PBML

The Prague Bulletin of Mathematical Linguistics
NUMBER 103 APRIL 2015 65–84

Scalable Reordering Models for SMT based on Multiclass SVM

Abdullah Alrajeh[a b], Mahesan Niranjan[a]

[a] School of Electronics and Computer Science, University of Southampton
[b] Computer Research Institute, King Abdulaziz City for Science and Technology (KACST)

Abstract

In state-of-the-art phrase-based statistical machine translation systems, modelling phrase reorderings is an important need to enhance naturalness of the translated outputs, particularly when the grammatical structures of the language pairs differ significantly. Posing phrase movements as a classification problem, we exploit recent developments in solving large-scale multiclass support vector machines. Using dual coordinate descent methods for learning, we provide a mechanism to shrink the amount of training data required for each iteration. Hence, we produce significant computational saving while preserving the accuracy of the models. Our approach is a couple of times faster than maximum entropy approach and more memory-efficient (50% reduction). Experiments were carried out on an Arabic-English corpus with more than a quarter of a billion words. We achieve BLEU score improvements on top of a strong baseline system with sparse reordering features.

1. Introduction

The mathematical basis of statistical machine translation (SMT) has its origins in the formulation due to Brown et al. (1988), who later introduced five statistical models widely known as the IBM models (Brown et al., 1993). While these early models were word-based, assuming the translation to take place on a word by word basis, in reality, groups of words (phrases) are recognised as better units of translation (Koehn, 2010).

Working at the phrase level helps resolve many ambiguities that occur at the word level. Since the IBM models allow one to many mappings of words, phrase can be automatically defined by training IBM word alignment models in both direction of source and target languages, and combining the two alignments (Och and Ney, 2004).

© 2015 PBML. Distributed under CC BY-NC-ND. Corresponding author: asar1a10@ecs.soton.ac.uk
Cite as: Abdullah Alrajeh, Mahesan Niranjan. Scalable Reordering Models for SMT based on Multiclass SVM. The Prague Bulletin of Mathematical Linguistics No. 103, 2015, pp. 65–84. doi: 10.1515/pralin-2015-0004.

Unauthentifiziert | Heruntergeladen 14.11.19 15:10 UTC

While such attempts at phrase level translation has shown improvement in translation performance, a further issue that has to be addressed is that of long range phrase reorderings (Galley and Manning, 2008). Such reorderings arise from differences in grammatical structures between language pairs and addressing this is important in achieving increased naturalness of the translated output (Koehn, 2010). This issue is particularly pronounced when language pairs separated by large evolutionary distances, or from different linguistic families, are considered such as Arabic and English.

Early work on handling phrase reorderings implemented a relaxation into the decoder which, instead of forcing phrases to be in synchrony, allowed a penalty function that penalised large movements proportionately (Koehn, 2004a). An alternative approach, adopted by several systems nowadays is lexicalised reordering modelling (Tillmann, 2004; Kumar and Byrne, 2005; Koehn et al., 2005), whereby the frequencies of relative positions of the phrase pairs are extracted from the training corpus and used as additional inputs to the decoder (see section 4).

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 modelled as weights of a linear classifier trained by entropic criteria. This maximum entropy-based approach (so called MaxEnt) is a popular choice (Zens and Ney, 2006; Xiong et al., 2006; Nguyen et al., 2009; Xiang et al., 2011).

However, if the underlying classification problem is not linearly separable, the MaxEnt classifier will not perform well and more advanced nonlinear methods will be needed. Kernel methods (such as support vector machines in the context of pattern recognition) are state-of-the-art approaches to capture nonlinear effects in datasets (Cristianini and Shawe-Taylor, 2000). They map the data into high dimensional spaces implicitly defined by properties of the chosen kernel, and achieve linear separability in the transformed space.

In many natural language processing problems, including the phrase reordering problem we address here, context information extracted from data are represented explicitly in very high dimensional spaces and linear separability in these spaces can be expected. Motivated by this, Ni et al. (2011) proposed the use of a structured perceptron approach to tackle long range phrase reorderings. While that system results in encouraging results on a Chinese-English translation task, dimensionality and the resulting computational complexity were noted as issues that needed to be tackled.

More recently, there have been extensive developments in the machine learning literature on scaling up support vector machines to problems with large data sizes. The underlying quadratic programming problem is being solved by stochastic gradient search type algorithms. Many researchers proposed fast learning techniques for linear SVM using a dual coordinate descent approach (Hsieh et al., 2008; Glasmachers and Dogan, 2013; Alrajeh et al., 2015). The method of Hsieh et al. (2008), for example, has linear complexity and reaches an ϵ-accurate solution in $O(\log(1/\epsilon))$ iterations. Later, Chang et al. (2010) took the approach a step further and applied linear SVM to

Unauthentifiziert | Heruntergeladen 14.11.19 15:10 UTC

the explicit form of low-degree polynomial kernel. Although, in many cases, kernel mapping is exponential to the input space or infinite as in the Gaussian kernel, the approach is shown to be useful for certain datasets such as NLP task on dependency parsing (Chang et al., 2010).

In this paper, we explore computationally fast and memory-efficient uses of multiclass SVM classifier as a model of long range phrase reorderings. Our results show significant improvement in the BLEU score over a lexicalised reordering model. Training multiclass SVM is shown to be faster than MaxEnt with 50% reduction in memory usage due to a shrinking heuristic we propose.

The remainder of this paper is organised as follows. Section 2 discusses previous work in the field and how it relates to our reordering model. Section 3 gives an overview of the baseline translation system. Section 4 and 5 briefly describe the lexicalised and maximum entropy-based reordering models. Section 6 introduces the proposed SVM-based reordering model. Starting from a brief introduction to the SVM formulation, we explain a fast learning technique for linear mulitclass SVM and how it is extended to nonlinear using kernel mapping. Section 7 evaluates multiclass SVM on benchmark datasets. Section 8 undertakes a comparison between our work and previously proposed models and reports the results evaluated as classification and translation problems. The experiments are based on a large-scale Arabic-English corpus. Finally, we end the paper with a summary of our conclusions and perspectives.

