Large-scale Reordering Model for Statistical Machine Translation using Dual Multinomial Logistic Regression

Abdullah Alrajeh[ab] **and Mahesan Niranjan**[b]
[a]Computer Research Institute, King Abdulaziz City for Science and Technology (KACST)
Riyadh, Saudi Arabia, asrajeh@kacst.edu.sa
[b]School of Electronics and Computer Science, University of Southampton
Southampton, United Kingdom, {asar1a10, mn}@ecs.soton.ac.uk

Abstract

Phrase reordering is a challenge for statistical machine translation systems. Posing phrase movements as a prediction problem using contextual features modeled by maximum entropy-based classifier is superior to the commonly used lexicalized reordering model. However, Training this discriminative model using large-scale parallel corpus might be computationally expensive. In this paper, we explore recent advancements in solving large-scale classification problems. Using the dual problem to multinomial logistic regression, we managed to shrink the training data while iterating and produce significant saving in computation and memory while preserving the accuracy.

1 Introduction

Phrase reordering is a common problem when translating between two grammatically different languages. Analogous to speech recognition systems, statistical machine translation (SMT) systems relied on language models to produce more fluent output. While early work penalized phrase movements without considering reorderings arising from vastly differing grammatical structures across language pairs like Arabic-English (Koehn, 2004a), many researchers considered lexicalized reordering models that attempted to learn orientation based on the training corpus (Tillmann, 2004; Kumar and Byrne, 2005; Koehn et al., 2005).

Building on this, some researchers have borrowed powerful ideas from the machine learning literature, to pose the phrase movement problem as a prediction problem using contextual input features whose importance is modeled as weights of a linear classifier trained by entropic criteria. The approach (so called maximum entropy classifier or simply MaxEnt) is a popular choice (Zens and Ney, 2006; Xiong et al., 2006; Nguyen et al., 2009; Xiang et al., 2011). Max-margin structure classifiers were also proposed (Ni et al., 2011). Alternatively, Cherry (2013) proposed recently using sparse features optimize the translation quality with the decoder instead of training a classifier independently.

While large-scale parallel corpus is advantageous for improving such reordering model, this improvement comes at a price of computational complexity. This issue is particularly pronounced when discriminative models are considered such as maximum entropy-based model due to the required iterative learning.

Advancements in solving large-scale classification problems have been shown to be effective such as dual coordinate descent method for linear support vector machines (Hsieh et al., 2008). Similarly, Yu et al. (2011) proposed a two-level dual coordinate descent method for maximum entropy classifier.

In this work we explore the dual problem to multinomial logistic regression for building large-scale reordering model (section 3). One of the main advantages of solving the dual problem is providing a mechanism to shrink the training data which is a serious issue in building such large-scale system. We present empirical results comparing between the primal and the dual problems (section 4). Our approach is shown to be fast and memory-efficient.

2 Baseline System

In statistical machine translation, the most likely translation $\mathbf{e}_{\text{best}}$ of an input sentence $\mathbf{f}$ can be found by maximizing the probability $p(\mathbf{e}|\mathbf{f})$, as follows:

$$\mathbf{e}_{\text{best}} = \arg\max_{\mathbf{e}} p(\mathbf{e}|\mathbf{f}). \quad (1)$$

Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 1758–1763,
October 25-29, 2014, Doha, Qatar.

A log-linear combination of different models (features) is used for direct modeling of the posterior probability $p(\mathbf{e}|\mathbf{f})$ (Papineni et al., 1998; Och and Ney, 2002):

$$\mathbf{e}_{\text{best}} = \arg\max_{\mathbf{e}} \sum_{i=1}^{n} \lambda_i h_i(\mathbf{f}, \mathbf{e}) \qquad (2)$$

where the feature $h_i(\mathbf{f}, \mathbf{e})$ is a score function over sentence pairs. The translation model and the language model are the main features in any system although additional features $h(.)$ can be integrated easily (such as word penalty). State-of-the-art systems usually have around ten features.

