Identification of Chinese Event Types Based on Local Feature Selection and Explicit Positive & Negative Feature Combination

Hongye Tan, Tiejun Zhao, Haochang Wang and Wanpyo Hong, *Member, KIMICS*

Abstract—An approach to identify Chinese event types is proposed in this paper which combines a good feature selection policy and a Maximum Entropy (ME) model. The approach not only effectively alleviates the problem that classifier performs poorly on the small and difficult types, but improve overall performance. Experiments on the ACE2005 corpus show that performance is satisfying with the 83.5% macro - average F measure. The main characters and ideas of the approach are: (1) Optimal feature set is built for each type according to local feature selection, which fully ensures the performance of each type. (2) Positive and negative features are explicitly discriminated and combined by using one - sided metrics, which makes use of both features' advantages. (3) Wrapper methods are used to search new features and evaluate the various feature subsets to obtain the optimal feature subset.

Index Terms—event types, local feature selection, positive feature, negative feature, one-sided metric.

I. INTRODUCTION

Now Event Detection and Recognition (VDR) has been defined as a fundamental task in Automatic Content Extraction (ACE) evaluation plan [1]. For example, the ACE2005 VDR task mainly requires identifying the events of some specified types, and extracting the selected information about these events including some attributes such as type, subtype etc., the event argument and the event mentions. In this paper, we focus on the identification of Chinese event and its types.

The closely related efforts on identifying event types were reported by Y.Y. Zhao etc. [2] and S. Bethard etc. [3]. In [2], the authors showed that an event trigger word motivated and machine learning algorithm could get the performance of 69.9% F-measure on the Chinese ACE2004 corpus. In [3], the authors viewed event identification as a classification task similar to the word-chunking task with the standard B-I-O formulation, and introduced a variety of linguistic features and trained a system, which can identify event types of the English TimeBank corpus with a precision of 67% and a recall of 71%. The above researches show that the identification of event types in both Chinese and English is not satisfying.

We have investigated the task of identifying Chinese event types and found that the difficulty level of each type varies. For example, the events of the Business and Justice types are easier to identify, while the Transaction type is more difficult to identify. Inevitably, the difficult types will hurt the final performance of the system.

This paper proposes an approach which combines the Maximum Entropy (ME) model with the local feature selection strategy and positive & negative feature combination. This approach alleviates the limitation that classifier performs poorly on the difficult categories to some degree. Experiments on the ACE2005 corpus show that the 83.5 macro-averaging F1-measure can be achieved. The main idea of the approach are: (1) according to the idea of local feature selection, features are chosen for each type, not for all types. This strategy pays adequate attention to small and difficult types and ensures their performances. (2) using one-sided metric, positive features and negative features are discriminated explicitly and the two kinds of features are combined. The strategy utilizes the various advantages of them and improves the system performance. (3) wrapper methods are used to evaluate various feature subsets and the optimal feature subset is obtained.

The remainder of the paper is organized as follows. Section 2 introduces some concepts of feature selection and gives the reasons of our approach. Section 3 describes the algorithms in detail. In Section 4, the related experimental results and their analysis are presented. The last section concludes with ideas for future work.

II. FEATURE SELECTION

So far, many feature selection methods have been

Manuscript received May 25, 2007.

Hongye Tan is with department of Computer Science and Technology, Harbin Institute of Technology, Harbin City , China, 150001), School of Computer and Information Technology, Shanxi University, Taiyuan City, China, 030006) Email: hytan@mtlab.hit.edu.cn)

Tiejun Zhao is with department of Computer Science and Technology, Harbin Institute of Technology, Harbin City , China, 150001) Email: tjzhao@mtlab.hit.edu.cn)

Haochang Wang is with department of Computer Science and Technology, Harbin Institute of Technology, Harbin City , China, 150001) Email: hcwang@mtlab.hit.edu.cn)

Wanpyo Hong is with IT Division of Hansei University, Gunpo City, Korea, 435-742 Email : wphong@hansei.ac.kr)

explored. Among them, wrapper methods are effective by applying general search mechanisms, such as sequential forward selection, to generate various feature subsets and evaluate the subsets with repeated calls to the induction algorithm. However, they usually involve great computational cost. A number of other feature scoring methods have also been proposed, in which each feature is scored according to the training data. In contrast to wrapper methods, these metrics are independent of the induction algorithm. Table2-1 shows the frequently used feature scoring metrics [4] [5] [7].

