A Framework for Tuning Posterior Entropy in Unsupervised Learning

Rajhans Samdani RSAMDAN2@ILLINOIS.EDU
Ming-Wei Chang MINCHANG@MICROSOFT.COM
Dan Roth DANR@ILLINOIS.EDU

Abstract

We present a general framework for unsupervised and semi-supervised learning containing a graded spectrum of Expectation Maximization (EM) algorithms. We call our framework Unified Expectation Maximization (UEM.) UEM allows us to tune the entropy of the inferred posterior distribution during the E-step to impact the quality of learning. Furthermore, UEM covers existing algorithms like standard EM and hard EM as well as constrained versions of EM such as Constraint-Driven Learning (Chang et al., 2007) and Posterior Regularization (Ganchev et al., 2010). Within the UEM framework, we can adapt the learning procedure to the data, initialization point, and supervision signals like constraints. Experiments on POS tagging, information extraction, and word-alignment show that often the best performing algorithm in the UEM family is a new algorithm that wasn't available earlier, exhibiting the benefits of the UEM framework.

1. Introduction

Expectation Maximization (EM) (Dempster et al., 1977) is inarguably the most widely used algorithm for unsupervised and semi-supervised learning (McCallum et al., 1998; Nigam et al., 2000; Brown et al., 1993; Klein & Manning, 2004). Recently, EM algorithms which incorporate constraints on structured output spaces have also been proposed (Chang et al., 2007; Ganchev et al., 2010; Mann & McCallum, 2008).

Several variations of EM (e.g. hard EM) exist in the literature and choosing a suitable variation is often very task-specific. In this paper, we focus on the variations of EM that learn different models via different ways of inferring the posterior distribution during the E-step. We observe that, while parameter regularization in supervised learning is very well studied, to the extent that tuning regularization penalty is now a standard practice, the effect of regularizing the inferred distribution in EM is relatively unexplored. The EM algorithms which try to achieve this are few and far between. Some works have shown that for certain tasks, hard EM is more suitable than regular EM (Spitkovsky et al., 2010). With constraints incorporated in changing the inference during the E-step, Posterior Regularization (PR) (Ganchev et al., 2010) corresponds to EM while Constraint-Driven Learning (CoDL) (Chang et al., 2007) corresponds to hard EM. The problem of choosing a good variation of EM remains elusive, along with the possibility of simple and better alternatives.

In this paper, we approach these concerns from a novel perspective. We present a unified framework for EM, Unified Expectation Maximization (UEM) (Samdani et al., 2012), that gives an explicit handle on the entropy of the distribution inferred during the E-step of learning. UEM provides a continuous spectrum of EM algorithms parameterized by a simple temperature-like tuning parameter. Furthermore, UEM covers existing versions of the EM algorithm like standard and hard EM, PR and CODL.

Using UEM, we can modulate the entropy of the inferred distribution to better fit the given data, initialization, and constraints. Existing EM codes can be very easily extended to implement UEM, which makes it a very easy way for practitioners to extract better performance out of their systems. We conduct experiments on unsupervised POS tagging, unsupervised word-alignment, and semi-supervised information extraction and show that choosing the right UEM variation outperforms existing EM algorithms by a significant margin.

Presented at the International Conference on Machine Learning (ICML) workshop on *Inferning: Interactions between Inference and Learning*, Edinburgh, Scotland, UK, 2012.

2. Preliminaries

Let $\mathbf{x}$ denote an input or observed features and $\mathbf{h}$ be a discrete output variable to be predicted from a finite set of possible outputs $\mathcal{H}(\mathbf{x})$. Let $P_\theta(\mathbf{x}, \mathbf{h})$ be a probability distribution over $(\mathbf{x}, \mathbf{h})$ parameterized by θ. Let $P_\theta(\mathbf{h}|\mathbf{x})$ refer to the conditional probability of $\mathbf{h}$ given $\mathbf{x}$. For instance, in part-of-speech tagging, $\mathbf{x}$ is a sentence, $\mathbf{h}$ the corresponding POS tags, and θ could be an HMM model; in word-alignment, $\mathbf{x}$ can be an English-French sentence pair, $\mathbf{h}$ the word alignment between the sentences, and θ the probabilistic alignment model. Let $\delta(\mathbf{h} = \mathbf{h}')$ be the Kronecker-Delta distribution centered at $\mathbf{h}'$, i.e., it puts a probability of 1 at $\mathbf{h}'$ and 0 elsewhere.