2. Related Work

Adding a lexicalised reordering model has been shown to consistently improve 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 lexicalised 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. This approach is shown to improve translation performance for several translation tasks. An additional appeal of the method is the low computing cost.

In addition to the fact that the lexicalised reordering model is always biased toward the most frequent orientation for such a phrase pair, it may suffer from a data sparseness problem since many phrase pairs occur only once (Nguyen et al., 2009). 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

Unauthentifiziert | Heruntergeladen 14.11.19 15:10 UTC

train a classifier for prediction. The maximum entropy classifier is a popular choice among many researchers.

Zens and Ney (2006) introduced the maximum entropy classifier for phrase reordering. Three different translation tasks were carried out: Arabic-English, Chinese-English and Japanese-English. Only two orientations were considered, left or right (i.e. monotone or swap). Although the proposed model outperforms the relative frequency model in terms of classification performance, they did not draw comparison between them in terms of translation performance. The translation results reported were between their model and the distance-based reordering model. We believe that such a comparison with a lexicalised reordering model is important because the model is faster to estimate (i.e. relative frequency) and also faster to use during translation since there is no overhead computation (i.e. retrieving probabilities from a table).

Xiong et al. (2006) also proposed a maximum entropy model to predicate reordering of neighbour blocks (i.e. phrase pairs) and considered straight or inverted orientations (i.e. monotone or swap). Their experiments were carried out on Chinese-English translation tasks. The reported results were only in terms of translation performance. Similar to Zens and Ney (2006), the authors compared their model with the distance-based reordering model although they did make reference to the lexicalised reordering model.

Nguyen et al. (2009) applied the maximum entropy model to learn orientations identified by the hierarchical reordering model proposed by Galley and Manning (2008). The previous work of Zens and Ney (2006) and Xiong et al. (2006) identified such an orientation without considering the hierarchical structure of previous phrases. The authors used a relatively small English-Vietnamese corpus (0.6 million words) collected from daily newspapers. The approach achieves translation improvements over the lexical hierarchical reordering model in a test set taken from the same corpus (i.e. not a benchmark).

Xiang et al. (2011) introduced a smoothed prior probability to maximum entropy model and used multiple features based on syntactic parsing. The smoothed prior is a combination of – through interpolation weight – a global distortion probability $p(o_k)$ and a local distortion probability $p(o_k|\bar{f}_n, \bar{e}_n)$ (i.e. lexicalised reordering model). The model predicts the jump distance (up to five words) from the previously translated source word to the current source word. This method does not capture the hierarchical structure of the sentence as explained by Galley and Manning (2008). The experiments were undertaken on a large-scale Chinese-English translation task (one million sentence pairs). The proposed model shows improvement over a distance-based reordering model. Like the findings of Zens and Ney (2006) and Xiong et al. (2006), there is no comparison with a lexicalised reordering model.

Ni et al. (2011) considered a variety of machine learning techniques including the maximum entropy model. They introduced a perceptron-based learning approach to modelling long-distance phrase movements. Similar to Xiang et al. (2011), their

Unauthentifiziert | Heruntergeladen 14.11.19 15:10 UTC

model predicts the jump distance (up to five words) from the previously translated source word to the current one. Differing from the previous works, training data were divided into small independent sets where all samples share the same source phrase. This method breaks down the learning complexity by having as many sub-models as source phrases. Although the number of parameters for each sub-model are small, the total number of parameters are larger than having just one model to incorporate all the data. Several learning techniques are compared and evaluated on a Chinese-English corpus (Hong Kong laws corpus). The perceptron-based learning approach outperforms both the lexicalised reordering model and the maximum entropy model. The reported results were based on a test set taken from the same corpus.

Alrajeh and Niranjan (2014b) explored generative learning approach to phrase reordering in Arabic to English namely Bayesian naive Bayes. We achieved an improvement over a lexicalised reordering model. Training time of the model is as fast as the lexicalised one. Its storage requirement is many times smaller, which makes it more efficient particularly for large-scale tasks. Previously proposed discriminative models might achieve higher score than the reported results. However, the model is scalable since parameter estimation requires only one pass over the data with limited memory (i.e. no iterative learning). This is a critical advantage over discriminative models.

Recently, Cherry (2013) proposed using sparse features to optimise BLEU with the decoder instead of training a classifier independently. The reported results shows that sparse decoder features are superior to maximum entropy classifier.

We distinguish our work from the previous ones in the following. We propose fast and memory-efficient reordering models using multiclass SVM. In this study, we undertake a comparison between our work and both lexicalised and maximum entropy-based reordering models.

3. Baseline System

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

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

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

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

where the feature $h_i(\mathbf{f}, \boldsymbol{e})$ 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).

Unauthentifiziert | Heruntergeladen 14.11.19 15:10 UTC

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, the model is not sufficient (Al-Onaizan and Papineni, 2006). It prefers sentences that are grammatically correct without considering their actual meaning (i.e. the dependence of the target sentence on the source sentence). Besides that, it has a bias towards short translations[1] (Koehn, 2010). Therefore, developing a specific reordering model will improve the accuracy particularly when translating between two grammatically different languages.

4. Lexicalised Reordering Model

Phrase reordering modelling 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:

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

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

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

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 model to recognise sentence hierarchical structure.

5. Maximum Entropy-based Reordering Model

As mentioned in the introduction, maximum entropy classifier is a popular choice to model phrase movements. It is also known as multinomial logistic regression or

[1] In Moses, it is balanced by the word/phrase count features as noted by one of the reviewers.

