A Combination of Topic Models with Max-margin Learning for Relation Detection

Dingcheng Li
University of Minnesota
Twin Cities, MN 55455
lixxx345@umn.edu

Swapna Somasundaran
Siemens Corporate Research
Princeton, NJ 08540
swapna.somasundaran@siemens.com

Amit Chakraborty
Siemens Corporate Research
Princeton, NJ 08540
amit.chakraborty@siemens.com

Abstract

This paper proposes a novel application of a supervised topic model to do entity relation detection (ERD). We adapt Maximum Entropy Discriminant Latent Dirichlet Allocation (MEDLDA) with mixed membership for relation detection. The ERD task is reformulated to fit into the topic modeling framework. Our approach combines the benefits of both, maximum-likelihood estimation (MLE) and max-margin estimation (MME), and the mixed membership formulation enables the system to incorporate heterogeneous features. We incorporate different features into the system and perform experiments on the ACE 2005 corpus. Our approach achieves better overall performance for precision, recall and Fmeasure metrics as compared to SVM-based and LLDA-based models.

1 Introduction

Entity relation detection (ERD) aims at finding relations between pairs of Named Entities (NEs) in text. Availability of annotated corpora (NIST, 2003; Doddington et al., 2004) and introduction of shared tasks (e.g. (Farkas et al., 2010; Carreras and Màrquez, 2005)) has spurred a large amount of research in this field in recent times. Researchers have used supervised and semi-supervised approaches (Hasegawa et al., 2004; Mintz et al., 2009; Jiang, 2009), and explored rich features (Kambhatla, 2004), kernel design (Culotta and Sorensen, 2004; Zhou et al., 2005; Bunescu and Mooney, 2005; Qian et al., 2008) and inference algorithms (Chan and Roth, 2011), to detect predefined relations between NEs.

In this work, we explore if and how the latent semantics of the text can help in detecting entity relations. For this, we adapt the Latent Dirichlet Allocation (LDA) approach to solve the ERD task. Specifically, we present a ERD system based on Maximum Entropy Discriminant Latent Dirichlet Allocation (MEDLDA). MEDLDA (Zhu et al., 2009), is an extension of Latent Dirichlet Allocation (LDA) that combines capability of capturing latent semantics with the discriminative capabilities of SVM.

There are a number of challenges in employing the LDA framework for ERD. Latent Dirichlet Allocation and its supervised extensions such as Labeled LDA (LLDA) (Ramage et al., 2009) and supervised LDA (sLDA) (Blei and McAuliffe, 2008) are powerful generative models that capture the underlying semantics of texts. However, they have trouble discovering marginal classes and easily employing rich feature sets, both of which are important for ERD. We overcome the first drawback by employing a MEDLDA framework, which integrates maximum likelihood estimation (MLE) and maximum margin estimation (MME). Specifically, it is a combination of sLDA and support vector machines (SVMs). Further, in order to employ rich and heterogeneous features we introduce a separate exponential family distribution for each feature, similar to (Shan et al., 2009), into our MEDLDA model.

We formulate the relation detection task within the topic model framework as follows. Pairs of NE mentions[1] and the text between them is considered

[1] Adopting the terminology used in the Automatic Context Extraction (ACE) program (NIST, 2003), specific NE instances are called mentions.

Proceedings of the TextGraphs-6 Workshop, pages 1–9,
Portland, Oregon, USA, 19-24 June 2011.

as *mini-document*. Each mini-document has a relation type (analogous to the response variable in the supervised topic model). The topic model infers the topic (relation type) distribution of the mini-documents. The supervised topic model discovers a latent topic representation of the mini-documents and a response parameter distribution. The topic representation is discovered with observed response variables during training. During testing, the topic distribution of each mini-document can form a prediction of the relation types.

We carry out experiments to measure the effectiveness of our approach and compare it to SVM-based and LLDA-based models, as well as to a previous work using the same corpora. We also measure and analyze the effectiveness of incorporating different features in our model relative to other models. Our approach exhibits better overall precision, recall and Fmeasure than baseline systems. We also find that the MEDLDA-based approach shows consistent capability for incorporation and improvement due to a variety of heterogeneous features.

