Natural Language Engineering **18** (2): 263–289. © Cambridge University Press 2012
doi:10.1017/S1351324912000058

ERD-MedLDA: Entity relation detection using supervised topic models with maximum margin learning

DINGCHENG LI[1], SWAPNA SOMASUNDARAN[2]
and AMIT CHAKRABORTY[2]

[1]*Liberal Arts–TC, University of Minnesota, Twin Cities, MN 55455, USA*
email: lixxx345@umn.edu

[2]*Siemens Corporate Research, Princeton, NJ 08540, USA*
email: swapna.somasundaran@siemens.com

(*Received 28 August 2011; accepted 30 December 2011*)

Abstract

This paper proposes a novel application of topic models to do entity relation detection (ERD). In order to make use of the latent semantics of text, we formulate the task of relation detection as a topic modeling problem. The motivation is to find underlying topics that are indicative of relations between named entities (NEs). Our approach considers pairs of NEs and features associated with them as mini documents, and aims to utilize the underlying topic distributions as indicators for the types of relations that may exist between the NE pair. Our system, ERD-MedLDA, adapts Maximum Entropy Discriminant Latent Dirichlet Allocation (MedLDA) with mixed membership for relation detection. By using supervision, ERD-MedLDA is able to learn topic distributions indicative of relation types. Further, ERD-MedLDA is a topic model that combines the benefits of both, maximum likelihood estimation (MLE) and maximum 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 F-measure metrics as compared to baseline SVM-based and LDA-based models. We also find that our system shows better and consistent improvements with the addition of complex informative features as compared to baseline systems.

1 Introduction

The entity relation detection (ERD) task aims at finding relationships between pairs of named entities (NEs) in text. NEs such as *persons* may be related to *organization* entities via *organization affiliation* relation (‘*John* is the founder of *Xyz Corp.*’). Organization entities may be related to *location* entities via *physical location* relation (‘*Xyz Corp* is located in *North Carolina*’) and persons may be related to one another via *social* relations (‘*John* and *Mary* were married last year’).

Availability of annotated corpora (Doddington *et al.* 2004) and introduction of shared tasks (e.g., Carreras and Màrquez 2005; Farkas *et al.* 2010) have spurred a large amount of research in this field in recent times. Researchers have used supervised and semi-supervised approaches (Hasegawa, Sekine and Grishman 2004; Jiang 2009; Mintz *et al.* 2009), and explored rich features (Kambhatla 2004), kernel design (Culotta and Sorensen 2004; Bunescu and Mooney 2005; Zhou *et al.* 2005; Qian *et al.* 2008), and joint inference (Chan and Roth 2011), to detect predefined relations between NEs.

In this work, we explore how the latent semantics of the text can help in detecting entity relations. Specifically, we are interested in the underlying topic distributions. Intuitively, topics such as marriage or birth could be indicative of a social relationship between the participating entities. Similarly, topics such as nationality or recruitment could be indicative of an affiliation relation between the participating entities. Thus in this paper, we investigate if hidden topic distributions indicative of entity relations can be effectively learned.

In order to achieve this, we adapt the Latent Dirichlet Allocation (LDA) approach to solve the ERD task. There are a number of challenges in employing the LDA framework for ERD. Primarily, the basic LDA model is unsupervised, and hence the discovered topics may not help classification. This problem has been solved in supervised models such as labeled LDA (LLDA) (Ramage *et al.* 2009) and supervised LDA (sLDA) (Blei and McAuliffe 2008). While LLDA and sLDA are powerful generative models that capture the underlying semantics of texts pertinent to classes of interest, we found during our preliminary experiments that they have trouble discovering marginal classes in ERD and do not naturally incorporate rich feature sets.

In order to incorporate the capabilities desirable for relation detection, we build our ERD system, *ERD-MedLDA*, based on Maximum Entropy Discriminant Latent Dirichlet Allocation (MedLDA) from Zhu, Ahmed and Xing (2009). MedLDA is a supervised extension of LDA that combines the capability of capturing latent semantics with maximum margin learning. Specifically, MedLDA is a combination of sLDA and support vector machines (SVMs); thus, it integrates maximum likelihood estimation (MLE) and maximum margin estimation (MME). Further, in order to employ rich and heterogeneous features we introduce a separate exponential family distribution for each feature, similar to Shan, Banerjee and Oza (2009), into our ERD-MedLDA model.

The relation detection task is formulated within the topic model framework in ERD-MedLDA as follows. Occurrences of pairs of NE mentions[1] in a document and the text between them is considered as a *mini-document*. Each mini-document has a relation type (analogous to the response variable in a supervised topic model). 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, which influences topic discovery towards the response variables. During prediction, the topic distribution of each mini-document can form a prediction of the relation types.

[1] Adopting terminology used in the Automatic Context Extraction (ACE) program (ACE 2000–2005), specific NE instances are called *Mentions*.

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 corpus. We also analyze the discovered topics and measure the effectiveness of incorporating different features in our model relative to other models.

Our approach exhibits better overall precision, recall, and F-measure than baseline systems. We also find that ERD-MedLDA 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 Section 3 describes the data and the task formulation. The features that we incorporate are explained in Section 4. Section 5 presents the experiments and Section 6 presents analyses. We discuss the related work in Section 7 before concluding in Section 8.

2 ERD-MedLDA

ERD-MedLDA is based on the principle of hierarchical Bayesian models. The basic LDA is an unsupervised model and the resulting topics may not help with classification tasks. However, with the explicit addition of supervised information (such as response variables), the resulting topic models have good predictive power for classification and regression. MedLDA model from Zhu *et al.* (2009) is one such extension of LDA, where the class information is added to the model. Further, MedLDA integrates maximum margin learning and topic models by optimizing a single objective function with a set of expected margin constraints.

In this section, we explain the development of ERD-MedLDA as follows: We first explain the MedLDA model from Zhu *et al.* (2009) and our modifications to it in Section 2.1. Section 2.2 describes further modifications that allow for the incorporation of heterogeneous features, Section 2.3 describes inference and estimation procedures, and, finally, Section 2.4 describes how relation detection is performed under the model.

The following notation is used in this paper:

K – The total number of topics
N – The total number of words and features in a document
C – The total number of relation types
$\alpha_{1:K}$ – K-dimensional parameter of a Dirichlet distribution
$\beta_{1:K}$ – Parameters for K component distribution over the words
$\theta_{1:K}$ – K-dimensional parameters of topic distribution variables over a
– document
$\eta_{1:C}$ – Parameters for C component distribution over the relation types
$z_{d1:dN}$ – Variables representing a sequence of topics in a specific document, d
$w_{d1:dN}$ – A finite set of observed variables that represent a specific document, d
$y_{1:C}$ – A finite set of observed variables which represents relation types
D – A collection of documents denoted by $D = \mathbf{w_1}, \mathbf{w_2}, ..., \mathbf{w_D}$

For brevity, we use bold case symbols for vectors and drop the subscript where dimensionality is unambiguous.

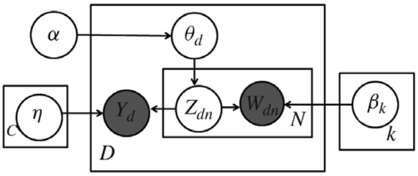

Fig. 1. MedLDA.

2.1 MedLDA

The MedLDA model described in Zhu *et al.* (2009) is illustrated in Figure 1. Here, $\boldsymbol{\theta}$ is a K-dimensional topic distribution variable, which is sampled from a Dirichlet distribution $\mathbf{Dirichlet}(\boldsymbol{\alpha})$. Like common LDAs, MedLDA uses independence assumption for a finite set of random variables $Z_1,\ldots,Z_n$ which are independent and identically distributed, conditioned on the parameter, θ. Like its predecessor, 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 useful for classification tasks. 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 $C \times K$-dimensional vector and the topic distribution $z_{1:N}$. 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. In regression, similar to sLDA, a normal distribution is used for generating Y, while in classification, MedLDA uses the maximum-margin principle to directly generate Y.

Based on the plate diagram, the joint distribution of latent and observable variables for our MedLDA-based relation detection is given by:

$$\begin{aligned} p(\boldsymbol{\theta},\mathbf{z},\mathbf{w},\mathbf{y}|\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\eta}) = \prod_{d=1}^{D} p(\theta_d|\boldsymbol{\alpha}) \times \left(\prod_{n=1}^{N} p(z_{dn}|\theta_d)p(w_{dn}|z_{dn},\boldsymbol{\beta})\right) \\ \times\, p(y_d|z_{d1:dN},\boldsymbol{\eta}) \end{aligned} \tag{1}$$

Another important difference from sLDA lies in the fact that MedLDA does joint learning with both MME and MLE. The joint learning for classification is done in two stages, unsupervised topic discovery and multi-class classification. During training, EM algorithms are utilized to infer the posterior distribution of the hidden variables $\boldsymbol{\theta}$, $\mathbf{z}$, and $\boldsymbol{\eta}$. During testing, the trained models are used to predict relation types $\mathbf{y}$.

2.2 Fine mixed-membership MedLDA

MedLDA is already a mixed-membership model and although the MedLDA model described above can be applied to detection and classification tasks, we felt a few modifications were necessary before it can be effective in predicting relation types. In particular, MedLDA is designed for using a homogeneous component distribution and we required it to use heterogeneous features.