Unauthentifiziert | Heruntergeladen 14.11.19 15:10 UTC

softmax regression, which is a probabilistic model for the multiclass problem. The model is an extension of logistic regression which is a binary classifier. The class probability is given by:

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

where $\phi(\bar{f}_i, \bar{e}_i)$ is a feature vector (see Table 3) and $\boldsymbol{w}_k$ is a weight vector measuring features' contribution to orientation o_k. The model's parameters are estimated by maximum likelihood. To do that, we write the function using the 1-of-K coding scheme in which $\mathbf{t}_i$ is a zero vector except where t_{ik} equals one, which indicates that an object is belonging to that class (Bishop, 2006). Then the likelihood is expressed as:

$$p(\mathbf{o}|\bar{\mathbf{f}}, \bar{\mathbf{e}}) = \prod_{i=1}^{N} \prod_{k=1}^{K} p(o_k|\bar{f}_i, \bar{e}_i)^{t_{ik}} \tag{6}$$

Now, taking the partial derivative of the log-likelihood we get (Bishop, 2006):

$$\frac{\partial \log L}{\partial \boldsymbol{w}_k} = \sum_{i=1}^{N} \left(t_{ik} - p(o_k|\bar{f}_i, \bar{e}_i)\right) \phi(\bar{f}_i, \bar{e}_i). \tag{7}$$

The solution is not closed-form but we can estimate $\boldsymbol{w}_k$ by the stochastic gradient descent. Similarly, MAP estimate can be used to impose regularisation on the parameters. In our experiments, we used a more advanced optimisation algorithm proposed by Andrew and Gao (2007)[2]. Their algorithm optimises L_1-regularised or L_2-regularised log-likelihood based on L-BFGS algorithm. The L_1 regularisation is equivalent to adding Laplacian prior over the model's parameters.

6. SVM-based Reordering Model

Phrase reordering problem is usually formulated as multiclass problem which can be solved as several binary problems in the standard SVM (Boser et al., 1992). One-versus-rest or one-versus-one are well known strategies.

In this work, we propose multiclass SVM to model phrase movements. We use dual coordinate method and a mechanism for pruning of the training samples, which allows us to train a reordering model efficiently. Before discussing our approach we briefly introduce multiclass SVM formulation.

Given a set $S = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_N, y_N)\}$ where $\mathbf{x}_i \in R^n$ and $y_i \in \{1, \ldots, K\}$, Crammer and Singer (2002) proposed a multiclass SVM formulation. Its dual optimisation problem is:

[2]We have used the authors' implementation of L-BFGS algorithm which is available at http://homes.cs.washington.edu/~galen/

Unauthentifiziert | Heruntergeladen 14.11.19 15:10 UTC

$$\begin{aligned}
\underset{\boldsymbol{\alpha}}{\text{minimise}} \quad & \mathcal{D}(\boldsymbol{\alpha}) = \frac{1}{2}\sum_{k=1}^{K}\sum_{i,j=1}^{N}\alpha_{ik}\alpha_{jk}\boldsymbol{x}_i^T\boldsymbol{x}_j + \sum_{i=1}^{N}\sum_{k=1}^{K}(1-\delta_{ik})\alpha_i, \\
\text{subject to} \quad & \sum_{k=1}^{K}\alpha_{ik} = 0 \quad \text{and} \quad \alpha_{ik} \leq C\delta_{ik} \quad \forall i,k, \\
& \delta_{ik} = \begin{cases} 0 & \text{if } y_i \neq y_k; \\ 1 & \text{if } y_i = y_k. \end{cases}
\end{aligned} \tag{8}$$

where the corresponding $\boldsymbol{w}_k = \sum_{i=1}^{N}\alpha_{ik}\boldsymbol{x}_i$. Here $C \geq 0$ is a penalty parameter for margin violation by each data point $\boldsymbol{x}_i$.

For the sake of clarity, we use $\boldsymbol{x}_i$ to represent data in our discussion on SVM, and the learning algorithms. In the context of our NLP problem, and previous discussion in this paper $\boldsymbol{x}_i = \phi(\bar{f}_i, \bar{e}_i)$. Table 3 shows how a phrase pair can be represented.

Note that SVM is not a probabilistic classifier but in our experiments we used softmax function to yield a probabilistic decision (Bishop, 2006).

6.1. Shrinking dual method for solving Multiclass SVM

Keerthi et al. (2008) propose a sequential dual method to solve the problem (8). The method sequentially picks $\boldsymbol{x}_i$ at a time and optimises its dual variable (i.e. $\alpha_{ik}\ \forall k$) while fixing all other variables. The sub-problem is given by:

$$\begin{aligned}
\underset{\boldsymbol{\alpha}_i}{\text{minimise}} \quad & \frac{1}{2}\sum_{k=1}^{K}\frac{1}{2}A\alpha_{ik}^2 + B_k\alpha_{ik}, \\
\text{subject to} \quad & \alpha_{ik} \leq C\delta_{ik} \quad \forall k,
\end{aligned} \tag{9}$$

where

$$\begin{aligned}
A = \boldsymbol{x}_i^T\boldsymbol{x}_i \quad \text{and} \quad & B_k = G_{ik} - A\alpha_{ik}, \\
& G_{ik} = \frac{\partial\mathcal{D}(\boldsymbol{\alpha})}{\partial\alpha_{ik}} = \boldsymbol{w}_k^T\boldsymbol{x}_i + 1 - \delta_{ik}.
\end{aligned} \tag{10}$$

Crammer and Singer (2002) provide $O(k \log k)$ algorithm to solve the sub-problem (9). Fan et al. (2008) present a simpler version given in Algorithm 1.