The rest of the paper is organized as follows. We describe the proposed model in Section 2 and the features that we explore in this work in Section 3. Section 4 describes the data, experiments, results and analyses. We discuss the related work in Section 5 before concluding in Section 6.

2 MEDLDA for Relation Detection

MEDLDA is an extension of LDA proposed by Zhu, Ahmed and Xing (2009). LDA is itself unsupervised and the results are often hard to interpret. However, with the addition of supervised information (such as response variables), the resulting topic models have much better predictive power for classification and regression. In our work, we use relation annotations from the ACE (ACE, 2000 2005) corpus to provide the supervision. NE pairs within a sentence, and the text between them are considered as a mini-document. Each mini-document is assumed to be composed of a set of topics. The topic model trained with these mini-documents given their relation type label can generate topics biased toward relation types. Thus, the trained topic model will have good predictive power on relation types.

We first describe the MEDLDA model from (Zhu et al., 2009) and then describe how we adapt it for relation detection using mixed membership extensions.

2.1 MEDLDA

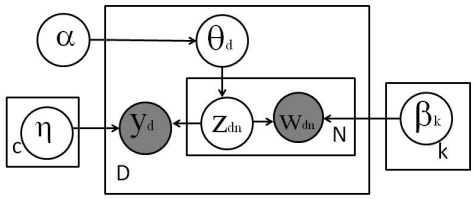

Figure 1: MEDLDA

The MEDLDA model described in (Zhu et al., 2009) is illustrated in Figure 1[2].

Here, α is a k-dimensional parameter of a Dirichlet distribution, $\beta_{1:k}$ are the parameters for k component distribution over the words. Each component refers to a topic. In a collection of documents D, each document $w_{1:N}$ is generated from a sequence of topics $z_{1:N}$. θ is a k-dimensional topic distribution variable, which is sampled from a Dirichlet distribution $Dir(\alpha)$. Like common LDAs, MEDLDA uses independence assumption for a finite set of random variables $z_1, ..., z_n$ which are independent and identically distributed, conditioned on the parameter θ. Like sLDA, MEDLDA is a supervised model. A response variable Y connected to each document is added for incorporating supervised side information. The supervised side information is expected to make MEDLDA topic discoveries more interpretable. Zhu, Ahmed and Xing's (2009) MEDLDA model can be used in both regression and classification. Concretely, Y is drawn from $\eta_{1:c}$, a c k-dimensional vector which can be derived from suitable statistical model. In our work, c is the number of relation types. Note that the plate diagram for MEDLDA is quite similar to sLDA (Blei and McAuliffe, 2008). But there is a difference – sLDA focuses on building regression models, and thus the response variable Y in sLDA is generated by a normal distribution.

Based on the plate diagram, the joint distribution of latent and observable variables for our MEDLDA-

[2](Zhu et al., 2009) do not have this plate digram in their paper; rather, we create this illustration from the description of their model.

based relation detection is given by

$$p(\theta, \mathbf{z}, \mathbf{w}, \mathbf{y}|\alpha, \beta_{1:k}, \eta_{1:c}) = \prod_{d=1}^{D} p(\theta_d|\alpha) \times \Big(\prod_{n=1}^{N} p(z_{dn}|\theta_d) p(w_{dn}|z_{dn}, \beta_{1:k}) \Big) \times p(y_d|z_{d1:dN}, \eta_{1:c}) \quad (1)$$

Another important difference from sLDA lies in the fact that MEDLDA does joint learning with both MME and MLE. The joint learning is done in two stages, unsupervised topic discovery and multi-class classification (we refer the reader to (Zhu et al., 2009) for details). During training, EM algorithms are utilized to infer the posterior distribution of the hidden variables θ, $\mathbf{z}$ and η. In testing, the trained models are used to predict relation types y.