The language model, which ensures fluent translation, plays an important role in reordering; however, it has a bias towards short translations (Koehn, 2010). Therefore, a need for developing a specific model for the reordering problem.

2.1 Lexicalized Reordering Model

Adding a lexicalized reordering model consistently improved the translation quality for several language pairs (Koehn et al., 2005). Reordering modeling involves formulating phrase movements as a classification problem where each phrase position considered as a class (Tillmann, 2004). Some researchers classified phrase movements into three categories (monotone, swap, and discontinuous) but the classes can be extended to any arbitrary number (Koehn and Monz, 2005). In general, the distribution of phrase orientation is:

$$p(o_k|\bar{f}_i, \bar{e}_i) = \frac{1}{Z}\, h(\bar{f}_i, \bar{e}_i, o_k)\,. \qquad (3)$$

This lexicalized reordering model is estimated by relative frequency where each phrase pair $(\bar{f}_i, \bar{e}_i)$ with such an orientation (o_k) is counted and then normalized to yield the probability as follows:

$$p(o_k|\bar{f}_i, \bar{e}_i) = \frac{\text{count}(\bar{f}_i, \bar{e}_i, o_k)}{\sum_o \text{count}(\bar{f}_i, \bar{e}_i, o)}. \qquad (4)$$

The orientation of a current phrase pair is defined with respect to the previous target phrase. Galley and Manning (2008) extended the model to tackle long-distance reorderings. Their hierarchical model enables phrase movements that are more complex than swaps between adjacent phrases.

3 Multinomial Logistic Regression

Multinomial logistic regression (MLR), also known as maximum entropy classifier (Zens and Ney, 2006), is a probabilistic model for the multiclass problem. The class probability is given by:

$$p(o_k|\bar{f}_i, \bar{e}_i) = \frac{\exp(\mathbf{w}_k^\top \phi(\bar{f}_i, \bar{e}_i))}{\sum_{k'} \exp(\mathbf{w}_{k'}^\top \phi(\bar{f}_i, \bar{e}_i))}, \qquad (5)$$

where $\phi(\bar{f}_i, \bar{e}_i)$ is the feature vector of the i-th phrase pair. An equivalent notation to $\mathbf{w}_k^\top \phi(\bar{f}_i, \bar{e}_i)$ is $\mathbf{w}^\top f(\phi(\bar{f}_i, \bar{e}_i), o_k)$ where $\mathbf{w}$ is a long vector composed of all classes parameters (i.e. $\mathbf{w}^\top = [\mathbf{w}_1^\top \dots \mathbf{w}_K^\top]$) and $f(.,.)$ is a joint feature vector decomposed via the orthogonal feature representation (Rousu et al., 2006). This representation simply means there is no crosstalk between two different feature vectors. For example, $f(\phi(\bar{f}_i, \bar{e}_i), o_1)^\top = [\phi(\bar{f}_i, \bar{e}_i)^\top\, 0 \dots 0]$.

The model's parameters can be estimated by minimizing the following regularized negative log-likelihood $\mathcal{P}(\mathbf{w})$ as follows (Bishop, 2006):

$$\min_{\mathbf{w}} \; \frac{1}{2\sigma^2} \sum_{k=1}^{K} \|\mathbf{w}_k\|^2 - \sum_{i=1}^{N} \sum_{k=1}^{K} \tilde{p}_{ik} \log p(o_k|\bar{f}_i, \bar{e}_i) \qquad (6)$$

Here σ is a penalty parameter and $\tilde{p}$ is the empirical distribution where $\tilde{p}_{ik}$ equals zero for all $o_k \neq o_i$.

