Bayesian Reordering Model with Feature Selection

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

In phrase-based statistical machine translation systems, variation in grammatical structures between source and target languages can cause large movements of phrases. Modeling such movements is crucial in achieving translations of long sentences that appear natural in the target language. We explore generative learning approach to phrase reordering in Arabic to English. Formulating the reordering problem as a classification problem and using naive Bayes with feature selection, we achieve an improvement in the BLEU score over a lexicalized reordering model. The proposed model is compact, fast and scalable to a large corpus.

1 Introduction

Currently, the dominant approach to machine translation is statistical, starting from the mathematical formulations and algorithms for parameter estimation (Brown et al., 1988), further extended in (Brown et al., 1993). These early models, widely known as the IBM models, were word-based. Recent extensions note that a better approach is to group collections of words, or phrases, for translation together, resulting in a significant focus these days on phrase-based statistical machine translation systems.

To deal with the alignment problem of one-to-many word alignments in the IBM model formulation, whereas phrase-based models may have many-to-many translation relationships, IBM models are trained in both directions, source to target and target to source, and their word alignments are combined (Och and Ney, 2004).

While phrase-based systems are a significant improvement over word-based approaches, a particular issue that emerges is long-range reorderings at the phrase level (Galley and Manning, 2008). Analogous to speech recognition systems, translation systems relied on language models to produce more fluent translation. While early work penalized phrase movements without considering reorderings arising from vastly differing grammatical structures across language pairs like Arabic-English, many researchers considered lexical reordering models that attempted to learn orientation based on content (Tillmann, 2004; Kumar and Byrne, 2005; Koehn et al., 2005). These approaches may suffer from the data sparseness problem since many phrase pairs occur only once (Nguyen et al., 2009).

As an alternative way of exploiting function approximation capabilities offered by machine learning methods, there is recent interest in formulating a learning problem that aims to predict reordering from linguistic features that capture their context. An example of this is the maximum entropy method used by (Xiang et al., 2011; Nguyen et al., 2009; Zens and Ney, 2006; Xiong et al., 2006).

In this work we apply a naive Bayes classifier, combined with feature selection to address the reordering problem. To the best of our knowledge, this simple model of classification has not been used in this context previously. We present empirical results comparing our work and previously proposed lexicalized reordering model. We show that our model is scalable to large corpora.

The remainder of this paper is organized as follows. Section 2 discusses previous work in the field and how that is related to our paper. Section 3 gives an overview of the baseline translation system. Section 4 introduces the Bayesian reordering model and gives details of different inference methods, while, Section 5 describes feature selection method. Section 6 presents the experiments and reports the results evaluated as classification and translation problems. Finally, we end the paper with a summary of our conclusions and perspectives.

Proceedings of the Ninth Workshop on Statistical Machine Translation, pages 477–485,
Baltimore, Maryland USA, June 26–27, 2014. ©2014 Association for Computational Linguistics

Symbol	Notation
f/**e**	a source / target sentence (string)
$\bar{\mathbf{f}}/\bar{\mathbf{e}}$	a source / target phrase sequence
N	the number of examples
K	the number of classes
$(\bar{f}_n, \bar{e}_n)$	the n-th phrase pair in $(\bar{\mathbf{f}}, \bar{\mathbf{e}})$
o_n	the orientation of $(\bar{f}_n, \bar{e}_n)$
$\phi(\bar{f}_n, \bar{e}_n)$	the feature vector of $(\bar{f}_n, \bar{e}_n)$

Table 1: Notation used in this paper.

2 Related Work

The phrase reordering model is a crucial component of any translation system, particularly between language pairs with different grammatical structures (e.g. Arabic-English). Adding a lexicalized reordering model consistently improved the translation quality for several language pairs (Koehn et al., 2005). The model tries to predict the orientation of a phrase pair with respect to the previous adjacent target words. Ideally, the reordering model would predict the right position in the target sentence given a source phrase, which is difficult to achieve. Therefore, positions are grouped into limited orientations or classes.The orientation probability for a phrase pair is simply based on the relative occurrences in the training corpus.

The lexicalized reordering model has been extended to tackle long-distance reorderings (Galley and Manning, 2008). This takes into account the hierarchical structure of the sentence when considering such an orientation. Certain examples are often used to motivate syntax-based systems were handled by this hierarchical model, and this approach is shown to improve translation performance for several translation tasks with small computational cost.