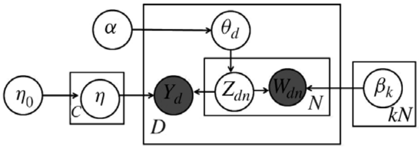

Fig. 2. Fine mixed-membership MedLDA.

As we can see from the plate model in Figure 1, each of $w_{1:N}$ is assumed to be generated from one of the discrete component distributions. In each document, bag of words are the same type of objects. The set of distributions remain the same across all words. Thus the original MedLDA is designed to handle data points with homogeneous features such as words. But previous work in relation detection has shown that it is important to incorporate part-of-speech tags, NEs, grammatical dependencies, and other linguistic features. We achieve this by introducing a separate exponential family distribution for each feature similar to (Shan *et al.* 2009). Thus, our relation detection model is really a mixed-member Bayesian network. Figure 2 illustrates our model with this extension.

Figure 2 is very similar to Figure 1. There are two differences: First, the topic component number k is now kN and second, there is another component η_0 before the response parameter η. We made the first modification in order to incorporate heterogeneous features. Note that now we have β_{ni}^{d} rather than only β_i^d since we have drawn separate distributions for each word (or feature) n.

In MedLDA, Zhu *et al.* (2009) have in fact introduced the idea that $\boldsymbol{\eta}$ is sampled from a prior $p_0(\boldsymbol{\eta})$. We too follow the same idea and show it as a hyper-parameter in Figure 2 for clarity. Like MedLDA, we assume that there are C classes and K topics. In our work, C is the number of relation types. However, unlike MedLDA, the response parameter $\eta_{1:C}$ in our model is a matrix with $C \times K$-dimensional *softmax* parameters as rows.

The generative process for each document in this model is similar to that given in (Shan *et al.* 2009) and is as follows:

(1) Sample a component proportion $\theta_d \sim \mathbf{Dirichlet}(\boldsymbol{\alpha})$,

(2) For each feature like word, part-of-speech, NE in the document,

   (a) For $n \in \{1,\dots,N\}$, sample $z_{dn} = i \sim \mathbf{Discrete}(\theta_d)$

   (b) For $n \in \{1,\dots,N\}$, sample $w_{dn} \sim P(w_{dn}|\beta_{ni}^{d}, z_{dn})$

(3) Sample the relation type label for a document from a *softmax* distribution, two steps are involved,

   (a) For $j \in \{1,\dots,C\}$, sample $\eta_j \sim \mathbf{Norm}(0,\eta_0)$

   (b) For $j \in \{1,\dots,C\}$, $y_d \sim softmax\left(\frac{\exp(\eta_j^T \bar{z})}{\sum_{j=1}^{C}\exp(\eta_j^T \bar{z})}\right)$

In step 2, index i is the number of the topic component which ranges from $1:K$. $P(w_{dn}|\beta_{ni}^{d}, z_{dn})$ in 2(b) is an exponential family distribution.

Like the original MedLDA, we make use of a prior $p_0(\eta)$ where $p_0 = N(0,I)$ to sample the response parameter η. That is, $\boldsymbol{\eta}$ is taken as a variable rather than a hyper-parameter like $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$. η_0 is the hyper-parameter for $\boldsymbol{\eta}$ and it is sampled

C times. Namely, for each relation type, a vector of response parameters will be sampled and they will in turn be used to sample relation types.

For the sampling in step 3(b), a *softmax* distribution is employed. *Softmax* is a variation of logistic regression. Logistic regression is usually used for binary classification while *softmax* is for multi-class classification. *Softmax* is a mature statistical method for classifications, easy to compute, and is appropriate for handling missing features. The input for the *softmax* distribution is $\bar{z}$, namely, the mean of $\mathbf{z}$ for all words/features.

Note that we do not directly use maximum margin principle for classification. This is another modification we make in the model. The original MedLDA's model for classification comprises of two separate parts – the first part is an unsupervised LDA that does MLE for modeling topics and the second part is an SVM that does MME for classification. Zhu, Ahmed and Xing's motivation of doing so is that calculations of partition factors or normalization factors are too hard and too slow. However, we believe that this separation cannot make full use of the advantages of the integration of MME and MLE. In the learning process, topic discoveries biased by supervised side information should be more helpful than topics discovered or learned in unsupervised fashion. Though the learning of normalization factor for classification will be slower, it is not intractable, and speed-ups can be achieved by selecting suitable sampling methods and by good optimizations and approximations.

Thus, a joint distribution can be written as:

$$p(\boldsymbol{\theta}, \mathbf{z}, \boldsymbol{\eta}, \mathbf{w}, \mathbf{y}|\boldsymbol{\alpha}, \boldsymbol{\beta}) = p_0(\boldsymbol{\eta})p(\boldsymbol{\theta}, \mathbf{z}, \mathbf{w}, \mathbf{y}|\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\eta}) \tag{2}$$

The second term is the same as that defined in (1). The first term, the prior distribution of $\boldsymbol{\eta}$, is defined as a normal distribution.

The density function for $\mathbf{w}, \mathbf{y}$, after $\boldsymbol{\theta}$ and $\mathbf{z}$ are integrated out, is given by:

$$\begin{aligned} & p(\mathbf{w}, \mathbf{y}|\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\eta}) \\ & = \prod_{d=1}^{D} \int_{\theta_d} p_0(\boldsymbol{\eta})p(\theta_d|\boldsymbol{\alpha})\Big(\prod_{n=1}^{N}\sum_{k=1}^{K}\big(p(z_{dn=k}|\theta_d)p(w_{dn}|\beta_{nk}, z_{dn} = k)\big)p(y_d|\mathbf{z}_d, \boldsymbol{\eta})\Big)d\theta_d \end{aligned} \tag{3}$$

Since relation detection only involves token features (even though these features are heterogeneous in nature, ranging from bag of words, parts-of-speech, chunk types, and so on), they are discrete symbols. Hence, before we find continuous features, such as numerical values, all features have discrete distributions.

With this extension, the distribution for generating w_{dn} not only depends on z_{dn}, but also on what kind of features are employed. Therefore, by choosing an appropriate exponential family distribution for each feature (in our relation detection, all features involve discrete distributions with diverse parameters), our ERD-MedLDA model can integrate diverse features of different types or the same features with different parameters.

2.3 Inference and estimation

In (Zhu *et al.* 2009), MedLDA integrates a Bayesian sLDA and a SVM for both MLE and MME. In our work, we follow the same overall approach. However, as shown in our generative process, instead of only using unsupervised LDA for topic modeling, we use fully supervised MedLDA for both topic modeling and the final classification. Yet, the inference is slower because of the normalization factor in the probability distribution of Y and also because we draw different distributions for each word. To remedy this, we make use of a strategy similar to Shan *et al.* (2009). Namely, for features of the same type, we sample them with the same exponential distributions using parameters averaged from the training data, which makes the learning much faster and efficient.

In our model, the learning task is to obtain an optimal set of parameters $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, and $p(\boldsymbol{\eta})$ such that the likelihood of observing the whole feature set and the relation types, $p(\mathbf{w}, \mathbf{y}|\boldsymbol{\alpha}, \boldsymbol{\beta}, p(\boldsymbol{\eta}))$, are maximized. However, the calculation of the likelihood function (3) is intractable. Following the generative process, parameter estimation and inferences can be made with either Gibbs sampling or EM-based variational methods. We use variational methods since we adapt MedLDA package[2] to mixed-membership ERD-MedLDA and train relation detection models.

Specifically, to obtain a tractable lower bound, we consider an entire family of parameterized lower bounds with a set of free variational parameters, and pick the best lower bounds by optimizing the lower bound with respect to the free variational parameters.

2.3.1 Variational inference

The EM-based variational method involves an E-step and a M-step. In the E-step, the latent variable distributions are computed while in the M-step, parameter estimation is done by maximizing the expectation of the complete likelihood distribution. As we know, in order to use MedLDA to make predictions on the relation types, the key inferential problem that we need to solve is that of computing the posterior distribution of the hidden variables given a document,

$$p(\boldsymbol{\theta}, \mathbf{z}, \boldsymbol{\eta}|\mathbf{w}, \mathbf{y}, \boldsymbol{\alpha}, \boldsymbol{\beta}) = \frac{p(\boldsymbol{\theta}, \mathbf{z}, \boldsymbol{\eta}, \mathbf{w}, \mathbf{y}|\boldsymbol{\alpha}, \boldsymbol{\beta})}{p(\mathbf{w}, \mathbf{y}|\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\eta})} \tag{4}$$

where, the denominator is (3), which is the probability of observed variables given all parameters; and the numerator is the probability of all observed variables plus sampled parameters given hyper-parameter for one document.

The partition function obtained from (3) and (4) is intractable; thus, it cannot be computed in closed form. Hence, the same approximation as in Zhu *et al.* (2009) is made. The upper bound (formalized as $L^{bs}(q)$) of the negative log-likelihood $-\log\ p(\mathbf{w}, \mathbf{y}|\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\eta})$ – is given as,

$$\begin{aligned} L^{bs}(q) &= -\mathbb{E}_q[\log\ p(\boldsymbol{\theta}, \mathbf{z}, \boldsymbol{\eta}, \mathbf{y}, \mathbf{w}|\boldsymbol{\alpha}, \boldsymbol{\beta}, p_0(\boldsymbol{\eta}))] - \mathscr{H}(q(\boldsymbol{\theta}, \mathbf{z}, \boldsymbol{\eta})) \\ &= KL(q(\boldsymbol{\eta})||p_0(\boldsymbol{\eta}) + \mathbb{E}_{q(\boldsymbol{\eta})}[L^s] \end{aligned} \tag{5}$$

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

where L^s refers to the likelihood function of sLDA and $KL(q(\boldsymbol{\eta})||p_0(\boldsymbol{\eta})) = \mathbb{E}_{p_0(\boldsymbol{\eta})}[\log \frac{p_0(\boldsymbol{\eta})}{q(\boldsymbol{\eta})}]$ is the Kullback-Leibler (KL) divergence; $\mathbb{E}[\cdot]$ is the expected value and $\mathscr{H}(\cdot)$ is the entropy.