Solving the primal optimization problem (6) using the gradient:

$$\frac{\partial \mathcal{P}(\mathbf{w})}{\partial \mathbf{w}_k} = \frac{\mathbf{w}_k}{\sigma^2} - \sum_{i=1}^{N} \big(\tilde{p}_{ik} - p(o_k|\bar{f}_i, \bar{e}_i)\big)\, \phi(\bar{f}_i, \bar{e}_i), \qquad (7)$$

do not constitute a closed-form solution. In our experiments, we used stochastic gradient decent method (i.e. online learning) to estimate $\mathbf{w}$ which is shown to be fast and effictive for large-scale problems (Bottou, 2010). The method approximates (7) by a gradient at a single randomly picked phrase pair. The update rule is:

$$\mathbf{w}'_k = \mathbf{w}_k - \eta_i \nabla_k \mathcal{P}_i(\mathbf{w}), \qquad (8)$$

where η_i is a positive learning rate.

3.1 The Dual Problem

Lebanon and Lafferty (2002) derived an equivalent dual problem to (6). Introducing Lagrange multipliers $\boldsymbol{\alpha}$, the dual becomes

$$\begin{aligned} \min_{\mathbf{w}} \quad & \frac{1}{2\sigma^2}\sum_{k=1}^{K}\|\mathbf{w}_k(\boldsymbol{\alpha})\|^2 + \sum_{i=1}^{N}\sum_{k=1}^{K}\alpha_{ik}\log\alpha_{ik}, \\ \text{s.t.} \quad & \sum_{k=1}^{K}\alpha_{ik} = 1 \ \text{ and } \ \alpha_{\text{ik}} \geq 0\,, \forall \text{i}, \text{k}, \end{aligned} \tag{9}$$

where

$$\mathbf{w}_k(\boldsymbol{\alpha}) = \sigma^2\sum_{i=1}^{N}(\tilde{p}_{ik} - \alpha_{ik})\phi(\bar{f}_i, \bar{e}_i) \tag{10}$$

As mentioned in the introduction, Yu et al. (2011) proposed a two-level dual coordinate descent method to minimize $\mathcal{D}(\boldsymbol{\alpha})$ in (9) but it has some numerical difficulties. Collins et al. (2008) proposed simple exponentiated gradient (EG) algorithm for Conditional Random Feild (CRF). The algorithm is applicable to our problem, a special case of CRF. The rule update is:

$$\alpha'_{ik} = \frac{\alpha_{ik}\exp(-\eta_i\,\nabla_{ik}\mathcal{D}(\boldsymbol{\alpha}))}{\sum_{k'}\alpha_{ik'}\exp(-\eta_i\,\nabla_{ik'}\mathcal{D}(\boldsymbol{\alpha}))} \tag{11}$$

where

$$\begin{aligned} \nabla_{ik}\mathcal{D}(\boldsymbol{\alpha}) &\equiv \frac{\partial\mathcal{D}(\boldsymbol{\alpha})}{\partial\alpha_{ik}} = 1 + \log\alpha_{ik} \\ &+ \left(\mathbf{w}_y(\boldsymbol{\alpha})^\top\phi(\bar{f}_i, \bar{e}_i) - \mathbf{w}_k(\boldsymbol{\alpha})^\top\phi(\bar{f}_i, \bar{e}_i)\right). \end{aligned} \tag{12}$$

Here y represents the true class (i.e. $o_y = o_i$). To improve the convergence, η_i is adaptively adjusted for each example. If the objective function (9) did not decrease, η_i is halved for number of trials (Collins et al., 2008). Calculating the function difference below is the main cost in EG algorithm,

$$\begin{aligned} \mathcal{D}(\boldsymbol{\alpha}') - \mathcal{D}(\boldsymbol{\alpha}) &= \sum_{k=1}^{K}\left(\alpha'_{ik}\log\alpha'_{ik} - \alpha_{ik}\log\alpha_{ik}\right) \\ &- \sum_{k=1}^{K}(\alpha'_{ik} - \alpha_{ik})\mathbf{w}_k(\boldsymbol{\alpha})^\top\phi(\bar{f}_i, \bar{e}_i) \\ &+ \frac{\sigma^2}{2}\|\phi(\bar{f}_i, \bar{e}_i)\|^2\sum_{k=1}^{K}(\alpha'_{ik} - \alpha_{ik})^2. \end{aligned} \tag{13}$$