Despite the fact that the lexicalized reordering model is always biased towards the most frequent orientation for such a phrase pair, it may suffer from a data sparseness problem since many phrase pairs occur only once. Moreover, the context of a phrase might affect its orientation, which is not considered as well.

Adopting the idea of predicting orientation based on content, it has been proposed to represent each phrase pair by linguistic features as reordering evidence, and then train a classifier for prediction. The maximum entropy classifier is a popular choice among many researchers (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). Recently, Cherry (2013) proposed using sparse features optimize BLEU with the decoder instead of training a classifier independently.

We distinguish our work from the previous ones in the following. We propose a fast reordering model using a naive Bayes classifier with feature selection. In this study, we undertake a comparison between our work and lexicalized reordering model.

3 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:

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

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):

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

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 (i.e. $n = 10$).

In phrase-based systems, the translation model can capture the local meaning for each source phrase. However, to capture the whole meaning of a sentence, its translated phrases need to be in the correct order. The language model, which ensures fluent translation, plays an important role in reordering; however, it prefers sentences that are grammatically correct without considering their actual meaning. Besides that, it has a bias towards short translations (Koehn, 2010). Therefore, developing a reordering model will improve the accuracy particularly when translating between two grammatically different languages.

3.1 Lexicalized Reordering Model

Phrase 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:

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

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

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

The orientation class of a current phrase pair is defined with respect to the previous target word or phrase (i.e. word-based classes or phrase-based classes). In the case of three categories (monotone, swap, and discontinuous): monotone is the previous source phrase (or word) that is previously adjacent to the current source phrase, swap is the previous source phrase (or word) that is next-adjacent to the current source phrase, and discontinuous is not monotone or swap.

Galley and Manning (2008) extended the lexicalized reordering mode to tackle long-distance phrase reorderings. Their hierarchical model enables phrase movements that are more complex than swaps between adjacent phrases.

4 Bayesian Reordering Model

Many feature-based reordering models have been proposed to replace the lexicalized reordering model. The reported results showed consistent improvement in terms of various translation metrics.

Naive Bayes method has been a popular classification model of choice in many natural language processing problems (e.g. text classification). Naive Bayes is a simple classifier that ignores correlation between features, but has the appeal of computational simplicity. It is a generative probabilistic model based on Bayes' theorem as below:

$$p(o_k|\bar{f}_n, \bar{e}_n) = \frac{p(\bar{f}_n, \bar{e}_n|o_k)p(o_k)}{\sum_o p(\bar{f}_n, \bar{e}_n|o)p(o)}. \tag{5}$$

The class prior can be estimated easily as a relative frequency (i.e. $p(o_k) = \frac{N_k}{N}$). The likelihood distribution $p(\bar{f}_n, \bar{e}_n|o_k)$ is defined based on the type of data. The classifier will be naive if we assume that feature variables are conditionally independent. The naive assumption simplifies our distribution and hence reduces the parameters that have to be estimated. In text processing, multinomial is used as a class-conditional distribution (Rogers and Girolami, 2011). The distribution is defined as:

$$p(\bar{f}_n, \bar{e}_n|\mathbf{q}) = C \prod_m q_m^{\phi_m(\bar{f}_n, \bar{e}_n)} \tag{6}$$

where C is a multinomial coefficient,

$$C = \frac{(\sum_m \phi_m(\bar{f}_n, \bar{e}_n))!}{\prod_m \phi_m(\bar{f}_n, \bar{e}_n)!}, \tag{7}$$

and $\mathbf{q}$ are a set of parameters, each of which is a probability. Estimating these parameters for each class by maximum likelihood,

$$\arg\max_{\mathbf{q}_k} \prod_n^{N_k} p(\bar{f}_n, \bar{e}_n|\mathbf{q}_k), \tag{8}$$

will result in (Rogers and Girolami, 2011):

$$q_{km} = \frac{\sum_n^{N_k} \phi_m(\bar{f}_n, \bar{e}_n)}{\sum_{m'}^M \sum_n^{N_k} \phi_{m'}(\bar{f}_n, \bar{e}_n)}. \tag{9}$$

MAP estimate It is clear that q_{km} might be zero which means the probability of a new phrase pair with nonzero feature $\phi_m(\bar{f}_n, \bar{e}_n)$ is always zero because of the product in (6). Putting a prior over $\mathbf{q}$ is one smoothing technique. A conjugate prior for the multinomial likelihood is the Dirichlet distribution and the MAP estimate for q_{km} is (Rogers and Girolami, 2011):

$$q_{km} = \frac{\alpha - 1 + \sum_n^{N_k} \phi_m(\bar{f}_n, \bar{e}_n)}{M(\alpha - 1) + \sum_{m'}^M \sum_n^{N_k} \phi_{m'}(\bar{f}_n, \bar{e}_n)} \tag{10}$$

where M is the feature vector's length or the feature dictionary size and α is a Dirichlet parameter with a value greater than one. The derivation is in Appendix A.