2.2 Mixed Membership MEDLDA

Although the MEDLDA model described above can be applied to the relation detection and classification task, a few modifications are necessary before it can be effective in predicting relation types. Mainly, a limitation of LDA or other existing topic models is the difficulty in incorporating rich features. This is because LDA is designed to handle data points with homogeneous features such as words. But for relation detection, like many other NLP tasks, it is important to have the flexibility of incorporating part-of-speech tags, named entities, grammatical dependencies and other linguistic features. We overcome this limitation by introducing a separate exponential family distribution for each feature similar to (Shan et al., 2009). Thus, our MEDLDA-based relation detection model is really a mixed-member Bayesian network. Figure 2 illustrates our model with this extension.

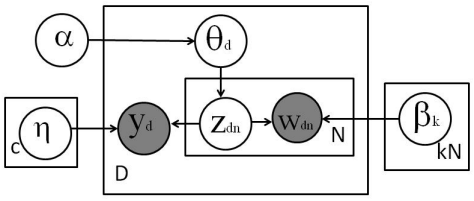

Figure 2: Mixed Membership MEDLDA

Figure 2 is very similar to Figure 1; the only difference is that the topic component number k is now kN. The generative process for each document this model is as follows:

1. Sample a component proportion $\theta_d \sim \mathbf{Dirichlet}(\alpha)$,
2. For each feature like word, part-of-speech, named entity in the document,
   (a) For $n \in \{1, ..., N\}$, sample $z_{dn} = i \sim \mathbf{Discrete}(\theta_d)$
   (b) For $n \in \{1,...,N\}$, sample $w_{dn} \sim P(w_{dn}|\beta^d_{ni})$
3. Sample the relation type label from a softmax($\bar{z}$,η) where $y_d \sim softmax(\frac{exp(\eta_h^T \bar{z})}{\sum_{h=1}^{c-1} exp(\eta_h^T \bar{z})})$

In the sampling, index i is the number of the topic component which ranges from $1:k$. $P(w_{dn}|\beta^d_{ni})$ in 2(b) is an exponential family distribution where i is from $1...k$. Note that now we have β^d_{ni} rather than only β^d_i since we have drawn separate distributions for each word (or feature) n.

Now, our MEDLDA-based relation-detection model can integrate diverse features of different types or the same features with different parameters.

Following the generative process, parameter estimation and inferences can be made with either Gibbs sampling or variational methods. We use variational methods since we adapt MEDLDA package[3] to mixed-membership MEDLDA and train relation detection models.

2.3 Relation Detection

With the generative process, inference and parameter estimation in place, we are ready to perform relation detection. The first step is to perform variational inference given the testing instances.

In classification, we estimate the probability of the relation type given topics and the response parameters, i.e. $p(y_d|z_{d1:dN}, \eta_{1:c-1})$. With variational approximation, we can derive the prediction rule as $F(y, z_{1:N}, \eta) = \eta^T f(y, \bar{z})$ where $f(y, \bar{z})$ is a feature vector. Now, SVM can be used to derive the

[3] this package is downloaded from http://www.cs.cmu.edu/~junzhu/medlda.htm

prediction rule. The final prediction can be generalized exactly the same as Zhu, Ahmed and Xing (Zhu et al., 2009):

$$\hat{y} = argmax_y E[\eta^T f(y, \bar{Z})|\alpha, \beta] \quad (2)$$

3 Features

We explore the effectiveness of incorporating features into our systems as well as the baselines. For this, we construct feature sets similar to Jiang and Zhai (2007) and Zhou (2005). Three kinds of features are employed:

1. **BOW** The Bag of Words (BOW) feature captures all the words in our mini-document. It comprises of the words of the two NE mentions and the words between them.

2. **SYN** The SYN features are constructed to capture syntactic, semantic and structural information of the mini-document. They include features such as HM1 (the head word of the first mention), HM2 (the head word of the second mention), ET1, ET2, M1 and M2 (Entity types and mention types of the two mentions involved), #MB (number of other mentions in between the two mentions), #WB (number of words in between the two mentions).

3. **COMP** The COMP features are composite features that are similar to SYN, but they additionally capture language order and dependencies between the features mentioned above. These include features such as HM1HM2 (combining head word of mention 1 and head word of mention 2) , ET12 (combinations of mention entity type), ML12 (combination of mention levels), M1InM2 or M2InM1 (flag indicating whether M2/M1 is included in M1/M2).