After each update, the corresponding weight vector for each class $\boldsymbol{w}_k$ is maintained as follows (Fan et al., 2008):

$$\boldsymbol{w}_k = \boldsymbol{w}_k + (\alpha_{ik}^{'} - \alpha_{ik})\boldsymbol{x}_i \tag{11}$$

Unauthentifiziert | Heruntergeladen 14.11.19 15:10 UTC

Algorithm 1 Solving the sub-problem of multiclass SVM

Require: A, $\mathbf{B}$ and a penalty parameter $C \geq 0$

$D_k \leftarrow B_k + AC\delta_{ik}$, $\forall k$
Sort $\mathbf{D}$ in decreasing order
$\beta \leftarrow D_1 - AC$, $r \leftarrow 2$
while $r \leq K$ and $\beta/(r-1) < D_r$ **do**
    $\beta \leftarrow \beta + D_r$, $r \leftarrow r + 1$
end while
$\beta \leftarrow \beta/(r-1)$
$\alpha'_{ik} \leftarrow \min(C\delta_{ik}, (\beta - B_k)/A)$, $\forall k$

The optimal dual variables are achieved when the following condition is satisfied for all samples (Keerthi et al., 2008):

$$\nu_i = 0\,, \forall i, \quad \text{where} \quad \nu_i = \max_k G_{ik} - \min_{k:\alpha_{ik} < C\delta_{ik}} G_{ik}. \tag{12}$$

We propose a shrinking heuristic based on this condition which is a key to accelerate our algorithm. The dual variables α_{ik} are associated with each sample (i.e. phrase pair) therefore a training sample can be disregarded once its optimal dual variables are obtained. More data shrinking can be achieved by tolerating a small difference between the two values in (12). Algorithm 2 presents the overall procedure (shrinking step is from line 6 to 8).

Algorithm 2 Shrinking dual method for training large-scale multiclass SVM

Require: training set $S = \{\boldsymbol{x}_i, y_i\}_{i=1}^N$

$\boldsymbol{\alpha} = 0$ and $\boldsymbol{w} = 0$
repeat
    Randomly pick i from S
    Calculate A_{ik} , B_{ik} , G_{ik} $\forall k$ by (10)
    Calculate ν_i by (12)
    if $\nu_i \leq \epsilon$ **then**
        Remove i from S
    else
        Calculate α'_{ik} $\forall k$ by Algorithm 1
        Update $\boldsymbol{\alpha}$ and $\boldsymbol{w}$ by (11)
    end if
until $\nu_i \leq \epsilon$ $\forall i$

Unauthentifiziert | Heruntergeladen 14.11.19 15:10 UTC

6.2. Kernel Mapping via Linear SVM

We have seen in the previous section that linear SVM can be scalable because of the advantage of accessing the feature space. On the other hand, kernel SVM is able to learn more complex patterns by working on high dimensional feature space, where the data might be linearly separable, without explicit mapping using the kernel trick.

An interesting technique to accelerate kernel SVM is to apply linear SVM to the explicit form. However, in many cases, kernel mapping is exponential to the input space or infinite as in the Gaussian kernel. Low-degree polynomial mapping is shown to be useful for certain datasets (Chang et al., 2010). All-subsets kernel is similar to polynomial kernel but has more flexibility in terms of the monomials' weightings (Shawe-Taylor and Cristianini, 2004). The mapping generates all combinations of input features and each monomial's coefficient equals one unlike polynomial mapping. Working with all monomials might be computationally expensive. Analysis of variance (ANOVA) kernel K_d, used in our experiments, restricts the mapping to subsets of cardinality d with $\binom{n}{d}$ dimensions (Shawe-Taylor and Cristianini, 2004). Table 3, in the next section, gives an example of AVOVA mapping.

7. Experiments

The Arabic-English parallel corpus used in our experiments is a combination of MultiUN, ISI and Ummah to set up a large-scale corpus. MultiUN is a large-scale parallel corpus extracted from the United Nations website[3] (Eisele and Chen, 2010). ISI and Ummah were taken from Linguistic Data Consortium[4] (LDC) with catalogue numbers (LDC2007T08) and (LDC2004T18), respectively. Table 1 shows general statistics of the corpora. Test sets are from NIST MT06 and MT08 where the Arabic sides are 1797 and 813 sentences, respectively. Each sentence has four English references.

Corpus	MultiUN		ISI		Ummah	
Statistics	Arabic	English	Arabic	English	Arabic	English
Sentence Pairs	9.7 M		1.1 M		80 K	
Running Words	255.5 M	285.7 M	30.5 M	34.4 M	2.7 M	2.9 M
Words/Line	22	25	27	31	33	36
Vocabulary Size	677 K	410 K	354 K	195 K	63 K	46 K
Vocabulary [%]	0.26	0.14	1.16	0.57	2.33	1.59

Table 1. General statistics of MultiUN, ISI and Ummah (M: million, K: thousand).

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

[4] http://ldc.upenn.edu/

Unauthentifiziert | Heruntergeladen 14.11.19 15:10 UTC

We compare our approach with previously proposed 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.

7.1. Classification

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

During the extraction process, each extracted phrase pair is represented by linguistic features. There are different feature representations in the literature as we have seen in Section 2. We explore a variety of feature sets as shown in Table 2. Each phrase pair is represented by all its words, its boundaries or its alignments. We have considered one or three words of context (i.e. occur before or after each phrase pair). Finally, one of ANOVA mappings were selected. Table 3 gives a generic example.

Feature	Phrase Pair			Context		ANOVA Mapping		
Set	all words	boundaries	alignments	size=1	size=3	d=1	d=2	d≤2
S1	●					✓		
S2		●				✓		
S3			●			✓		
S4		●		●		✓		
S5		●		●			✓	
S6		●		●				✓
S7			●	●		✓		
S8			●	●			✓	
S9			●	●				✓
S10		●			●	✓		
S11		●			●		✓	
S12		●			●			✓
S13			●		●	✓		
S14			●		●		✓	
S15			●		●			✓

Table 2. A variety of feature sets to represent a phrase pair.