Equation (5) can be expanded as follows (please refer to the Appendix A for details).

$$\begin{aligned} L(\gamma_d, \phi_d; \boldsymbol{\alpha}, \boldsymbol{\beta}, q(\boldsymbol{\eta})) = {} & \mathbb{E}_q[\log\ p(\theta_d|\boldsymbol{\alpha})] + \mathbb{E}_q[\log\ p(z_d|\boldsymbol{\theta}_d)] + \mathbb{E}_q[\log\ p(w_d|z_d, \boldsymbol{\beta})] \\ & - \mathbb{E}_q[\log\ q(\theta_d|\gamma_d)] - \mathbb{E}_q[\log q(z_d|\phi_d)] + \mathbb{E}_q[\log p(y_d|\bar{\mathbf{z}}, q(\boldsymbol{\eta}))] \end{aligned} \tag{6}$$

Then, with exactly the same idea of integration of MLE and MME as in Zhu *et al* (2009), we can define the integrated training of ERD-MedLDA as a constrained optimization problem in (7).

$$\begin{aligned} & \min_{q, q(\boldsymbol{\eta}), \boldsymbol{\alpha}, \boldsymbol{\beta}, \xi} \mathbb{E}_{q(\boldsymbol{\eta})}[L^s] + KL(q(\boldsymbol{\eta})||p_0(\eta)) + C \sum_{d=1}^{D} \xi_d \\ & \text{s.t. } \mathbb{E}[\boldsymbol{\eta}^T(\mathbf{f}(y_d, \bar{z}) - \mathbf{f}(y, \bar{z}))] \geq 1 - \xi_d;\ \mu_d \\ & \quad \xi_d \geq 0;\ v_d \\ & \quad \forall\ d \in [1, D];\ y \neq y_d \end{aligned} \tag{7}$$

where $\boldsymbol{\mu}, \mathbf{v}$ are Lagrange multipliers, $\boldsymbol{\xi}$ is the slack variable tolerating errors in training data. y_d is the true label while y is the prediction. So, $\mathbf{f}(y_d, \bar{z}_d) - \mathbf{f}(y, \bar{z}_d)$ is the difference between truth and prediction, which we call expected margin. Namely, the former is the average boundary composed of true labels and the latter is the average boundary of predicted labels. The smaller the difference, the closer it is to the true labels. This is essentially the advantage of ERD-MedLDA over other MLE or maximum-entropy models. Further, since ERD-MedLDA also employs regular MLE for data generated from sampling, it enjoys advantages of both kinds of learning. That is, it takes care of sufficient statistics as well as handles examples that are around the decision boundary with support vectors. When Lagrange multipliers are not zeros, terms related to them act as a regularizer, biasing the model towards discovering a latent representation. Thus, more accurate predictions will be obtained on difficult examples located at decision boundaries. These latent representations are fixed for words in a document and therefore yield better discriminant power.

Then, the Lagrangian $\mathscr{L}$ of (7) is similar to the one for classification as in (Zhu *et al.* 2009) except that the first term is supervised rather than unsupervised. This equation is reproduced below for readability.

$$\begin{aligned} \mathscr{L} = {} & L(q)^s + KL\big(q(\boldsymbol{\eta})||p_o(\boldsymbol{\eta})\big) + C \sum_{d=1}^{D} \xi_d - \sum_{d=1}^{D} v_d \xi_d \\ & - \sum_{d=1, y \neq y_d}^{D} \mu_d\big(\mathbb{E}[\boldsymbol{\eta}^T \Delta \mathbf{f}_d] + \xi_d - 1\big) - \sum_{d=1}^{D} \sum_{i=1}^{N} c_{di} \Bigg(\sum_{j=1}^{K} \phi_{dij} - 1 \Bigg) \end{aligned} \tag{8}$$

where $\Delta \mathbf{f}_d = \mathbf{f}(y_d, \bar{z}) - \mathbf{f}(y, \bar{z})$ and the last term is from the normalization condition $\sum_{j=1}^{K} \phi_{dij} = 1,\ \forall\, i, d$.

2.3.2 Parameter estimation

In the last section, we constructed variational parameters for approximating hidden variables $\boldsymbol{\theta}$, $\mathbf{z}$, and $\boldsymbol{\eta}$. Following this, the EM algorithm will iteratively solve the approximation problem with two steps:

(1) $E-step$: infer the posterior distribution of the hidden variables $\boldsymbol{\theta}$, $\mathbf{z}$, and $\boldsymbol{\eta}$, where, for $\boldsymbol{\theta}$ and $\mathbf{z}$, inferring the posterior distributions are in fact to fit the variational parameters $\boldsymbol{\phi}$ and γ; while for $\boldsymbol{\eta}$, more complex issues will be involved.
(2) $M-step$: estimate the unknown model parameters $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$.

The update rules are in fact done by sequentially taking partial derivative of (8) over related variables. We optimize the Lagrangian $\mathscr{L}$ over each of the related variables in the following order γ, $\boldsymbol{\phi}$, $q(\boldsymbol{\eta})$, $\boldsymbol{\alpha}$, and $\boldsymbol{\beta}$. Since the constraints in (8) are not on $\boldsymbol{\theta}$ (its variational γ), $\boldsymbol{\alpha}$, or $\boldsymbol{\beta}$, the update rules are the same as LDA or sLDA.

The update of $\boldsymbol{\phi}$ is the same as MedLDA in Zhu *et al* (2009) where the product of Lagrangian multiplier $\boldsymbol{\mu}$ and $\mathbb{E}[\eta^T \Delta \mathbf{f}_d]$ makes the update rule of $\boldsymbol{\phi}$. But since we will add more variational parameters to approximate $q(\boldsymbol{\eta})$, the final update rule of $\boldsymbol{\phi}$ is different.

The hardest part is the update of $q(\boldsymbol{\eta})$. This is because $\boldsymbol{\eta}$, as the parameter of the response variable y (or the relation type y), is coupled with $\bar{\mathbf{z}}$. In our general process, the relation type label y_d is sampled from a *softmax* or a multi-class logistic regression. Namely, y_d is generated from a discrete distribution $[p_1,p_2,\ldots,p_C, 1-\sum_{j=1}^{C} p_j]$ with $p_j = \frac{\exp(\boldsymbol{\eta}_j^T \bar{z})}{\sum_{j=1}^{C}(\boldsymbol{\eta}_j^T \bar{z})}$. The corresponding term in the posterior distribution is $\mathbb{E}_q[\log p(y_d|\mathbf{z_d}, q(\boldsymbol{\eta}))]$. This term cannot be solved even after introducing the variational distribution q. Consequently, another variational distribution parameter δ is introduced for further approximation. We give the final inference form as follows. The details are given in Appendix B.

$$\delta_d = 1 + \sum_{j=1}^{C} \sum_{k=1}^{K} \phi_{dk} \exp(\eta_{jk}) \tag{9}$$

$$\phi_{di} \propto \exp\left(\mathbb{E}[\log p(\boldsymbol{\theta}|\gamma]) + \mathbb{E}[\log p(w_{di}|\boldsymbol{\beta})] + \frac{1}{N} \sum_{y \neq y_d} \mu_d(y) \mathbb{E}[(\eta_{yd} - \eta_y)/\delta_d] \right) \tag{10}$$

The optimization of Lagrangian $\mathscr{L}$ over $q(\boldsymbol{\eta})$ is obtained by setting $\partial \mathscr{L} / \partial q(\boldsymbol{\eta}) = 0$. Next, the optimization problem (7) can be converted into its dual by plugging the resultant $q(\boldsymbol{\eta})$ into $\mathscr{L}$. For the standard normal prior for $p_0(\eta) = N(0, I)$, the dual problem will be quadratic and standard QP solvers can be used to solve it. However, according to the principle of conjugate prior, $q(\boldsymbol{\eta})$ is a normal distribution with a shifted mean as $q(\boldsymbol{\eta}) = N(\lambda, I)$, where $\lambda = \sum_{d=1, y \neq y_d}^{D} \mathbb{E}[\bar{z}] + \mathbb{E}[\Delta \mathbf{f}_d(y)]$. This can be taken into consideration to reformat the primal form of dual and it is solved by using existing multi-class SVM methods to get Lagrange multiplier μ_d and its duals (Zhu *et al.* 2009).