The main intuitions behind employing composite features, COMP, are as follows. First, they capture the ordering information. The ordering of words are not captured by BOW. That is, BOW features assume exchangeability. This works for models based on random or seeded sampling (e.g. LDA) – as long as words sampled are associated with a topic, the hidden topics of the documents can be discovered. In the case of ERD, this assumption might work with symmetric relations. However, when the relations are asymmetric, ordering information is important. Composite features such as HM1HM2 encodes what mention head word precedes the other. Second, features such as M1InM2 or M2InM1 capture token dependencies. Besides exchangeability, LDA-based models also assume that words are conditionally independent. Consequently, the system cannot capture the knowledge that some mentions may be included in other mentions. By constructing features such as M1InM2 or M2InM1, we encode the dependency information explicitly.

4 Experiments

As MEDLDA is a combination of maximum margin principle with maximum likelihood estimation for topic modes, we compare it with two baseline systems. The first, *SVM*, uses only the maximum margin principle, while the second, *LLDA*, uses only maximum likelihood estimation for topic modeling.

4.1 Data

We use the ACE corpus (Phase 2, 2005) for evaluation. The ACE corpus has annotations for both entities and relations. The corpus has six major relations types, 23 subtypes and 7 entity types. In this work, we focus only on the six high-level relation types listed in Table 1. In addition to the the 6 major types, we have an additional category, no relation (NO-REL), that exists between entities that are not related.

The data for our experiments consists of pairs of NEs from a sentence, and the gold standard annotation of their relation type (or NO-REL). All relations in the ACE corpus are intra-sentential and hence we do not create NE pairs that cross sentence boundaries. Also, almost all positive instances are within two mentions of each other. Hence, we create NE pairs for only those NEs that have at most 2 intervening NEs in between. This gives us a total of 38,342 relation instances of which 32,640 are negative instances (NO-REL) and 5702 are positive relation instances belonging to one of the 6 categories.

4.2 Experimental Setup

We use 80% of the instances for training and 20% for testing. The topic numbers and the penalty parameter of the cost function C are first determined

Major Type	Definition	Example
ART artifact	User, owner, inventor or manufacturer	the makers of the Kursk
GEN-AFF	citizen, resident, religion, ethnicity and organization-location	U.S. Companies
ORG-AFF (Org-affiliation)	employment, founder, ownership, sports-affiliation, investor-shareholder student-alumni and membership	The CEO of Siemens
PART-WHOLE	geographical, subsidiary and so on	a branch of U.S bank
PER-SOC (person-social)	business, family and lasting personal relationship	a spokesman for the senator
PHYS (physical)	located or near	a military base in Germany

Table 1: Relation types for ACE 05 corpus

for each of the models (wherever applicable) using the training data. Best parameters are determined for the three conditions: 1) BOW features alone *BOW*, 2) BOW plus SYN features (*PlusSYN*) and 3) BOW plus SYN and COMP features (*PlusCOMP*). All systems achieved their overall best performance with PlusCOMP features (see Section 4.4 for a detailed analysis).

4.2.1 MEDLDA

The number of topics are determined using the equation $2K_0 + K_1$ following Zhu, Ahmed and Xing (2009) and $K_1 = 2K_0$. K_0 is the number of topics per class and K_1 is the number of topics shared by all relation types. The choice of topics is based on the intuition that the shared component K_1 should use all class labels to model common latent structure while non-overlapping components should model specific characteristics data from each class. The ratio of topics is based on the understanding that shared topics may be more than topics of each class. The specific numbers do not produce much variation in the final results. We experimented with the following number of topics: 20, 40, 70, 80, 90, 100, 110. BOW, PlusSYN, and PlusCOMP configurations obtain the best performance for 90 topics, 80 topics, and 70 topics respectively.

Since SVMs are employed in the MEDLDA implementation, we need to determine the penalty parameter of the cost function, C. We used 5 fold cross-validation to locate the parameter C. The best values for C are 25, 28, 30 respectively for BOW, PlusSYN and PlusCOMP configurations. We used a linear kernel as it is the most commonly used kernel for text classification tasks. Since MEDLDA is run by sampling, the result may be different each time. We ran it 5 times for each setting and took the average as the final results.

