.39%	85.89%	81.97%	85.39%	88.89%	87.10%

validity of the proposed method, we only select some inherent basic features in the document regardless of any excessive rules and more contexts (e.g. 2 or more gram model). Experimental result shows that using only local and short distance dependences features, experiment result of MLNs is better than CRF. When long distances dependency features are integrated into MLNs, experiments result both in open domain and restricted domain are all improved and more obvious in restricted domain. Increase of precision, Recall and F1 in open domain and restricted domain are respectively (5.11%, 4.27%, 4.66%) and (4.11%, 6.27%, 5.16%). This is because entities have more repetitions in restricted domain and then, long distances dependence features contribute more to improve the experimental result in restricted domain.

Acknowledgement. This paper is supported by National Nature Science Foundation (No.60863011, 61175068), and the Key Project of Yunnan Nature Science Foundation (No. 2008CC023), and the National Innovation Fund for Technology based Firms (No.11C26215305905), and the Open Fund of Software Engineering Key Laboratory of Yunnan Province (No.2011SE14), and the Ministry of Education of Returned Overseas Students to Start Research and Fund Projects.

References

1. Grishman, R.: Information Extraction. The Oxford Handbook of Computational Linguistics (2003)
2. Zhao, J.: A Survey on Named Entity Recognition, Disambiguation and Cross-Lingual Co-reference Resolution. Journal of Chinese Information Processing 23(2) (2009)
3. Bidel, D.M., Schwarta, R., Weischedel, R.M.: An Algorithm that learns what's in a Name. Machine Learning Journal Special Issue on Natural Language Learning 34(1-3), 211–231 (1999)

4. Liao, W., Veeramachanei, S.: A Simple Semi-supervised Algorithm for Name Entity Recognition. In: Proceedings of the NAACL HLT 2009 Workshop on Semi-Supervised Learning for Natural Language Processing, pp. 58–65 (2009)
5. Liu, J., Huang, M., Zhu, X.: Recognizing biomedical named entities using skip-chain conditional random fields. In: Proceedings of the 2010 Workshop on Biomedical Natural Language Processing, BioNLP 2010, pp. 10–18 (2010)
6. Xu, C.-F., Hao, C.-L., Su, B.-J., Lou, J.-J.: Research on Markov Logic Networks. Journal of Software 22(8), 1699–1713 (2011)
7. Bunescu, Mooney: Statistical Relational Learning for Natural Language Information Extraction. Introduction to Statistical Relational Learning, pp. 535–552. MIT Press, Cambridge (2007)
8. Poon, Domingos: Joint Inference in Information Extraction. In: Proceedings of the Twenty-Second National Conference on Artificial Intelligence, pp. 913–918. AAAI Press, Vancouver (2007)
9. Domingos, Richardson: Markov Logic: A Unifying Framework for Statistical Relational Learning. Introduction to Statistical Relational Learning, pp. 339–371. MIT Press, Cambridge (2007)
10. Domingos, P., Lowd, D.: Markov Logic: An Interface Layer for Artificial Intelligence. Morgan and Claypool, San Rafael (2009)
11. Poon, Domingos: Sound and Efficient Inference with Probabilistic and Deterministic Dependencies. In: Proceedings of the Twenty-First National Conference on Artificial Intelligence, pp. 458–463. AAAI Press, Boston (2006)
12. Sutton, McCallum: An Introduction to Conditional Random Fields for Relational Learning. Introduction to Statistical Relational Learning. MIT Press, Cambridge (2007)