Journal of Machine Learning Research 12 (2011) 1-30 Submitted 5/10; Revised 11/10; Published 1/11

Exploitation of Machine Learning Techniques in Modelling Phrase Movements for Machine Translation

Yizhao Ni ENXYN@BRISTOL.AC.UK
Pattern Analysis and Intelligent Systems Group
Department of Engineering Mathematics, University of Bristol
Bristol, BS8 1UB, United Kindom

Craig Saunders CRAIG.SAUNDERS@XRCE.XEROX.COM
Xerox Research Centre Europe
6 Chemin de Maupertuis
38240 Meylan, France

Sandor Szedmak SS03V@ECS.SOTON.AC.UK
Mahesan Niranjan MN@ECS.SOTON.AC.UK
ISIS Group, School of Electronics and Computer Science
University of Southampton, Southampton
SO17 1BJ, United Kingdom

Editor: John Shawe-Taylor

Abstract

We propose a distance phrase reordering model (DPR) for statistical machine translation (SMT), where the aim is to learn the grammatical rules and context dependent changes using a phrase reordering classification framework. We consider a variety of machine learning techniques, including state-of-the-art structured prediction methods. Techniques are compared and evaluated on a Chinese-English corpus, a language pair known for the high reordering characteristics which cannot be adequately captured with current models. In the reordering classification task, the method significantly outperforms the baseline against which it was tested, and further, when integrated as a component of the state-of-the-art machine translation system, MOSES, it achieves improvement in translation results.

Keywords: statistical machine translation (SMT), phrase reordering, lexicalized reordering (LR), maximum entropy (ME), support vector machine (SVM), maximum margin regression (MMR) , max-margin structure learning (MMS)

1. Introduction

Machine translation (MT) is a challenging problem in artificial intelligence. Natural languages are characterised by large variabilities of expressions, exceptions to grammatical rules and context dependent changes. Differences in these across different languages make automatic translation a very difficult task. While early work in machine translation was dominated by rule based approaches (Bennett and Slocum, 1985), the availability of large corpora, and the ease with which they can be processed on computers has, similar to developments in other areas of artificial intelligence, paved the way for statistical methods to be applied. The process of translation from a *source* language to a *target* language is considered equivalent to a problem of retrieving a target message from the

Symbol	Notation
$\mathbf{f}$	the source sentence (string)
$\mathbf{e}$	the target sentence (string)
f_j	the j-th word in the source sentence
e_i	the i-th word in the target sentence
$\overline{\mathbf{f}}^I$	the source phrase sequence
$\overline{\mathbf{e}}^I$	the target phrase sequence
$\overline{f}_j$	the source phrase where $\overline{f}$ denotes the sequence of words $[f_{j_l}, \ldots, f_{j_r}]$ and j denotes that $\overline{f}_j$ is the j-th phrase in $\overline{\mathbf{f}}^I$
$\overline{e}_i$	the target phrase where $\overline{e}$ denotes the sequence of words $[e_{i_l}, \ldots, e_{i_r}]$ and i denotes that $\overline{e}_i$ is the i-th phrase in $\overline{\mathbf{e}}^I$
Υ	the set of phrase pairs $(\overline{f}_j, \overline{e}_i) \in \Upsilon$
N	the number of examples in Υ
$(\overline{f}_j^n, \overline{e}_i^n)$	the n-th example in Υ that is also abbreviated as $(\overline{f}^n, \overline{e}^n)$
$\phi(\overline{f}_j, \overline{e}_i)$	the feature vector of phrase pair $(\overline{f}_j, \overline{e}_i)$
d	the phrase reordering distance
o	the phrase orientation class
O	the set of phrase orientations $o \in O$
C_O	the number of phrase orientations in O
φ	embedding function to map the orientation set to an output space $\varphi : O \to \mathbb{R}$
$\mathbf{w}_o$	weight vector measuring features' contribution to an orientation o
$\{\mathbf{w}_o\}_{o \in O}$	The set of weight vectors for the phrase reordering model
dim	the dimension of

Table 1: Notation used in this paper.

"source code" (Weaver, 1949). This view enables a probabilistic formulation in which the task becomes the maximisation of the posterior probability over all the phrase sequences in the target language. Principled approaches to designing the different components of such a system, shown in Figure 1, have been developed in recent years (Koehn et al., 2005).

Phrase-based *statistical machine translation* (SMT) is a task where each source sentence $\mathbf{f}$ is segmented into a sequence of I phrases $\overline{\mathbf{f}}^I$ and translated into a target sequence $\overline{\mathbf{e}}^I$, often by means of a stochastic process that maximises the posterior probability $\overline{\mathbf{e}}^I = \arg\max_{\hat{\mathbf{e}}^I \in E} \left\{ P(\hat{\mathbf{e}}^I | \overline{\mathbf{f}}^I) \right\}$. Usually the posterior probability $P(\hat{\mathbf{e}}^I | \overline{\mathbf{f}}^I)$ is modelled in a log-linear maximum entropy framework (Berger et al., 1996) which permits easy integration of additional models, and is given by

$$P(\hat{\mathbf{e}}^I | \overline{\mathbf{f}}^I) = \frac{\exp\left(\sum_m \lambda_m h_m(\overline{\mathbf{f}}^I, \hat{\mathbf{e}}^I) \right)}{\sum_{I', \hat{\mathbf{e}}^{I'}} \exp\left(\sum_m \lambda_m h_m(\overline{\mathbf{f}}^{I'}, \hat{\mathbf{e}}^{I'}) \right)},$$

where $\{h_m\}$ represent sub-models with scaling factors $\{\lambda_m\}$. As the denominator only depends on the source phrase sequence $\overline{\mathbf{f}}^I$, it is usually discarded and the solution is also represented as $\overline{\mathbf{e}}^I = \arg\max_{\hat{\mathbf{e}}^I \in E} \left\{ \exp\left(\sum_m \lambda_m h_m(\overline{\mathbf{f}}^I, \hat{\mathbf{e}}^I) \right) \right\}$.

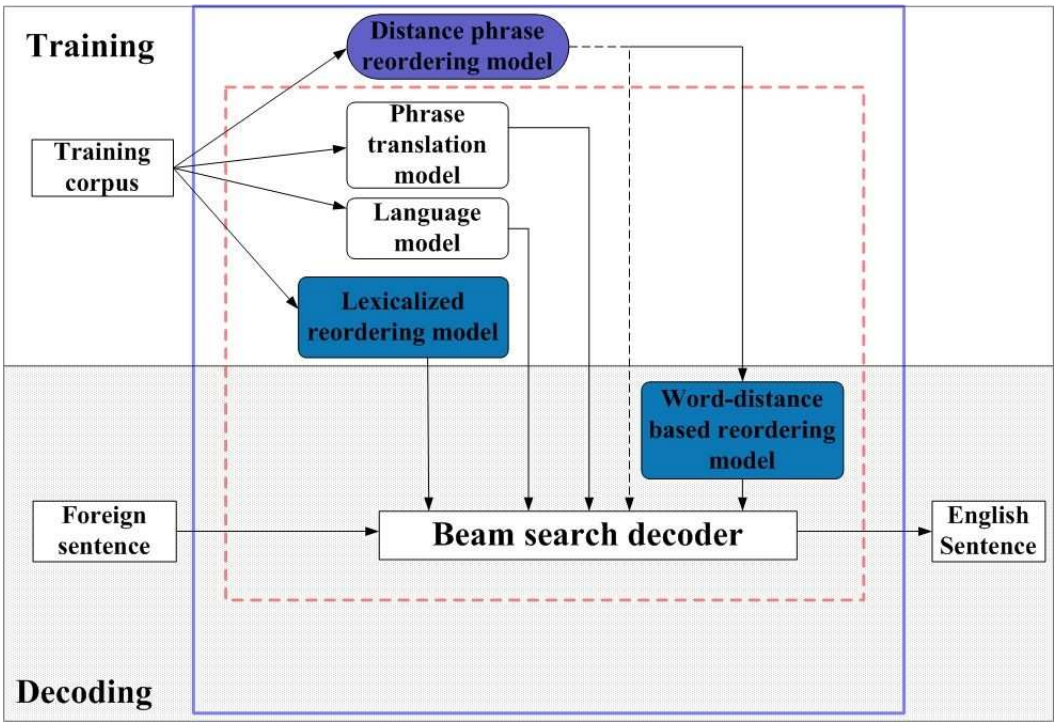


Figure 1: Training (top box) and decoding (bottom box) procedures for a state-of-the-art SMT systems (dotted line box) and our MT system (solid line box).

