A Semantic Feature for Relation Recognition Using A Web-based Corpus

Chen-Ming Hung
Institute of Information Science
Academia Sinica, Taipei, Taiwan
rglly@iis.sinica.edu.tw

Abstract

Selecting appropriate features to represent an entity pair plays a key role in the task of relation recognition. However, existing syntactic features or lexical features cannot capture the interaction between two entities because of the dearth of annotated relational corpus specialized for relation recognition. In this paper, we propose a semantic feature, called the $latent\ topic\ feature$, which is topic-based and represents an entity pair at the semantic level instead of the word level. Moreover, to address the problem of insufficiently annotated corpora, we propose an algorithm for compiling a training corpus from the Web. Experiment results demonstrate that $latent\ topic\ features$ are as effective as syntactic or lexical features. Moreover, the Web-based corpus can resolve the problems caused by insufficiently annotated relational corpora.

1 INTRODUCTION

Relation recognition is a challenging task because finding appropriate features to represent the relationship between two entities is difficult and limited by the scarcity of annotated corpora. Prior works on relation recognition have focused on syntactic features, e.g., parsing trees (Culotta and Sorensen, 2004; Zelenko et al., 2003), and on lexical features, e.g., Part-Of-Speech (POS) features. These approaches show that syntactic features and lexical features outperform bag-of-words (BOW) on existing annotated corpora such as the RDC corpus of the ACE project. The superior performance achieved by syntactic and lexical features is due to their ability to capture the grammatical relations between two entities and the characteristics of the entities. For example, (Culotta and Sorensen, 2004) add hypernyms of entities to features derived from WordNet. However, neither syntactic nor lexical features can capture the interaction between two entities at the semantic level.

Another issue in the task of relation recognition is insufficiently annotated corpora. For example, given a pair $\{$*the U.N. body, Kosovo*$\}$, we can only find three sentences containing both entities in the RDC corpus, which is commonly used corpus in the relation recognition task. The problem of an insufficiently annotated corpus biases feature vectors and distorts the prediction of entity pairs. However, (Huang et al., 2004; Hung and Chien, 2007) have shown that the Web can be used as an alternative source of documents related to a given query. That is possibly because of the increasing size of the Web and the efficiency in commercial search engines, e.g., Google and Yahoo!.

To resolve the above problems, we propose a semantic feature called the *latent topic feature*, which is extracted by exploiting the *Latent Dirichlet Allocation* (LDA) algorithm. Unlike syntactic features or lexical features, *latent topic features* represent entity pairs as random mixtures of latent topics, where each topic is characterized by a distribution of words. We prove experimentally that *latent topic features* are as effective as syntactic features or lexical features in capturing the interaction between two entities. The experiment results are predictable. In

the above {*the U.N. body, Kosovo*} example, it may be difficult to determine the relationship between *U.N. body* and *Kosovo* straightforwardly. However, making the right guess about the relationship is easier if *the U.N. body* is grouped with *army* and *government*. Therefore, the $right\ guess$ in this example is $management$.

To overcome the problems caused by an insufficiently annotated corpus, we exploit the Web as a source of training data for the relation recognition task. Given an entity pair, documents describing the entity pair are extracted from the Web via commercial search engines using both entities as the query. In other words, snippets returned from the Web are treated as documents related to the query. Our assumption, which has been proved in previously published works, is that returned snippets can capture the interaction between two entities. After the *latent topic features* extracted from returned snippets using the Web as the corpus, an SVM classifier is trained as the relation recognition classifier for use in the later experiments.

The remainder of the paper is organized as follows. in Section 2, we discuss works related to feature selection in the relation recognition task as well as using the Web as a corpus. The concept of *latent topic features* is presented in Section 3. We also explain how we represent a document in the vector space of a *latent topic feature*. Section 4contains an evaluation of the *latent topic feature*. We then present our conclusions in Section 5.