Bayesian inference Instead of using a point estimate of $\mathbf{q}$ as shown previously in equation (10), Bayesian inference is based on the whole parameter space in order to incorporate uncertainty into our multinomial model. This requires a posterior probability distribution over $\mathbf{q}$ as follows:

$$p(\bar{f}_n, \bar{e}_n|o_k) = \int p(\bar{f}_n, \bar{e}_n|\mathbf{q}_k) p(\mathbf{q}_k|\boldsymbol{\alpha}_k)\, \mathrm{d}\mathbf{q}_k$$
$$= C\, \frac{\Gamma\left(\sum_m \alpha_{km}\right)}{\prod_m \Gamma(\alpha_{km})} \frac{\prod_m \Gamma(\alpha_{km} + \phi_m(\bar{f}_n, \bar{e}_n))}{\Gamma\left(\sum_m \alpha_{km} + \phi_m(\bar{f}_n, \bar{e}_n)\right)}. \tag{11}$$

Here $\boldsymbol{\alpha}_k$ are new hyperparameters of the posterior derived by means of Bayes theorem as follows:

$$p(\mathbf{q}_k|\boldsymbol{\alpha}_k) = \frac{p(\mathbf{q}_k|\boldsymbol{\alpha}) \prod_n^{N_k} p(\bar{f}_n, \bar{e}_n|\mathbf{q}_k)}{\int p(\mathbf{q}_k|\boldsymbol{\alpha}) \prod_n^{N_k} p(\bar{f}_n, \bar{e}_n|\mathbf{q}_k) \mathrm{d}\mathbf{q}_k}. \tag{12}$$

The solution of (11) will result in:

$$\boldsymbol{\alpha}_k = \boldsymbol{\alpha} + \sum_n^{N_k} \Phi(\bar{f}_n, \bar{e}_n). \tag{13}$$

For completeness we give a summary of derivations of equations (11) and (13) in Appendix B, more detailed discussions can be found in (Barber, 2012).

5 Feature Selection

In several high dimensional pattern classification problems, there is increasing evidence that the discriminant information may be in small subspaces, motivating feature selection (Li and Niranjan, 2013). Having irrelevant or redundant features could affect the classification performance (Liu and Motoda, 1998). They might mislead the learning algorithms or overfit them to the data and thus have less accuracy.

The aim of feature selection is to find the optimal subset features which maximize the ability of prediction, which is the main concern, or simplify the learned results to be more understandable. There are many ways to measure the goodness of a feature or a subset of features; however the criterion will be discussed is mutual information.

5.1 Mutual Information

Information criteria are based on the concept of entropy which is the amount of randomness. The distribution of a fair coin, for example, is completely random so the entropy of the coin is very high. The following equation calculates the entropy of a variable X (MacKay, 2002):

$$H\left(X\right) = -\sum_x p(x) \log p(x). \tag{14}$$

The mutual information of a feature X can be measured by calculating the difference between the prior uncertainty of the class variable Y and the posterior uncertainty after using the feature as follows (MacKay, 2002):

$$I(X;Y) = H(Y) - H(Y|X) \tag{15}$$
$$= \sum_{x,y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)}.$$

The advantage of mutual Information over other criteria is the ability to detect nonlinear patterns. The disadvantage is its bias towards higher arbitrary features; however this problem can be solved by normalizing the information as follows (Estévez et al., 2009):

$$I_{norm}(X;Y) = \frac{I(X;Y)}{\min(H(X), H(Y))}. \tag{16}$$

6 Experiments

The corpus used in our experiments is MultiUN which is a large-scale parallel corpus extracted from the United Nations website[1] (Eisele and Chen, 2010). We have used Arabic and English portion of MultiUN. Table 2 shows the general statistics.