A combination of several sub-models $\{h_m\}$ (see Figure 1), including a phrase translation probability model, a language model and a phrase reordering model are commonly used. Each sub-model is trained individually and then weighted by a scale factor λ_m tuned to achieve good final translation quality (Och, 2003). Finally, the decoder searches a Viterbi-best string path given the joint decoding information. The reader is referred to Ni (2010) for detailed discussions on these models.

1.1 Modelling Phrase Movements

In this paper, we focus on developing a crucial component in statistical machine translation—the *phrase reordering model*. Word or phrase reordering is a common problem in bilingual translations arising from different grammatical structures. For example, the Chinese "NP_1 DEG NP_2" sequence is analogous to the English possessive structure of "NP_1's NP_2" and does not require reordering (see Figure 2 (a)). However, due to different linguistic environment it may come from, this Chinese possessive structure can express more sophisticated relationships which are inappropriate for the "NP_1's NP_2" expression, for example, the "NP_2 of NP_1" sequence which requires phrase swapping (see Figure 2 (b)). In general, if the decoder "knows" the orders of phrase translations in the target

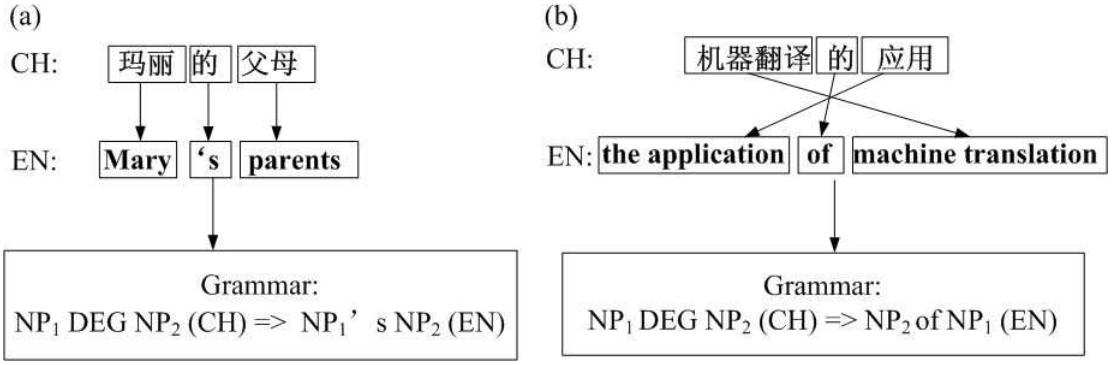


Figure 2: Example: the distance phrase reordering in Chinese-to-English bilingual translation.

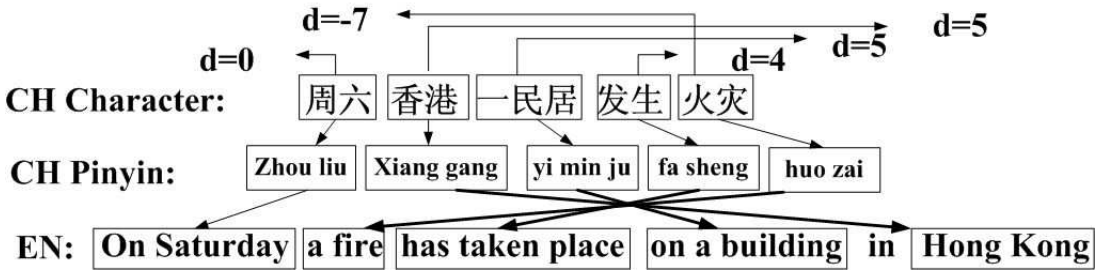


Figure 3: The phrase reordering distance d.

language, the fluency of machine translation can be greatly improved. This motivates investigations into, and development of models for, phrase reordering.

Now taking a Chinese-to-English translation (see Figure 3) for example, obviously not all words are translated one by one and some words are translated far behind after its preceding words are translated (e.g., phrase "a fire"). Therefore, an ideal phrase reordering model should be able to handle arbitrary distance phrase movements. However, handling such movements is a computationally expensive problem (Knight, 1999). Within recently developed SMT systems, a simple phrase reordering model, named *word distance-based reordering model* (WDR), is commonly used (Och et al., 1999; Koehn, 2004; Zens et al., 2005). This model defines a reordering distance for the j-th source phrase $\bar{f}_j$ as (see Figure 3 for an illustration of this.)

$$d_j := \begin{array}{l}\text{abs}(\text{last source word position of previously translated phrase} + 1 \\ \quad -\text{first source word position of newly translated phrase } \bar{f}_j)\end{array}, \tag{1}$$

and the total cost of phrase movements for a sentence pair $(\mathbf{f},\mathbf{e})$ is linear proportional to these reordering distances $h_d(\bar{\mathbf{f}}^J,\bar{\mathbf{e}}^I) = -\alpha\sum_j d_j$ with a tuning parameter α. Although computationally efficient, this model has been shown to be weak due to its content independence. A content-based extension to WDR is the *lexicalized reordering model* (LR) (Tillmann, 2004; Koehn et al., 2005), which splits the distance space into several segments, each of which represents a phrase reordering orientation o (see Figure 4). Then the phrase reordering probability for a phrase pair $(\bar{f}_j,\bar{e}_i)$ is

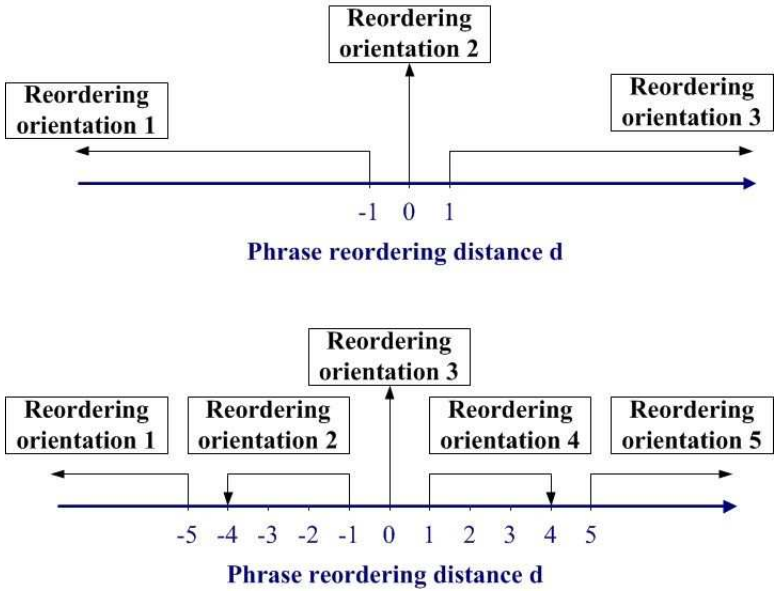


Figure 4: The phrase reordering orientations: the three-class setup (top) and the five-class setup (bottom).

predicted using *maximum likelihood estimation* (MLE)

$$p\big(o|(\bar{f}_j,\bar{e}_i)\big) = \frac{count(o,(\bar{f}_j,\bar{e}_i))}{\sum_{o'} count(o',(\bar{f}_j,\bar{e}_i))},$$

where $h_d(\bar{\mathbf{f}}^J,\bar{\mathbf{e}}^I) = \sum_{(\bar{f}_j,\bar{e}_i)\in(\bar{\mathbf{f}}^J,\bar{\mathbf{e}}^I)} p(o|(\bar{f}_j,\bar{e}_i))$ is used to represent the cumulative cost of phrase movements. Although the overall performance is better than WDR, it usually suffers from data sparseness, and some heuristics have to be employed to make the approach effective.