small and difficult classes are omitted by the global feature selection, which does not pay adequate attention to difficult and small classes. For the second issue, we believe that negative features are valuable for classification. Now let TP, FP, FN and TN denotes the number of true positive, false positive, false negative and true negative examples respectively. And the precision P and the recall R can be computed by the formulas $P = \frac{TP}{TP+FP}$ and $R = \frac{TP}{TP+FN}$ respectively. We can see that positive features are useful for increasing TP and

Table 2-1 Feature scoring metrics being requently used

Information Gain (IG)	$IG(t_k, c_i) = \sum_{c \in \{c_i, \bar{c}_i\}} \sum_{t \in \{t_k, \bar{t}_k\}} p(t,c) \cdot \log \frac{p(t,c)}{p(t) \cdot p(c)}$
Chi-Square (x^2)	$x^2(t_k, c_i) = \frac{N[p(t_k, c_i)p(\bar{t}_k, \bar{c}_i) - p(t_k, \bar{c}_i)p(\bar{t}_k, c_i)]^2}{p(t_k)p(\bar{t}_k)p(c_i)p(\bar{c}_i)}$
Correlation Coefficcient (CC)	$CC(t_k, c_i) = \frac{\sqrt{N}[p(t_k, c_i)p(\bar{t}_k, \bar{c}_i) - p(t_k, \bar{c}_i)p(\bar{t}_k, c_i)]}{\sqrt{p(t_k)p(\bar{t}_k)p(c_i)p(\bar{c}_i)}}$
Notes : $p(t_k, c_i)$ denotes the probability of an instance including the feature t_k belongs to the category c_i. $p(\bar{t}_k, c_i)$ denotes the probability of an instance not including the feature t_k belongs to the category c_i N denotes the total number of instances.	

Since there exists multi-class classification task, features can be chosen globally under all categories or locally per category, i.e. global feature selection (GFS) or local feature selection (LFS) [4] [5].

Features can be categorized into positive features (PF) and negative features (NF). For positive features, their presence in an instance highly indicates its relevance to a certain category. And for negative features, their appearance in an instance highly indicates its non-relevance to a certain category [5].

A feature selection metric is regarded as one-sided if its positive and negative values correspond to positive and negative features respectively [5]. In contrast, it is considered as a two-sided feature selection metric if its values are non-negative. A one sided metric can be changed into its corresponding two-sided format, if the signs of feature score are omitted. And a two-sided metric can be changed into its one-sided counterpart by using a certain strategy to recover the signs of feature scores. For example, in Table2-1, feature selection metrics like CC and OR are one-sided metrics, while IG and Chi-square metrics are two-sided metrics. And CC metric is also the corresponding two-sided metric of Chi-square.

There are two main issues in feature selection. One is that features are chosen locally or globally? The other is whether to explicitly utilize negative features? For the first issue, Forman [3] argues that features available in global selection distribute uneven for all categories. Most of the "best" global features are for the easy and large classes, while the strong predictive local features for the

decreasing FN, while negative features are helpful for increasing TN and decreasing FP. The two kinds of features have different influences on the performance and are beneficial to the classifier. Our hypothesis is similar to that in [5].

III. ALGORITHM

A. Maximum entropy model

The Maximum Entropy (ME) method is selecting the model which has maximum entropy, satisfies the known constraints and assumes nothing about what is unknown. The ME model enjoys the advantages that it belongs to discriminative learning models avoiding the difficulty to model the generative component, can combine all kinks of features together conveniently, and need not any parameter smoothing and any independent hypothesis. So far, the ME model has been successfully used in a lot of tasks in NLP, such as word segmentation, POS tagging, parsing, and text categorization etc. and all achieved nearly the best performance[8].

The implementation we used in this paper is ME toolkit of ZhangLe[1], and we train the model with the default training parameters, that is, with 30 iterations of L-BFGS parameter estimation method.

B. Feature selection method

The main idea of our feature selection method is: (1)

[1]The toolkit can be downloaded at http://homepages.inf.ed.ac.uk/s0450736/maxent_toolkit.html.

