

Unsupervised Feature Selection for Relation Extraction

Jinxiu Chen¹

Donghong Ji¹

Chew Lim Tan²

Zhengyu Niu¹

¹Institute for Infocomm Research

21 Heng Mui Keng Terrace

119613 Singapore

{jinxiu,dhji,zniu}@i2r.a-star.edu.sg

²Department of Computer Science

National University of Singapore

117543 Singapore

tancl@comp.nus.edu.sg

Abstract

This paper presents an unsupervised relation extraction algorithm, which induces relations between entity pairs by grouping them into a “natural” number of clusters based on the similarity of their contexts. Stability-based criterion is used to automatically estimate the number of clusters. For removing noisy feature words in clustering procedure, feature selection is conducted by optimizing a trace based criterion subject to some constraint in an unsupervised manner. After relation clustering procedure, we employ a discriminative category matching (DCM) to find typical and discriminative words to represent different relations. Experimental results show the effectiveness of our algorithm.

1 Introduction

Relation extraction is the task of finding relationships between two entities from text contents. There has been considerable work on supervised learning of relation patterns, using corpora which have been annotated to indicate the information to be extracted (e.g. (Califf and Mooney, 1999; Zelenko et al., 2002)). A range of extraction models have been used, including both symbolic rules and statistical rules such as HMMs or Kernels. These methods have been particularly successful in some specific domains. However, manually tagging of large amounts of training data is

very time-consuming; furthermore, it is difficult for one extraction system to be ported across different domains.

Due to the limitation of supervised methods, some weakly supervised (or semi-supervised) approaches have been suggested (Brin, 1998; Eugene and Luis, 2000; Sudo et al., 2003). One common characteristic of these algorithms is that they need to pre-define some initial seeds for any particular relation, then bootstrap from the seeds to acquire the relation. However, it is not easy to select representative seeds for obtaining good results.

Hasegawa, et al. put forward an unsupervised approach for relation extraction from large text corpora (Hasegawa et al., 2004). First, they adopted a hierarchical clustering method to cluster the contexts of entity pairs. Second, after context clustering, they selected the most frequent words in the contexts to represent the relation that holds between the entities. However, the approach exists its limitation. Firstly, the similarity threshold for the clusters, like the appropriate number of clusters, is somewhat difficult to pre-defined. Secondly, the representative words selected by frequency tends to obscure the clusters.

For solving the above problems, we present a novel unsupervised method based on model order selection and discriminative label identification. For achieving model order identification, stability-based criterion is used to automatically estimate the number of clusters. For removing noisy feature words in clustering procedure, feature selection is conducted by optimizing a trace based criterion subject to some constraint in an

unsupervised manner. Furthermore, after relation clustering, we employ a discriminative category matching (DCM) to find typical and discriminative words to represent different relations types.

2 Proposed Method

Feature selection for relation extraction is the task of finding important contextual words which will help to discriminate relation types. Unlike supervised learning, where class labels can guide feature search, in unsupervised learning, it is expected to define a criterion to assess the importance of the feature subsets. Due to the interplay between feature selection and clustering solution, we should define an objective function to evaluate both feature subset and model order.

In this paper, the model selection capability is achieved by resampling based stability analysis, which has been successfully applied to several unsupervised learning problems (e.g. (Levine and Domany, 2001), (Lange et al., 2002), (Roth and Lange et al., 2003), (Niu et al., 2004)). We extend the cluster validation strategy further to address both feature selection and model order identification.

Table 1 presents our model selection algorithm. The objective function $M_{F_k, k}$ is relevant with both feature subset and model order. Clustering solution that is stable against resampling will give rise to a local optimum of $M_{F_k, k}$, which indicates both important feature subset and the true cluster number.

2.1 Entropy-based Feature Ranking

Let $P = \{p_1, p_2, \dots, p_N\}$ be a set of local context vectors of co-occurrences of entity pair E_1 and E_2 . Here, the context includes the words occurring between, before and after the entity pair. Let $W = \{w_1, w_2, \dots, w_M\}$ represent all the words occurred in P . To select a subset of important features from W , words are first ranked according to their importance on clustering. The importance can be assessed by the entropy criterion. Entropy-based feature ranking is based on the assumption that a feature is irrelevant if the presence of it obscures the separability of data set (Dash et al., 2000).