Adopting the idea of predicting phrase reordering orientations, researchers started exploiting context or grammatical content which may relate to phrase movements (Tillmann and Zhang, 2005; Xiong et al., 2006; Zens and Ney, 2006). In general, the distribution of phrase reorderings is expressed with a log-linear form

$$p(o|(\bar{f}_j,\bar{e}_i),\mathbf{w}_o) = \frac{h(\mathbf{w}_o^T\phi(\bar{f}_j,\bar{e}_i))}{Z(\bar{f}_j,\bar{e}_i)} \tag{2}$$

with the normalisation term $Z(\bar{f}_j,\bar{e}_i) = \sum_{o\in O} h(\mathbf{w}_o^T\phi(\bar{f}_j,\bar{e}_i))$. The feature parameters $\{\mathbf{w}_o\}_{o\in O}$ are then tuned by different discriminative models, among which the *maximum entropy* (ME) framework is a popular candidate. To characterise phrase movements, a variety of linguistic features are proposed

- Context features – word sequence (n-gram) features in (or around) the phrases. These indicator functions are the basic features used in Zens and Ney (2006) and also used in other MT experiments such as the word-sense disambiguation of Vickrey et al. (2005).

- Shallow syntactic features – part-of-speech (POS) tags or word-class features in (or around) the phrases. These indicator features are also used in the models above, and also in the context-aware phrase selection model of Giménez and Màrquez (2007).

- Statistical features – features such as the lexicalized reordering probability (Koehn et al., 2005) and the language model probability, etc. These real-value features are introduced by Tillmann and Zhang (2005) and are shown to be beneficial in capturing the local phrase reordering information.

Many other feature sets, such as lemma features and syntactic relationships in POS tags have also been investigated, posing a feature selection problem for any learning algorithm. Instead of investigating features sets, in this paper we concentrate on exploiting a limited set of linguistic features with different learning agents. We propose a *distance phrase reordering model* (DPR) that is also inspired by the orientation prediction framework (Koehn et al., 2005). Unlike Xiong et al. (2006) and Zens and Ney (2006) we regard phrase movements as a classification problem and use three multi-class learning agents—*support vector machine* (SVM), *maximum margin regression* (MMR) and *max-margin structure learning* (MMS) to perform the classification. Our goal is to find a learning agent that provides good tradeoff between classification accuracy with a limited feature set and computational efficiency. Furthermore, we also integrate the DPR model in a traditional SMT system, and the resulting MT system (solid line box in Figure 1) is compared with a state-of-the-art SMT system (dotted line box in Figure 1) on a Chinese-to-English MT task so as to demonstrate the effectiveness of the proposed DPR model.

1.2 Contribution and Structure

This paper makes two significant contributions. The first is a comparison, in terms of classification accuracy and computational efficiency, between different machine learning techniques for distance phrase movements in machine translation. This is mainly in the paradigm of structured learning, including maximum margin structure learning (MMS) and maximum margin regression (MMR), which is seen as a powerful framework that takes advantage of output structures in supervised learning problems, in modern machine learning literature. Our second contribution is the demonstration that this paradigm is effective in the task of phrase movements, which is acknowledged as a challenging task in machine translation. This turns out to be true, both in stand-alone translation tasks and when the method is integrated into a complete end-to-end statistical machine translation system. This is sufficiently encouraging that we have made our work available as a public domain software package[1] (Ni et al., 2010a) in a form that it can be integrated into the widely used MOSES system.[2]

The remainder of the paper is organised as follows: a general framework of the DPR model is given in Section 2, which specifies the modelling of phrase movements and describes the motivations of using the three learning agents. Then in Section 3 we demonstrate the linguistic features used and the training procedure for the DPR model. Section 4 evaluates the performance of the DPR model with both phrase reordering classification and machine translation experiments. Finally, we draw conclusions and mention areas for future work in Section 5.

1. The software is available at `http://patterns.enm.bris.ac.uk/distance-phrase-reordering-for-moses`.
2. MOSES is available at `http://www.statmt.org/moses/`.

2. Distance Phrase Reordering (DPR)

We adopt a discriminative model to capture the frequently occurring distance reorderings (e.g., Figure 2). An ideal model would consider every word position as a class and predict the start position of the next phrase, although in practice this is rather difficult to achieve. Hence, we consider a limited set of classes.

2.1 Orientation Class Definition

Following Koehn's lexicalized reordering model, we use the phrase reordering distance d in (1) to measure phrase movements. The distance space $d \in \mathbb{Z}$ is then split into C_O segments (i.e., C_O classes) and the possible start positions of phrases are grouped to make up a phrase orientation set O. Note that the more orientation classes a model has, the closer it is to the ideal model, but the smaller amount of training samples it would receive for each class. Therefore we consider two setups: a three-class approach $O = \{d < 0, d = 0, d > 0\}$ and one with five classes $O = \{d \leq -5, -5 < d < 0, d = 0, 0 < d < 5, d \geq 5\}$[3] (see Figure 4).

2.2 Reordering Probability Model and Learning Agents

Given a (source, target) phrase pair $(\bar{f}_j^n, \bar{e}_i^n) \in \Upsilon$ with $\bar{f}_j = [f_{j_l}, \ldots, f_{j_r}]$ and $\bar{e}_i = [e_{i_l}, \ldots, e_{i_r}]$, the *distance phrase reordering probability* has the form

$$p(o|(\bar{f}_j^n, \bar{e}_i^n), \{\mathbf{w}_o\}) := \frac{h\big(\mathbf{w}_o^T \phi(\bar{f}_j^n, \bar{e}_i^n)\big)}{\sum\limits_{o' \in O} h\big(\mathbf{w}_{o'}^T \phi(\bar{f}_j^n, \bar{e}_i^n)\big)}, \tag{3}$$

where $\mathbf{w}_o = [w_{o,0}, \ldots, w_{o,dim(\phi)}]^T$ is the weight vector measuring features' contribution to an orientation $o \in O$, ϕ is the feature vector and h is a pre-defined monotonic function.

Equation (3) is analogous to the well-known maximum entropy framework of Equation (2). In contrast to learning $\{\mathbf{w}_o\}_{o \in O}$ by maximising the entropy over all phrase pairs' orientations

$$\max_{\{\mathbf{w}_o \in O\}} \Big\{ - \sum_{(\bar{f}_j^n, \bar{e}_i^n) \in \Upsilon} \sum_{o \in O} p(o|\bar{f}_j^n, \bar{e}_i^n, \{\mathbf{w}_o\}) \log p(o|\bar{f}_j^n, \bar{e}_i^n, \{\mathbf{w}_o\}) \Big\},$$

we propose using maximum-margin based approaches to learn $\{\mathbf{w}_o\}_{o \in O}$. Under this framework, three discriminative models are introduced, for different purposes of capturing phrase movements. We now describe each of these in the following subsections.

2.2.1 SUPPORT VECTOR MACHINE (SVM) LEARNING

Support vector machines (SVMs) is a learning method which has become very popular in many application areas over recent years (see, e.g., Cristianini and Shawe-Taylor, 2000 for details). The basic SVM is a binary classifier, and we learn each $\mathbf{w}_o$ with a separated SVM that solves the following convex optimisation problem

3. The five-word parameter setting is designed specifically for the MT experiments, which enables each class to have similar sizes of samples.

$$\begin{array}{ll}\min\limits_{\mathbf{w}_o,\xi} & \frac{1}{2}\mathbf{w}_o^T\mathbf{w}_o + C\sum\limits_{(\bar{f}^n,\bar{e}^n)\in\Upsilon}\xi(\bar{f}^n,\bar{e}^n) \\ s.t. & \varphi(o_n,o)\left(\mathbf{w}_o^T\phi(\bar{f}^n,\bar{e}^n)\right) \geq 1-\xi(\bar{f}^n,\bar{e}^n)\ , \\ & \xi(\bar{f}^n,\bar{e}^n)\geq 0,\ \forall(\bar{f}^n,\bar{e}^n)\in\Upsilon\end{array}$$

where $\varphi(o_n,o)$ is an embedding function for the phrase orientation o_n, which is assigned 1 if $o_n = o$ and -1 otherwise.