Table 1. *Relation types for ACE 05 corpus*

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

2.4 Relation detection

With the generative process, inference, and parameter estimation in place, ERD-MedLDA is 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., $p(y_d|z_{d1:dN}, \boldsymbol{\eta})$. Using variational approximation described in the previous section, we can derive the prediction rule as $F(\mathbf{y}, \mathbf{z}, \boldsymbol{\eta}) = \boldsymbol{\eta}^T \mathbf{f}(\mathbf{y}, \bar{z})$ where $\mathbf{f}(\mathbf{y}, \bar{z})$ is a feature vector. Now, SVM can be used to derive the prediction rule (Zhu *et al.* 2009).

$$\hat{y} = \arg\max_{y} \mathbb{E}[\boldsymbol{\eta}^T \mathbf{f}(\mathbf{y}, \bar{z})|\boldsymbol{\alpha}, \boldsymbol{\beta}] \tag{11}$$

Recall that we make use of *softmax* regression in sampling relation types; consequently, the derivation rules are given as:

$$\mathbb{E}[\log p(y = h|\mathbf{w}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\eta})] = \eta_h^T \mathbb{E}[\bar{z}] - \mathbb{E}\left[\log\left(\sum_{h=1}^{C} \exp(\eta_h^T \bar{z})\right)\right];\ \forall h \in [1, C] \tag{12}$$

The term $\mathbb{E}[\bar{z}]$, like in model learning, is intractable. Therefore, similar variational distributions are introduced. Namely, we use $\mathbb{E}_q(\bar{z}) = q(\boldsymbol{\theta}, \mathbf{z})$ to approximate $\mathbb{E}[\bar{z}]$.

3 Data

We use the ACE corpus (Phase 2, 2005) for training and evaluation. The ACE corpus has annotations for both entities and relations. The corpus has seven entity types, six major relations types, and twenty-three relation subtypes . In this work, we focus only on the six high-level relation types listed in Table 1. In addition to

Table 2. *NE pairs, mini-documents, and labels for the sentence 'John is married to Liz who works in Xyz Corp located in New York'*

NE pair	Mini-document	Label
John-Liz	John is married to Liz	PER-SOC
Liz-Xyz Corp	Liz who works in Xyz Corp	ORG-AFF
Liz-New York	Liz who works in Xyz Corp located in New York	GEN-AFF
Xyz Corp- New York	Xyz Corp located in New York	PHYS
John-Xyz Corp	John is married to Liz who works in Xyz Corp	NO-REL
John-NewYork	John is married to Liz who works in Xyz Corp located in New York	NO-REL

Table 3. *Distribution of relation types in the ACE corpus*

Relation type	Count	Distribution
ART	536	0.09
GEN-AFF	746	0.126
ORG-AFF	1762	0.298
PART-WHOLE	926	0.157
PER-SOC	911	0.15
PHYS	1031	0.174

the six major types, we have an additional category – no relation (NO-REL) that exists between entities that are not related.

NEs within a sentence are paired, and all text in between and including the NEs is considered as a mini-document. The gold standard annotation of their ACE relation type, or NO-REL if no relation exists, forms the mini-document's label. For instance, consider the following sentence (all NEs are shown in *italics*):

> *John* is married to *Liz* who works in *Xyz Corp* located in *New York*.

All NE pairs, the corresponding mini-documents, and their labels, constructed from this sentence, are listed in Table 2.

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 two intervening NEs in between. This gives us a total of 38,342 relation instances of which 32,640 are negative instances and 5912 are positive relation instances belonging to one of the six categories. The distribution of the six ACE categories is given in Table 3.

4 Features

We explore the effectiveness of incorporating features into our system as well as the baselines. For this, we construct feature sets similar to Jiang and Zhai (2007) and Zhou *et al.* (2005). Three sets of features are employed: Bag of Words (BOW), Syntactic (SYN), and Composite (COMP).

4.1 Bag of words features (BOW)

The BOW feature captures all the words in our mini-document. Consider, for example, the following text snippet that reveals an affiliation relation (ORG-AFF) between X and the United States.

(1) *X*, the president of the *United States [ACE relation type: ORG-AFF]*

Notice that words such as 'of' can be indicative of the ORG-AFF relation. BOW features capture this lexical information. However, compared to traditional classification settings, there is a difference in how ERD-MedLDA employs these features. BOW features are trained as topics and then the discovered topics will be employed as features. Thus, words such as 'of' may fall into topic(s) that, in turn, would eventually contribute to the recognition of ORG-AFF class.

4.2 Syntactic features (SYN)

The SYN features are constructed to capture syntactic, semantic, and structural information of the mini-document. Let us consider the following text snippet where there is no relation between X and the United States:

(2) *X*, the president, visited the *United States [ACE relation type: no relation]*

Notice that even though examples 1 and 2 exhibit different relation types, they share a large number of words. Now, observe that the words that indeed differ between the two text snippets, 'of' and 'visited,' also differ in their part of speech: one is a preposition (IN), while the other is a verb (VBD). Thus, we include part of speech (POS) information of the words between two NEs to additionally clue the system to the type of relation that might exist between them.

We observed that the syntactic roles are also indicative of relationships between entities. For instance, in Example 1, *X*, *the president*, is the subject of *the United States*. We encode syntactic features to capture this information. These include:

- *HM1:* The head word of the first mention
- *HM2:* The head word of the second mention
- *ET1:* Entity type of the first entity
- *ET2:* Entity type of the second entity
- *M1:* Mention type of the first entity
- *M2:* Mention type of the second entity
- *#MB:* Number of other mentions in between the two mentions under consideration
- *#WB:* Number of words in between the two mentions

4.3 Composite features (COMP)

The composite features (COMP) are similar to SYN, but they additionally capture order and dependencies between the features mentioned above. In particular, they capture structural information. Ordering of words are not captured by BOW or SYN. This feature exchangeability 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. 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.

We overcome these limitations by explicitly encoding these information as COMP features. Composite features comprise of feature pairs indicating which of the two occurs first. These include:

- *HM1HM2:* Head word of mention 1 and head word of mention 2. This encodes what mention head word occurs first.
- *ET12 :* Ordered pair of mention entity type.
- *ML12:* Combination of mention levels.
- *M1InM2:* Flag indicating whether M1 is included in M2. This feature captures mention dependencies.
- *M2InM1:* Flag indicating whether M2 is included in M1.

5 Experiments

ERD-MedLDA is a LDA-based framework that uses maximum-margin learning. Thus, we need to verify if our MedLDA formulation does better than topic modeling alone and maximum-margin methods alone. For this, we compare ERD-MedLDA to both LDA and SVM. Comparison with SVM is straightforward as it is a supervised framework. However, as the basic LDA is unsupervised, the discovered topics are not tied to any particular class. Thus, for a fair comparison, we use a labeled variant of LDA, the LLDA (Ramage *et al.* 2009), as the baseline topic model system.

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 for each of the models (wherever applicable) using the training data (the training data is split into training and validation sets). 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 6.1 for a detailed analysis).

5.1 ERD-MedLDA setup

The number of topics for the LDA-based models are determined using the equation $2K_0 + K_1$ following Zhu *et al.* (2009) and $K_1 = 2K_0$. K_0 is the number of topics

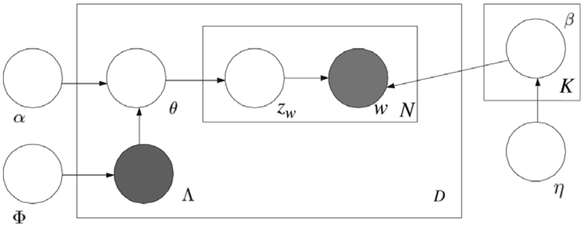

Fig. 3. Graphical model of LLDA.

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 data characteristics 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, and 110. BOW, PlusSYN, and PlusCOMP configurations obtain the best performance for ninety, eighty, and seventy topics respectively.

Since SVMs are employed in the ERD-MedLDA implementation, we need to determine the penalty parameter of the cost function, C. We used five-fold cross-validation to locate the parameter C. The best values for C are 25, 28, and 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 ERD-MedLDA is run by sampling, the result may be different each time. We ran it five times for each setting and took the average as the final results.

5.2 Baselines

We employ the same features and the same settings for ERD-MedLDA, LLDA, and SVM wherever possible.

5.2.1 LLDA

LLDA (Ramage *et al.* 2009) is a variation of sLDA. While the original LLDA model was designed for recognizing credit attribution, it can be easily used for relation detection. We recreate the plate diagram from the original paper in Figure 3 and briefly explain the LLDA model for relation detection as follows.

The relation types (Λ) generate the topic distribution parameter θ with α. Each document d is represented by a tuple consisting of a list of word indices $w^d = (w_1,\dots,w_{N_d})$ and a list of binary relation type presence/absence indicators $\Lambda^{(d)} = (l_1,\dots,l_K)$, where each $w_i \in \{1,\dots,V\}$ and each $l_k \in \{0,1\}$. N_d is the document length, V is the vocabulary size, and K the total number of unique topic labels in the corpus.

Drawing the multinomial topic distributions over vocabulary β_k for each topic k, from a Dirichlet prior η, is the same as that for traditional LDA. However, LLDA

Table 4. *Overall performance of the three systems*

	Precision %	Recall %	F-measure %
SVM	53.2	35.2	40.3
LLDA	28.3	51.6	36.6
ERD-MedLDA	**57.8**	**53.2**	**55.4**

restricts θ^d to be defined only over the topics that correspond to its labels $\Lambda^{(d)}$. Since the word-topic assignments z_i are drawn from this distribution, this restriction ensures that all the topic assignments are limited to the document's label. We refer the reader to the original paper for further details.

The setting of topics for LLDA is similar to ERD-MedLDA. As LLDA is also run by sampling, we ran it five times for each setting and took the average as the final results.