2 RELATED WORK

In the field of information extraction (IE), the goal of relation recognition is to find the relationship between two entities. Without considering entity detection, relation recognition depends heavily on the representation of entity pairs. (Zelenko et al., 2003) showed how to extract relations by computing the kernel functions between the kernels of shallow parse trees. The kernels are defined over a shallow parse representation of the text and used in conjunction with a Support Vector Machine (SVM) learning algorithm to extract person-affiliation and organization-location relations. (Culotta and Sorensen, 2004) extended this work to estimate kernel functions between augmented dependency trees, while (Kambhatla, 2004) combined lexical features, syntactic features, and semantic features in a maximum entropy model. However, the semantic features discussed in (Kambhatla, 2004) still focus on the word level instead of the conceptual level.

LDA is an aspect model that represents documents as a set of $topics$ instead of a bag-of-words. *Latent semantic indexing (LSI)* (Deerwester et al., 1990) and *probabilistic latent semantic indexing (PLSI)* (Hofmann, 1999) are also aspect models and have been widely used in the field of information retrieval. LSI simply assumes that each document is generated from single latent topic, while PLSI attempts to relax the assumption by using a mixture of latent topics for each document. However, PLSI is highly dependent on training documents; in other words, it cannot handle the probability of latent topics in a previously unseen document. In addition, the number of parameters that must be estimated in PLSI grows linearly with the number of training documents. (Blei et al., 2003; Blei and Jordan, 2003) proposed LDA to resolve the abovementioned limitations. It can easily generate an unseen document under controllable parameters.

A number of works, e.g., (Huang et al., 2004), have investigated using the Web to acquire a training corpus or acquire additional information not provided by existing annotated corpora. (Huang et al., 2004) exploited the Web as a training corpus to train a classifier with user-defined categories. However, it is widely recognized that when using documents on the Web users must spend a great deal of time filtering out unrelated contents. (Hung and Chien, 2007) designed a bootstrapping method that adapts an existing corpus with an automatic verification algorithm in order to control the quality of returned snippets in each iteration. (Matsuo et al., 2006) used the Web to construct a social network system, called *POLYPHONET*, which visualizes the relationship between two personal names.

3 LATENT TOPIC FEATURE

In this section, we introduce the concept of using the Web to augment an insufficiently annotated corpus for relation recognition. Then we apply the LDA algorithm to the corpus to extract the $latent\ topic$

$features$ to represent entity pairs in the corpus for relation recognition.

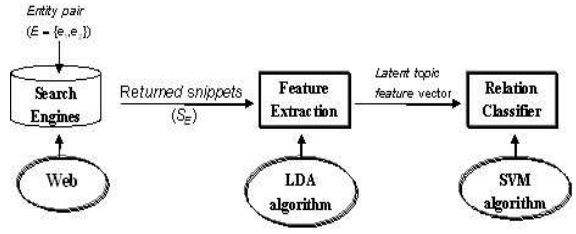

Figure 1: The framework of the proposed approach.

3.1 Compiling a Web-based Relational Corpus

For an entity pair, $E = \{e_1, e_2\}$, where e_1 and e_2 are named entities, it is difficult to find sufficient sentences to describe their relationship from existing annotated corpora. In other words, given an entity pair without a relation label, users cannot recognize the pair. Even with a widely used thesaurus, like WordNet, we can only obtain hypernyms or synonyms of given entities. It is not possible to obtain knowledge about the interaction between two entities.

To capture the interaction between two entities, we send both entities, $e_1\ AND\ e_2$, to commercial search engines and collect returned snippets as training documents for an entity pair, E. Snippets of returned search results are defined as the surrounding contexts of queries highlighted by commercial search engines. In other words, the full texts of search results are not considered in the collected corpus when filtering noisy information in full documents. Let R be the relation label of entity pairs $\{E_1, \ldots, E_M\}$; then, the training corpus for R is the collection of all returned snippets for $\{E_1, \ldots, E_M\}$. Through effective commercial search engines such as Google and Yahoo!, sentences describing the interaction between two entities can easily be retrieved. Returning to the example $\{$*the U.N. body, Kosovo*$\}$, almost two million sentences with co-occurrences of the two entities are retrieved by Google. Another advantage of using the Web to retrieve relevant documents is the $auto-correction$ ability of commercial search engines. The feature can correct a misspelled query or replace an uncommon word with a synonym or a common word which is correct, so that more related information about entity pairs can be retrieved from the Web as returned snippets. For example, Google can automatically link $the\ U.N.\ body$ to $United\ Nations$, which is used more frequently in searching. Clearly, the number of returned snippets must be considered. Actually, based on experiment results in Section 4, we set the number as five, which achieves the best performance.