This approach has been successfully used for many tasks. However, for N training examples (phrase pairs) the computation complexity of the SVM model is somewhere between $O(C_O N + N^2 dim(\phi))$ and $O(C_O N^2 + N^2 dim(\phi))$ (Bishop, 2006). The dependence on C_O may cause computational problems, especially when the number of phrase orientations increase.

2.2.2 MAXIMUM MARGIN REGRESSION (MMR) LEARNING

A good agent for learning $\{\mathbf{w}_o\}_{o\in O}$ should adapt to the number of phrase orientations C_O, enabling Equation (3) to extend to more classes in the future. In this sense, we introduce the maximum margin regression (MMR) technique, that acquires $\{\mathbf{w}_o\}_{o\in O}$ by solving the following optimisation problem (Szedmak et al., 2006)

$$\begin{array}{ll}\min\limits_{\{\mathbf{w}_o\}_{o\in O}} & \frac{1}{2}\sum\limits_{o\in O}\mathbf{w}_o^T\mathbf{w}_o + C\sum\limits_{(\bar{f}^n,\bar{e}^n)\in\Upsilon}\xi(\bar{f}^n,\bar{e}^n) \\ s.t. & \sum\limits_{o\in O}\varphi(o_n,o)\mathbf{w}_o^T\phi(\bar{f}^n,\bar{e}^n) \geq 1-\xi(\bar{f}^n,\bar{e}^n)\ , \\ & \xi(\bar{f}^n,\bar{e}^n)\geq 0,\ \forall(\bar{f}^n,\bar{e}^n)\in\Upsilon\end{array}$$

where $\varphi(o_n,o)$ is an indicator function, which is assigned 1 if the phrase reordering orientation satisfies $o_n = o$ and 0 otherwise.

The computational complexity of MMR is the complexity of a binary SVM (Szedmak et al., 2006), which is independent to the output structure (i.e., number of classes). This allows the orientation class approach presented here to be extended, say to tree structured models, whilst not increasing the computational complexity. Furthermore, it allows the use of non-linear functions, going beyond the approach presented in Zens and Ney (2006), and is expected to provide more flexibility in the expression of phrase features. The reader is referred to Appendix B for further description of MMR.

2.2.3 MAX-MARGIN STRUCTURE (MMS) LEARNING

The two techniques above only consider a fixed margin to separate one orientation class from the others. However, as the phrase reordering orientations tend to be interdependent, introducing flexible margins to separate different orientations sounds more reasonable. Take the five-class setup for example, if an example in class $d \leq -5$ is classified in class $-5 < d < 5$, intuitively the loss should be smaller than when it is classified in class $d > 5$. Therefore, learning $\{\mathbf{w}_o\}_{o\in O}$ is more than a multi-class classification problem: the output (orientation) domain has an inherent structure and the model should respect it. By this motivation, we introduce the max-margin learning framework proposed in Taskar et al. (2003) which is equivalent to minimising the sum of all classification errors

$$\min_{\{\mathbf{w}_o\}_{o\in O}} \frac{1}{N}\sum_{n=1}^{N}\rho(o_n,\bar{f}^n,\bar{e}^n,\{\mathbf{w}_o\}_{o\in O}) + \frac{\lambda}{2}\sum_{o\in O}\|\mathbf{w}_o\|^2, \tag{4}$$

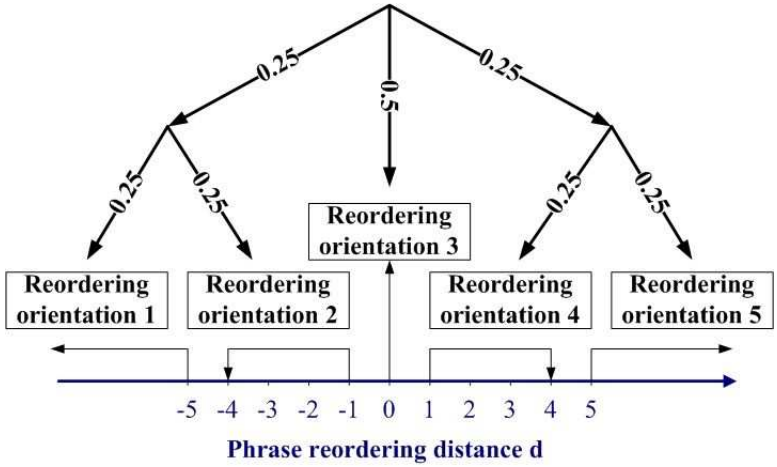


Figure 5: The tree structure constructed by the distance matrix $\triangle(o_n, o')$.

where $\lambda \geq 0$ is a regularisation parameter,

$$\rho(o_n, \overline{f}^n, \overline{e}^n, \{\mathbf{w}_o\}_{o\in O}) = \max\left\{0, \max_{o'\neq o_n}[\triangle(o_n, o') + \mathbf{w}_{o'}^T \phi(\overline{f}^n, \overline{e}^n)] - \mathbf{w}_{o_n}^T \phi(\overline{f}^n, \overline{e}^n)\right\} \quad (5)$$

is a structured margin loss and function $\triangle(o_n, o')$ is applied to measure the "distance" between a pseudo-orientation o' and the correct one o_n. In the experiments, the distance matrix is pre-defined as

$$\triangle(o_n, o') = \begin{cases} 0 & \text{if } o' = o_n \\ 0.5 & \text{if } o' \text{ and } o_n \text{ are close in } O \\ 1 & \text{else} \end{cases} .$$

As shown in Figure 5, this is equivalent to constructing a heuristic tree structure in the orientation domain.

Theoretically, the structured loss (5) requires that the orientation o' which is "far away" from the true orientation o_n must be classified with a large margin $\triangle(o_n, o')$, while nearby candidates are allowed to be classified with a smaller margin. This is an extension of that provided by Collins (2002) where no distance between classes is considered (i.e., $\triangle(o_n, o') = 1, \forall o'$), and it has been applied successfully to phrase translation tasks (Ni et al., 2010b).

Considering the training time, we ignored the regularisation term (i.e., $\lambda = 0$) and used a perceptron-based structured learning (PSL) algorithm to tune the parameters $\{\mathbf{w}_o\}_{o\in O}$, the pseudo-code is demonstrated in Table 2.

Table 2 indicates that the computational complexity of PSL is $O(Ndim(\phi)C_O)$, which still depends on the number of classes. However, compared with the previous SVM and even the MMR models, PSL is substantially faster as in practice the number of classes C_O is much smaller than the number of examples N. This time efficiency is also verified by the experiment results shown in Figure 15.