We assume p_n , $1 \leq n \leq N$, lies in feature space W , and the dimension of feature space is

Table 1: Model Selection Algorithm for Relation Extraction

	Input: Corpus D tagged with Entities (E_1, E_2); Output: Feature subset and Model Order (number of relation types);
1.	Collect the contexts of all entity pairs in the document corpus D , namely P ;
2.	Rank features using entropy-based method described in section 2.1;
3.	Set the range (K_l, K_h) for the possible number of relation clusters;
4.	Set estimated model order $k = K_l$;
5.	Conduct feature selection using the algorithm presented in section 2.2;
6.	Record $\hat{F}_{k, k}$ and the score of the merit of both of them, namely $M_{F, k}$;
7.	If $k < K_h$, $k = k + 1$, go to step 5; otherwise, go to Step 7;
8.	Select k and feature subset \hat{F}_k which maximizes the score of the merit $M_{F, k}$;

M . Then the similarity between i -th data point p_i and j -th data point p_j is given by the equation: $S_{i, j} = \exp(-\alpha * D_{i, j})$, where $D_{i, j}$ is the Euclidean distance between p_i and p_j , and α is a positive constant, its value is $-\frac{\ln 0.5}{\bar{D}}$, where \bar{D} is the average distance among the data points. Then the entropy of data set P with N data points is defined as:

$$E = - \sum_{i=1}^N \sum_{j=1}^N (S_{i, j} \log S_{i, j} + (1 - S_{i, j}) \log(1 - S_{i, j})) \quad (1)$$

For ranking of features, the importance of each word $I(w_k)$ is defined as entropy of the data after discarding feature w_k . It is calculated in this way: remove each word in turn from the feature space and calculate E of the data in the new feature space using the Equation 1. Based on the observation that a feature is the least important if the removal of it results in minimum E, we can obtain the rankings of the features.

2.2 Feature Subset Selection and Model Order Identification

In this paper, for each specified cluster number, firstly we perform K-means clustering analysis on each feature subset and adopts a scattering criterion "Invariant Criterion" to select an optimal feature subset F from the feature subset space. Here, $trace(P_W^{-1} P_B)$ is used to compare the cluster quality for different feature subsets ¹, which

¹ $trace(P_W^{-1} P_B)$ is $trace$ of a matrix which is the sum of its diagonal elements. P_W is the within-cluster scatter

Table 2: Unsupervised Algorithm for Evaluation of Feature Subset and Model Order

Function: $\text{criterion}(F, k, P, q)$
Input: feature subset F , cluster number k , entity pairs set P , and sampling frequency q ;
Output: the score of the merit of F and k ;
1. With the cluster number k as input, perform k -means clustering analysis on pairs set P^F ;
2. Construct connectivity matrix $C_{F,k}$ based on above clustering solution on full pairs set P^F ;
3. Use a random predictor ρ_k to assign uniformly drawn labels to each entity pair in P^F ;
4. Construct connectivity matrix C_{F,ρ_k} based on above clustering solution on full pairs set P^F ;
5. Construct q sub sets of the full pairs set, by randomly selecting αN of the N original pairs, $0 \leq \alpha \leq 1$;
6. For each sub set, perform the clustering analysis in Step 2, 3, 4, and result $C_{F,k}^\mu, C_{F,\rho_k}^\mu$;
7. Compute $M_{F,k}$ to evaluate the merit of k using Equation 3;
8. Return $M_{F,k}$;

measures the ratio of between-cluster to within-cluster scatter. The higher the $\text{trace}(P_W^{-1}P_B)$, the higher the cluster quality.

To improve searching efficiency, features are first ranked according to their importance. Assume $W_r = \{f_1, \dots, f_M\}$ is the sorted feature list. The task of searching can be seen in the feature subset space: $\{(f_1, \dots, f_k), 1 \leq k \leq M\}$.