In the rest of this section, we review EM and constraints-based learning with EM.

2.1. EM Algorithm

To obtain the parameter θ in an unsupervised way, one maximizes log-likelihood of the observed data:

$$\mathcal{L}(\theta) = \log P_\theta(\mathbf{x}) = \log \sum_{\mathbf{h} \in \mathcal{H}(\mathbf{x})} P_\theta(\mathbf{x}, \mathbf{h}) \ . \quad (1)$$

EM (Dempster et al., 1977) is the most common technique for learning θ, which maximizes a tight lower bound on $\mathcal{L}(\theta)$. While there are a few different styles of expressing EM, we follow the style of Neal & Hinton (1998) which views EM as a block coordinate ascent algorithm which infers a distribution q over the outputs during the E-step and estimates θ during the M-step. In particular, the E-step for EM can be written as

$$q = \arg\min_{q' \in \mathcal{Q}} KL(q', P_\theta(\mathbf{h}|\mathbf{x})) \ , \quad (2)$$

where $\mathcal{Q}$ is the space of all distributions.

While EM infers a distribution in the E-step, hard EM is thought of as predicting a single output given by

$$\mathbf{h}^* = \arg\max_{\mathbf{h} \in \mathcal{H}(\mathbf{x})} P_\theta(\mathbf{h}|\mathbf{x}) \ . \quad (3)$$

However, one can also think of hard EM as producing a distribution given by $q = \delta(\mathbf{h} = \mathbf{h}^*)$. In this paper, we pursue this distributional view of both EM and hard EM and show its benefits.

EM for Discriminative Models EM-like algorithms can also be used in discriminative settings (Bellare et al., 2009; Ganchev et al., 2010) specifically for semi-supervised learning (SSL.) Given some labeled and unlabeled data, such algorithms maximize

$$F(\theta, q) = L_c(\theta) - c_1\|\theta\|^2 - c_2 KL(q, P_\theta(\mathbf{h}|\mathbf{x})) \ , \quad (4)$$

where, q, as before, is a probability distribution over $\mathcal{H}(\mathbf{x})$, $L_c(\theta)$ is the conditional log-likelihood of the labels given the features for the labeled data, and c_1 and c_2 are constants specified by the user; the KL divergence is measured only over the unlabeled data.

The EM algorithm in this case has the same E-step as unsupervised EM, but the M-step is different. The M-step is similar to supervised learning as it finds θ by maximizing a regularized conditional likelihood of the data w.r.t. the labels — true labels are used for labeled data and "soft" pseudo labels based on q are used for unlabeled data.

2.2. Constraints in EM

It has become a common practice, especially in NLP, to use constraints on output variables to guide inference during "test-time" (Roth & Yih, 2004; Clarke & Lapata, 2006; Koo et al., 2010) and during unsupervised and semisupervised learning (Chang et al., 2007; Ganchev et al., 2008; 2010).

In this paper, we focus on linear constraints over $\mathbf{h}$ (potentially non-linear over $\mathbf{x}$.) Assume that we have m *linear* constraints on outputs where the k^{th} constraint can be written as

$$\mathbf{u_k}^T \mathbf{h} \leq b_k \ .$$

Defining a matrix U as $\mathbf{U}^T = \begin{bmatrix} \mathbf{u_1}^T & \dots & \mathbf{u_m}^T \end{bmatrix}$ and a vector $\mathbf{b}$ as $\mathbf{b}^T = [b_1, \dots, b_m]$, we write down the set of all *feasible* structures as $\{\mathbf{h} \mid \mathbf{h} \in \mathcal{H}(\mathbf{x}), \mathbf{U}\mathbf{h} \leq \mathbf{b}\}$.