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

Unauthentifiziert | Heruntergeladen 14.11.19 15:10 UTC

Sentence pair:
Foreign sentence $\mathbf{f}$: f_1 f_2 $_1$ f_3 f_4 f_5 $_2$ f_6 $_3$.
English sentence $\mathbf{e}$: e_1 $_1$ e_2 e_3 $_3$ e_4 e_5 $_2$.

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

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

Feature Representation:
a phrase pair is represented as a vector ϕ where each feature is a discrete number (0=not exist). Below is a representation of $\phi(\bar{f}_2, \bar{e}_2)$ in different feature sets:

S1 : f_3, f_4, f_5, e_4, e_5
S2 : f_3, f_5, e_4, e_5
S3 : f_3&e_5, f_5&e_4
S4 : f_3, f_5, e_4, e_5, f_2^-, f_6^+
S5 : f_3_f_5, f_3_e_4, f_3_e_5, f_3_f_2^-, f_3_f_6^+, f_5_e_4, f_5_e_5, f_5_f_2^-, f_5_f_6^+, e_4_e_5, e_4_f_2^-, e_4_f_6^+, f_2^-_f_6^+
S6 : f_3, f_5, e_4, e_5, f_2^-, f_6^+, f_3_f_5, f_3_e_4, f_3_e_5, f_3_f_2^-, f_3_f_6^+, f_5_e_4, f_5_e_5, f_5_f_2^-, f_5_f_6^+, e_4_e_5, e_4_f_2^-, e_4_f_6^+, f_2^-_f_6^+

Table 3. A generic example of the process of phrase pair extraction and representation in different feature sets

Firstly, we present the performance of lexicalised reordering model in Table 4. Then, we compare MaxEnt and multiclass SVM under all feature sets in Table 2. It is not hard to see that using MaxEnt with an alternate feature set that enumerates all conjunctions of size d is equal to ANOVA mapping. Tables 5 and 6 report the results.

Orientation	Confusion Matrix			Accuracy	Precision	Recall	F_1 score
	Mono	Swap	Disc.	all classes			
Monotone	68.9	0.9	1.3		97.0	77.0	85.9
Swap	6.4	2.6	0.8	75.9	26.8	63.5	37.7
Discontinuous	14.2	0.6	4.4		23.0	68.4	34.5

Table 4. The performance of lexicalised reordering model.

Unauthentifiziert | Heruntergeladen 14.11.19 15:10 UTC

Feature Set	Time	Acc.	Precision			Recall			F_1 score		
			M	S	D	M	S	D	M	S	D
S1	1h26m	74.1	96.6	17.3	19.7	75.5	57.7	61.4	84.8	26.6	29.8
S2	1h10m	74.0	**97.1**	13.9	19.1	75.1	**61.2**	**62.0**	84.7	22.7	29.2
S3	1h40m	**76.1**	93.5	**32.1**	**34.0**	**79.8**	55.5	58.6	**86.1**	**40.7**	**43.0**
S4	1h50m	77.0	**95.5**	37.3	29.1	78.9	69.6	63.0	86.4	48.5	39.9
S5	5h59m	80.7	94.9	49.2	44.2	83.3	71.9	**68.4**	88.7	58.4	53.7
S6	6h21m	**81.3**	94.4	**53.3**	**46.9**	**84.3**	**72.5**	67.6	**89.1**	**61.4**	**55.4**
S7	3h10m	78.7	93.8	45.2	39.7	82.2	67.1	61.7	87.6	54.0	48.3
S8	4h32m	81.4	**93.9**	51.6	50.7	85.0	**72.7**	66.7	89.3	60.4	57.6
S9	4h43m	**82.5**	93.4	**59.5**	**53.9**	**86.7**	72.1	**67.3**	**89.9**	**65.2**	**59.9**
S10	2h45m	76.2	94.1	34.0	31.3	79.2	61.9	58.8	86.0	43.9	40.9
S11	15h11m	82.4	**95.2**	56.4	47.3	**84.8**	**75.0**	**74.0**	**89.7**	64.4	57.7
S12	16h04m	**82.6**	94.9	**58.1**	**48.9**	84.7	73.3	71.2	89.5	**64.8**	**58.0**
S13	3h24m	78.8	92.3	46.3	45.1	83.8	62.2	59.8	87.9	53.1	51.4
S14	13h03m	82.2	**93.9**	50.0	45.4	83.5	**78.6**	68.8	88.4	61.1	54.7
S15	15h12m	**82.9**	93.4	**59.8**	**54.8**	**88.3**	72.8	**69.9**	**90.8**	**65.7**	**61.4**

Table 5. Maximum entropy-based reordering model's performance (M is monotone, S is swap, D is discontinuous). The reported time is in hours (h) and minutes (m).