Then the selected feature subset F is evaluated with the cluster number using the objective function, which can be formulated as: $\hat{F}_k = \arg \max_{F \subseteq W_r} \{\text{criterion}(F, k)\}$, subject to $\text{coverage}(P, F) \geq \tau^2$. Here, \hat{F}_k is the optimal feature subset, F and k are the feature subset and the value of cluster number under evaluation, and the criterion is set up based on resampling-based stability, as Table 2 shows.

Let P^μ be a subset sampled from full entity pairs set P with size $\alpha|P|$ (α set as 0.9 in this paper.), $C(C^\mu)$ be $|P| \times |P|(|P^\mu| \times |P^\mu|)$ connectivity matrix based on the clustering results on $P(P^\mu)$. Each entry $c_{ij}(c_{ij}^\mu)$ of $C(C^\mu)$ is calculated in the following: if the entity pair $p_i \in P(P^\mu)$, $p_j \in P(P^\mu)$ belong to the same cluster, then $c_{ij}(c_{ij}^\mu)$ equals 1, else 0. Then the stability is de-

matrix as: $P_W = \sum_{j=1}^c \sum_{X_i \in X_j} (X_i - m_j)(X_j - m_j)^t$ and P_B is the between-cluster scatter matrix as: $P_B = \sum_{j=1}^c (m_j - m)(m_j - m)^t$, where m is the total mean vector and m_j is the mean vector for j^{th} cluster and $(X_j - m_j)^t$ is the matrix transpose of the column vector $(X_j - m_j)$.

²let $\text{coverage}(P, F)$ be the coverage rate of the feature set F with respect to P . In practice, we set $\tau = 0.9$.

finied in Equation 2:

$$M(C^\mu, C) = \frac{\sum_{i,j} 1\{C_{i,j}^\mu = C_{i,j} = 1, p_i \in P^\mu, p_j \in P^\mu\}}{\sum_{i,j} 1\{C_{i,j} = 1, p_i \in P^\mu, p_j \in P^\mu\}} \quad (2)$$

Intuitively, $M(C^\mu, C)$ denotes the consistency between the clustering results on C^μ and C . The assumption is that if the cluster number k is actually the ‘‘natural’’ number of relation types, then clustering results on subsets P^μ generated by sampling should be similar to the clustering result on full entity pair set P . Obviously, the above function satisfies $0 \leq M \leq 1$.

It is noticed that $M(C^\mu, C)$ tends to decrease when increasing the value of k . Therefore for avoiding the bias that small value of k is to be selected as cluster number, we use the cluster validity of a random predictor ρ_k to normalize $M(C^\mu, C)$. The random predictor ρ_k achieved the stability value by assigning uniformly drawn labels to objects, that is, splitting the data into k clusters randomly. Furthermore, for each k , we tried q times. So, in the step 7 of the algorithm of Table 2, the objective function $M(C_{F,k}^\mu, C_{F,k})$ can be normalized as equations 3:

$$M_{F,k}^{\text{norm}} = \frac{1}{q} \sum_{i=1}^q M(C_{F,k}^{\mu_i}, C_{F,k}) - \frac{1}{q} \sum_{i=1}^q M(C_{F,\rho_k}^{\mu_i}, C_{F,\rho_k}) \quad (3)$$

Normalizing $M(C^\mu, C)$ by the stability of the random predictor can yield values independent of k .

After the number of optimal clusters and the feature subset has been chosen, we adopted the K-means algorithm for the clustering phase. The output of context clustering is a set of context clusters, each of them is supposed to denote one relation type.

2.3 Discriminative Feature identification

For labelling each relation type, we use DCM (discriminative category matching) scheme to identify discriminative label, which is also used in document classification (Gabriel et al., 2002) and weights the importance of a feature based on their distribution. In this scheme, a feature is not important if the feature appears in many clusters and is evenly distributed in these clusters, otherwise it will be assigned higher importance.