Statistics	Arabic	English
Sentence Pairs	9.7 M	
Running Words	255.5 M	285.7 M
Word/Line	22	25
Vocabulary Size	677 K	410 K

Table 2: General statistics of Arabic-English MultiUN (M: million, K: thousand).

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 [2] toolkit (Koehn et al., 2007) in order to extract phrase pairs along with their orientation classes.

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.

[1] http://www.ods.un.org/ods/

[2] Moses is an open source toolkit for statistical machine translation (www.statmt.org/moses/).

- Words within a window around the source phrase to capture the context. We choose adjacent words of the phrase boundary.

Most researchers build one reordering model for the whole training set (Zens and Ney, 2006; Xiong et al., 2006; Nguyen et al., 2009; Xiang et al., 2011). Ni et al. (Ni et al., 2011) simplified the learning problem to have as many sub-models as source phrases. Training data were divided into small independent sets where samples having the same source phrase are considered a training set. In our experiments, we have chosen the first method.

We compare lexicalized and Bayesian reordering models in two phases. In the classification phase, we see the performance of the models as a classification problem. In the translation phase, we test the actual impact of these reordering models in a translation system.

6.1 Classification

We built naive Bayes classifier with both MAP estimate and Bayesian inference. We also used mutual Information in order to select the most informative features for our classification task.

Table 3 reports the error rate of the reordering models compared to the lexicalized reordering model. All experiments reported here were repeated three times to evaluate the uncertainties in our results. The results shows that there is no advantage to using Bayesian inference instead of MAP estimate.

Classifier	Error Rate
Lexicalized model	25.2%
Bayes-MAP estimate	19.53%
Bayes-Bayesian inference	20.13%

Table 3: Classification error rate of both lexicalized and Bayesian models.

The feature selection process reveals that many features have low mutual information. Hence they are not related to the classification task and can be excluded from the model. Figure 1 shows the normalized mutual information for all extracted features.

A ranking threshold for selecting features based on their mutual information is specified experimentally. In Figure 2, we tried different thresholds ranging from 0.001 to 0.05 and measure the error rate after each reduction. Although there is no much gain in terms of performance but the Bayesian model maintains low error rate when the proportion of selected features is low. The model with almost half of the feature space is as good as the one with full feature space.

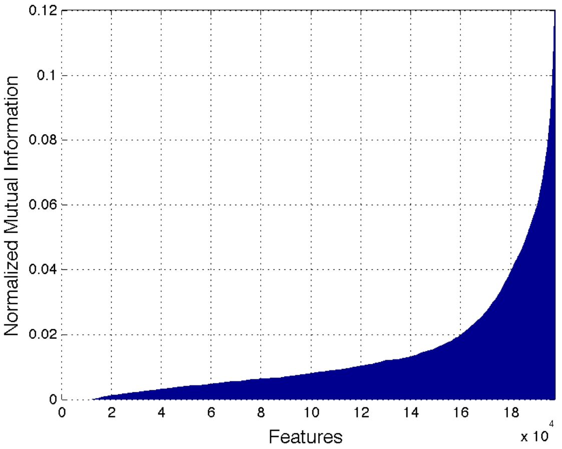

Figure 1: Normalized mutual information for all extracted features (ranked from lowest to highest).

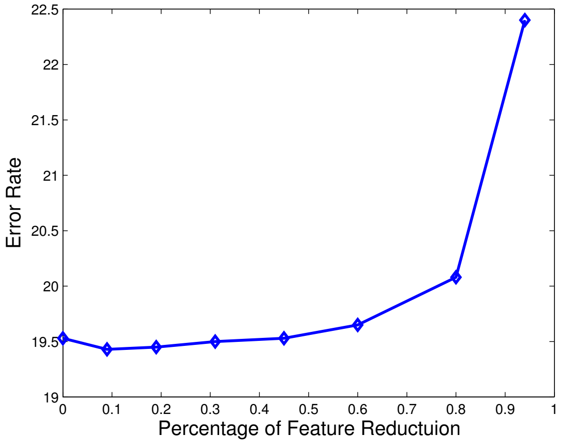

Figure 2: Classification error rate of the Baysien model with different levels of feature reduction.

6.2 Translation

6.2.1 Experimental Design

We used the Moses toolkit (Koehn et al., 2007) with its default settings. 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).