Constraint-Driven Learning (CoDL) (Chang et al., 2007) augments the E-step of hard EM (3) by doing inference with respect to these constraints:

$$\mathbf{h}^* = \arg\max_{\mathbf{h} \in \mathcal{H}(\mathbf{x}) | \mathbf{U}\mathbf{h} \leq \mathbf{b}} P_\theta(\mathbf{h}|\mathbf{x}) \ . \quad (5)$$

Constraints on structures can be relaxed to *expectation constraints* by requiring the distribution q to satisfy them only in expectation. Define expectation w.r.t. a distribution q over $\mathcal{H}(\mathbf{x})$ as $E_q[\mathbf{U}\mathbf{h}] = \sum_{\mathbf{h} \in \mathcal{H}(\mathbf{x})} q(\mathbf{h}) \mathbf{U}\mathbf{h}$. In the expectation constraints setting, q is required to satisfy:

$$E_q[\mathbf{U}\mathbf{h}] \leq \mathbf{b} \ .$$

The space of distributions $\mathcal{Q}$ can be modified as:

$$\mathcal{Q} = \{q \mid q(\mathbf{h}) \geq 0, E_q[\mathbf{U}\mathbf{h}] \leq \mathbf{b}, \sum_{\mathbf{h} \in \mathcal{H}(\mathbf{x})} q(\mathbf{h}) = 1\}.$$

Augmenting these constraints into the E-step of EM (2), gives the Posterior Regularization (PR) framework (Ganchev et al., 2010). In this paper, we adopt the expectation constraint setting.

3. Unified Expectation Maximization

We now present the Unified Expectation Maximization (UEM) framework which changes the E-step by regularizing the entropy of the inferred posterior distribution. A key observation underlying the development of UEM is that during inference, hard EM (or CoDL) finds a distribution with zero entropy while EM (or PR) finds a distribution with the same entropy as P_θ (or close to it). Specifically, we obtain the posterior distribution q by modifying the objective of the E-step of EM (2) as

$$q = \arg\min_{q' \in \mathcal{Q}} KL(q', P_\theta(\mathbf{h}|\mathbf{x}); \gamma) \ , \tag{6}$$

where $KL(q, p; \gamma)$ is a modified KL divergence:

$$KL(q, p; \gamma) = \sum_{\mathbf{h} \in \mathcal{H}(\mathbf{x})} \gamma q(\mathbf{h}) \log q(\mathbf{h}) - q(\mathbf{h}) \log p(\mathbf{h}). \tag{7}$$

In other words, UEM projects $P_\theta(\mathbf{h}|\mathbf{x})$ on the space of feasible distributions $\mathcal{Q}$ w.r.t. $KL(\cdot, \cdot; \gamma)$ to infer the posterior q. By simply varying γ, UEM changes the projection and obtains different variations of EM including EM (PR, in the presence of constraints) and hard EM (CoDL.) The M-step for UEM is exactly the same as EM (or discriminative EM.)

3.1. Relationship between UEM and Other EM Algorithms

In order to better understand different UEM variations, we write the UEM E-step (6) explicitly as an optimization problem:

$$\begin{aligned} \min_{q} \quad & \sum_{\mathbf{h} \in \mathcal{H}(\mathbf{x})} \gamma q(\mathbf{h}) \log q(\mathbf{h}) - q(\mathbf{h}) \log P_\theta(\mathbf{h}|\mathbf{x}) \\ \text{s.t.} \quad & E_q[\mathbf{U}\mathbf{h}] \leq \mathbf{b}, \\ & q(\mathbf{h}) \geq 0, \forall \mathbf{h} \in \mathcal{H}(\mathbf{x}), \\ & \textstyle\sum_{\mathbf{h} \in \mathcal{H}(\mathbf{x})} q(\mathbf{h}) = 1 \ . \end{aligned} \tag{8}$$

The relation between unconstrained versions of EM has been mentioned before (Ueda & Nakano, 1998; Smith & Eisner, 2004). Notably, in the unconstrained case, by setting $\gamma = 1$, we obtain EM and by setting $\gamma = 0$, we obtain hard EM. We show that the relationship takes novel aspects in the presence of constraints.

UEM and Posterior Regularization ($\gamma = 1$) For $\gamma = 1$, UEM solves $\arg\min_{q \in \mathcal{Q}} KL\left(q, P_\theta(\mathbf{h}|\mathbf{x})\right)$ which is the same as Posterior Regularization (Ganchev et al., 2010).