Clearly, the cost is affordable because $\mathbf{w}_k(\boldsymbol{\alpha})$ is maintained throughout the algorithm as follows:

$$\mathbf{w}_k(\boldsymbol{\alpha}') = \mathbf{w}_k(\boldsymbol{\alpha}) - \sigma^2(\alpha'_{ik} - \alpha_{ik})\phi(\bar{f}_i, \bar{e}_i) \tag{14}$$

Following Yu et al. (2011), we initialize α_{ik} as follows:

$$\alpha_{ik} = \begin{cases} (1-\epsilon) & \text{if } o_k = o_i; \\ \frac{\epsilon}{K-1} & \text{else.} \end{cases} \tag{15}$$

where ϵ is a small positive value. This is because the objective function (9) is not well defined at $\alpha_{ik} = 0$ due to the logarithm appearance.

Finally, the optimal dual variables are achieved when the following condition is satisfied for all examples (Yu et al., 2011):

$$\max_k \nabla_{ik}\mathcal{D}(\boldsymbol{\alpha}) = \min_k \nabla_{ik}\mathcal{D}(\boldsymbol{\alpha}) \tag{16}$$

This condition is the key to accelerate EG algorithm. Unlike the primal problem (6), the dual variables α_{ik} are associated with each example (i.e. phrase pair) therefore a training example can be disregarded once its optimal dual variables obtained. More data shrinking can be achieved by tolerating a small difference between the two values in (16). Algorithm 1 presents the overall procedure (shrinking step is from line 6 to 9).

Algorithm 1 Shrinking stochastic exponentiated gradient method for training the dual problem

Require: training set $S = \{\phi(\bar{f}_i, \bar{e}_i), o_i\}_{i=1}^N$

1: Given $\boldsymbol{\alpha}$ and the corresponding $\mathbf{w}(\boldsymbol{\alpha})$
2: **repeat**
3: Randomly pick i from S
4: Claculate $\nabla_{ik}\mathcal{D}(\boldsymbol{\alpha})\ \forall k$ by (12)
5: $v_i = \max_k \nabla_{ik}\mathcal{D}(\boldsymbol{\alpha}) - \min_k \nabla_{ik}\mathcal{D}(\boldsymbol{\alpha})$
6: **if** $v_i \leq \epsilon$ **then**
7: Remove i from S
8: Continue from line 3
9: **end if**
10: $\eta = 0.5$
11: **for** $t = 1$ **to** maxTrial **do**
12: Calculate $\alpha'_{ik}\ \forall k$ by (11)
13: **if** $\mathcal{D}(\boldsymbol{\alpha}') - \mathcal{D}(\boldsymbol{\alpha}) \leq 0$ **then**
14: Update $\boldsymbol{\alpha}$ and $\mathbf{w}(\boldsymbol{\alpha})$ by (14)
15: Break
16: **end if**
17: $\eta = 0.5\,\eta$
18: **end for**
19: **until** $v_i \leq \epsilon \quad \forall i$

4 Experiments

We used MultiUN which is a large-scale parallel corpus extracted from the United Nations website (Eisele and Chen, 2010). We have used Arabic and English portion of MultiUN where the English side is about 300 million words.

We simplify the problem by classifying phrase movements into three categories (monotone, swap, discontinuous). To train the reordering models, we used GIZA++ to produce word alignments (Och and Ney, 2000). Then, we used the `extract` tool that comes with the Moses toolkit (Koehn et al., 2007) in order to extract phrase pairs along with their orientation classes.

As shown in Table 1, each extracted phrase pair is represented by linguistic features as follows:

- Aligned source and target words in a phrase pair. Each word alignment is a feature.
- Words within a window around the source phrase to capture the context. We choose adjacent words of the phrase boundary.