To weight a feature f_i within a category, we take into account the following information:

Table 3: Three domains of entity pairs: frequency distribution for different relation types

PER-ORG		ORG-GPE		ORG-ORG	
# of pairs:786		# of pairs:262		# of pairs:580	
Relation types	Percentage	Relation types	Percentage	Relation types	Percentage
Management	36.39%	Based-In	46.56%	Member	27.76%
General-staff	29.90%	Located	35.11%	Subsidiary	19.83%
Member	19.34%	Member	11.07%	Part-Of	18.79%
Owner	4.45%	Affiliate-Partner	3.44%	Affiliate-Partner	17.93%
Located	3.28%	Part-Of	2.29%	Owner	8.79%
Client	1.91%	Owner	1.53%	Client	2.59%
Other	1.91%			Management	2.59%
Affiliate-Partner	1.53%			Other	1.21%
Founder	0.76%			Other	0.52%

- The relative importance of f_i within a cluster is defined as: $WC_{i,k} = \frac{\log_2(pf_{i,k}+1)}{\log_2(N_k+1)}$, where $pf_{i,k}$ is the number of those entity pairs which contain feature f_i in cluster k . N_k is the total number of term pairs in cluster k .
- The relative importance of f_i across clusters is given by: $CC_i = \log \frac{N \cdot \max_{k \in C_i} \{WC_{i,k}\}}{\sum_{k=1}^N WC_{i,k}} \cdot \frac{1}{\log N}$, where C_i is the set of clusters which contain feature f_i . N is the total number of clusters.

Here, $WC_{i,k}$ and CC_i are designed to capture both local information within a cluster and global information about the feature distribution across clusters respectively. Combining both $WC_{i,k}$ and CC_i we define the weight $W_{i,k}$ of f_i in cluster k as: $W_{i,k} = \frac{WC_{i,k}^2 \cdot CC_i^2}{\sqrt{WC_{i,k}^2 + CC_i^2}} \cdot \sqrt{2}$, $0 \leq W_{i,k} \leq 1$.

3 Experiments and Results

3.1 Data

We constructed three subsets for domains PER-ORG, ORG-GPE and ORG-ORG respectively from ACE corpus³ The details of these subsets are given in Table 3, which are broken down by different relation types. To verify our proposed method, we only extracted those pairs of entity mentions which have been tagged relation types. And the relation type tags were used as ground truth classes to evaluate.

3.2 Evaluation method for clustering result

Since there was no relation type tags for each cluster in our clustering results, we adopted a permutation procedure to assign different relation type tags to only $\min(|EC|, |TC|)$ clusters, where $|EC|$ is the estimated number of clusters, and $|TC|$ is the number of ground truth classes

(relation types). This procedure aims to find an one-to-one mapping function Ω from the TC to EC . To perform the mapping, we construct a contingency table T , where each entry $t_{i,j}$ gives the number of the instances that belong to both the i -th cluster and j -th ground truth class. Then the mapping procedure can be formulated as: $\hat{\Omega} = \arg \max_{\Omega} \sum_{j=1}^{|TC|} t_{\Omega(j),j}$, where $\Omega(j)$ is the index of the estimated cluster associated with the j -th class.

Given the result of one-to-one mapping, we can define the evaluation measure as follows: $Accuracy(P) = \frac{\sum_j t_{\hat{\Omega}(j),j}}{\sum_{i,j} t_{i,j}}$. Intuitively, it reflects the accuracy of the clustering result.

3.3 Evaluation method for relation labelling

For evaluation of the relation labeling, we need to explore the relatedness between the identified labels and the pre-defined relation names. To do this, we use one information-content based measure (Lin, 1997), which is provided in Wordnet-Similarity package (Pedersen et al., 2004) to evaluate the similarity between two concepts in Wordnet. Intuitively, the relatedness between two concepts in Wordnet is captured by the information content of their lowest common subsumer (lcs) and the information content of the two concepts themselves, which can be formalized as follows: $Relatedness_{lin}(c_1, c_2) = \frac{2 \times IC(lcs(c_1, c_2))}{IC(c_1) + IC(c_2)}$. This measure depends upon the corpus to estimate information content. We carried out the experiments using the British National Corpus (BNC) as the source of information content.

3.4 Experiments and Results

For comparison of the effect of the outer and within contexts of entity pairs, we used five dif-

³<http://www ldc.upenn.edu/Projects/ACE/>

Table 4: Automatically determined the number of relation types using different feature ranking methods.