UEM and CoDL ($\gamma = -\infty$) When $\gamma \to -\infty$ then due to an infinite penalty on the entropy of the posterior, the entropy must become zero. Thus, now the E-step, as expressed by Eq. (8), can be written as $q = \delta(\mathbf{h} = \mathbf{h}^*)$ where $\mathbf{h}^*$ is obtained as

$$\begin{aligned} \arg\max_{\mathbf{h} \in \mathcal{H}(\mathbf{x})} \quad & \log P_\theta(\mathbf{h}|\mathbf{x}) \\ \text{s.t.} \quad & \mathbf{U}\mathbf{h} \leq \mathbf{b} \ , \end{aligned} \tag{9}$$

which is the same as CoDL (5).

3.2. UEM with $\gamma \in [0, 1]$

The set of EM algorithms for $\gamma \in (0, 1)$ has not been explored before. This paper focuses on values of $\gamma \in [0, 1]$ for the following reasons. First, the E-step (8) is convex for $\gamma \geq 0$, (8), which can be solved exactly and efficiently. Second, for $\gamma = 0$, the E-step solves

$$\begin{aligned} \max_{q} \quad & \textstyle\sum_{\mathbf{h} \in \mathcal{H}(\mathbf{x})} q(\mathbf{h}) \log P_\theta(\mathbf{h}|\mathbf{x}) \\ \text{s.t.} \quad & E_q[\mathbf{U}\mathbf{h}] \leq \mathbf{b}, \\ & q(\mathbf{h}) \geq 0, \forall \mathbf{h} \in \mathcal{H}(\mathbf{x}), \\ & \textstyle\sum_{\mathbf{h} \in \mathcal{H}(\mathbf{x})} q(\mathbf{h}) = 1 \ , \end{aligned} \tag{10}$$

which is an **LP-relaxation** of CODL (Eq. (9)) and in the absence of constraints is exactly the same as hard EM. LP relaxations often provides a decent approximation to ILP (Roth & Yih, 2004; Martins et al., 2009). Third, $\gamma \in [0, 1]$ covers standard EM/PR.

3.3. Role of γ in Learning

The modified KL divergence can be related to standard KL divergence as $KL(q, P_\theta(\mathbf{h}|\mathbf{x}); \gamma) = KL(q, P_\theta(\mathbf{y}|\mathbf{x})) + (1 - \gamma)H(q)$ — UEM (6) minimizes the former during the E-step, while Standard EM (2) minimizes the latter. The additional term $(1-\gamma)H(q)$ is essentially an entropic prior on the posterior distribution q. This term gives a handle on changing the quality of learning by changing the entropy of the distribution q infered during the E-step.

For $\gamma < 1$, the regularization term penalizes the entropy of the posterior thus inferring distributions with lesser mass on the tail. This is significant, for instance, in unsupervised structured prediction where the tail, in total, can carry a substantial amount of probability mass as the output space is massive. This notion aligns with the observation of Spitkovsky et al. (2010) who criticize EM for frittering away too much probability mass on unimportant outputs while showing that hard EM does much better in PCFG parsing.

3.4. Solving the Constrained E-step

The E-step of UEM becomes complicated with the introduction of constraints. To solve this step, we propose a dual projected subgradient-ascent algorithm

which reduces to dual decomposition (Bertsekas, 1999; Rush et al., 2010; Koo et al., 2010) in several cases. For $\gamma = 0$, our algorithm naturally reduces to an efficient Lagrange-relaxation (Rush & Collins, 2011) algorithm which approximates the exact hard inference used in CODL (9). An interested reader can refer to Samdani et al. (2012) for more details.

4. Experiments

We conduct experiments to study the effect of changing the entropy of the inferred distribution on the quality of learning. In particular, we tune the value of γ over the set $\{0.0, 0.1, \ldots, 1.0\}$ as a way to obtain gains over EM and hard EM in the constrained and unconstrained cases. We conduct experiments on POS-tagging, word-alignment, and information extraction. We omit some of the details in this paper; refer to Samdani et al. (2012) for more details.

As we vary γ over $[0, 1]$, we regularizes the entropy of the posterior in a "smooth" way thus circumventing much of the debate over EM vs hard EM (Spitkovsky et al., 2010). Furthermore, in the case of POS tagging, we study the relation between the quality of model initialization and the impact of entropy modulation. This is inspired by a general "research wisdom" that hard EM is a better choice than EM with a good initialization point whereas the opposite is true with an "uninformed" initialization.