Notice that in PSL $\mathbf{w}_{o,k+1}$ is tested on the example $(o_n, \phi(\overline{f}^n, \overline{e}^n))$ which is not available for training $\mathbf{w}_{o,k}$, so if we can guarantee a low cumulative loss we are already guarding against over-fitting. If one wished to add regularisation to further guard against over-fitting, one could apply methods

Input of the learner: The samples $\Upsilon = \{o_n, \phi(\bar{f}^n, \bar{e}^n)\}_{n=1}^N$, learning rate η
Initialization: $k = 0$; $\mathbf{w}_{o,k} = \mathbf{0} \quad \forall o \in O$;
Repeat
randomly sample $(\bar{f}^n, \bar{e}^n) \in \Upsilon$
$V = \max_{o'} \{\triangle(o_n, o') + \mathbf{w}_{o',k}^T \phi(\bar{f}^n, \bar{e}^n)\}$
$o^* = \arg\max_{o'} \{\triangle(o_n, o') + \mathbf{w}_{o',k}^T \phi(\bar{f}^n, \bar{e}^n)\}$
if $\mathbf{w}_{o_n,k}^T \phi(\bar{f}^n, \bar{e}^n) < V$ **then**
$\mathbf{w}_{o,k+1}|_{o=o_n} = \mathbf{w}_{o,k}|_{o=o_n} + \eta\phi(\bar{f}^n, \bar{e}^n)$
$\mathbf{w}_{o,k+1}|_{o=o^*} = \mathbf{w}_{o,k}|_{o=o^*} - \eta\phi(\bar{f}^n, \bar{e}^n)$
$k = k + 1$
until converge
Output of the learner: $\mathbf{w}_{o,k+1} \in \mathbb{R}^{dim(\phi)} \quad \forall o \in O$

Table 2: Pseudo-code of perceptron-based structured learning (PSL).

such as ALMA (Gentile, 2001) or NORMA (Kivinen et al., 2004). However, the requirement of normalising $\mathbf{w}_o$ at each step makes the implementation intractable for a large structured learning problem. As an alternative, the risk function (4) can be reformulated as a joint convex optimisation problem[4]

$$\min_{\{\|\mathbf{w}_o\| \le R\}} \max_{\{\mathbf{z}_o \in Z\}} L(\{\mathbf{w}_o\}_{o \in O}, \{\mathbf{z}_o\}_{o \in O}) \tag{6}$$

with

$$L(\{\mathbf{w}_o\}_{o \in O}, \{\mathbf{z}_o\}_{o \in O}) = \sum_{n=1}^{N} \max\left\{0, \sum_{o \in O} z_o^n \left(\triangle(o_n, o) + \mathbf{w}_o^T \phi(\bar{f}^n, \bar{e}^n)\right)\right\}$$
$$s.t. \quad \begin{cases} z_o^n = -1 & o = o_n \\ z_o^n \ge 0 & o \ne o_n \\ \sum_{o \in O} z_o^n = 0 \end{cases} \quad n = 1, \ldots, N \quad .$$

This min-max problem can be solved by the *extra-gradient* algorithm, which is guaranteed to converge linearly to a solution of $\{\mathbf{w}_o^*\}_{o \in O}$ and $\{\mathbf{z}_o^*\}_{o \in O}$ under mild conditions (Taskar et al., 2006).

3. Feature Extraction and Application

In this section, we describe two key steps for the method: feature extraction and model training.

3.1 Feature Extraction

Following (Vickrey et al., 2005; Zens and Ney, 2006), we consider different kinds of information extracted from the phrase environment (see Table 3). To capture unknown grammars and syntactic structures, some of the features would depend on the word-class[5] information. Mathematically, given a sequence s from the feature environment (e.g., $s = [f_{j_l - d_l}, \ldots, f_{j_l}]$ in Figure 6), the features

4. The reader is referred to Appendix A for detailed inference.
5. The word-class tags are provided by the state-of-the-art SMT system (MOSES).

	features for source phrase f_j	features for target phrase $\bar{e_i}$
Context features	Word n-grams within a window (length d) around the source phrase edge $[j_l]$ and $[j_r]$	Word n-grams (subphrases) of the target phrase $[e_{i_l}, \ldots, e_{i_r}]$
Syntactic features	Word-class n-grams within a window (length d) around the source phrase edge $[j_l]$ and $[j_r]$	Word-class n-grams (subphrases) of the target phrase $[e_{i_l}, \ldots, e_{i_r}]$

Table 3: Features extracted from the phrase environment. n-gram indicates a word sequence of length n.

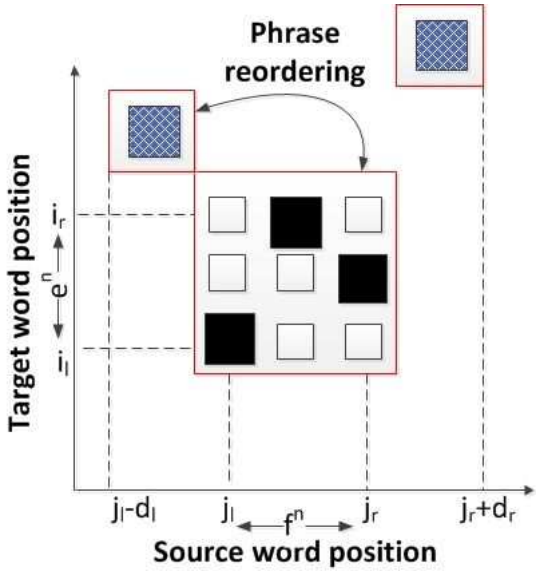


Figure 6: Illustration of the phrase pair $(\bar{f_j^n}, \bar{e_i^n})$ (the word alignments are in black rectangle). The linguistic features are extracted from the target phrase and a window environment (blue shadow boxes) around the source phrase.

extracted are of the form

$$\phi_u(s_p^{|u|}) = \delta(s_p^{|u|}, u),$$

with the indicator function $\delta(\cdot,\cdot)$, $p = \{j_l - d_l, \ldots, j_l, j_r, \ldots, j_r + d_r\}$ and string $s_p^{|u|} = [f_p, \ldots, f_{p+|u|}]$. In this way, the phrase features are distinguished by both the content u and its start position p.

This *position-dependent linguistic* feature expression creates a very high dimensional feature space where each example $(\bar{f_j^n}, \bar{e_i^n})$ is assigned a sparse feature vector. Figure 7 shows the context feature space created for all five phrase pairs in Figure 3 and the non-zero features for the phrase pair ("Xiang gang", "Hong Kong"). The whole feature space contains 180 features and only 9

features are non-zero for this phrase pair. The advantage of this feature expression is the collection of comprehensive linguistic information which may relate to phrase movements. However, the side effect it brings in is a large set of free parameters which may cause over-fitting on the training data.

3.2 Training and Application

The training samples $\{o_n, (\bar{f}^n, \bar{e}^n)\}_{n=1}^N$ (phrase pairs up to length 8) for the DPR model are derived from a general phrase pair extraction procedure described in Koehn et al. (2005). At translation time, we follow the idea of Giménez and Màrquez (2007), where the samples having the same source phrase $\bar{f}$ are considered to be from the same cluster (cf., Figure 8 (a)). A sub-model using the above learning agents is then trained for each cluster. In our largest experiment, this framework results in training approximately $70,000$ sub-DPR models (Figure 8 (b)). A statistics of the number of free parameters (features) against the number of training examples for each cluster is depicted in Figure 8 (c), implying a potential over-fitting risk. To avoid the over-fitting, a prior of $\{\mathbf{w}_o\}_{o\in O}$ is applied to the maximum entropy (ME) model as used in Zens and Ney (2006) and for the MMS model, the *early stopping* strategy[6] is used which involves the careful design of the maximum number of iterations.

During the decoding, the DPR model finds the corresponding sub-model for a source phrase $\bar{f}_j$ and generates the phrase reordering probability for each orientation class with Equation (3). In particular, for the classification experiments, the most-confident orientation is selected as the predicted class.

4. Experiments

Experiments used the *parallel texts of Hong Kong laws*.[7] This bilingual *Chinese-English* corpus consists of mainly legal and documentary texts from Hong Kong which is aligned at the sentence level. The sizes of the corpus are shown in Figure 9. As the vocabulary sizes of the corpus are very small, the content information is relatively easy to learn. However, due to many differences in word order (grammar) occurring for Chinese-English, this corpus contains many long distance phrase movements (see Figure 9). In this case, the phrase reordering model is expected to have more influence on the translation results, which makes this a suitable data set to analyse and demonstrate the effectiveness of our proposed DPR model.