Feature Set	Time	Acc.	Precision			Recall			F_1 score		
			M	S	D	M	S	D	M	S	D
S1	0h30m	70.8	92.7	7.4	22.3	76.5	31.9	36.8	83.8	12.0	27.8
S2	0h28m	71.7	**96.2**	13.3	10.7	74.9	30.9	44.5	84.2	18.6	17.3
S3	0h40m	**75.8**	93.3	**36.3**	**31.1**	**80.0**	**50.5**	**58.8**	**86.1**	**42.2**	**40.7**
S4	0h33m	75.6	**95.9**	35.9	21.0	77.7	59.9	62.2	85.8	44.9	31.4
S5	1h45m	82.1	95.8	55.8	44.8	84.1	**73.7**	**73.3**	89.6	63.5	55.6
S6	2h07m	**82.5**	95.1	**60.0**	**47.4**	**85.2**	72.1	72.4	**89.9**	**65.5**	**57.3**
S7	0h47m	79.3	93.7	49.5	41.3	82.8	69.0	62 .7	87.9	57.7	49.8
S8	1h24m	81.0	**95.3**	50.4	42.9	83.3	69.0	**71.0**	88.9	58.2	53.5
S9	1h41m	**82.1**	92.5	**61.0**	**54.1**	**86.8**	**69.7**	65.6	**89.6**	**65.1**	**59.3**
S10	0h44m	74.0	95.6	24.5	18.7	76.8	49.2	53.4	85.2	32.7	27.7
S11	4h33m	**82.7**	**95.9**	56.2	47.4	84.7	**75.0**	**74.2**	89.8	64.3	57.9
S12	4h51m	82.6	94.9	**57.9**	**49.7**	**85.4**	73.3	71.8	**89.9**	**64.7**	**58.7**
S13	0h59m	78.0	94.2	46.3	35.0	81.6	58.4	62.2	87.4	49.7	44.8
S14	3h32m	82.0	**96.8**	49.1	44.0	83.2	**77.3**	**75.8**	89.5	60.0	55.6
S15	4h04m	**82.8**	95.5	**55.8**	**49.8**	**85.4**	73.6	72.8	**90.2**	**63.5**	**59.1**

Table 6. Multiclass SVM-based reordering model's performance.

Unauthentifiziert | Heruntergeladen 14.11.19 15:10 UTC

Four observations can be drawn from the results in Table 5 and Table 6. First, the performance of multiclass SVM is similar to MaxEnt in most feature sets. Second, our classifier is a couple of times faster than MaxEnt (around 4-fold) due to the shrinking method. Third, context around phrase pairs is important to achieve high accuracy and only one word before and after is enough. Finally, alignment features usually have higher F_1 score than boundary features in both MaxEnt and multiclass SVM.

Alrajeh and Niranjan (2014a) propose a dual multinomial logistic regression (Dual MLR) with a shrinking heuristic to model phrase movements. We compare their shrinking approach with multiclass SVM in Figure 1.

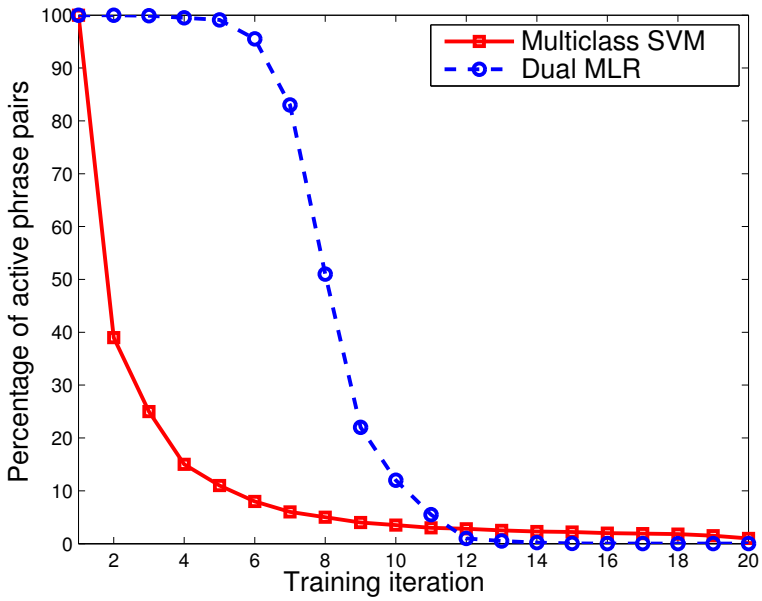

Figure 1. Comparison between multiclass SVM and dual multinomial logistic regression (MLR) in terms of active phrase pairs during training.

In Figure 2, we show training time and memory usage for each classifier (Multiclass SVM, Dual MLR, MaxEnt) when the number of phrase pairs increases. The results show that multiclass SVM consumes mush less memory (nearly half) than MaxEnt due to the shrinking technique discussed in Section 6.1.