3.2 Modified LDA for the Relational Corpus

LDA is an aspect model with three levels, namely, the corpus level, the document level, and the word level. Given a document, variables of the corpus it belongs to are sampled first, after which the variable of the document is sampled once. Finally, variables for words in the document are sampled.

For a document d in a corpus D, the modeling process is as follows:

1. Sample $\theta \sim Dir(\theta|\alpha)$.
2. For each word w_n in d, $n \in \{1, \ldots, N\}$:
   (a) sample $z_n \sim Mult(\theta)$,
   (b) sample a word $w_n \sim p(w_n|z_n, \beta)$ from a multivariate Gaussian distribution conditioned on the topic z_n.

Note that α is a vector of corpus-level variables whose dimensionality is equal to the number of latent topics; θ is a variable of the document and is assumed to follow Dirichlet distribution for the given corpus; and β is a word-level variable. In addition, $Z = \{z_1, z_2, \ldots, z_N\}$ are latent factors that generate the document, and z_n is the latent topic that w_n is generated from. Finally, N is the length of the document d.

An entity pair E in the relational corpus is similar to a document d in the text corpus. In other words, the corpus D^R is comprised of returned snippets for all entity pairs E^R with the same relation label R. Therefore, given the parameters α and β, we obtain the distribution of entity pair E as follows:

$$p(E|\alpha, \beta) = \int \sum_{z_n} p(\theta, z_n, S_E|\alpha, \beta) d\theta,$$

where

$$p(\theta, z_n, S_E|\alpha, \beta) = p(\theta|\alpha) \prod_{n=1}^{N_E} p(z_n|\theta) p(w_n|z_n, \beta),$$

N_E is the number of words in the returned snippets for E; and w_n is the n^{th} word in S_E, the returned snippets of E. Table 1 summarizes notations used in the paper.

Table 1: Notations used in this paper.

SYMBOL	DESCRIPTION
R	relation label
D^R	corpus for R
E_j^R	j^{th} entity pair in the relation label R
S_E	returned snippets for an entity pair E
$\|E^R\|$	number of entity pairs in the relation R
N_E	number of words in S_E
w_n	n_{th} word in S_E
z_n	latent topic that w_n is generated from

In Section 3.1, we discussed the advantages of using the Web as a corpus to model entity pairs. In the modeling process, we estimate the probability of w_n conditioned on z_n, $p(w_n|z_n, \beta)$, to maximize the probability of the entire corpus of R. The probability that we try to maximize is

$$p(D^R|\alpha, \beta) = \prod_{j=1}^{|E^R|} p(E_j^R|\alpha, \beta).$$

3.3 Latent Topic Feature

In different corpora, z obtains a different distribution to maximize the likelihood of the given corpus. In this section, we describe how to exploit z as features to represent a snippet for an entity pair E, i.e., S_E. In Section 3.2, we noted that the parameters to be estimated in the aspect model are all probabilities of words in each latent topic z. Thus, we let the expected number of words generated from latent topics be features of each entity pair. In other words, an entity pair E is represented as a feature vector whose length is equal to the number of latent topics and whose i^{th} attribute is equal to

$$\alpha_i + \sum_{n=1}^{N_E} |w_n| \times p(w_n|z_n, \beta),$$

where α_i is the i^{th} prior Dirichlet parameter.

In addition, because there is no solution good enough to determine the dimensionality of the feature vector or the number of latent topics, we set the number of topics at thirty because it probably minimize the computation cost without significantly affecting the performance.

4 EXPERIMENTS

In this section, we evaluate the performance of the $latent\ topic\ feature$ in representing entity pairs extracted from the Relation Detection and Characterization (RDC) corpus of the Automatic Content Extraction 2003 model (ACE 2003)[1].