For the experiments, sentences of lengths between 1 and 100 words were extracted and the ratio of source/target lengths was no more than $2:1$. The training set was taken among $\{20K, 50K, 100K, 150K, 185K\}$ sentences while the test set was fixed at $1K$ sentences.

4.1 Classification Experiments

We used *GIZA++* to produce word alignments, enabling us to compare a DPR model against a baseline LR model (Koehn et al., 2005) that uses MLE orientation prediction and a discriminative model that uses an ME framework (Zens and Ney, 2006). In addition, we also compared the clas-

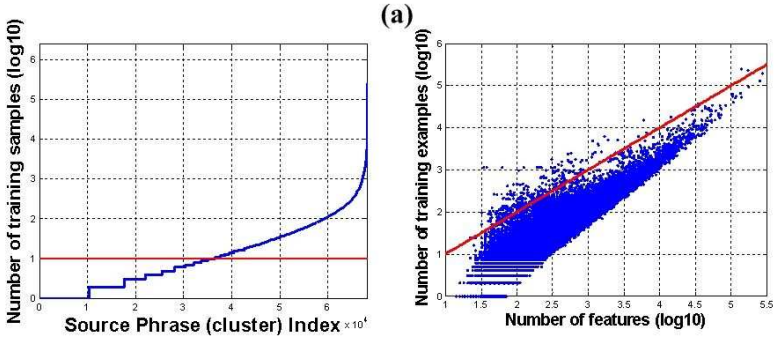
6. The strategy selects the maximum number of iterations and the learning rate η by cross-validating on a validation set. In our experiments, this was done on the $185K$-sentence Chinese-to-English MT task and the (max-iteration, learning rate) with the best performance was chosen for all other MT experiments.
7. The original corpus is available at `http://projects.ldc.upenn.edu/Chinese/hklaws.htm`, which however contains some sentence alignment errors. The corpus has been further cleaned up and aligned at the sentence level by the authors. This refined corpus is now available upon request.

Context (source) / Position	-3	-2	-1	1	2	3
Zhou	0	**1**	0	0	0	0
liu	0	0	**1**	0	0	0
xiang	0	0	0	0	0	0
gang	0	0	0	0	0	0
yi	0	0	0	**1**	0	0
min	0	0	0	0	**1**	0
ju	0	0	0	0	0	**1**
fa	0	0	0	0	0	0
sheng	0	0	0	0	0	0
huo	0	0	0	0	0	0
zai	0	0	0	0	0	0
Zhou liu	0	**1**	0	0	0	0
liu Xiang	0	0	0	0	0	0
Xiang gang	0	0	0	0	0	0
gang yi	0	0	0	0	0	0
yi min	0	0	0	**1**	0	0
min ju	0	0	0	0	**1**	0
ju fa	0	0	0	0	0	0
fa sheng	0	0	0	0	0	0
sheng huo	0	0	0	0	0	0
huo zai	0	0	0	0	0	0
Zhou liu Xiang	0	0	0	0	0	0
liu Xiang gang	0	0	0	0	0	0
Xiang gang yi	0	0	0	0	0	0
gang yi min	0	0	0	0	0	0
yi min ju	0	0	0	**1**	0	0
min ju fa	0	0	0	0	0	0
ju fa sheng	0	0	0	0	0	0
fa sheng huo	0	0	0	0	0	0
sheng huo zai	0	0	0	0	0	0

Figure 7: An example of the linguistic feature space created for all phrases in Figure 3 and the non-zero features for the phrase pair (“Xiang gang”, “Hong Kong”). Due to space limitation, this example only demonstrates the context features for the source phrases (i.e., the top left block in Table 3).

Cluster label	安全 (Chinese character) an quan (Chinese pinyin)
Training Samples N=4	an quan – safety an quan – safely an quan – safety of an quan – safe

(a)



(b) (c)

Figure 8: (a) A cluster for the source phrase "an quan" and its training samples (phrase pairs). Note that the linguistic features for the samples are not demonstrated in this example. (b) The number of training samples for each cluster (phrases are extracted from 185K Chinese-English sentence pairs). (c) The statistics of the number of features against the number of training samples (phrases are extracted from 185K Chinese-English sentence pairs).

Statistics	Chinese	English
Sentence Pairs	216,250	
Running Words	9.76M	7.00M
Vocabulary Size	8,129	23,025

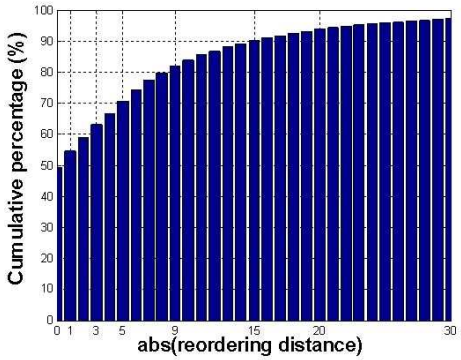


Figure 9: The data statistics for the *parallel texts of Hong Kong laws* corpus (left) and the statistics of phrase reordering distance d for all consistent phrase pairs (up to length 8) extracted from the corpus (right). The word alignments are provided by the word alignment toolkit $GIZA++$. The right figure shows that short distance phrase movements (i.e., $d < 4$) only take up 62% of the whole phrase movements.

Chinese-to-English task										
Orientations	Training set					Test set				
	20K	50K	100K	150K	185K	20K	50K	100K	150K	185K
$d<0$	0.17M	0.45M	0.82M	1.25M	1.63M	13K	16K	16K	17K	17K
$d=0$	0.41M	1.11M	2.10M	3.30M	4.04M	28K	33K	34K	38K	38K
$d>0$	0.12M	0.32M	0.61M	0.90M	1.11M	9K	10K	11K	11K	11K
$d\leq -5$	80K	0.20M	0.38M	0.56M	0.70M	6.0K	6.5K	7.3K	7.5K	7.4K
$-5<d<0$	90K	0.25M	0.44M	0.69M	0.83M	7.0K	9.5K	8.7K	9.5K	9.6K
$d=0$	4.1M	1.11M	2.10M	3.30M	4.04M	28K	33K	34K	38K	38K
$0<d<5$	40K	0.10M	0.20M	0.27M	0.31M	2.5K	2.8K	2.5K	2.4K	2.2K
$d\geq 5$	80K	0.22M	0.41M	0.63M	0.80M	6.5K	7.2K	8.5K	8.6K	8.8K

Table 4: The training and the test sizes (phrase pairs) for three-class setup (top) and five-class setup (bottom), where "K" indicates thousand and "M" indicates million.

sification performance and the computational efficiency among the three learning agents for DPR: SVM,[8] MMR and MMS, where the goal was to find the best learning agent for the MT experiments.

Two orientation classification tasks were carried out: one with three-class setup and one with five-class setup. We discarded points that had long distance reordering ($|d|>15$, representing less than 8% of the data) to avoid some alignment errors caused by *GIZA++*. This results in data sizes shown in Table 4. The classification performance was measured by an overall precision across all orientation classes and the class-specific F1 measures and the experiments were repeated three times to asses variance.

4.1.1 COMPARISON OF OVERALL PRECISIONS AND THE CLASS-SPECIFIC F1-SCORES

Figure 10 shows classification accuracies at different sizes of training sets, and we observed a monotonic increase with the amount of training data used. In addition, all discriminative models perform better than the generative LR model. The MMS approach achieves the best classification performance, with an absolute 8.5% average improvement with three-class setup and an absolute 8.7% average improvement with five classes. Similar improvements are observed when examining class-specific F1 scores on Table 5 and Table 6; the DPR model with the MMS learning agent achieves the best results. However, the DPR models with SVM and MMR techniques do not perform very well in the experiments, possibly due to the feature expression we used. Since constructing a kernel using the sparse feature expression usually results in a very sparse kernel matrix where little similarity between samples is presented, SVM and MMR might not extract adequate information for modelling phrase movements.