We built four Arabic-English translation systems. Three systems differ in how their reordering models were estimated and the fourth system is a baseline system without reordering model. In all cases, orientation extraction is hierarchical-based since it is the best approach while orientations are monotone, swap and discontinuous. The model is trained in Moses by specifying the configuration string `hier-msd-backward-fe`.

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 NIST MT08. Table 4 shows statistics of development and test sets. We also computed statistical significance for the proposed models using the *paired bootstrap resampling* method (Koehn, 2004).

Evaluation Set		Arabic	English
Development	sentences	696	696
	words	19 K	21 K
NIST MT06	sentences	1797	7188
	words	49 K	223 K
NIST MT08	sentences	813	3252
	words	25 K	117 K

Table 4: Statistics of development and test sets. The English side in NIST is larger because there are four translations for each Arabic sentence.

6.2.2 Results

We first demonstrate in Table 5 a general comparison of the proposed model and the lexicalized model in terms of disc size and average speed in a translation system. The size of Bayesian model is far smaller. The lexicalized model is slightly faster than the Bayesian model because we have overhead computational cost to extract features and compute the orientation probabilities. However, the disc size of our model is much smaller which makes it more efficient practically for large-scale tasks.

Model	Size (MB)	Speed (s/sent)
Lexicalized model	604	2.2
Bayesian model	18	2.6

Table 5: Disc size and average speed of the reordering models in a translation system.

Table 6 shows the BLEU scores for the translation systems according to two test sets. The baseline system has no reordering model. In the two test sets, our Bayesian reordering model is better than the lexicalized one with at least 95% statistical significance. As we have seen in the classification section, Bayes classifier with Bayesian inference has no advantage over MAP estimate.

Translation System	MT06	MT08
Baseline	28.92	32.13
BL+ Lexicalized model	30.86	34.22
BL+ Bayes-MAP estimate	**31.21***	**34.72***
BL+ Bayes-Baysien inference	31.20	34.69

Table 6: BLEU scores for Arabic-English translation systems (*: better than the baseline with at least 95% statistical significance).

7 Conclusion

In this paper, we have presented generative modeling approach to phrase reordering in machine translation. We have experimented with translation from Arabic to English and shown improvements over the lexicalized model of estimating probabilities as relative frequencies of phrase movements. Our proposed Bayesian model with feature selection is shown to be superior. The training time of the model is as fast as the lexicalized model. Its storage requirement is many times smaller which makes it more efficient practically for large-scale tasks.

The feature selection process reveals that many features have low mutual information. Hence they are not related to the classification task and can be excluded from the model. The model with almost half of the feature space is as good as the one with full feature space.

Previously proposed discriminative models might achieve higher score than the reported results. However, our model is scalable to large-scale systems since parameter estimation require only one pass over the data with limited memory (i.e. no iterative learning). This is a critical advantage over discriminative models.

Our current work focuses on three issues. The first is improving the translation speed of the proposed model. The lexicalized model is slightly faster. The second is using more informative features. We plan to explore part-of-speech information, which is more accurate in capturing content. Finally, we will explore different feature selection methods. In our experiments, feature reduction is based on univariate ranking which is riskier than multivariate ranking. This is because useless feature can be useful with others.

References

D. Barber. 2012. *Bayesian Reasoning and Machine Learning*. Cambridge University Press.

P. Brown, J. Cocke, S. Della Pietra, V. Della Pietra, F. Jelinek, R. Mercer, and P. Roossin. 1988. A statistical approach to language translation. In *12th International Conference on Computational Linguistics (COLING)*, pages 71–76.

P. Brown, V. Pietra, S. Pietra, and R. Mercer. 1993. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19(2):263–311.

C. 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.

A. Eisele and Y. 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.

P. Estévez, M. Tesmer, C. Perez, and J. Zurada. 2009. Normalized mutual information feature selection. *Trans. Neur. Netw.*, 20(2):189–201, February.

M. Galley and C. 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.

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

P. Koehn and C. 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.

P. Koehn, A. Axelrod, A. Mayne, C. Callison-Burch, M. Osborne, and D. Talbot. 2005. Edinburgh system description for the 2005 IWSLT speech translation evaluation. In *Proceedings of International Workshop on Spoken Language Translation*, Pittsburgh, PA.

P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, and E. Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the ACL 2007 Demo and Poster Sessions*, pages 177–180.