Features are locally chosen for each type, that is, each type has its own optimal feature subset. (2) Positive and negative features are discriminated by using one-sided metrics. (3) Wrapper methods are used to get and evaluate the various feature subsets. In order to alleviate the great computational cost caused by general search mechanism, we take the feature score as heuristics to guide the search. The higher scores correspond to the better positive features, while the smaller scores correspond to the stronger negative features. In order to determine the optimal feature subset, the induction algorithm is called repeatedly to evaluate the various subsets. (4) Feature selection is implemented in two stages. The first stage focuses on positive feature subset generation. After the optimal positive subset is determined, the second stage, aiming at negative feature subset generation, will be started. The detailed algorithm is shown in Figure3-1.

```
input :  F_full (f1,f2, . . . . ,fn)      // the full feature set including n features
         F0+                              // the initial positive feature subset
         γ+                               // the stopping criteria for positive feature
         γ−                               // the stopping criteria for negative feature
output :  F_best                          // the best feature subset
   begin
        initialize: F_best = F0+ ;
        P_best = eval(F0+)                // evaluate F0+
        do begin                          // start positive feature subset generation
              F+=Choos(F_full)            // choose positive features
        F_full = F_full - F+
              F = F_best ∪ F+             // generate a new feature subset
              P = eval(F)                 // evaluate the current subset
              if ( P is better than P_best )
                    P_best = P
                    F_best = F
        end until (γ+ is reached )
        do begin                          // start negative feature subset generation
              F−=choose(F_full)           // choose negative features
        F_full = F_full − F−
              F = F_best ∪ F−             // generate a new feature
              P = eval(F)                 // evaluate the current subset
              if ( P is better than P_best )
                    P_best = P
                    F_best = F
        end until (γ− is reached )
     return F_best
  end
```

Fig. 3-1 Detailed feature selection procedure

The following key elements are involved in Figure3-1: (1) **Full feature set** we extract all the words appearing in the sentences of each event type and take them as the full feature set of each type[2]. (2) **Initial positive feature set** the initial positive feature set contains only one feature, which is with the highest score. (3) **Feature subset generation** in the first stages, one positive feature with the highest value will be added to the subset in each iteration, and the chosen feature will be deleted from the full feature set. Likewise, in the second stage, one negative feature with the smallest value will be added to the subset in each iteration, and then will be deleted from the full feature set. (4) **Evaluation criteria** F1-measure is used as the evaluation criteria, which can be obtained by evaluating the classifier based on ME model on the current feature subset (5) **Stopping criteria** in both stages, the circles will be stopped when all the positive or negative features are added to the feature subsets.

IV. EXPRIMENTS

The sources of ACE2005 corpus are Broadcast News、Newswire and Weblog. The VDR task involves 8 types and 33 subtypes of events. The 8 types are Justice, Conflict, Contact, Life, Movement, Business, Personnel and Transaction [1], and this paper aims at the identification of the 8 types of events. We use 3/5, 1/5, 1/5 of the ACE2005 corpus as training set, testing set and validation set respectively[3]. And the validation set is used to determine the optimal feature subset. The task is a multi-label classification problem, and we transform it into 8 independent binary

[2] The full feature set of the Movement type has 3581 features, with the most size, while that of the Business type has 1127 features, with the fewest size. The total size of the total features of all the event types is 6674.

[3] We count the sentence number of every type and every subtype. There are 6325 sentences totally, of which, 3084 sentences do not belong to any one of the above 8 types. Example number of each type in ACE2005 corpus are: Contact(291), Personnel(242), Life(488), Conflict(593), Justice(545), Business(168), Transaction(173), Movement(741).

classification problem.

For evaluating the performance, the F1- measure is used. There are two ways to get the overall performance over all types. One is micro-averaging, the other is macro-averaging. We adopt the latter to evaluate the overall performance since macro-averaging is more affected by the ability of a classifier on small or difficult classes while micro-averaging is more affected by the classifier's performance on common classes.

In order to explore the effect of our approach, several feature selection metrics are respectively used, including the CC, Chi-Square, signed-IG and IG metrics. Note that CC is the one-sided metric counterpart of Chi-square, while signed-IG[4] is IG's one-sided counterpart.

Table4-1 and Table4-2 show the results of the system using various feature selection mechanism where the macro-averaging is boldfaced. And the best macro-averaging 83.5% is achieved in the case of using LFS,explicitly discriminating and combining positive and negative features by utilizing the one-sided CC metric.