5.2.2 SVM

The SVM (Cortes and Vapnik 1995) baseline is a straightforward supervised system. We use the SVMlight implementation from (Joachims 1999). For for our task, labels are relation types and the training vector comprises of the features discussed in Section 4. Binary (presence/absence) features are used to overcome sparsity problems. In SVMlight, a grid search tool is provided to locate the best value for parameter C. The best C value for all three feature conditions, BOW, PlusSYN, and PlusCOMP, was found to be 1. All other settings are similar to those of ERD-MedLDA, including the linear kernel.

5.3 Results

We present the results of the three systems built using PlusCOMP, as all systems achieved their best overall performance using these features. Table 4 reports the precision, recall, and F-measure of the three systems averaged across all seven categories (the best numbers for each metric are highlighted in **bold**). Among the baselines, SVM has better precision, while LLDA has better overall recall. Here, we see that ERD-MedLDA outperforms LLDA and SVM across all metrics. Specifically, there is a four percentage point improvement in precision, two percentage point improvement in recall, and fifteen percentage point improvement in F-measure over the best performing baseline. This result indicates that our approach of combining topic models and maximum-margin learning is effective for relation detection.

We performed t-tests to verify the significance of the improvements obtained by our system. The improvements obtained by ERD-MedLDA over SVM for recall and F-measure are highly significant ($p < 0.001$), while the improvement for precision is significant at $p < 0.05$. Comparing with LLDA, ERD-MedLDA's improvement for recall is significant at $p < 0.1$ while its improvements in precision and F-measure are significant at $p < 0.001$.

Table 5. *Multi-class classification results with PlusCOMP for SVM, LLDA, and ERD-MedLDA for the six ACE 05 categories and NO-REL*

	SVM			LLDA			ERD-MedLDA		
Labels	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

Now, looking at the results for each individual relationship category (see Table 5; the best numbers for each category and metric are highlighted in **bold**) we see that the F-measure for ERD-MedLDA is better than that for SVM for four out of the six ACE relation types; and better than the F-measure obtained by LLDA for all relation types except ORG-AFF. Specifically, comparing with the best performing baseline, ERD-MedLDA produces a F-measure improvement of twenty-seven percentage points for ART, three percentage points for PART-WHOLE, and fifty percentage points for PER-SOC. Also, for four of the six ACE relation types, ERD-MedLDA achieves the best precision. Even in the cases where ERD-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).

Notice that LLDA has more difficulty with marginal classes such as ART and GEN-AFF. 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 MLE. Notice that ERD-MedLDA achieves performance close to SVM for this category. This is because, even though both LLDA and ERD-MedLDA model hidden topics and then employ the discovered hidden topics to predict relation types, ERD-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 results of previous research. Much of the previous mainstream research has been carried out on previous versions of the ACE corpus (e.g., Bunescu and Mooney 2005; Jiang and Zhai 2007; Qian *et al.* 2008; Nguyen, Moschitti and Riccardi 2009; Chan and Roth 2011). Due to the difference in corpora, our results are not directly comparable to theirs. Relation extraction work by Khayyamian, Mirroshandel and Abolhassani (2009) on the ACE 2005 corpus is closest to our work. 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.

Table 6. *F-measures for every kernel in Khayyamian, Mirroshandel and Abolhassani (2009) and ERD-MedLDA*

Labels	CD'01	AAP	AAPD	TSAAPD-0	TSAAPD-01	ERD-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**

Khayyamian *et al.* (2009) employ classical kernel methods developed by Collins and Duffy (2002) and only report F-measures over the six ACE relation types. For clarity, we reproduce their results in Table 6 and repeat ERD-MedLDA F-measures from Table 5 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 ERD-MedLDA outperforms all kernels. ERD-MedLDA also performs better than the best kernel for four of the six relation types.

6 Analysis

We incorporated an exponential family distribution into ERD-MedLDA model in order to make use of rich features. In this section, we analyze if indeed ERD-MedLDA effectively utilizes the variety of features listed in Section 4 in comparison with baseline methods. Additionally, as supervised topic models are designed to infer topic distributions indicative of the class, we inspect the topics learned by ERD-MedLDA to see if the inclusion of supervision has created topics biased towards relation types.

6.1 Feature incorporation

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 informative features consistently across all relation types. Figures 4–6 illustrate the F-measures for SVM, LLDA, and ERD-MedLDA respectively for the three conditions: BOW, PlusSYN, and PlusCOMP.

Let us first look at the best systems (based on F-measure) for each of the six ACE relation types in Table 5, and look at what feature set produces the best result for that system and relation. ERD-MedLDA is the best performer for ART, PART-WHOLE, and PER-SOC in Table 5. Figure 6 reveals that ERD-MedLDA's best performance for these relation types are obtained using PlusCOMP features. Similarly, SVM obtains the best F-measure for GEN-AFF and PHYS relations and Figure 4 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

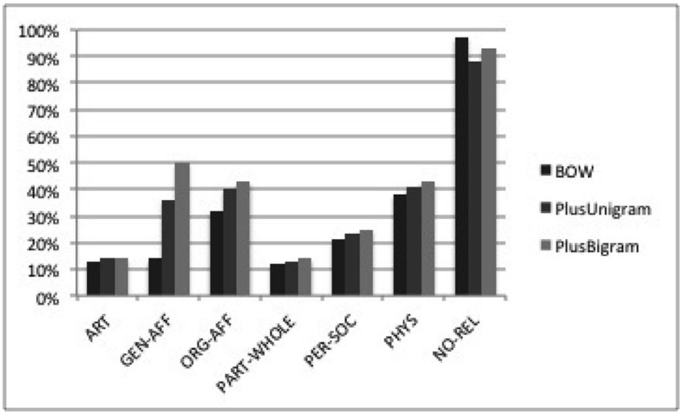

Fig. 4. SVM F-meausres for three feature conditions.

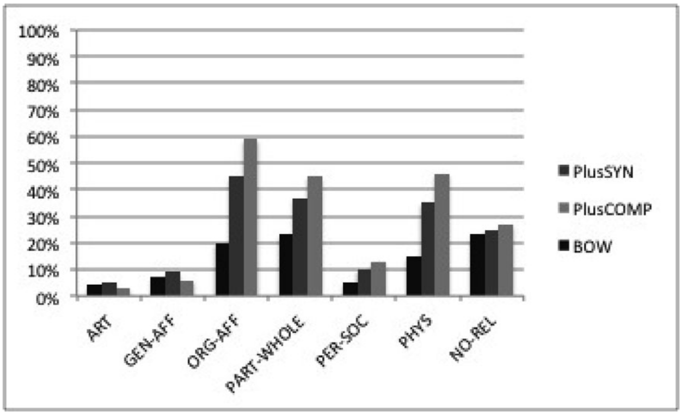

Fig. 5. LLDA F-meausres for three feature conditions.

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 4, 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 5). 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, ERD-MedLDA (see Figure 6) shows more consistent improvement for all relation types with the addition of more complex features. Also,

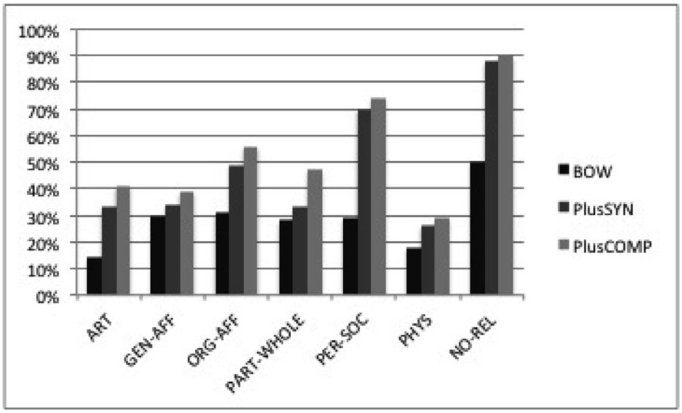

Fig. 6. MedLDA F-meausres for three feature conditions.

the gains are more substantial. This is encouraging and opens up avenues for further exploration of richer features for ERD-MedLDA.

6.2 Topic discovery

Our goal in employing a supervised framework was to guide the topic discovery to topics useful for relation detection. In this section, we examine the topics discovered for each relation type and discuss the effects of varying the number of topics in our model.

Figure 7 illustrates the topic distribution in ERD-MedLDA for the different relation types when twenty topics are used. The distributions are computed by averaging the expected latent representation of documents in each class. For all six ACE relation categories, we observe a sharp, sparse, and fast-decaying distribution over topics, indicating an affinity of the relation types for certain indicative topics. The NO-REL class, on the other hand, does not show this characteristic. Here, we see an almost uniform distribution over (almost) all topics. This is not surprising, as the NO-REL is a catch-all category comprising of various topics that do not correspond to the ACE relation types.

Notice that Topics 0 and 10 are prominent across some relation types. Specifically, Topic 10 is prominent for all six ACE relation categories, and Topic 0 is prominent for GEN-AFF and PHYS categories. When we inspected the top features belonging to Topic 10, we discovered that it covered person-related and pronominal features such as *I − I*, *I − you* or *I − my*, *he*, *he − he*, *they − he*, *them − I*, and *that*. Intuitively, these features could indicate any one of the relation types. Topic 10 is in fact an indicator of a *presence* of an ACE relation – which also explains why Topic 10 is not seen in the NO-REL category.