Unsupervised POS Tagging We conduct experiments on unsupervised POS learning experiment using a first order (bigram) HMM model. We consider initialization points of varying quality which are constructed as follows. The "posterior uniform" initialization is created by spreading the probability uniformly over all possible tags for each token. To construct better initialization points, we train a supervised HMM tagger on hold-out labeled data. The quality of the initialization points is varied by varying the size of the labeled data over $\{5, 10, 20, 40, 80\}$. Those initialization points are then fed into different UEM algorithms.

Results The results are summarized in Figure 1. Note that when we use the "posterior uniform" initialization, EM wins by a significant margin. Surprisingly, with the initialization point constructed with merely 5 or 10 examples, EM is not the best algorithm anymore. The best result for most cases is obtained at γ somewhere between 0 (hard EM) and 1 (EM). The results not only indicate that a measure of "hardness" of inference during EM is closely related to the quality of the initialization point, but also elicit a more fine-grained relationship between initialization and UEM.

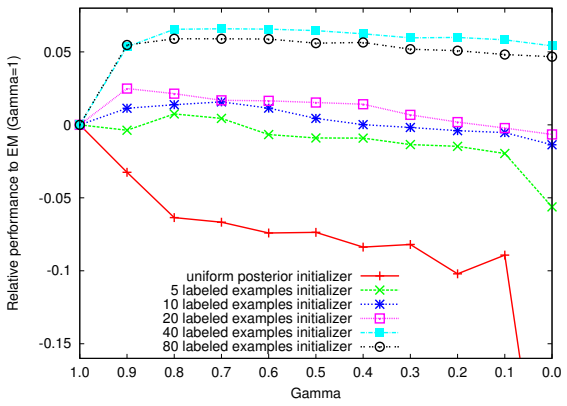

Figure 1. POS Experiments showing the relation between the quality of initialization and the tuning parameter γ. The results show that the value of the best γ is sensitive to the initialization point, and EM ($\text{UEM}_{1.0}$) and hard EM ($\text{UEM}_{0.0}$) are often not the best choice. We report the relative performance compared to EM given by $\frac{Acc(\text{UEM}) - Acc(\text{EM})}{Acc(\text{EM})}$ where *Acc* represents the accuracy as evaluated on the ambiguous words of the given data. The number of labeled examples indicates the size of the training set used to create the initial $\mathbf{w}$ (greater size means better initialization.) The posterior uniform initialization does not use any labeled example.

This experiment agrees with Merialdo (1994), who show that EM performs poorly in the semi-supervised setting. In Spitkovsky et al. (2010), the authors show that hard EM (Viterbi EM) works better than standard EM. We extend these results by showing that this issue can be overcome with the UEM framework by picking appropriate γ based on the amount of available labeled data.

Semi-Supervised Entity-Relation Extraction We conduct semi-supervised learning (SSL) experiments on entity and relation type prediction assuming that we are given mention boundaries. We borrow the data and the setting from (Roth & Yih, 2004).

We train two log linear models for entity type and relation type prediction, respectively via discriminative UEM. For our experiments, we use 20% of data for testing, a small amount, κ%, as labeled training data (we vary κ), and the remaining as unlabeled training data. We initialize with a classifier trained on the given labeled data.

We add two kinds of constraints during inference — entity and relation type compatibility constraints mentioned in Roth & Yih (2007) and expected count constraints similar to the *label regularization* technique mentioned in Mann & McCallum (2010) (Samdani et al. (2012) has more details.)

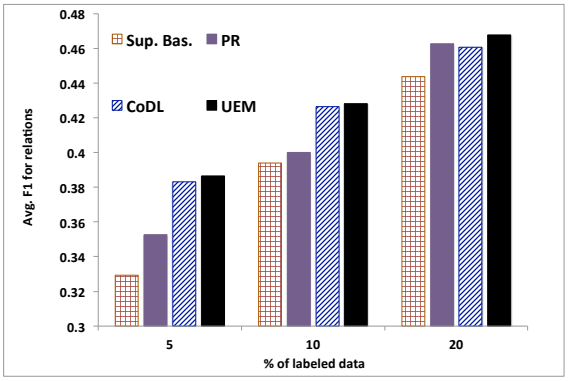

Figure 2. Average F1 for relation prediction for varying sizes of labeled data comparing the supervised baseline, PR, CoDL, and UEM. UEM is statistically significantly better than supervised baseline and PR in all the cases.