4.2.2 LLDA and SVM

The setting of topics for LLDA is similar to MEDLDA. As LLDA is also run by sampling, we ran it 5 times for each setting and took the average as the final results. In SVMlight, a grid search tool is provided to locate the the best value for parameter C. The best C for all three conditions was found to be 1. All other settings for the two models are similar to those of MEDLDA.

4.3 Results

	Prec%	Rec%	F%
SVM	53.2	35.2	40.3
LLDA	28.3	51.6	36.6
MEDLDA	**57.8**	**53.2**	**55.4**

Table 2: Overall performance of the 3 systems

We present the results of the three systems built using PlusCOMP, as all systems achieved their best overall performance using these features. Table 2 reports the precision, recall and Fmeasure of the three systems averaged across all 7 categories (the best numbers for each metric are highlighted in **bold**). Here we see that MEDLDA outperforms LLDA and

Labels	SVM			LLDA			MEDLDA		
	Pre%	Rec%	F%	Pre%	Rec%	F%	Pre%	Rec%	F%
ART	30	8	14	1.5	33	3	**49**	**36**	**41**
GEN-AFF	**53**	**48**	**50**	3	32	6	40	39	40
ORG-AFF	55	35	43	**59**	58	**59**	53	**59**	56
PART-WHOLE	39	08	14	31	**82**	45	**44**	52	**48**
PER-SOC	50	17	25	7	**92**	13	**73**	76	**75**
PHYS	55	35	**43**	26	**47**	33	**56**	19	29
NO-REL	**90**	**95**	**93**	70	17	27	89	91	90

Table 3: Multi-class Classification Results with PlusCOMP for SVM, LLDA and MEDLDA for the six ACE 05 categories and NO-REL

SVM across all metrics. Specifically, there is a 15 percentage point improvement in Fmeasure over the best performing baseline. This result indicates that our approach of combining topic model with max-margin learning is effective for relation detection.

Now, looking at the results for each individual relationship category (see Table 3; the best numbers for each category and metric are highlighted in **bold**) we see that the Fmeasure for MEDLDA is better than that for SVM for 4 out of the 6 ACE relation types; and better than the Fmeasure obtained by LLDA for all relation types except ORG-AFF. Specifically, comparing with the best performing baseline, MEDLDA produces a Fmeasure improvement 27 percentage points for ART, 3 percentage points for PART-WHOLE and 50 percentage points for PER-SOC. Also, for four of the six ACE relation types, MEDLDA achieves the best precision. Even in the cases where MEDLDA is not the best performer for a relation category, its performance is not very poor (unlike, for example, SVM for PART-WHOLE and LLDA for ART, respectively).

Interestingly, the NO-REL category reveals a sharp contrast in the performance of SVM and LLDA. NO-REL is a difficult, catch-all category that is a mixture of data with diverse distributions. This is a category where maximum-margin learning is more effective than maximum-likelihood estimation. Notice that MEDLDA achieves performance close to SVM for this category. This is because, even though both LLDA and MEDLDA model hidden topics and then employ discovered hidden topics to predict relation types, MEDLDA does joint inference of MLE and MME. This joint inference helps to improve the detection of NO-REL.

Finally, we also compare our system's results (using PlusCOMP features) with the results of previous research on the same corpus (Khayyamian et al., 2009). They use similar experimental settings: every pair of entities within a sentence is regarded to involve a negative relation instance unless it is annotated as positive in the corpus. A similar filter (they use a distance filter) is used to sift out unrelated negative instances. Their train/test ratio of data split is also the same as ours.

Khayyamian, Mirroshandel and Abolhassani (2009) employ state-of-art kernel methods developed by Collins and Duffy (2002) and only report Fmeasures over the six ACE relation types. For clarity, we reproduce their results in Table 4 and repeat MEDLDA Fmeasures from Table 3 in the last column. The last row (Overall) reports the macro-averages computed over all relation types for each system. Here we see that overall, MEDLDA outperforms all kernels. MEDLDA also performs better than the best kernel for four of the six relation types.