Unauthentifiziert | Heruntergeladen 14.11.19 15:10 UTC

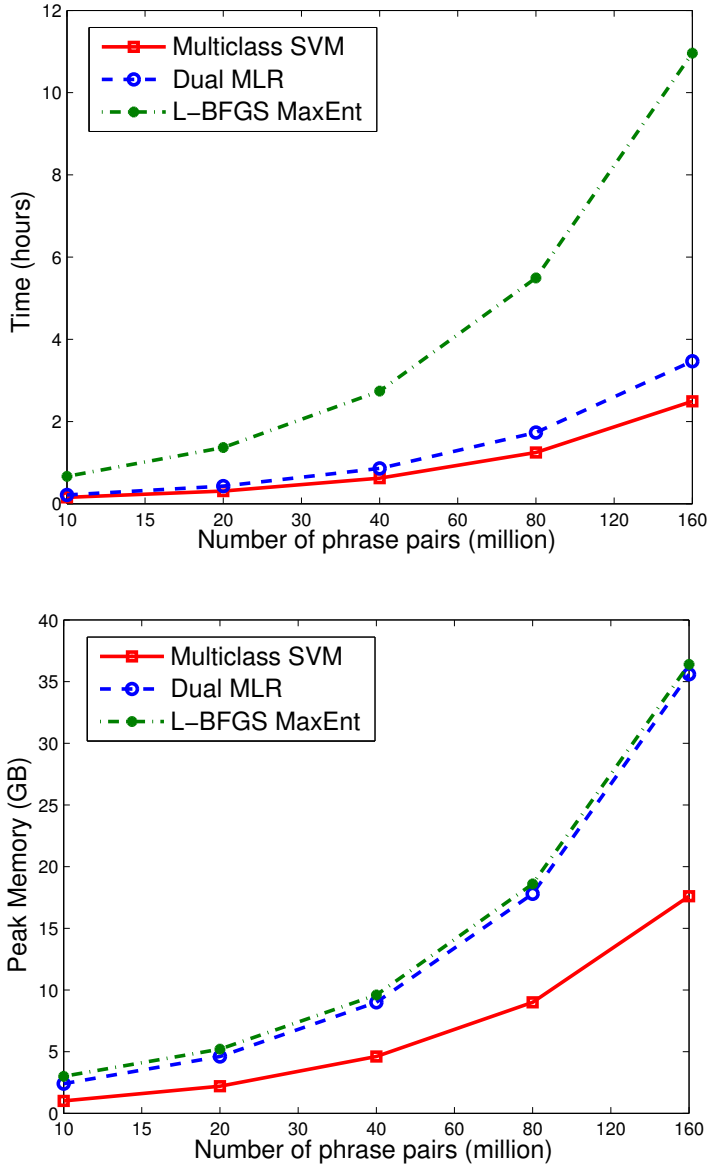

Figure 2. Training time (above) and memory usage (below) for each classifier when the number of phrase pairs increases.

Unauthentifiziert | Heruntergeladen 14.11.19 15:10 UTC

7.2. Translation

We used the Moses toolkit (Koehn et al., 2007) with its default settings. The language model is a 5-gram built from the English side with interpolation and Kneser-Ney smoothing (Kneser and Ney, 1995). We tuned the system using PRO technique (Hopkins and May, 2011). We built seven Arabic-English translation systems. The first system has no reordering model, only a distortion penalty. The second system has a hierarchical lexicalised reordering model that is built by specifying the configuration string `hier-msd-backward-fe`. Sparse reordering features (Cherry, 2013) are added in the third system. We only used '`sparse-phrase=1`' option with top 200 words. The last four systems have SVM-based or MaxEnt-based reordering models.

As commonly used in statistical machine translation, we evaluated the translation performance by BLEU score (Papineni et al., 2002) and NIST (Doddington, 2002). We also computed statistical significance for the proposed models using a *paired bootstrap resampling* method (Koehn, 2004b).

Table 7 reports the size of each reordering model. Note that there is a big difference between the lexicalised model and the discriminative ones.

Reordering Model	Lexicalised	Multiclass SVM (S6)	Multiclass SVM (S7)
Parameters (million)	73.2	17.1	2.4
Disk Storage (GB)	5.9	0.7	0.1

Table 7. Comparison of problem sizes for the different reordering models.

Table 8 presents NIST and BLEU scores for five translation systems in MT06 and MT08 test sets. Our models achieve improvements on top of a strong baseline system with sparse reordering features. Note that feature sets (S6) and (S7) have similar scores although (S6) has higher classification accuracy in Table 6.

	MT06				MT08			
Phrase-based SMT	NIST	Δ	BLEU	Δ	NIST	Δ	BLEU	Δ
No Reordering Model	9.1	-0.3	35.5	-1.6	9.9	-0.2	41.2	-1.7
LexicalRM (**baseline**)	9.4	–	37.1	–	10.1	–	42.9	–
LexicalRM + sparseRM	9.5	+0.1	37.6	+0.5	10.2	+0.2	43.8	+0.9
SVM-RM (S6) + sparseRM	9.6	+0.2	38.1	+1.0	10.4	+0.3	44.4	+1.5
SVM-RM (S7) + sparseRM	9.6	+0.2	38.0	+0.9	10.4	+0.3	44.3	+1.4
MaxEnt-RM (S6) + sparseRM	9.6	+0.2	38.1	+1.0	10.4	+0.3	44.4	+1.5
MaxEnt-RM (S7) + sparseRM	9.6	+0.2	38.1	+1.0	10.4	+0.3	44.4	+1.5

Table 8. NIST and BLEU [%] scores for two evaluation sets (RM: Reordering Model).

Unauthentifiziert | Heruntergeladen 14.11.19 15:10 UTC

8. Conclusion

Posing phrase movements as a classification problem, we exploit recent developments in solving large-scale multiclass support vector machines using stochastic gradient learning algorithm and show significant advantages in Arabic-English systems. The algorithms we propose are shown to be computationally fast and memory-efficient. In terms of evaluating translation quality using the BLEU score, we achieve 1.0 point in MT06 and 1.5 in MT08 over a lexicalised reordering model with at least 95% statistical significance. Our SVM-based model is shown to be superior to the maximum entropy-based model. It is a couple of times faster (nearly 4-fold) and more memory-efficient (50% reduction).

The expanded space due to ANOVA mapping can be reduced significantly by removing less frequent features. We found that a reordering model based on alignments features (S7) is more compact than using boundaries features (S6).

Our current work focuses on two issues. The first issue is exploring higher degrees of ANOVA kernels and others in order to reduce the classification error rate. The second issue is reducing feature space by using limited but informative features such as part-of-speech tags.

Acknowledgements

The first author was funded by a scholarship from King Abdulaziz City for Science and Technology (KACST).

Bibliography

Al-Onaizan, Yaser and Kishore Papineni. Distortion models for statistical machine translation. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 529–536, Sydney, Australia, July 2006. Association for Computational Linguistics. URL `http://www.aclweb.org/anthology/P/P06/P06-1067`.

Alrajeh, Abdullah and Mahesan Niranjan. Large-scale reordering model for statistical machine translation using dual multinomial logistic regression. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1758–1763. Association for Computational Linguistics, 2014a. URL `http://aclweb.org/anthology/D14-1183`.

Alrajeh, Abdullah and Mahesan Niranjan. Bayesian reordering model with feature selection. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 477–485, Baltimore, Maryland, USA, June 2014b. Association for Computational Linguistics. URL `http://www.aclweb.org/anthology/W/W14/W14-3361`.