Results We use Alg. 2 in Samdani et al. (2012) for solving the constrained E-step. We report results averaged over 10 random splits of the data and measure statistical significance using paired t-test with $p = 0.05$. The results for relation prediction are shown in Fig. 2. For relation extraction, UEM is always statistically significantly better than the baseline and PR. The difference between UEM and CoDL is small which is not very surprising because hard EM approaches like CoDL are known to work very well for discriminative SSL. We omit the graph for entity prediction because EM-based approaches do not outperform the supervised baseline there. However, notably, for entities, for $\kappa = 10\%$, UEM outperforms CoDL and PR and for 20%, the supervised baseline outperforms PR statistically significantly.

Word Alignment Statistical word alignment is a well known application of unsupervised learning and is a key step towards machine translation from a source language S to a target language T. We experiment with two language-pairs: English-French and English-Spanish. We use Hansards corpus for French-English translation (Och & Ney, 2000) and Europarl corpus (Koehn, 2002) for Spanish-English translation with EPPS (Lambert et al., 2005) annotation.

We use an HMM-based model for word-alignment (Vogel et al., 1996) and add agreement constraints (Liang et al., 2008; Ganchev et al., 2008) to constrain alignment probabilities in one direction (P_{θ_1}: from S to T) to agree with the alignment probabilities in the other direction (P_{θ_2}: from T to S.) We use a small development set of size 50 to tune the model.

Results We compare UEM with EM, PR, and CoDL on the basis of Alignment Error Rate (AER) for different sizes of unlabeled data (See Tab. 1.) See Och & Ney (2003) for the definition of AER. UEM consistently outperforms EM, PR, and CoDL with a wide margin.

Size	EM	PR	CoDL	UEM	EM	PR	CoDL	UEM
	En-Fr				Fr-En			
10k	23.54	10.63	14.76	**9.10**	19.63	10.71	14.68	**9.21**
50k	18.02	8.30	10.08	**7.34**	16.17	8.40	10.09	**7.40**
100k	16.31	8.16	9.17	**7.05**	15.03	8.09	8.93	**6.87**
	En-Es				Es-En			
10k	33.92	22.24	28.19	**20.80**	31.94	22.00	28.13	**20.83**
50k	25.31	19.84	22.99	**18.93**	24.46	20.08	23.01	**18.95**
100k	24.48	19.49	21.62	**18.75**	23.78	19.70	21.60	**18.64**

Table 1. AER (Alignment Error Rate) comparisons for French-English (above) and Spanish-English (below) alignment for various data sizes. For French-English setting, tuned γ for all data-sizes is either 0.5 or 0.6. For Spanish-English, tuned γ for all data-sizes is 0.7.

5. Conclusion

We proposed a continuum of EM algorithms called UEM that is parameterized by a single parameter. UEM enables us to study the impact of the entropy of the distribution inferred during the E-step on the quality of learning. Our framework naturally incorporates constraints on output variables and generalizes existing constrained and unconstrained EM algorithms like standard and hard EM, PR, and CoDL. Using experiments, we showed how important it is to regularize the entropy of the posterior distribution using UEM.

Our technique is a simple and principled way to further analyze learning in unsupervised scenarios. More work is needed to establish an explicit relationship between the data, the initialization point, the parameter γ, and the output of UEM. Furthmore, our technique is amenable to be combined with many existing variations of EM (Berg-Kirkpatrick et al., 2010). We leave these questions as future work.

Acknowledgments: This research is sponsored by the Army Research Laboratory (ARL) under agreement W911NF-09-2-0053, Defense Advanced Research Projects Agency (DARPA) Machine Reading Program under Air Force Research Laboratory (AFRL) prime contract no. FA8750-09-C-018, and an ONR Award on Guiding Learning and Decision Making in the Presence of Multiple Forms of Information. Any opinions, findings, conclusions or recommendations are those of the authors and do not necessarily reflect the views of the funding agencies.

References

Bellare, K., Druck, G., and McCallum, A. Alternating projections for learning with expectation constraints. In *UAI*, 2009.

Berg-Kirkpatrick, T., Bouchard-Côté, A., DeNero, J., and Klein, D. Painless unsupervised learning with features. In *ACL*, HLT '10, 2010.