Most relation types have distributions over additional topics that help in distinguishing them. For example, in addition to Topic 10, ART has a substantial percentage of distribution over Topics 14, 15, and 17. Similarly, ORG-AFF has noticeable distributions over Topics 1 and 11. The only exception is the PHYS category. PHYS has prominent distributions only over Topics 0 and 10, both of

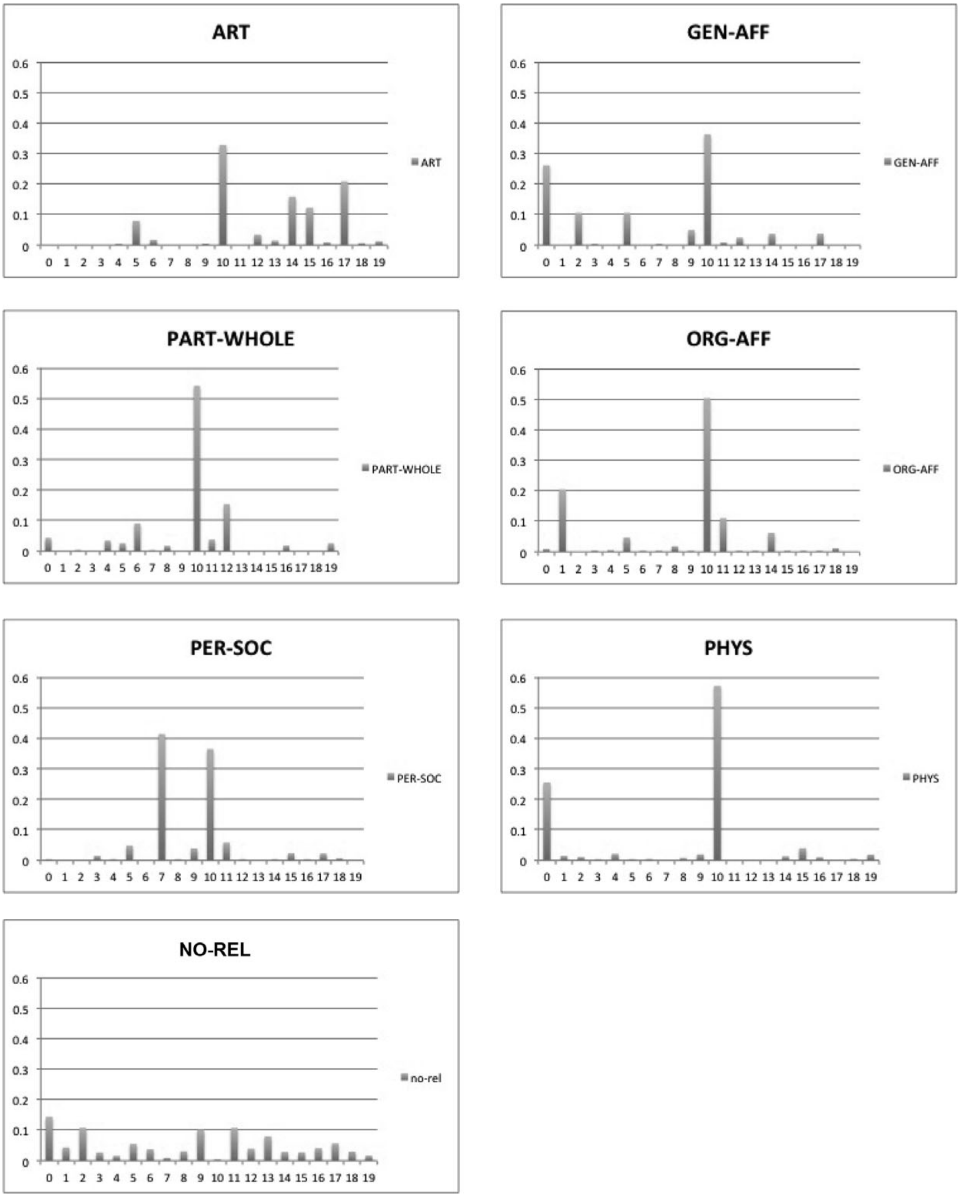

Fig. 7. Topic distribution for all relation types and NO-REL with twenty topics.

which are not unique to it alone. Not surprisingly, this affects the recognition of this category, as evidenced by the relatively low F-measures in Table 5. On the other hand, PER-SOC has a strong component of Topic 7, which is not seen in any other category. Consequently, it is recognized well by our system.

We observed that increasing the number of topics helps the system to obtain more distinct distributions for the relation types. Figure 8 illustrates the topic distribution when 110 topics are used. Notice that the topic distributions are more distinct for the ACE relation categories. However, we observed that if the number of topics is

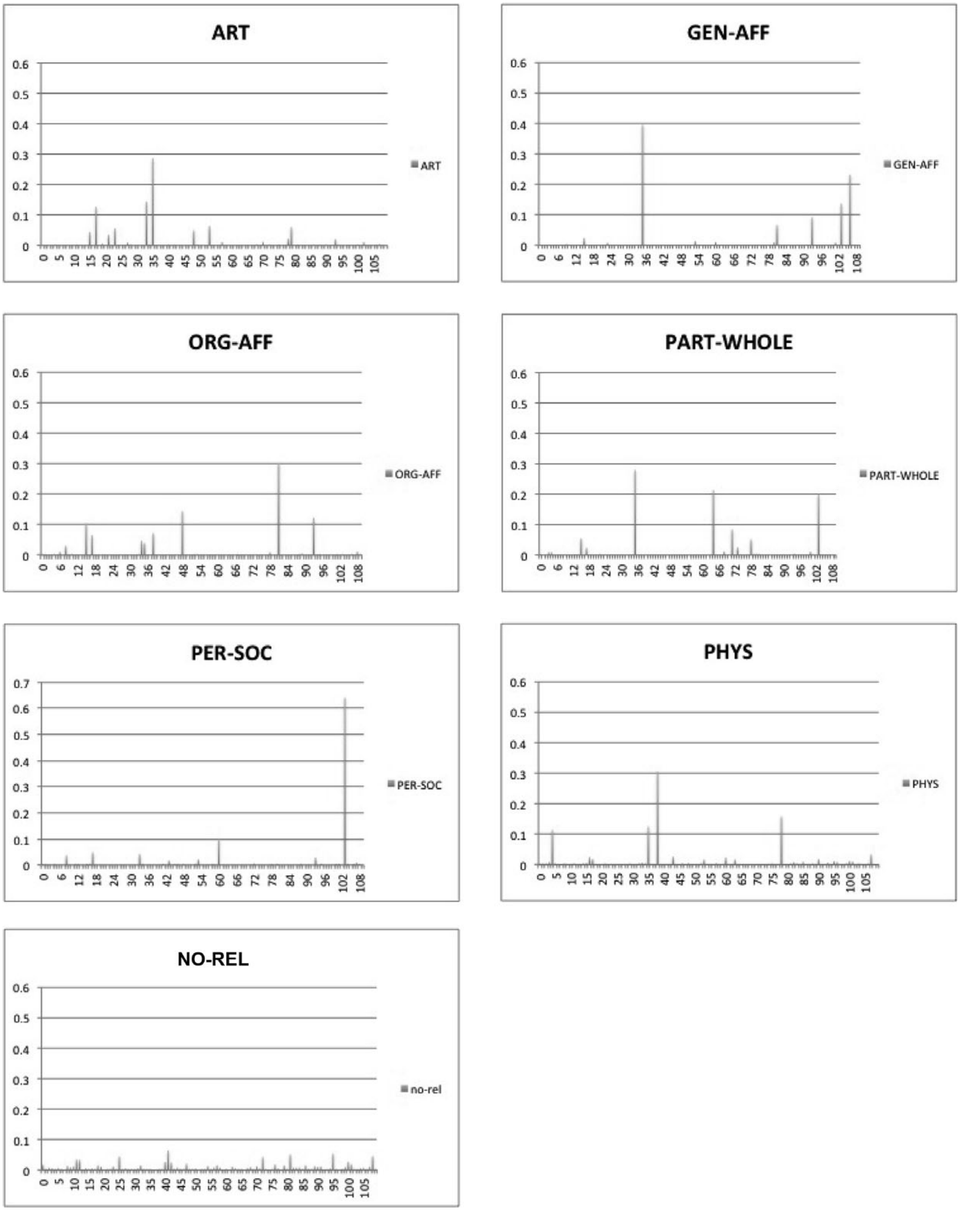

Fig. 8. Topic distribution for all relation types and NO-REL with 110 topics.

made too large, it can lead to overlapping topic distributions again. Finally, notice that NO-REL for 110 topics shows similar trend as that for twenty topics: it shows an almost uniform distribution over all topics.

We also inspected the topics discovered by LLDA and basic LDA. LLDA assumes that the topic discovered is the relation type. Here too, supervision helped with tying the topics to the relation types. For example, *family* and *wife* were among the top features for PER-SOC, and *weapons* was a top feature for ART. However, we observed that the topics discovered in LLDA are relatively more general, and show more overlap between ACE relation categories. Additionally, it shows an even more broad topic distribution for NO-REL. In contrast, in ERD-MedLDA, though NO-REL has a broad topic distribution, not many topics overlap between NO-REL and the true relation types. Intuitively, this results in better NO-REL recognition in ERD-MedLDA as compared to LLDA.

Finally, topics discovered (and consequently, the features for those topics) by the basic LDA model were not clearly interpretable as indicative of the relation types. This is not surprising, as basic LDA does not incorporate supervision.