Domain	Context Window Size	# of real relation types	Model Order Base-line	Model Order with χ^2	Model Order with Freq	Model Order with Entropy
PER-ORG	0-5-0	9	7	7	7	7
	2-5-2	9	8	6	7	8
	0-10-0	9	8	6	8	8
	2-10-2	9	6	7	6	8
	5-10-5	9	5	5	6	7
ORG-GPE	0-5-0	6	3	3	3	4
	2-5-2	6	2	3	4	4
	0-10-0	6	6	4	5	6
	2-10-2	6	4	3	4	5
	5-10-5	6	2	3	3	3
ORG-ORG	0-5-0	9	7	7	7	7
	2-5-2	9	7	5	6	7
	0-10-0	9	9	8	9	9
	2-10-2	9	6	6	6	7
	5-10-5	9	8	5	7	9

ferent settings of context window size (WIN_{pre} - WIN_{mid} - WIN_{post}) for each domain.

Table 4 shows the results of model order identification without feature selection (Baseline) and with feature selection based on different feature ranking criterion (χ^2 , Frequency and Entropy). The results show that the model order identification algorithm with feature selection based on entropy achieve best results: estimate cluster numbers which are very close to the true values. In addition, we can find that with the context setting, 0-10-0, the estimated number of the clusters is equal or close to the ground truth value. It demonstrates that the intervening words less than 10 are appropriate features to reflect the structure behind the contexts, while the intervening words less than 5 are not enough to infer the structure. For the contextual words beyond (before or after) the entities, they tend to be noisy features for the relation estimation, as can be seen that the performance deteriorates when taking them into consideration, especially for the case without feature selection.

Table 5 gives a comparison of the average accuracy over five different context window size settings for different clustering settings. For each domain, we conducted five clustering procedures: Hasegawa’s method, $RL_{Baseline}$, $RLFS_{\chi^2}$, $RLFS_{Freq}$ and $RLFS_{Entropy}$. For Hasegawa’s method (Hasegawa et al., 2004), we set the cluster number to be identical with the number of ground truth classes. For $RL_{Baseline}$, we use the estimated cluster number to clus-

ter contexts without feature selection. For $RLFS_{\chi^2}$, $RLFS_{Freq}$ and $RLFS_{Entropy}$, we use the selected feature subset and the estimated cluster number to cluster the contexts, where the feature subset comes from χ^2 , frequency and entropy criterion respectively. Comparing the average accuracy of these clustering methods, we can find that the performance of feature selection methods is better than or comparable with the baseline system without feature selection. Furthermore, it is noted that $RLFS_{Entropy}$ achieves the highest average accuracy in three domains, which indicates that entropy based feature pre-ranking provides useful heuristic information for the selection of important feature subset.

Table 6 gives the automatically estimated labels for relation types for the domain PER-ORG. We select two features as labels of each relation type according to their DCM scores and calculate the average (and maximum) relatedness between our selected labels (E) and the predefined labels (H). Following the same strategy, we also extracted relation labels (T) from the ground truth classes and provided the relatedness between T and H. From the column of relatedness (E-H), we can see that it is not easy to find the hand-tagged relation labels exactly, furthermore, the identified labels from the ground-truth classes are either not always comparable to the pre-defined labels in most cases (T-H). The reason may be that the pre-defined relation names tend to be some abstract labels over the features, e.g., ‘management’ vs. ‘president’,

Table 5: Performance of the clustering algorithms over three domains: the average accuracy over 5 different context window size.

Domain	Hasegawa's method	RL _{Baseline}	RLFS _{χ^2}	RLFS _{Freq}	RLFS _{Entropy}
PER-ORG	32.4%	34.3%	33.9%	36.6%	41.3%
ORG-GPE	43.7%	47.4%	47.1%	48.4%	50.6%
ORG-ORG	26.5%	36.2%	36.0%	38.7%	42.4%

Table 6: Relation Labelling using DCM strategy for the domain PER-ORG. Here, (T) denotes the identified relation labels from ground truth classes. (E) is the identified relation labels from our estimated clusters. 'Ave (T-H)' denotes the average relatedness between (T) and (H). 'Max (T-H)' denotes the maximum relatedness between (T) and (H).