Bertsekas, D. P. *Nonlinear Programming*. Athena Scientific, 2nd edition, 1999.

Brown, Peter, Pietra, Stephen Della, Pietra, Vincent Della, and Mercer, Robert. The mathematics of statistical machine translation: parameter estimation. *Computational Linguistics*, 1993.

Chang, M., Ratinov, L., and Roth, D. Guiding semi-supervision with constraint-driven learning. In *ACL*, 2007.

Clarke, James and Lapata, Mirella. Constraint-based sentence compression: An integer programming approach. In *ACL*, 2006.

Dempster, A. P., Laird, N. M., and Rubin, D. B. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, 1977.

Ganchev, K., Graca, J.V., and Taskar, B. Better alignments = better translations. In *ACL*, 2008.

Ganchev, K., Graça, J., Gillenwater, J., and Taskar, B. Posterior regularization for structured latent variable models. *Journal of Machine Learning Research*, 2010.

Klein, D. and Manning, C. D. Corpus-based induction of syntactic structure: models of dependency and constituency. In *ACL*, 2004.

Koehn, P. Europarl: A multilingual corpus for evaluation of machine translation. 2002.

Koo, T., Rush, A. M., Collins, M., Jaakkola, T., and Sontag, D. Dual decomposition for parsing with non-projective head automata. In *EMNLP*, 2010.

Lambert, P., De Gispert, A., Banchs, R., and Marino, J. Guidelines for word alignment evaluation and manual alignment. *Language Resources and Evaluation*, 2005.

Liang, P., Klein, D., and Jordan, M. I. Agreement-based learning. In *NIPS*, 2008.

Mann, G. S. and McCallum, A. Generalized expectation criteria for semi-supervised learning with weakly labeled data. *JMLR*, 11, 2010.

Mann, Gideon and McCallum, Andrew. Generalized expectation criteria for semi-supervised learning of conditional random fields. In *ACL*, 2008.

Martins, A., Smith, N. A., and Xing, E. Concise integer linear programming formulations for dependency parsing. In *ACL*, 2009.

McCallum, Andrew K., Rosenfeld, Ronald, Mitchell, Tom M., and Ng, Andrew Y. Improving text classification by shrinkage in a hierarchy of classes. In *ICML*, 1998.

Merialdo, Bernard. Tagging text with a probabilistic model. *Computational Linguistics*, 1994.

Neal, R. M. and Hinton, G. E. A new view of the EM algorithm that justifies incremental, sparse and other variants. In Jordan, M. I. (ed.), *Learning in Graphical Models*. 1998.

Nigam, Kamal, Mccallum, Andrew Kachites, Thrun, Sebastian, and Mitchell, Tom. Text classification from labeled and unlabeled documents using EM. *Machine Learning*, 2000.

Och, F. J. and Ney, H. Improved statistical alignment models. In *ACL*, 2000.

Och, F. J. and Ney, H. A systematic comparison of various statistical alignment models. *CL*, 29, 2003.

Roth, D. and Yih, W. A linear programming formulation for global inference in natural language tasks. In Ng, Hwee Tou and Riloff, Ellen (eds.), *CoNLL*, 2004.

Roth, D. and Yih, W. Global inference for entity and relation identification via a linear programming formulation. In Getoor, Lise and Taskar, Ben (eds.), *Introduction to Statistical Relational Learning*, 2007.

Rush, A. M. and Collins, M. Exact decoding of syntactic translation models through lagrangian relaxation. In *ACL*, 2011.

Rush, A. M., Sontag, D., Collins, M., and Jaakkola, T. On dual decomposition and linear programming relaxations for natural language processing. In *EMNLP*, 2010.

Samdani, Rajhans, Chang, Ming-Wei, and Roth, Dan. Unified expectation maximization. In *NAACL*, 2012.

Smith, N. A. and Eisner, J. Annealing techniques for unsupervised statistical language learning. In *ACL*, 2004.

Spitkovsky, V. I., Alshawi, H., Jurafsky, D., and Manning, C. D. Viterbi training improves unsupervised dependency parsing. In *CoNLL*, 2010.

Ueda, N. and Nakano, R. Deterministic annealing em algorithm. *Neural Network*, 1998.

Vogel, S., Ney, H., and Tillmann, C. Hmm-based word alignment in statistical translation. In *COLING*, 1996.