7 Related work

There has been significant research in relation detection and topic modeling. Our work straddles the two areas; we perform ERD within a topic modeling framework and our topic model is adapted to fit the requirements of the ERD task. The following sections discuss literature related to our work.

7.1 Relation detection

One of the popular approaches for ERD is kernel methods. The main advantage of kernel methods is the ability of exploiting large feature sets without an explicit feature representation. Some examples are dependency tree kernels (Culotta and Sorensen 2004) and shortest dependency path kernels (Bunescu and Mooney 2005). Other research in kernel methods include work form Zelenko, Aone and Richardella (2003) that has exploited similarity measures over diverse features, 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). Besides independent kernel functions, combinations of different kernel functions have also been explored. Nguyen *et al.* (2009) combine constituent and dependency trees and sequential structures with kernel methods. To fully exploit the potential of dependency tree, they also applied the partial tree kernel proposed by (Moschitti 2006), and investigated the incorporation of dependency structure into rich sequence kernels.

Our method focuses on addressing the underlying semantics more directly than typical kernel-based methods. We try to capture structural information captured by kernel methods via explicit encoding of features. Further, the combination of MLE-based sLDA and MME-based SVM makes it possible for our model to directly employ rich kernel functions. The use of kernel functions in ERD-MedLDA is one of the directions for future work.

Chan and Roth (2011) employ constraints for relation detection 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 (Miller *et al.* 2000; Kambhatla 2004; Zhou *et al.* 2005; Jiang and Zhai 2007; Zhou *et al.* 2008). 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. For instance, Roth and Yih (2002) have applied a probabilistic approach to solve the problems of NE and relation extraction with the incorporation of all these features. Kambhatla (2004) has employed maximum entropy models with diverse features including words, entity, and mention types and the number of words (if any) separating the two entities.

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 a topic for future work.

A closely related task is that of relation mining and discovery, where unsupervised and semi-supervised approaches have been effectively employed (Hasegawa *et al.* 2004; Jiang 2009; Mintz *et al.* 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 ACE-05 corpus.

7.2 Topic models

Many researchers have explored extensions to the original LDA from Blei, Ng and Jordan (2003), such as correlated topic models (Blei and Jordan 2006), LLDA (Ramage *et al.* 2009), sLDA (Blei and McAuliffe 2008), and discLDA (Lacoste-Julien, Sha and Jordan 2008). LDA-based models have been adapted and employed for a wide variety of tasks such as genomic profiling (Flaherty *et al.* 2005) and image annotation (Wang, Blei and Li 2009). In NLP, they have been employed for review mining (Titov and McDonald 2008; Zhao *et al.* 2010), perspective analysis (Lin, Xing and Hauptmann 2008), and key phrase extraction (Zhao *et al.* 2011). In this work, we adapt the MedLDA topic model from (Zhu and Xing 2010) and (Shan *et al.* 2009) to incorporate rich features for the task of relation extraction.

The work on relation discovery from Hachey (2006) is close to our work. However, Hachey uses LDA as a module in his system to reduce feature dimensions only, while our approach employs the LDA-based framework for full relation detection.

8 Conclusion and future work

In this work we presented ERD-MedLDA, a system for detecting entity relations based on topic models. Our approach was motivated by the idea that latent semantics of text, as discovered by LDA-based models, are useful for relation detection. For this, we reformulated ERD as a topic modeling task. To the best of our knowledge, this is the first work to make full use of topic models for relation detection.

Starting with MedLDA and mixed-membership models, we adapted them to the relation detection task. Specifically, we modified the optimization problem and incorporated an exponential family distribution for each feature. The resulting ERD-MedLDA model has the advantages of both maximum-margin and maximum likelihood methods and is also able to benefit from rich features.

Our experiments show that ERD-MedLDA achieves better overall performance than SVM-based and LLDA-based approaches across all metrics. Comparing with previous work from (Khayyamian *et al.* 2009), which have employed diverse kernels, we showed that ERD-MedLDA has better overall performance.

We also experimented with different features and investigated the effectiveness of ERD-MedLDA in harnessing these features as compared to baseline methods. Our analysis showed that ERD-MedLDA is able to effectively and consistently incorporate informative features. Examination of the topic distribution obtained by our system shows that supervision indeed helps the model to learn topics biased towards relation types.

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

Appendix A: Variational inferences

We can derive the variational inferences for (4) as follows. Based on Figure 2, for each data point $\mathbf{w}_d, \mathbf{y}_d$, the expected value of joint probability of L^s can be factored as:

$$\begin{aligned}\mathbb{E}_q[\log p(\boldsymbol{\theta}, \mathbf{z}, \boldsymbol{\eta}, \mathbf{y}, \mathbf{w}|\boldsymbol{\alpha}, \boldsymbol{\beta}, p_0(\boldsymbol{\eta}))] &= \mathbb{E}_q[\log p(\theta_d|\boldsymbol{\alpha})] + \mathbb{E}_q[\log p(z_d|\theta_d)] \\ &\quad + \mathbb{E}_q[\log p(w_d|z_d, \boldsymbol{\beta})] + \mathbb{E}_q[\log p(y_d|\bar{z}, q(\boldsymbol{\eta})] \quad \text{(A 1)}\end{aligned}$$

The second term of (5), $\mathbb{E}_{q(\eta)}L^s = q(\boldsymbol{\theta}, \mathbf{z}, \boldsymbol{\eta}|\gamma, \boldsymbol{\phi})$ is the variational form of the original log-likelihood form. The expanded form of that term is:

$$\mathbb{E}_{q(\boldsymbol{\eta})}[L^s] = q(\boldsymbol{\eta}) \prod_{d=1}^{D} q(\theta_d|\gamma_d) \prod_{n=1}^{N} q(z_{dn}|\phi_{dn}), \quad \text{(A 2)}$$

where γ_d is a K-dimensional vector of Dirichlet parameters, and each ϕ_{dn} is a categorical distribution over K topics. Namely, γ_d is the approximation of θ, and ϕ_{dn} is the approximation of z_{dn}. Denoting the lower bound for each data point (w_d, y_d) with $L(\gamma_d, \phi_d; \boldsymbol{\alpha}, \boldsymbol{\beta}, q(\boldsymbol{\eta}))$, (5) can be expanded as:

$$\begin{aligned}L(\gamma_d, \phi_d; \boldsymbol{\alpha}, \boldsymbol{\beta}, q(\boldsymbol{\eta})) &= \mathbb{E}_q[\log\ p(\theta_d|\boldsymbol{\alpha})] + \mathbb{E}_q[\log\ p(z_d|\boldsymbol{\theta}_d)] + \mathbb{E}_q[\log\ p(w_d|z_d, \boldsymbol{\beta})] \\ &\quad - \mathbb{E}_q[\log\ q(\theta_d|\gamma_d)] - \mathbb{E}_q[\log q(z_d|\phi_d)] + \mathbb{E}_q[\log p(y_d|\bar{\mathbf{z}}, q(\boldsymbol{\eta}))] \quad \text{(A 3)}\end{aligned}$$

Appendix B: Parameter estimation

Without loss of generality, the following equations are for *softmax* distribution. The results remain similar for multi-class logistic regression after taking into account the number of elements in series summations. Hence, the probability mass of relation type y_d given z_d and $\boldsymbol{\eta}$ is:

$$p(y_d|\mathbf{z}_d, \boldsymbol{\eta}) = \exp\Big(\sum_{j=1}^{C} \boldsymbol{\eta}_j^T \bar{z} y_j - \log\Big(\sum_{j=1}^{C} \exp\left(\eta_j^T \bar{z}_d\right)\Big)\Big) \tag{B 1}$$

Accordingly, the last term in (A 3) is:

$$\mathbb{E}_q[\log p(y_d|\mathbf{z_d}, q(\boldsymbol{\eta}))] = \sum_{j=1}^{C}\sum_{k=1}^{K} \eta_{jk} E_q[\bar{z}_k] y_{dj} - \mathbb{E}_q\left[\log\left(\sum_{j=1}^{C} \exp\left(\boldsymbol{\eta}_j^T \bar{z}\right)\right)\right] \tag{B 2}$$

Even after introducing the variational distribution q, the above equation cannot be efficiently computed. Consequently, further approximation must be done. Thus, a new variational parameter δ is introduced to obtain a further lower bound for it. Specifically, besides Jensen inequality, another inequality, namely, $-\log(x) \geq 1 - \frac{x}{\delta} - \log(\delta)$ (Minka 2003; Shan *et al.* 2009), is used here to lower bound the term as:

$$\mathbb{E}_q[\log \ p(y_d|\bar{\mathbf{z}}, q(\boldsymbol{\eta}))] \geq \sum_{k=1}^{K} \mathbb{E}_q[\bar{z}_k] \sum_{j=1}^{C}\left(\eta_{jk} y_{dj} - \frac{1}{\delta}\exp(\eta_{jk})\right) + (1 - \log\delta) \tag{B 3}$$

Thus, with the addition of another variational parameter, $\boldsymbol{\delta}$, maximizing the lower-bound Lagrange function (8) with respect to the variational parameters will now give the update equation of γ_d, ϕ_d, η, and δ_d . The update rule of δ_d is similar to γ_d and ϕ_d. A simple partial derivation of (8) over δ will give its update equation as:

$$\delta_d = 1 + \sum_{j=1}^{C}\sum_{k=1}^{K} \phi_{dk} \exp(\eta_{jk}) \tag{B 4}$$

As we see from (B 3), $\boldsymbol{\delta}$ and $\boldsymbol{\phi}$ are coupled together now. Hence, in the update rule of $\boldsymbol{\delta}$, $\boldsymbol{\phi}$ is part of it. Consequently, the update rule of $\boldsymbol{\phi}$ is not separate from $\boldsymbol{\delta}$.

$$\phi_{di} \propto \exp\Big(\mathbb{E}[\log p(\boldsymbol{\theta}|\gamma]) + \mathbb{E}[\log p(w_{di}|\boldsymbol{\beta})] + \frac{1}{N}\sum_{y \neq y_d} \mu_d(y)\mathbb{E}[(\eta_{yd} - \eta_y)/\delta_d]\Big) \tag{B 5}$$

References

ACE. 2000–2005. Automatic content extraction. http://www.ldc.upenn.edu/Projects/ACE/

Blei, D. M., and Jordan, M. I. 2006. Variational inference for Dirichlet process mixtures. *Bayesian Analysis* **1**(1): 121–44.