Hand-tagged Label (H)	Identified Label (T)	Identified Label (E)	Ave (T-H)	Max (T-H)	Ave (E-H)	Max (E-H)	Ave (E-T)	Max (E-T)
management	head,president	president,control	0.3703	0.4515	0.3148	0.3406	0.7443	1.0000
general-staff	work,fire	work,charge	0.6254	0.7823	0.6411	0.7823	0.6900	1.0000
member	join,communist	become,join	0.394	0.4519	0.1681	0.3360	0.3366	1.0000
owner	bond,bought	belong,house	0.1351	0.2702	0.0804	0.1608	0.2489	0.4978
located	appear,include	lobby,appear	0.0000	0.0000	0.1606	0.3213	0.2500	1.0000
client	hire,reader	bought,consult	0.4378	0.8755	0.0000	0.0000	0.1417	0.5666
affiliate-partner	affiliate,associate	assist,affiliate	0.9118	1.0000	0.5000	1.0000	0.5000	1.0000
founder	form,found	invest,set	0.1516	0.3048	0.3437	0.6875	0.4376	0.6932

'head' or 'control'; 'member' vs. 'join', 'become', etc., while the abstract words and the features are located far away in Wordnet. Table 6 also lists the relatedness between (E) and (T). We can see that the labels are comparable by their maximum relatedness(E-T).

4 Conclusion and Future work

In this paper, we presented an unsupervised approach for relation extraction from corpus. The advantages of the proposed approach includes that it doesn't need any manual labelling of the relation instances, it can identify an important feature subset and the number of the context clusters automatically, and it can avoid extracting those common words as characterization of relations.

References

- Mary Elaine Califf and Raymond J.Mooney. 1999. *Relational Learning of Pattern-Match Rules for Information Extraction*, AAAI99.
- Sergey Brin. 1998. *Extracting patterns and relations from world wide web*. In *Proc. of WebDB'98*. pages 172-183.
- Kiyoshi Sudo, Satoshi Sekine and Ralph Grishman. 2003. *An Improved Extraction Pattern Representation Model for Automatic IE Pattern Acquisition*. *Proceedings of ACL 2003; Sapporo, Japan*.
- Eugene Agichtein and Luis Gravano. 2000. *Snowball: Extracting Relations from large Plain-Text Collections*, In *Proc. of the 5th ACM International Conference on Digital Libraries (ACMDL'00)*.
- Takaaki Hasegawa, Satoshi Sekine and Ralph Grishman. 2004. *Discovering Relations among Named Entities from Large Corpora*, *ACL2004*. Barcelona, Spain.
- Dmitry Zelenko, Chinatsu Aone and Anthony Richardella. 2002. *Kernel Methods for Relation Extraction*, *EMNLP2002*. Philadelphia.
- Lange,T., Braun,M.,Roth, V., and Buhmann,J.M.. 2002. *Stability-Based Model Selection*, *Advances in Neural Information Processing Systems 15*.
- Levine,E. and Domany,E.. 2001. *Resampling Method for Unsupervised Estimation of Cluster Calidity*, *Neural Computation*, Vol.13, 2573-2593.
- Zhengyu Niu, Donghong Ji and Chew Lim Tan. 2004. *Document Clustering Based on Cluster Validation*, *CIKM'04*. November 8-13, 2004, Washington, DC, USA.
- Volker Roth and Tilman Lange. 2003. *Feature Selection in Clustering Problems*, *NIPS2003 workshop*.
- Manoranjan Dash and Huan Liu. 2000. *Feature Selection for Clustering*, *Proceedings of Pacific-Asia Conference on Knowledge Discovery and Data Mining*.
- Gabriel Pui Cheong Fung, Jeffrey Xu Yu and Hongjun Lu. 2002. *Discriminative Category Matching: Efficient Text Classification for Huge Document Collections*, *ICDM2002*. December 09-12, 2002, Japan.
- D.Lin. 1997. *Using syntactic dependency as a local context to resolve word sense ambiguity*. In *Proceedings of the 35th Annual Meeting of ACL*,. Madrid, July 1997.
- Ted Pedersen, Siddharth Patwardhan and Jason Michelizzi. 2004. *WordNet::Similarity-Measuring the Relatedness of Concepts*, *AAAI2004*.