P. Koehn. 2004. 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.

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

S. Kumar and W. 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.

Hongyu Li and M. Niranjan. 2013. Discriminant subspaces of some high dimensional pattern classification problems. In *IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*, pages 27–32.

H. Liu and H. Motoda. 1998. *Feature Selection for Knowledge Discovery and Data Mining*. Kluwer Academic Publishers, Norwell, MA, USA.

D. MacKay. 2002. *Information Theory, Inference & Learning Algorithms*. Cambridge University Press, New York, NY, USA.

V. Nguyen, A. Shimazu, M. Nguyen, and T. 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.

Y. Ni, C. Saunders, S. Szedmak, and M. Niranjan. 2011. Exploitation of machine learning techniques in modelling phrase movements for machine translation. *Journal of Machine Learning Research*, 12:1–30, February.

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

F. Och and H. 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)*.

F. Och and H. Ney. 2004. The alignment template approach to statistical machine translation. *Computational Linguistics*, 30(4):417–449.

F. 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.

K. Papineni, S. Roukos, and T. Ward. 1998. Maximum likelihood and discriminative training of direct translation models. In *Proceedings of ICASSP*, pages 189–192.

K. Papineni, S. Roukos, T. Ward, and W. 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.

S. Rogers and M. Girolami. 2011. *A First Course in Machine Learning*. Chapman & Hall/CRC, 1st edition.

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

B. Xiang, N. Ge, and A. 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.

D. Xiong, Q. Liu, and S. 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.

R. Zens and H. 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.

A MAP Estimate Derivation

Multinomial distribution is defined as:

$$p(\mathbf{x}|\mathbf{q}) = C\prod_m q_m^{x_m} \tag{17}$$

where C is a multinomial coefficient,

$$C = \frac{(\sum_m x_m)!}{\prod_m x_m!}, \tag{18}$$

and q_m is an event probability ($\sum_m q_m = 1$).

A maximum a posteriori probability (MAP) estimate requires a prior over $\mathbf{q}$. Dirichlet distribution is a conjugate prior and is defined as:

$$p(\mathbf{q}|\boldsymbol{\alpha}) = \frac{\Gamma\left(\sum_m \alpha_m\right)}{\prod_m \Gamma(\alpha_m)}\prod_m q_m^{\alpha_m - 1} \tag{19}$$

where α_m is is a parameter with a positive value.

Finding the MAP estimate for $\mathbf{q}$ given a data is as follows:

$$\begin{aligned}
\mathbf{q}^* &= \arg\max_{\mathbf{q}} p(\mathbf{q}|\boldsymbol{\alpha}, \mathbf{X}) \\
&= \arg\max_{\mathbf{q}} \{p(\mathbf{q}|\boldsymbol{\alpha})p(\mathbf{X}|\mathbf{q})\} \\
&= \arg\max_{\mathbf{q}} \left\{p(\mathbf{q}|\boldsymbol{\alpha})\prod_n p(\mathbf{x}_n|\mathbf{q})\right\} \\
&= \arg\max_{\mathbf{q}} \left\{\prod_m q_m^{\alpha_m-1}\prod_{n,m} q_m^{x_{nm}}\right\} \\
&= \arg\max_{\mathbf{q}} \left\{\sum_m \log q_m^{\alpha_m-1} + \sum_{n,m}\log q_m^{x_{nm}}\right\}.
\end{aligned} \tag{20}$$

Since our function is subject to constraints ($\sum_m q_m = 1$), we introduce Lagrange multiplier as follows:

$$f(\mathbf{q}) = \sum_m \log q_m^{\alpha_m-1} + \sum_{n,m}\log q_m^{x_{nm}} - \lambda(\sum_m q_m - 1). \tag{21}$$

Now we can find $\mathbf{q}^*$ by taking the partial derivative with respect to one variable q_m:

$$\begin{aligned}
\frac{\partial f(\mathbf{q})}{\partial q_m} &= \frac{\alpha_m - 1 + \sum_n x_{nm}}{q_m} - \lambda \\
q_m &= \frac{\alpha_m - 1 + \sum_n x_{nm}}{\lambda}.
\end{aligned} \tag{22}$$

Finally, we sum both sides over M to find λ :

$$\begin{aligned}
\lambda\sum_m q_m &= \sum_m\left(\alpha_m - 1 + \sum_n x_{nm}\right) \\
\lambda &= \sum_m(\alpha_m - 1) + \sum_{n,m} x_{nm}.
\end{aligned} \tag{23}$$