4.1 The RDC Corpus

In the RDC corpus, five relation types, AT, $NEAR$, $PART$, $ROLE$, and SOC, are defined; each relation type has extended sub-relations. Table 2 summarizes the relations in the RDC corpus for ACE 2003. Based on Table 2, we find that the distribution of the number relations is very unbalanced, ranging from 2 to 773. In the following experiments, we only consider the $Role$ relation because it has the largest numbers of sub-relations and it is easier to verify the recognition results manually. Note that a relation is dropped if it has less than ten sub-relations in order to avoid the bias of learned classifiers. Therefore, the sub-relation $founder$ in $Role$ is dropped in the following experiments because it occurs less than ten times. $Other$ is also dropped because its definition is unclear.

Table 2: Distribution over relation types in the RDC corpus (ACE 2003).

Relations	Sub-Relations(Size)	
AT	Based-in(78)	Located(773)
	Residence(186)	
NEAR	Relative-location(73)	
PART	Part-of(242)	Subsidiary(172)
	Other(2)	
ROLE	Affiliate-partner(34)	
	Citizen-of(93)	Client(33)
	Founder(6)	General-Staff(460)
	Management(294)	Member(398)
	Owner(41)	Other(98)
SOC	Associate(25)	Grandparent(3)
	Parent(23)	Sibling(5)
	Spouse(22)	Other-relative(24)
	Other-Personal(10)	
	Other-Professional(88)	

[1] http://projects.ldc.upenn.edu/ace/

4.2 Setting and Measurement

We used the package of (Chang and Lin, 2001) to design the following experiments. In addition, ν-SVM with a radial kernel function was used to learn the relation classifier. To determine the parameters in ν-SVM, i.e., γ and ν, we observed the performance of the ν-SVM classifier by randomly selecting 80% of the sentences in the RDC corpus as training data and the remaining 20% as test data. In other words, we applied five-fold cross validation to build a temporary model for parameter estimation. Furthermore, it is well known that parameters in the SVM model must be optimized manually; therefore, we estimate ν first and then estimate γ. γ is fixed while ν is being estimated and vice versa. After estimation, the best result is achieved at the point that γ is equal to 2.5×10^{-4} and ν is equal to 0.05. We summarize the results in Figure 2. The top graph in Figure 2 is the accuracy curve, where fixed $\gamma = 2.5 \times 10^{-4}$ and flexible ν; the bottom graph is the accuracy curve with fixed $\nu = 0.05$ and flexible γ.

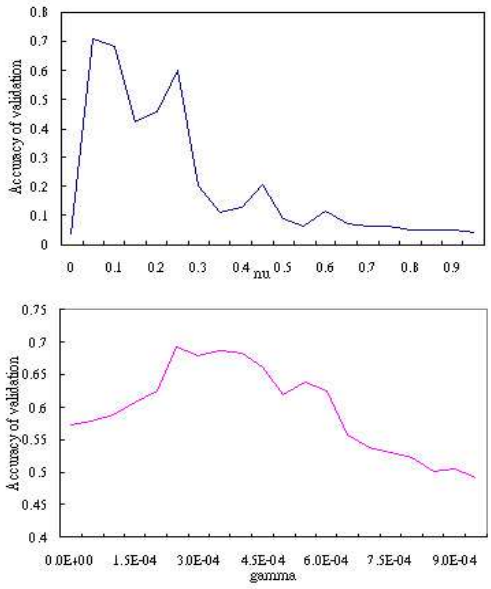

Figure 2: Accuracy of five-fold cross validation using bigram features. Top: ν with $\gamma = 2.5 \times 10^{-4}$. Bottom: γ with $\nu = 0.05$.

For each sub-relation in $Role$, binary classification is used in the experiments and the F-measure of each sub-relation is used as the metric for assessing the performance of $latent\ topic\ features$.

$$
\begin{aligned}
F - value &= \frac{2 \times Precision \times Recall}{Precision + Recall} \\
Recall &= \frac{\sharp \text{ of correct positive predictions}}{\sharp \text{ of positive examples}} \\
Precision &= \frac{\sharp \text{ of correct positive predictions}}{\sharp \text{ of positive predictions}}
\end{aligned}
$$

4.3 Web-based Corpus vs. Annotated Corpus

We now evaluate the performance of relational classifier on a Web-based corpus and on an annotated corpus. To assess the performance of on the annotated corpus, sentences in the RDC corpus containing a co-occurrence of both given entities were extracted as training data to learn a benchmark relation classifier. On the other hand, the Web-based corpus is compiled from snippets retrieved by using both entities as a query. The $latent\ topic\ feature$ is applied on both the Web-based corpus and the RDC corpus using the procedure described in Section 4.2. In addition, to analyze the effect of the number of returned snippets, we increased the number of snippets from 3 to 45 in increments of three and then summarized the relationship between the number of returned snippets and the achieved accuracy curve shown in Figure 3. In the figure, the training data is comprised of snippets of information returned by querying 80% of the entity pairs selected at random in the RDC corpus. The test data comprises snippets returned by querying the remaining 20% of entity pairs in the corpus.