And the performance per type is quiet satisfying, since even for the most difficult Transaction type, its F1-measure can achieve 73.7%.

Table4-1 shows the results of LFS and GLS[5] based on the Chi-square and IG metrics. It is obvious that both metrics of LFS outperforms those of GFS. For example, the F1-measure of the Business type, with the fewest training examples, has been improved 4.7% and 6.6% respectively, and the F1-measure of the difficult Transaction type has been improved 18% and 2.9%

Table 4-1 Results of GFS and LFS based on Chi-Square & CC metric

Types	GFS of Chi-Square	LFS of Chi-Square		GFS of IG	LFS of IG	
	F1(%) (size of optimal feature subset is 2800)	F1(%)	size of optimal feature subset	F1(%) (size of optimal feature subset is 5650)	F1(%)	size of optimal feature subset
Contact	71.3	74.5	1300	70.4	73.5	200
Personnel	73.6	78.1	300	69.2	77.7	300
Life	81.5	86.6	950	83.1	87.4	1200
Conflict	76	83.8	700	74	81.6	200
Justice	80.2	83.8	600	80.5	83.3	700
Business	83.7	88.4	300	81.8	88.4	250
Transaction	41	59.3	900	47.1	50	1000
Movement	69.9	75.3	1750	71.5	74.2	1750
MacroAverage	72.2	78.7		72.2	77	

Table 4-2 Results without and with NF based on CC & signedIG metrics

Types	CC metric				signedIG metric			
	without NF		with NF		without NF		with NF	
	F1 (%)	Size of optimal positive feature subset	F1 (%)	Size of optimal positive feature subset	F1 (%)	Size of optimal positive feature subset	F1 (%)	Size of optimal positive feature subset
Contact	75	600	-	-	78.4	200	-	-
Personnel	84.1	250	-	-	80.8	200	-	-
Life	90.9	750	91.6	5	77.9	850	87.4	5
Conflict	83.7	850	-	-	81.8	950	-	-
Justice	83.9	1250	84.7	5	84.9	900	85.1	10
Business	84	450	88.4	5	95.4	350	-	-
Transaction	60	400	73.7	10	34	700	55.2	10
Movement	86.9	2700	-	-	86.9	2100	-	-
MacroAverage	81.1		83.5		77.5		81.4	

[4] Here, we used the signed-IG metric proposed in [5] to recover the sign of the feature score given by IG. The formula is $SIG(t_k,c_i)=sign(AD-BC)\cdot IG(t_k,c_i)$, where A、B、C、D denotes the number of four tuples (t_k,c_i)、$(t_k,\bar{c}_i)$、$(\bar{t}_k,c_i)$、$(\bar{t}_k,\bar{c}_i)$ respectively; the tuple (t_k,c_i) denotes the co-occurrence number of the feature t_k and the category c_i

[5] Features are globally chosen according to the global score available in the formula: $f_{max}(t_k)=\max_{i=1}^{m} f(t_k,c_i)$, where $f(t_k,c_i)$ is the score of feature t_k on the category c_i, and is obtained by the metrics shown in Table2-1.

respectively. And the overall macro-averaging has also been improved 6.5% and 4.8% respectively. The reason for the improvement is that the local mechanism let each type choose its own optimal feature subset and ensure the performance pet type (especially the small and difficult classes), while in global selection case only one feature subset is used under different types, which can not ensure its optimization for each type.

Table4-2 gives the results of the system without and with negative features based on the CC and signedIG metrics. The performances are further improved after positive and negative features are explicitly discriminated and combined by utilizing one-sided metrics. The overall macro-averaging has been improved 2.4% and 3.9%, and the performances of some types are also improved. For example, the F1-measure of Transaction type has been improved 13.7 and 21.2%. This shows that for those types sensitive to negative features, their performances can be significantly improved by introducing a few negative features. This also demonstrates that negative features are beneficial to the classifier.

In Table4-2, the lines including the symbol "-", correspond to the types which performance can not be improved by the introduction of NFs on the validation set. And therefore, we do not introduce any NFs under these types in the testing set. The reason for this phenomenon is that some types are not sensitive to NFs, or the NFs obtained in the current training data are not strong enough to predict non-relevance.

From the above analysis, we conclude that (1) LFS is very useful for ensuring the performance of each type. (2) explicit discrimination of positive & negative features, and the moderate introduction of NFs can significantly improve the performance of some types.