4.4 Analysis

As mentioned previously, all three systems achieved their overall best performance with PlusCOMP features. Here, we analyze if informative features are consistently useful and if the systems can harness the informative features consistently across all relation types. Figures 3, 4 and 5 illustrate the F-measures for SVM, LLDA and MEDLDA respectively for the three conditions: BOW, PlusSYN and PlusCOMP.

Labels	CD'01	AAP	AAPD	TSAAPD-0	TSAAPD-01	MEDLDA
ART%	**51**	49	50	48	47	41
GEN-AFF %	9	10	12	11	11	**40**
ORG-AFF %	43	43	43	43	45	**56**
PART-WHOLE %	30	28	29	30	28	**48**
PER-SOC %	62	58	70	63	73	**75**
PHYS %	32	**36**	29	33	33	29
Overall (Avg)	38	37	39	38	40	**48**

Table 4: F-measures for every kernel in (Khayyamian et al., 2009) and MEDLDA

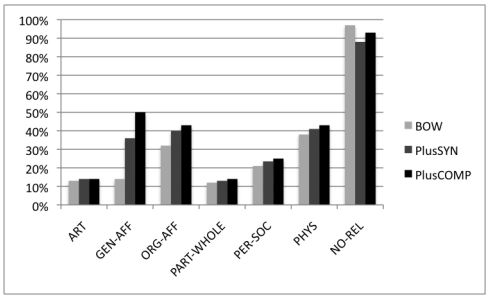

Figure 3: SVM Fmeausres for 3 feature conditions

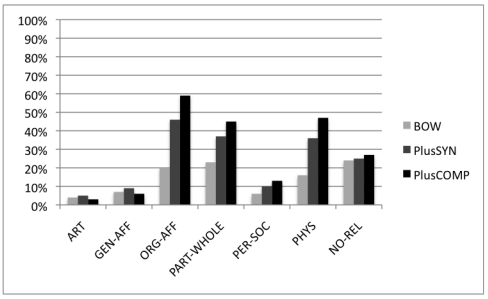

Figure 4: LLDA Fmeausres for 3 feature conditions

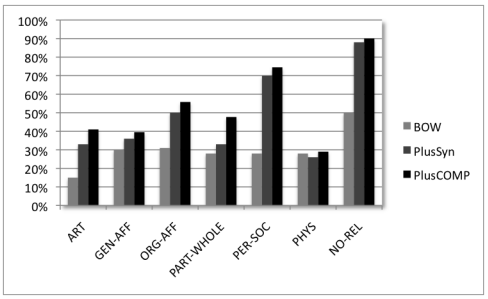

Figure 5: MEDLDA Fmeausres for 3 feature conditions

Let us first look at the best systems (based on Fmeasure) for each of the six ACE relation types in Table 3, and look at what feature set produces the best result for that system and relation. MEDLDA is the best performer for ART, PART-WHOLE and PER-SOC in Table 3. Figure 5 reveals that MEDLDA's best performance for these relation types are obtained using PlusCOMP features. Similarly SVM obtains the best Fmeasure for GEN-AFF and PHYS relations and Figure 3 shows that SVM achieves its best performance for these categories using PlusCOMP. We also see a similar trend with LLDA and the ORG-AFF relation type. These results corroborate intuition from previous research that informative features are important for relation type recognition. The only exception to this is the performance of SVM for NO-REL. This is not surprising, as the features we use are focused on determining true relation types and NO-REL is a mixture of all cases (and features) where relations do not exist.

Further analysis of the figures reveal that even though there is a general trend towards better performance with addition of more informative features, not all systems show consistent improvements across all relation types with the addition of composite features. That is, some systems get degraded performance due to feature addition. For example, in Figure 3, we see that the SVM with PlusCOMP features is outperformed by SVM with PlusSYN for ART and SVM with BOW for NO-REL. The gains from features are also inconsistent in the case of LLDA (Figure 4). While the LLDA system with PlusSYN features always improves over the one using BOW, the performance drops considerably when using PlusCOMP features for ART and GEN-AFF. On the other hand, MEDLDA (see Figure 5) shows more consistent improvement for all relation types with the addition of more complex features. Also,

the gains are more substantial. This is encouraging and opens up avenues for further exploration.