From Figure 3, we observe that using five returned snippets for each entity pair achieves the best accuracy (0.85), which is substantially higher than the accuracy achieved by using annotated corpus (0.69). Note that using more returned snippets does not guarantee higher accuracy. For example, when 39 returned snippets are used for each entity pair, the accuracy (0.56) is almost the same as that (0.55) achieved by using only 3 returned snippets. Moreover, it is significantly less than the accuracy (0.69) achieved by using the RDC corpus. This is reasonable because the greater the number of returned snippets, the larger the amount of noisy information introduced to the classifier, which degrades its performance.

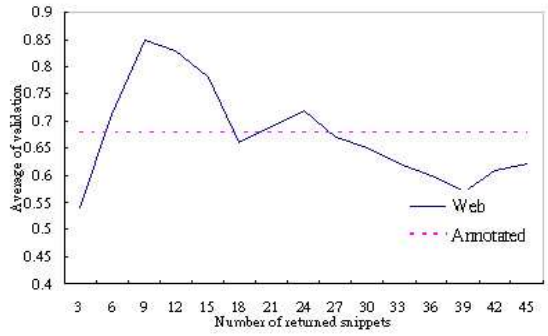

Figure 3: Accuracy of five-fold cross validation using the Web-based corpus and the annotated corpus.

4.4 Latent Topic Feature vs. Other Features

In this section, we compare the performance of $latent\ topic\ features$ with that of syntactic features and lexical features, i.e., $bag - of - words$ or $parts - of - speech$. Because of the superior performance achieved by using the Web-based corpus described in Section 4.3, we extracted features from the training corpus compiled from that corpus rather than the annotated corpus.

Based on the results reported in Section 4.3, five snippets were returned by the Web-based corpus for each entity pair. For each sub-relation, a one-class SVM was trained to perform binary classification.

Each sub-relation of $Role$ in Table 3 is applied with binary classification using a one-class SVM. Table 5 summarizes the results of a comparison between the $latent\ topic\ feature$ and the features used by (Culotta and Sorensen, 2004). The latter depends on dependency tree kernels, which represent the grammatical dependencies in a sentence and are considered as syntactical features. In Table 5, BOW denotes bag-of-word, $sparse$ represents a sparse kernel, and $contiguous$ represents a contiguous kernel.

Surprisingly, for every sub sub-relation in Table 3, the $latent\ topic\ feature$ consistently achieves a significantly higher average recall rate, but a lower average precision rate. This may be due to the $latent\ topic\ feature's$ ability to capture information at the semantic level precisely, but it cannot distinguish the information at the word level easily. In other words, the $latent\ topic\ feature$ can capture the common semantic information, probably the $Role$, of all sub-relations, but it cannot tell the difference between $citizen - of$ and $founder$. Table 4 shows the results of applying binary classification to five relations in the RDC corpus. Although the precision rate for each relation is still low, the recall rate has been increased significantly. This demostrates the ability of the $latent\ topic\ feature$ to capture semantic information.

Table 3: Binary classification results for each sub-relation of $Role$.

	Latent Topic Feature		
	F	**Prec.**	**Rec.**
Aff.-Part.	0.30	0.18	1.00
Client	0.40	0.28	0.71
Citizen-Of	0.62	0.47	0.91
Gen.-Staff	0.78	0.64	0.99
Manage.	0.56	0.39	1.00
Member	0.62	0.46	0.93
Owner	0.45	0.29	0.98

Table 4: Binary classification results for each relation in the RDC corpus.