The extracted phrase pairs after filtering are 47,227,789. The features that occur more than 10 times are 670,154.

Sentence pair:

$\mathbf{f}$: $f_1\ f_2\ {}_1\ f_3\ f_4\ f_5\ {}_2\ f_6\ {}_3$.

$\mathbf{e}$: $e_1\ {}_1\ e_2\ e_3\ {}_3\ e_4\ e_5\ {}_2$.

Extracted phrase pairs $(\bar{\mathbf{f}}, \bar{\mathbf{e}})$:

$\bar{f}_i$	$\bar{e}_i$	o_i	alignment	context
$f_1\ f_2$	e_1	mono	0-0 1-0	f_3
$f_3\ f_4\ f_5$	$e_4\ e_5$	swap	0-1 2-0	$f_2\ f_6$
f_6	$e_2\ e_3$	other	0-0 0-1	f_5

All linguistic features:

1. f_1&e_1 2. f_2&e_1 3. f_3 4. f_3&e_5 5. f_5&e_4
6. f_2 7. f_6 8. f_6&e_2 9. f_6&e_3 10. f_5

Bag-of-words representation:

a phrase pair is represented as a vector where each feature is a discrete number (0=not exist).

$\phi(\bar{f}_i, \bar{e}_i)$	1	2	3	4	5	6	7	8	9	10
$\phi(\bar{f}_1, \bar{e}_1) =$	1	1	1	0	0	0	0	0	0	0
$\phi(\bar{f}_2, \bar{e}_2) =$	0	0	0	1	1	1	1	0	0	0
$\phi(\bar{f}_3, \bar{e}_3) =$	0	0	0	0	0	0	1	1	1	1

Table 1: A generic example of the process of phrase pair extraction and representation.

4.1 Classification

We trained our reordering models by both primal and dual classifiers for 100 iterations. For the dual MLR, different shrinking levels have been tried by varying the parameter (ϵ) in Algorithm 1. Table 2 reports the training time and classification error rate of these models.

Training the dual MLR with moderate shrinking level (i.e. $\epsilon = 0.1$) is almost four times faster than training the primal one. Choosing larger value for (ϵ) leads to faster training but might harm the performance as shown below.

Classifier	Training Time	Error Rate
Primal MLR	1 hour 9 mins	17.81%
Dual MLR ϵ:0.1	18 minutes	17.95%
Dual MLR ϵ:1.0	13 minutes	21.13%
Dual MLR ϵ:0.01	22 minutes	17.89%

Table 2: Performance of the primal and dual MLR based on held-out data.

Figure 1 shows the percentage of active set during training dual MLR with various shrinking levels. Interestingly, the dual MLR could disregard more than 99% of the data after a couple of iterations. For very large corpus, the data might not fit in memory and training primal MLR will take long time due to severe disk-swapping. In this situation, using dual MLR is very beneficial.

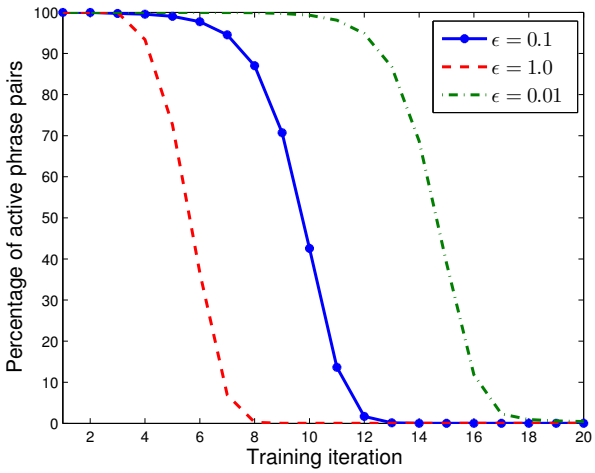

Figure 1: Percentage of active set in dual MLR. As the data size decreases, each iteration takes far less computation time (see Table 2 for total time).