5 Related Work

Previous research has explored various methods and features for relationship detection and mining. Kernel methods have been popularly used for relation detection. Some examples are are dependency tree kernels (Culotta and Sorensen, 2004), shortest dependency path kernels (Bunescu and Mooney, 2005), and more recently, convolution tree kernels (Zhao and Grishman, 2005; Zhang et al., 2006) context-sensitive convolution tree kernels (Zhou et al., 2007) and dynamic syntax tree kernels (Qian et al., 2008). Kernel methods for relation extraction focus on representing and capturing the structured information of the text between the entities. In our MEDLDA model, instead of computing distances between subtrees, we sample topics based on their distributions. The sampling is not only on the (mini) document level, but also on the word level or on the syntactic or semantic level. Our model focuses on addressing the underlying semantics more directly than typical kernel-based methods.

Chan and Roth (2011) employ constraints using an integer linear programming (ILP) framework. Using this, they apply rich linguistic and knowledge-based constraints based on coreference annotations, a hierarchy of relations, syntacto-semantic structure, and knowledge from Wikipedia. In our work, we focus on capturing the latent semantics of the text between the NEs.

A variety of features have been explored for ERD in previous research (Zhou et al., 2005; Zhou et al., 2008; Jiang and Zhai, 2007; Miller et al., 2000). Syntactic features such as POS tags and dependency path between entities; semantic features such as Word-Net relations, semantic parse trees and types of NEs; and structural features such as which entity came first in the sentence have been found useful for ERD. We too observe the utility of informative features for this task. However, exploration of the feature space is not the main focus of this work. Rather, our focus is on whether the models are capable of incorporating rich features. A fuller exploration of rich heterogeneous features is the focus of our future work.

A closely related task is that of relation mining and discovery, where unsupervised, semi-supervised approaches have been effectively employed (Hasegawa et al., 2004; Mintz et al., 2009; Jiang, 2009). For example, Hasegawa et al. (2004) use clustering and entity type information, while Mintz et al. (2009) employ distant supervision. Our ERD task is different from these as we focus on classifying the relation types into predefined relation types in the ACE05 corpus.

Topic models have been applied previously for a number of NLP tasks (e.g. (Lin et al., 2006; Titov and McDonald, 2008). LDAs have also been employed to reduce feature dimensions in relation detection systems (Hachey, 2006). However, to the best of our knowledge, this is the first work to make use of topic models to perform relation detection.

6 Conclusion and Future Work

In this work, we presented a system for entity relation detection based on mixed-membership MEDLDA. Our approach was motivated by the idea that combination of max margin and maximum likelihood can help to improve relation detection task. For this, we adapted the existing work on MEDLDA and mixed membership models and formulated ERD as a topic detection task. To the best of our knowledge, this is the first work to make full use of topic models for relation detection.

Our experiments show that the proposed approach achieves better overall performance than SVM-based and LLDA-based approaches across all metrics. We also experimented with different features and the effectiveness of the different models for harnessing these features. Our analysis show that our MEDLDA-based approach is able to effectively and consistently incorporate informative features.

As a model that incorporates maximum-likelihood, maximum-margin and mixed membership learning, MEDLDA has the potential of incorporating rich kernel functions or conditional topic random fields (CTRF) (Zhu and Xing, 2010). These are some of the promising directions for our future exploration.

References

ACE. 2000-2005. Automatic Content Extraction. *http://www.ldc.upenn.edu/Projects/ACE/*.

D.M. Blei and J. McAuliffe. 2008. Supervised topic models. *Advances in Neural Information Processing Systems*, 20:121–128.

R.C. Bunescu and R.J. Mooney. 2005. A shortest path dependency kernel for relation extraction. In *HLT & EMNLP*.

X. Carreras and L. Màrquez. 2005. Introduction to the CoNLL-2005 shared task: Semantic role labeling. In *CONLL*, pages 152–164. ACL.

Y. Chan and D. Roth. 2011. Exploiting syntactico-semantic structures for relation extraction. In *ACL*.

M. Collins and N. Duffy. 2002. Convolution kernels for natural language. *Advances in neural information processing systems*, 1:625–632.