Alrajeh, Abdullah, Akiko Takeda, and Mahesan Niranjan. Memory-efficient large-scale linear support vector machine. In *Proceedings of SPIE: Seventh International Conference on Machine Vision (ICMV 2014)*, volume 9445, pages 944527–944527–6, Milan, Italy, February 2015. SPIE. doi: 10.1117/12.2180925. URL `http://dx.doi.org/10.1117/12.2180925`.

Unauthentifiziert | Heruntergeladen 14.11.19 15:10 UTC

Andrew, Galen and Jianfeng Gao. Scalable training of L1-regularized log-linear models. In *Proceedings of the 24th International Conference on Machine Learning*, ICML '07, pages 33–40. ACM, 2007. ISBN 978-1-59593-793-3.

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

Boser, Bernhard E., Isabelle M. Guyon, and Vladimir N. Vapnik. A training algorithm for optimal margin classifiers. In *Proceedings of the Fifth Annual Workshop on Computational Learning Theory*, COLT '92, pages 144–152, 1992.

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

Brown, Peter F., John Cocke, Stephen A. Della-Pietra, Vincent J. Della-Pietra, Frederick Jelinek, Robert L. Mercer, and Paul Rossin. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19(2):263–311, 1993.

Chang, Yin-Wen, Cho-Jui Hsieh, Kai-Wei Chang, Michael Ringgaard, and Chih-Jen Lin. Training and testing low-degree polynomial data mappings via linear SVM. *Journal of Machine Learning Research*, 2:1471–1490, Apr. 2010.

Cherry, Colin. 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 2013. Association for Computational Linguistics. URL `http://www.aclweb.org/anthology/N13-1003`.

Crammer, Koby and Yoram Singer. On the algorithmic implementation of multiclass kernel-based vector machines. *Journal of Machine Learning Research*, 2:265–292, Mar. 2002.

Cristianini, Nello and John Shawe-Taylor. *An Introduction to Support Vector Machines: And Other Kernel-based Learning Methods*. Cambridge University Press, New York, NY, USA, 2000. ISBN 0-521-78019-5.

Doddington, George. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In *Proceedings of the Second International Conference on Human Language Technology Research*, HLT '02, pages 138–145, San Francisco, CA, USA, 2002. Morgan Kaufmann Publishers Inc.

Eisele, A. and Y. Chen. MultiUN: A multilingual corpus from united nation documents. In Tapias, Daniel, 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 2010.

Fan, Rong-En, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. LIBLINEAR: A library for large linear classification. *Journal of Machine Learning Research*, 9:1871–1874, 2008.

Galley, Michel and Christopher D. Manning. 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 2008. Association for Computational Linguistics.

Unauthentifiziert | Heruntergeladen 14.11.19 15:10 UTC

Glasmachers, Tobias and Ürün Dogan. Accelerated coordinate descent with adaptive coordinate frequencies. In Ong, Cheng Soon and Tu Bao Ho, editors, *Asian Conference on Machine Learning, ACML*, volume 29 of *JMLR Proceedings*, pages 72–86. JMLR.org, 2013.

Hopkins, Mark and Jonathan May. Tuning as ranking. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 1352–1362, Edinburgh, Scotland, UK., July 2011. Association for Computational Linguistics. URL `http://www.aclweb.org/anthology/D11-1125`.

Hsieh, Cho-Jui, Kai-Wei Chang, Chih-Jen Lin, S. Sathiya Keerthi, and S. Sundararajan. 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, 2008.

Keerthi, S. Sathiya, Sellamanickam Sundararajan, Kai-Wei Chang, Cho-Jui Hsieh, and Chih-Jen Lin. A sequential dual method for large scale multi-class linear SVMs. In *Proceedings of the Forteenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 408–416, 2008. URL `http://www.csie.ntu.edu.tw/~cjlin/papers/sdm_kdd.pdf`.

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

Koehn, Philipp. 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, 2004a.

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

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

Koehn, Philipp and Christof Monz. 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, 2005.

Koehn, Philipp, Amittai Axelrod, Alexandra Birch Mayne, Chris Callison-Burch, Miles Osborne, and David Talbot. Edinburgh system description for the 2005 IWSLT speech translation evaluation. In *Proceedings of International Workshop on Spoken Language Translation*, Pittsburgh, PA, 2005.

Koehn, Philipp, 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. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the ACL 2007 Demo and Poster Sessions*, pages 177–180, 2007.

Kumar, Shankar and William Byrne. 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 2005. Association for Computational Linguistics.

Nguyen, Vinh Van, Akira Shimazu, Minh Le Nguyen, and Thai Phuong Nguyen. 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, 2009.

Unauthentifiziert | Heruntergeladen 14.11.19 15:10 UTC

Ni, Y., C. Saunders, S. Szedmak, and M. Niranjan. Exploitation of machine learning techniques in modelling phrase movements for machine translation. *Journal of Machine Learning Research*, 12:1–30, Feb. 2011. ISSN 1532-4435.

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

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

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

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

Papineni, K., S. Roukos, T. Ward, and W. Zhu. 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, 2002. ACL.

Shawe-Taylor, John and Nello Cristianini. *Kernel Methods for Pattern Analysis*. Cambridge University Press, New York, NY, USA, 2004.

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

Xiang, Bing, Niyu Ge, and Abraham Ittycheriah. 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, 2011. Association for Computational Linguistics.

Xiong, Deyi, Qun Liu, and Shouxun Lin. 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 2006. Association for Computational Linguistics.

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

Address for correspondence:
Abdullah Alrajeh
asar1a10@ecs.soton.ac.uk
School of Electronics and Computer Science
University of Southampton
Southampton, United Kingdom, SO17 1BJ

Unauthentifiziert | Heruntergeladen 14.11.19 15:10 UTC