4.2 Translation

We used the Moses toolkit (Koehn et al., 2007) with its default settings to build three phrase-based translation systems. They differ in how their reordering models were estimated. The language model is a 5-gram with interpolation and Kneser-Ney smoothing (Kneser and Ney, 1995). We tuned the system by using MERT technique (Och, 2003).

As commonly used in statistical machine translation, we evaluated the translation performance by BLEU score (Papineni et al., 2002). The test sets are NIST MT06 and MT08 where the English sides are 35,481 words (1056 sentences) and 116,840 words (3252 sentences), respectively. Table 3 shows the BLEU scores for the translation systems. We also computed statistical significance for the models using the *paired bootstrap resampling* method (Koehn, 2004b).

Translation System	MT06	MT08
Baseline + Lexical. model	30.86	34.22
Baseline + Primal MLR	31.37*	34.85*
Baseline + Dual MLR ϵ:0.1	31.36*	34.87*

Table 3: BLEU scores for Arabic-English translation systems with different reordering models (*: better than the lexicalized model with at least 95% statistical significance).

5 Conclusion

In training such system with large data sizes and big dimensionality, computational complexity become a serious issue. In SMT, maximum entropy-based reordering model is often introduced as a better alternative to the commonly used lexicalized one. However, training this discriminative model using large-scale corpus might be computationally expensive due to the iterative learning.

In this paper, we propose training the model using the dual MLR with shrinking method. It is almost four times faster than the primal MLR (also know as MaxEnt) and much more memory-efficient. For very large corpus, the data might not fit in memory and training primal MLR will take long time due to severe disk-swapping. In this situation, using dual MLR is very beneficial. The proposed method is also useful for many classification problems in natural language processing that require large-scale data.

References

Christopher M. Bishop. 2006. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag New York, Inc., Secaucus, NJ, USA.

Léon Bottou. 2010. Large-scale machine learning with stochastic gradient descent. In Yves Lechevallier and Gilbert Saporta, editors, *Proceedings of the 19th International Conference on Computational Statistics (COMPSTAT'2010)*, pages 177–187, Paris, France, August. Springer.

Colin Cherry. 2013. Improved reordering for phrase-based translation using sparse features. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 22–31, Atlanta, Georgia, June. Association for Computational Linguistics.

Michael Collins, Amir Globerson, Terry Koo, Xavier Carreras, and Peter L. Bartlett. 2008. Exponentiated gradient algorithms for conditional random fields and max-margin markov networks. *Journal of Machine Learning Research*, 9:1775–1822, June.

Andreas Eisele and Yu Chen. 2010. Multiun: A multilingual corpus from united nation documents. In Daniel Tapias, Mike Rosner, Stelios Piperidis, Jan Odjik, Joseph Mariani, Bente Maegaard, Khalid Choukri, and Nicoletta Calzolari (Conference Chair), editors, *Proceedings of the Seventh conference on International Language Resources and Evaluation*, pages 2868–2872. European Language Resources Association (ELRA), 5.

Michel Galley and Christopher D. Manning. 2008. A simple and effective hierarchical phrase reordering model. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 848–856, Hawaii, October. Association for Computational Linguistics.

Cho-Jui Hsieh, Kai-Wei Chang, Chih-Jen Lin, S. Sathiya Keerthi, and S. Sundararajan. 2008. A dual coordinate descent method for large-scale linear svm. In *Proceedings of the 25th International Conference on Machine Learning*, ICML '08, pages 408–415.

Reinhard Kneser and Hermann Ney. 1995. Improved backing-off for m-gram language modeling. *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 181–184.

Philipp Koehn and Christof Monz. 2005. Shared task: Statistical machine translation between european languages. In *Proceedings of ACL Workshop on Building and Using Parallel Texts*, pages 119–124. Association for Computational Linguistics.