A. Culotta and J. Sorensen. 2004. Dependency tree kernels for relation extraction. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, page 423. ACL.

G. Doddington, A. Mitchell, M. Przybocki, L. Ramshaw, S. Strassel, and R. Weischedel. 2004. The automatic content extraction (ACE) program–tasks, data, and evaluation. In *Proceedings of LREC*, volume 4, pages 837–840.

R. Farkas, V. Vincze, G. Móra, J. Csirik, and G. Szarvas. 2010. The CoNLL-2010 Shared Task: Learning to Detect Hedges and their Scope in Natural Language Text. In *CoNLL-2010*, pages 1–12.

B. Hachey. 2006. Comparison of similarity models for the relation discovery task. In *COLING & ACL 2006*, page 25.

T Hasegawa, S Sekine, and Ralph Grishman. 2004. Discovering relations among named entities from large corpora. In *42nd ACL*.

J. Jiang and C.X. Zhai. 2007. A systematic exploration of the feature space for relation extraction. In *NAACL/HLT*, pages 113–120.

J. Jiang. 2009. Multi-task transfer learning for weakly-supervised relation extraction. In *47th ACL & 4th AFNLP*, pages 1012–1020. ACL.

N. Kambhatla. 2004. Combining lexical, syntactic, and semantic features with maximum entropy models for extracting relations. In *ACL 2004 Interactive poster and demonstration sessions*.

M. Khayyamian, S.A. Mirroshandel, and H. Abolhassani. 2009. Syntactic tree-based relation extraction using a generalization of Collins and Duffy convolution tree kernel. In *HLT/NAACL,: Student Research Workshop*.

Wei-Hao Lin, Theresa Wilson, Janyce Wiebe, and Alexander Hauptmann. 2006. Which side are you on? Identifying perspectives at the document and sentence levels. In *CoNLL-2006*.

S. Miller, H. Fox, L. Ramshaw, and R. Weischedel. 2000. A novel use of statistical parsing to extract information from text. In *NAACL*.

M Mintz, S Bills, R Snow, and D Jurafsky. 2009. Distant supervision for relation extraction without labeled data. In *47th ACL & 4th AFNLP*.

US NIST. 2003. The ACE 2003 Evaluation Plan. *US National Institute for Standards and Technology (NIST)*, pages 2003–08.

L. Qian, G. Zhou, F. Kong, Q. Zhu, and P. Qian. 2008. Exploiting constituent dependencies for tree kernel-based semantic relation extraction. In *22nd ACL*.

D. Ramage, D. Hall, R. Nallapati, and C.D. Manning. 2009. Labeled LDA: A supervised topic model for credit attribution in multi-labeled corpora. In *EMNLP*.

H. Shan, A. Banerjee, and N.C. Oza. 2009. Discriminative Mixed-membership Models. In *ICDM*, pages 466–475. IEEE.

Ivan Titov and Ryan McDonald. 2008. A joint model of text and aspect ratings for sentiment summarization. In *ACL-08: HLT*.

M. Zhang, J. Zhang, J. Su, and G. Zhou. 2006. A composite kernel to extract relations between entities with both flat and structured features. In *21st ICCL & 44th ACL*.

S. Zhao and R. Grishman. 2005. Extracting relations with integrated information using kernel methods. In *43rd ACL*.

G Zhou, S. Jian, Z. Jie, and Z. Min. 2005. Exploring various knowledge in relation extraction. In *In 43rd ACL*.

G Zhou, M. Zhang, D.H. Ji, and Q Zhu. 2007. Tree kernel-based relation extraction with context-sensitive structured parse tree information. In *EMNLP/CoNLL-2007*, pages 728–736.

G.D. Zhou, M. Zhang, D.H. Ji, and Q.M. Zhu. 2008. Hierarchical learning strategy in semantic relation extraction. *Information Processing & Management*, 44(3):1008–1021.

J. Zhu and E.P. Xing. 2010. Conditional Topic Random Fields. In *ICML*. ACM.

J. Zhu, A. Ahmed, and E.P. Xing. 2009. MedLDA: maximum margin supervised topic models for regression and classification. In *ICML*, pages 1257–1264. ACM.