Blei, D. M., and McAuliffe, J. 2008. Supervised topic models. *Advances in Neural Information Processing Systems* **20**: 121–8.

Blei, D. M. , Ng, A. Y., and Jordan, M. I. 2003. Latent Dirichlet allocation. *Journal of Machine Learning Research* **3**: 993–1022.

Bunescu, R. C., and Mooney, R. J. 2005. A shortest path dependency kernel for relation extraction. In *HLT & EMNLP Proceedings*, pp. 724–31, Vancouver, Canada.

Carreras, X., and Màrquez, L. 2005. Introduction to the CoNLL-2005 shared task: semantic role labeling. In *Proceedings of the 9th Conference on Computational Natural Language Learning*, pp. 152–64, Ann Arbor, MI.

Chan, Y., and Roth, D. 2011. Exploiting syntactico-semantic structures for relation extraction. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, Portland, OR.

Collins, M., and Duffy, N. 2002. Convolution kernels for natural language. *Advances in Neural Information Processing Systems* **1**: 625–32.

Cortes, C., and Vapnik, V. 1995. Support-vector networks. *Machine Learning* **20**(3): 273–97.

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

Doddington, G., Mitchell, A., Przybocki, M., Ramshaw, L., Strassel, S., and Weischedel, R. 2004. The automatic content extraction (ACE) program: tasks, data, and evaluation. *Proceedings of LREC* **4**: 837–40.

Farkas, R., Vincze, V., Móra, G., Csirik, J., and Szarvas, G. 2010. The CoNLL-2010 shared task: learning to detect hedges and their scope in natural language text. In *Proceedings of the 14th Conference on Computational Natural Language Learning (CoNLL-2010): Shared Task*, pp. 1–12, Uppsala, Sweden.

Flaherty, P., Giaever, G., Kumm, J., Jordan, M. I., and Arkin, A. P. 2005. A latent variable model for chemogenomic profiling. *Bioinformatics* **21**(15): 3286–93.

Hachey, B. 2006. Comparison of similarity models for the relation discovery task. In *Proceedings of the Workshop on Linguistic Distances*, p. 25, Sydney, Australia.

Hasegawa, T., Sekine, S., and Grishman, R. 2004. Discovering relations among named entities from large corpora. In *Proceedings of the 42nd Meeting of the Association for Computational Linguistics (ACL'04), Main Volume*, pp. 415–22, Barcelona, Spain.

Jiang, J. 2009. Multi-task transfer learning for weakly-supervised relation extraction. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2*, pp. 1012–20, Suntec, Singapore.

Jiang, J., and Zhai, C. X. 2007. A systematic exploration of the feature space for relation extraction. In *Proceedings of NAACL/HLT*, pp. 113–20, Rochester, NY.

Joachims, T. 1999. Making large scale SVM learning practical. In *Advances in Kernel Methods: Support Vector Learning*, pp. 169–184. Cambridge, MA: MIT Press.

Kambhatla, N. 2004. Combining lexical, syntactic, and semantic features with maximum entropy models for extracting relations. In *Proceedings of the ACL 2004 on Interactive Poster and Demonstration Sessions*, p. 22, Barcelona, Spain.

Khayyamian, M., Mirroshandel, S. A., and Abolhassani, H. 2009. Syntactic tree-based relation extraction using a generalization of Collins and Duffy convolution tree kernel. In *Proceedings of the HLT/NAACL Student Research Workshop and Doctoral Consortium*, pp. 66–71, Boulder, CO.

Lacoste-Julien, S., Sha, F., and Jordan, M. I. 2008. DiscLDA: discriminative learning for dimensionality reduction and classification. In *Advances in Neural Information Processing Systems 21: Proceedings of the 22nd Annual Conference on Neural Information Processing Systems*, Vancouver, Canada.

Lin, W. H., Xing, E., and Hauptmann, A. 2008. A joint topic and perspective model for ideological discourse. In W. Daelemans, B. Goethals, and K. Morik (eds.), *Machine Learning and Knowledge Discovery in Databases*, pp. 17–32. Berlin: Springer-Verlag.

Miller, S., Fox, H., Ramshaw, L., and Weischedel, R. 2000. A novel use of statistical parsing to extract information from text. In *Proceedings of the 1st North American Chapter of the Association for Computational Linguistics Conference*, pp. 226–33, Seattle, WA.

Minka, T. P. 2003. A comparison of numerical optimizers for logistic regression. Technical Report, Department of Statistics, Carnegie Mellon University.

Mintz, M., Bills, S., Snow, R., and Jurafsky, D. 2009. Distant supervision for relation extraction without labeled data. In *47th ACL & 4th AFNLP Proceedings*, pp. 1003–11, Suntec, Singapore.

Moschitti, A. 2006. Efficient convolution kernels for dependency and constituent syntactic trees. In *17th ECML Proceedings*, pp. 318–29, Berlin, Germany.

Nguyen, T. V. T., Moschitti, A., and Riccardi, G. 2009. Convolution kernels on constituent, dependency and sequential structures for relation extraction. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pp. 1378–87, Singapore.

Qian, L., Zhou, G., Kong, F., Zhu, Q., and Qian, P. 2008. Exploiting constituent dependencies for tree kernel-based semantic relation extraction. In *Proceedings of the 22nd ACL Conference*, pp. 697–704, Manchester.

Ramage, D., Hall, D., Nallapati, R., and Manning, C. D. 2009. Labeled LDA: a supervised topic model for credit attribution in multi-labeled corpora. In *Proceedings of the 2009 EMNLP Conference*, pp. 248–56, Singapore.

Roth, D., and Yih, W. 2002. Probabilistic reasoning for entity & relation recognition. In *Proceedings of the 19th International Conference on Computational Linguistics (COLING)*, p. 7, Morristown, NJ.

Shan, H., Banerjee, A., and Oza, N. C. 2009. Discriminative mixed-membership models. In *Proceedings of the 9th IEEE International Conference on Data Mining*, pp. 466–75, Miami, FL.

Titov, I., and McDonald, R. 2008. Modeling online reviews with multi-grain topic models. In *Proceeding of the 17th International Conference on World Wide Web*, pp. 111–20, New York.

Wang, C., Blei, D., and Li, F. F. 2009. Simultaneous image classification and annotation. In *Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1903–10, Miami, FL.

Zelenko, D., Aone, C., and Richardella, A. 2003. Kernel methods for relation extraction. *Journal of Machine Learning Research* **3**: 1083–106.

Zhang, M., Zhang, J., Su, J., and Zhou, G. 2006. A composite kernel to extract relations between entities with both flat and structured features. In *21st ICCL & 44th ACL Proceedings*, pp. 825–32, Sydney, Australia.

Zhao, S., and Grishman, R. 2005. Extracting relations with integrated information using kernel methods. In *43rd ACL Proceedings*, p. 426, Ann Arbor, MI.

Zhao, W. X., Jiang, J., Yan, H., and Li, X. 2010. Jointly modeling aspects and opinions with a MaxEnt-LDA hybrid. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pp. 56–65, MIT Stata Center, MA.

Zhao, X., Jiang, J., He, J., Song, Y., Achananuparp, P., LIM, E. P., and Li, X. 2011. Topical keyphrase extraction from twitter. *Proceedings of the 49th Annual ACL-HLT Meeting*, Portland, OR.

Zhou, G., Jian, S., Jie, Z., and Min, Z. 2005. Exploring various knowledge in relation extraction. In *Proceedings of the 43rd Annual Meeting of the ACL*, pp. 427–34, Ann Arbor, MI.

Zhou, G., Zhang, M., Ji, D. H., and Zhu, Q. 2007. Tree kernel-based relation extraction with context-sensitive structured parse tree information. In *Proceedings of the EMNLP/CoNLL-2007 Conference*, pp. 728–36, Prague, Czech Republic.

Zhou, G. D., Zhang, M., Ji, D. H., and Zhu, Q. M. 2008. Hierarchical learning strategy in semantic relation extraction. *Information Processing & Management* **44**(3): 1008–21.

Zhu, J., Ahmed, A., and Xing, E. P. 2009. MedLDA: maximum margin supervised topic models for regression and classification. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pp. 1257–64, Montreal, Canada.

Zhu, J., and Xing, E. P. 2010. Conditional topic random fields. In *Proceedings of the 27th International Conference on Machine Learning*, Haifa, Israel.