When the training sample size is large, the ME model performs better than all other learning agents except MMS, showing its good ability in exploiting features. But when the training sample size is small (e.g., 50K-sentence task), its results are worse than that of SVM, possibly due to the over-fitting on the training data. This reveals the importance of choosing the priors for the ME models: a simple prior may not be helpful while a complicated prior usually makes the training

8. The multi-class SVM model is trained by SVM-Multiclass (Tsochantaridis et al., 2004).

Orientations	Training Data	Generative model	Discriminative models			
		LR	MMR	SVM	ME	MMS
$d<0$	20K	57.2 ± 0.8	63.7 ± 0.6	64.1 ± 0.9	63.9 ± 0.5	$\mathbf{64.7}\pm0.6$
	50K	58.5 ± 0.1	65.6 ± 0.6	65.8 ± 0.7	65.9 ± 0.5	$\mathbf{67.4}\pm0.1$
	100K	61.6 ± 1.1	69.6 ± 1.4	70.6 ± 1.3	71.8 ± 1.3	$\mathbf{74.2}\pm0.3$
	150K	63.8 ± 0.6	72.3 ± 0.8	73.0 ± 0.6	75.3 ± 1.3	$\mathbf{76.5}\pm1.0$
	185K	63.3 ± 0.8	72.2 ± 1.2	73.1 ± 0.8	75.7 ± 1.0	$\mathbf{76.8}\pm1.0$
$d=0$	20K	80.1 ± 0.3	83.6 ± 0.1	84.3 ± 0.2	83.7 ± 0.2	$\mathbf{84.7}\pm0.2$
	50K	80.0 ± 0.1	83.4 ± 0.5	84.5 ± 0.2	84.5 ± 0.3	$\mathbf{85.5}\pm0.2$
	100K	81.7 ± 0.2	85.7 ± 0.6	87.0 ± 0.3	87.8 ± 0.3	$\mathbf{88.6}\pm0.3$
	150K	83.0 ± 0.3	86.8 ± 0.4	88.1 ± 0.3	89.0 ± 0.4	$\mathbf{89.9}\pm0.4$
	185K	82.9 ± 0.2	86.9 ± 0.2	88.2 ± 0.3	89.5 ± 0.3	$\mathbf{90.3}\pm0.2$
$d>0$	20K	44.2 ± 0.8	55.9 ± 0.7	56.6 ± 0.8	55.6 ± 0.6	$\mathbf{58.1}\pm1.0$
	50K	44.3 ± 0.3	54.9 ± 0.5	56.7 ± 0.2	56.1 ± 0.2	$\mathbf{59.3}\pm0.5$
	100K	48.4 ± 2.0	63.6 ± 0.6	65.1 ± 0.2	66.5 ± 0.1	$\mathbf{68.7}\pm0.1$
	150K	51.4 ± 0.6	64.7 ± 0.3	66.5 ± 0.2	68.5 ± 0.5	$\mathbf{70.8}\pm0.3$
	185K	49.2 ± 1.0	64.9 ± 1.5	66.5 ± 0.3	69.6 ± 1.5	$\mathbf{71.5}\pm1.6$

P-value in T-test				
Orientations	Generative model	Discriminative models		
	LR	MMR	SVM	ME
$d<0$	$1.02e-8$	$2.75e-5$	$7.27e-5$	$1.00e-4$
$d=0$	$8.66e-10$	$8.39e-7$	$1.55e-5$	$2.19e-6$
$d>0$	$1.97e-9$	$1.45e-6$	$7.19e-6$	$3.84e-9$

Table 5: Classification performance on the Chinese-English corpus: the class-specific F1-scores [%] for three-class setup. Bold numbers refer to the best results. P-values of T-test for statistical significance in the differences between MMS and other models are shown in the lower table.

time increase dramatically. Hence, how to choose the appropriate priors for ME in order to balance training speed and performance is often difficult. Alternatively, using the early stopping strategy DPR with MMS does not over-fit the training data, indicating that the PSL algorithm in company with early stopping already guards against over-fitting.

Figure 11 further demonstrates the average precision for each reordering distance d on the 185K-sentence task (five-class setup), using the results provided by LR, ME and DPR with MMS respectively. It shows that even for long distance reordering, the DPR model still performs well, while the LR baseline usually performs badly (more than half examples are classified incorrectly). With so many classification errors, the effect of this baseline in an SMT system is in doubt, even with a powerful language model. Meanwhile, we observed that results for forward phrase movements (i.e., $d<0$) are better than those for backward reorderings (i.e., $d>0$). We postulate this is because the reordering patterns for backward reorderings also depend on the orientation classes of the phrases nearby. For example, in Figure 3, the phrase "on a building" would be in "forward reordering" if it does not meet another "forward" phrase "a fire has taken place". This observation shows that a richer feature set including a potential orientation class of nearby phrases may help the reordering classification and will be investigated in our future work.

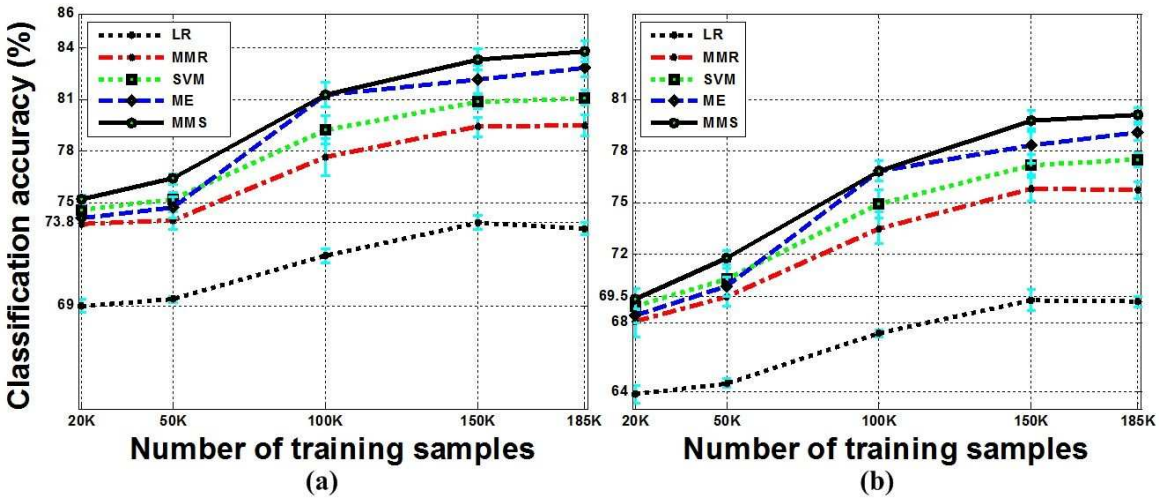


Figure 10: The overall classification precision of three-class setup (Figure (a)) and five-class setup (Figure (b)), where "*K*" indicates thousand and the error bars show the variances.

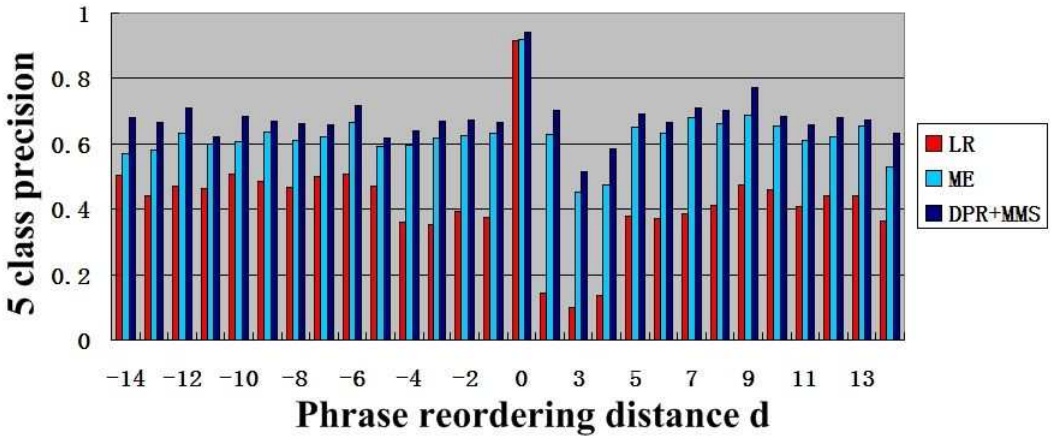


Figure 11: Five-class classification precision with respect to d on the 185K-sentence task. A similar trend is also observed on the three-class classification precision.