	Latent Topic Feature		
	F	**Prec.**	**Rec.**
At	0.61	0.48	0.84
NEAR	0.36	0.23	0.88
PART	0.58	0.46	0.80
ROLE	0.71	0.64	0.79
SOC	0.59	0.45	0.87

In Table 5, although the recall rate using the $latent\ topic\ feature$ is much higher than that achieved by the other features, unfortunately, the $F - score$ of the $latent\ topic\ feature$ cannot be redeemed because of the much lower precision rate. Moreover, the $latent\ topic\ feature$ is comparable to the sparse kernel method in a different way because it has a low precision rate but a high recall rate. Finally, the $latent\ topic\ feature$ achieves a higher average F-score than the bag-of-words feature, which proves the assumption that the $latent\ topic\ feature$ can better capture the interaction between two entities than features at the word level.

5 CONCLUSION

We have proposed a concept called the $latent\ topic\ feature$ for the task of relation recognition and evaluated it on the RDC of the ACE project. The feature captures the interaction between two entities

Table 5: Comparison between $Latent\ topic\ feature$ and other features.

	Average		
	F	**Prec.**	**Rec.**
$Latent\ Topic$	0.58	0.45	0.84
Sparse	0.59	0.83	0.46
Contiguous	0.62	0.85	0.49
BOW	0.52	0.73	0.40
Sparse+BOW	0.62	0.80	0.50
Cont.+BOW	0.63	0.81	0.52

at the semantic level rather than at the word level. Therefore, combining the $latent\ topic\ feature$ with syntactic features and lexical features should achieve a better performance than using the features separately. In our future work, we will devise an appropriate way of combining $latent\ topic\ features$ with syntactical and lexical features.

Because of the lack of a sufficiently annotated corpus for relation corpus for relation recognition, we have also proposed using a Web-based corpus to train classifiers for the purpose. Our experiment results demonstrates that Web documents can accurately capture information about the interaction between two named entities in the absence of an annotated corpus. By using a Web-based corpus, the time cost to manually annotating a corpus for relation recognition is expected to be significantly reduced if the quality of returned snippetscan be controlled.

References

David M. Blei and Michael I. Jordan. 2003. Modeling annotated data. In *Proceedings of the 26th SIGIR*, pages 127–134. ACM Press.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022.

Chih-Chung Chang and Chih-Jen Lin, 2001. *LIBSVM: a library for support vector machines*.

A. Culotta and J. Sorensen. 2004. Dependency tree kernels for relation extraction. *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics*.

Scott C. Deerwester, Susan T. Dumais, Thomas K. Landauer, George W. Furnas, and Richard A. Harshman. 1990. Indexing by latent semantic analysis. *Journal of the American Society of Information Science*, 41(6):391–407.

Thomas Hofmann. 1999. Probabilistic latent semantic analysis. In *Proc. of Uncertainty in Artificial Intelligence, UAI'99*, Stockholm.

Chien-Chung Huang, Shui-Lung Chuang, and Lee-Feng Chien. 2004. Liveclassifier: creating hierarchical text classifiers through web corpora. In *WWW '04: Proceedings of the 13th international conference on World Wide Web*, pages 184–192, New York, NY, USA.

Chen-Ming Hung and Leeq-Feng Chien. 2007. Web-based text classification in the absence of manually labeled training documents. *J. Am. Soc. Inf. Sci. Technol.*, 58(1):88–96.

Nanda Kambhatla. 2004. Combining lexical, syntactic, and semantic features with maximum entropy models for information extraction. In *The Companion Volume to the Proceedings of 42st Annual Meeting of the Association for Computational Linguistics*, pages 178–181, Barcelona, Spain, July. Association for Computational Linguistics.

Yutaka Matsuo, Junichiro Mori, Masahiro Hamasaki, Keisuke Ishida, Takuichi Nishimura, Hideaki Takeda, Koiti Hasida, and Mitsuru Ishizuka. 2006. Polyphonet: an advanced social network extraction system from the web. In *WWW '06: Proceedings of the 15th international conference on World Wide Web*, pages 397–406, New York, NY, USA.

Dmitry Zelenko, Chinatsu Aone, and Anthony Richardella. 2003. Kernel methods for relation extraction. *J. Mach. Learn. Res.*, 3:1083–1106.