IV. CONCLUSIONS

In this paper, we studied an approach to identify Chinese event and their types, which uses local feature selection, discriminates negative and positive features, and introduces negative features into feature subset. The approach not only effectively alleviates the problem that classifier performs poorly on the difficult types, but improve overall performance.

Currently, experiments on using more negative features available in all training examples are in progress. And we will try to further investigate when and why the negative features will be useful, and try to use some new wrapper method to find the optimal feature subset. Besides that, we will try to explore some new strategies to find features, not appearing in training data, to further improve the performance.

ACKNOWLEDGMENT

Supported by the NSFC(60575041, 60473139), 863HTFC(2006AA01Z150) and ShanxiYSF (20051018).

REFERENCES

[1] The ACE 2005 Evaluation Plan, http://www.ldc.upenn.edu/Projects/ACE/Annotation.

[2] Yanyan Zhao, Xiaoyin Wang, Bin Qin etal., Automatic Event Type Extraction in Chinese Event Extraction, In: Proceeding of the 3rd Student Workshop of Computational Linguistics(SWCL-2006), Shenyang city, China, 2006. 240-245.

[3] Steven Bethard, James H. Martin, Identification of Event Mentions and their Semantic Class, In: Proceeding of 2006 Conference on Empirical Methods in Natural Language Processing (EMNLP2006), Sydney, Australia, 2006. 146-154.

[4] Forman G, a Pitfall and Solution in Multi-Class Feature Selection for Text Classification, In: Proceedings of the 21st International Conference on Machine Learning (ICML2004), Banff, Canada, Morgan Kaufmann Publishers, 2004. 38.

[5] Zheng Z H, Wu X Y, Srihari R, Feature Selection for Text Categorization on Imbalanced Data, In: Proceeding s of SIGKDD2004, vol.6, Issue 1, 2004. 80-89.

[6] Huan Liu, Lei Yu, Toward Integrating Feature Selection Algorithm for Classification and Clustering, IEEE Transaction on Knowledge and Data Engineering, 2005, 17(4): 491-502.

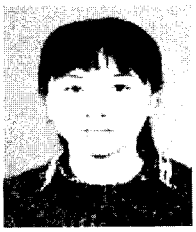
[7] Fabrizio Sebastiani, Machine Learning in Automated Text Categorization, ACM Computing Surveys, 2002, 34(1): 1-47.

[8] Adwait Ratnaparkhi, A Simple Introduction to Maximum Entropy Models for Natural Language Processing, Technical Report 97-08, Institute for Research in Cognitive Science, University of Pennsylvania, http://citeseer.ist.psu.edu/128751.html.

Hongye Tan
Born in 1971. She is currently a Ph.D. candidate of Dept. computer science and technology at Harbin Institute of Technology, and she is also an associate professor of School of Computer and Information Technology at Shanxi University. Her research interests are in natural language processing and artificial intelligence. Email: hytan@mtlab.hit.edu.cn.

Tiejun Zhao
Born in 1962. He is a professor of Dept. computer science and technology at Harbin Institute of Technology. He received the M.S. and PH.D degrees in computer science and technology at Harbin Institute of Technology. His research interests

are in natural language processing, machine translation and artificial intelligence.
Email: tjzhao@mtlab.hit.edu.cn



Haochang Wang
Born in 1974. She is currently a Ph.D. candidate of Dept. computer science and technology at Harbin Institute of Technology. Her research interests are in natural language processing and artificial intelligence.
Email: hcwang@mtlab.hit.edu.cn

Wan-Pyo Hong
Born in 1955. Received the B.S. and M.S. and Ph.D degrees degrees in electronics engineering in 1991, 1993 and 1999 from the National Seoul Polytechnic University, Yonsei and Kwangwoon University, Seoul, Korea respectively. He was a deputy director in the headquarter of Ministry of Information and Communication from 1984 to 1997, a chief manager of the transmission equipment marketing group, Samsung Electronics Co., Ltd from 1997 to 1999 and a Research Professor of Information and Telecommunication Institute in Kwangwoon University from 1999 to 2002. He is currently a associate Professor of Hansei University, Kyungki Province, Korea. He also was a chairman of the Information & Communication Professional Engineer Association. His research interests include RF devices and satellite broadcasting / communications.