Philipp Koehn, Amittai Axelrod, Alexandra Birch Mayne, Chris Callison-Burch, Miles Osborne, and David Talbot. 2005. Edinburgh system description

for the 2005 IWSLT speech translation evaluation. In *Proceedings of International Workshop on Spoken Language Translation*, Pittsburgh, PA.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Christopher J. Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the ACL 2007 Demo and Poster Sessions*, pages 177–180.

Philipp Koehn. 2004a. Pharaoh: a beam search decoder for phrase-based statistical machine translation models. In *Proceedings of 6th Conference of the Association for Machine Translation in the Americas (AMTA)*, pages 115—124, Washington DC.

Philipp Koehn. 2004b. Statistical significance tests for machine translation evaluation. In Dekang Lin and Dekai Wu, editors, *Proceedings of EMNLP 2004*, pages 388–395, Barcelona, Spain, July. Association for Computational Linguistics.

Philipp Koehn. 2010. *Statistical Machine Translation*. Cambridge University Press.

Shankar Kumar and William Byrne. 2005. Local phrase reordering models for statistical machine translation. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 161–168, Vancouver, British Columbia, Canada, October. Association for Computational Linguistics.

Guy Lebanon and John D. Lafferty. 2002. Boosting and maximum likelihood for exponential models. In T.G. Dietterich, S. Becker, and Z. Ghahramani, editors, *Advances in Neural Information Processing Systems 14*, pages 447–454. MIT Press.

Vinh Van Nguyen, Akira Shimazu, Minh Le Nguyen, and Thai Phuong Nguyen. 2009. Improving a lexicalized hierarchical reordering model using maximum entropy. In *Proceedings of the Twelfth Machine Translation Summit (MT Summit XII)*. International Association for Machine Translation.

Yizhao Ni, Craig Saunders, Sandor Szedmak, and Mahesan Niranjan. 2011. Exploitation of machine learning techniques in modelling phrase movements for machine translation. *Journal of Machine Learning Research*, 12:1–30, February.

Franz Josef Och and Hermann Ney. 2000. Improved statistical alignment models. In *Proceedings of the 38th Annual Meeting of the Association of Computational Linguistics (ACL)*.

Franz Josef Och and Hermann Ney. 2002. Discriminative training and maximum entropy models for statistical machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*.

Franz Josef Och. 2003. Minimum error rate training in statistical machine translation. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics - Volume 1*, ACL '03, pages 160–167, Stroudsburg, PA, USA. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, and Todd Ward. 1998. Maximum likelihood and discriminative training of direct translation models. In *Proceedings of ICASSP*, pages 189–192.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pages 311–318, Stroudsburg, PA, USA. Association for Computational Linguistics.

Juho Rousu, Craig Saunders, Sandor Szedmak, and John Shawe-Taylor. 2006. Kernel-based learning of hierarchical multilabel classification models. *Journal of Machine Learning Research*, pages 1601–1626.

Christoph Tillmann. 2004. A unigram orientation model for statistical machine translation. In *Proceedings of HLT-NAACL: Short Papers*, pages 101–104.

Bing Xiang, Niyu Ge, and Abraham Ittycheriah. 2011. Improving reordering for statistical machine translation with smoothed priors and syntactic features. In *Proceedings of SSST-5, Fifth Workshop on Syntax, Semantics and Structure in Statistical Translation*, pages 61–69, Portland, Oregon, USA. Association for Computational Linguistics.

Deyi Xiong, Qun Liu, and Shouxun Lin. 2006. Maximum entropy based phrase reordering model for statistical machine translation. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the ACL*, pages 521–528, Sydney, July. Association for Computational Linguistics.

Hsiang-Fu Yu, Fang-Lan Huang, and Chih-Jen Lin. 2011. Dual coordinate descent methods for logistic regression and maximum entropy models. *Machine Learning*, 85(1-2):41–75, October.

Richard Zens and Hermann Ney. 2006. Discriminative reordering models for statistical machine translation. In *Proceedings on the Workshop on Statistical Machine Translation*, pages 55–63, New York City, June. Association for Computational Linguistics.