The solution can be simplified by choosing the same value for each α_m which will result in:

$$q_m = \frac{\alpha - 1 + \sum_n x_{nm}}{M(\alpha - 1) + \sum_{n,m'} x_{nm'}}. \tag{24}$$

B Bayesian Inference Derivation

In Appendix A, the inference is based on a single point estimate of $\mathbf{q}$ that has the highest posterior probability. However, it can be based on the whole parameter space to incorporate uncertainty. The probability of a new data point marginalized over the posterior as follows:

$$p(\mathbf{x}|\boldsymbol{\alpha}, \mathbf{X}) = \int p(\mathbf{x}|\mathbf{q}) p(\mathbf{q}|\boldsymbol{\alpha}, \mathbf{X}) \, \mathrm{d}\mathbf{q}, \tag{25}$$

$$p(\mathbf{q}|\boldsymbol{\alpha}, \mathbf{X}) = \frac{p(\mathbf{q}|\boldsymbol{\alpha}) p(\mathbf{X}|\mathbf{q})}{\int p(\mathbf{q}|\boldsymbol{\alpha}) p(\mathbf{X}|\mathbf{q}) \mathrm{d}\mathbf{q}}. \tag{26}$$

Since Dirichlet and Multinomial distributions are conjugate pairs, they form the same density as the prior. Therefore the posterior is also Dirichlet. Now we can expand the posterior expression and re-arrange it to look like a Dirichlet as follows:

$$\begin{aligned} p(\mathbf{q}|\boldsymbol{\alpha}, \mathbf{X}) &\propto p(\mathbf{q}|\boldsymbol{\alpha}) \prod_n p(\mathbf{x}_n|\mathbf{q}) \\ &\propto \prod_m q_m^{\alpha_m - 1} \prod_n \prod_m q_m^{x_{nm}} \\ &\propto \prod_m q_m^{(\alpha_m + \sum_n x_{nm}) - 1}. \end{aligned} \tag{27}$$

The new hyperparameters of the posterior is:

$$\alpha^*_m = \alpha_m + \sum_n x_{nm}. \tag{28}$$

Finally, we expand and re-arrange Dirichlet and multinomial distributions inside the integral in (25) as follows:

$$\begin{aligned} &p(\mathbf{x}|\boldsymbol{\alpha}, \mathbf{X}) = \\ &\quad \int C \prod_m q_m^{x_m} \frac{\Gamma\left(\sum_m \alpha^*_m\right)}{\prod_m \Gamma(\alpha^*_m)} \prod_m q_m^{\alpha^*_m - 1} \, \mathrm{d}\mathbf{q} \\ &= C \frac{\Gamma\left(\sum_m \alpha^*_m\right)}{\prod_m \Gamma(\alpha^*_m)} \int \prod_m q_m^{\alpha^*_m + x_m - 1} \, \mathrm{d}\mathbf{q}. \end{aligned} \tag{29}$$

Note that inside the integral looks a Dirichlet without a normalizing constant. If we multiply and divide by its normalizing constant (i.e. Beta function), the integral is going to be one because it is a density function, resulting in:

$$\begin{aligned} &p(\mathbf{x}|\boldsymbol{\alpha}, \mathbf{X}) = C \frac{\Gamma\left(\sum_m \alpha^*_m\right)}{\prod_m \Gamma(\alpha^*_m)} \\ &\quad \mathrm{B}(\boldsymbol{\alpha}^* + \mathbf{x}) \int \frac{1}{\mathrm{B}(\boldsymbol{\alpha}^* + \mathbf{x})} \prod_m q_m^{\alpha^*_m + x_m - 1} \, \mathrm{d}\mathbf{q}_c \\ &= C \frac{\Gamma\left(\sum_m \alpha^*_m\right)}{\prod_m \Gamma(\alpha^*_m)} \mathrm{B}(\boldsymbol{\alpha}^* + \mathbf{x}) \\ &= C \frac{\Gamma\left(\sum_m \alpha^*_m\right)}{\prod_m \Gamma(\alpha^*_m)} \frac{\prod_m \Gamma(\alpha^*_m + x_m)}{\Gamma\left(\sum_m (\alpha^*_m + x_m)\right)}. \end{aligned} \tag{30}$$