Orientations	Training Data	Generative model	Discriminative models			
		LR	MMR	SVM	ME	MMS
$d \leq -5$	20K	40.9 ± 2.4	46.2 ± 1.8	$\mathbf{47.2} \pm 2.4$	45.6 ± 1.9	47.0 ± 1.5
	50K	41.0 ± 0.2	46.5 ± 0.6	48.1 ± 0.4	47.5 ± 0.3	$\mathbf{49.6} \pm 0.7$
	100K	46.9 ± 0.1	54.7 ± 1.5	56.3 ± 0.8	57.3 ± 0.5	$\mathbf{58.7} \pm 0.8$
	150K	47.6 ± 0.9	57.1 ± 1.1	58.9 ± 1.4	60.5 ± 1.3	$\mathbf{62.1} \pm 1.1$
	185K	47.8 ± 0.3	57.6 ± 0.4	59.3 ± 0.3	61.8 ± 0.7	$\mathbf{63.4} \pm 0.7$
$-5 < d < 0$	20K	35.0 ± 1.5	44.6 ± 1.6	45.2 ± 1.3	46.6 ± 1.1	$\mathbf{47.6} \pm 1.0$
	50K	40.8 ± 1.5	52.3 ± 1.2	52.4 ± 0.8	53.8 ± 0.7	$\mathbf{55.5} \pm 0.2$
	100K	43.3 ± 0.5	55.3 ± 1.2	56.1 ± 1.5	58.7 ± 1.4	$\mathbf{60.9} \pm 0.5$
	150K	47.8 ± 1.7	60.8 ± 2.0	61.8 ± 2.0	65.1 ± 2.5	$\mathbf{66.1} \pm 2.0$
	185K	45.7 ± 1.5	59.2 ± 1.5	61.0 ± 1.5	64.8 ± 1.6	$\mathbf{66.0} \pm 1.5$
$d = 0$	20K	79.9 ± 0.3	83.6 ± 0.2	84.0 ± 0.2	83.9 ± 0.2	$\mathbf{84.7} \pm 0.3$
	50K	80.0 ± 0.1	83.7 ± 0.2	84.3 ± 0.2	84.4 ± 0.2	$\mathbf{85.5} \pm 0.2$
	100K	81.4 ± 0.1	86.0 ± 0.6	86.8 ± 0.5	87.6 ± 0.4	$\mathbf{88.6} \pm 0.4$
	150K	82.7 ± 0.3	87.2 ± 0.3	87.9 ± 0.3	88.8 ± 0.5	$\mathbf{89.8} \pm 0.3$
	185K	82.7 ± 0.2	87.2 ± 0.1	88.2 ± 0.2	89.5 ± 0.3	$\mathbf{90.4} \pm 0.2$
$0 < d < 5$	20K	13.4 ± 1.8	39.2 ± 3.0	42.8 ± 3.5	41.0 ± 3.0	$\mathbf{46.3} \pm 2.5$
	50K	22.0 ± 1.7	44.5 ± 1.0	47.6 ± 0.6	45.5 ± 0.4	$\mathbf{50.8} \pm 0.6$
	100K	19.2 ± 2.4	50.9 ± 0.9	53.6 ± 1.5	54.6 ± 1.3	$\mathbf{58.1} \pm 1.1$
	150K	23.8 ± 0.7	50.2 ± 0.9	54.4 ± 0.7	56.8 ± 1.7	$\mathbf{60.4} \pm 1.1$
	185K	19.6 ± 2.8	47.8 ± 2.8	51.4 ± 3.0	56.2 ± 3.7	$\mathbf{60.0} \pm 3.0$
$d \geq 5$	20K	41.4 ± 0.9	47.9 ± 3.5	50.7 ± 1.1	49.9 ± 1.0	$\mathbf{50.8} \pm 2.1$
	50K	39.4 ± 0.8	49.5 ± 0.2	50.9 ± 0.5	50.8 ± 0.4	$\mathbf{55.4} \pm 0.5$
	100K	47.0 ± 1.3	59.9 ± 0.1	61.2 ± 0.6	62.7 ± 0.6	$\mathbf{64.5} \pm 0.8$
	150K	48.8 ± 0.5	62.0 ± 0.1	63.8 ± 0.2	65.2 ± 0.6	$\mathbf{67.1} \pm 0.2$
	185K	49.4 ± 0.6	62.9 ± 1.3	64.9 ± 1.3	67.2 ± 1.4	$\mathbf{68.8} \pm 1.2$

P-value in T-test				
Orientations	Generative model	Discriminative models		
	LR	MMR	SVM	ME
$d \leq -5$	$6.19e-7$	$3.93e-5$	$9.30e-3$	$7.89e-10$
$-5 < d < 0$	$7.77e-10$	$3.07e-7$	$3.69e-7$	$2.02e-5$
$d = 0$	$1.14e-9$	$1.19e-6$	$3.50e-6$	$8.99e-11$
$0 < d < 5$	$2.36e-11$	$5.21e-8$	$8.50e-5$	$9.51e-9$
$d \geq 5$	$4.76e-8$	$4.25e-7$	$8.56e-4$	$1.00e-3$

Table 6: Classification performance on the Chinese-English corpus: the class-specific F1-scores [%] for five-class setup. Bold numbers refer to the best results. P-values of T-test for statistical significance in the differences between MMS and other models are shown in the lower table.

4.1.2 EXPLORING DPR WITH MMS

With the above general view, the DPR model with the MMS learning agent has shown to be the best classifier. Here we further explore its advantages by analysing more detailed results.

Figure 12 first illustrates the relative improvements of DPR with MMS over LR, ME and DPRs with MMR and SVM on the switching classes (i.e., $d \neq 0$), where we observed that the relative

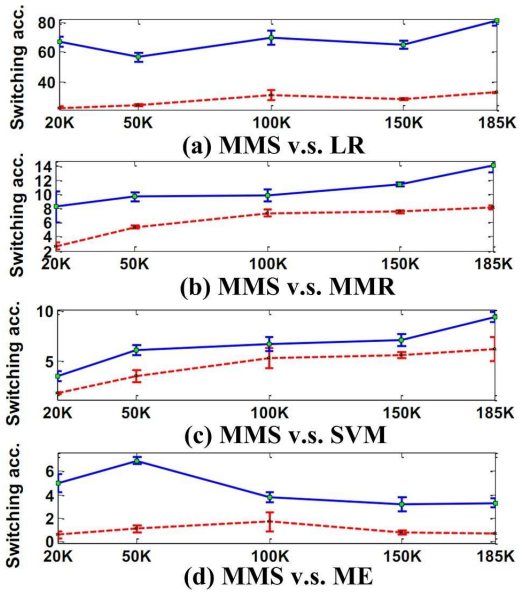


Figure 12: The average relative improvements of DPR with MMS over (a) LR, (b) DPR with MMR, (c) DPR with SVM and (d) ME on the switching classes (i.e., $d \neq 0$) for three-class setup (Red dashed lines) and five-class setup (Blue solid lines). "*K*" indicates thousand and the error bars show the variances.

improvement with five-class setup is usually greater than that with three-class setup. This implies the more orientation classes DPR has, the better performance MMS achieves compared with other models. This observation makes MMS the most promising learning agent in our future work where we expect to extend the orientation set further.

We then compared the DPR model with MMS with the LR and the ME models according to the overall precision of each cluster on Figure 13. Compared with the generative model LR, DPR performs better in many of the clusters, especially when given enough training samples (the black lines in the figure). This verifies the advantage of discriminative models. In particular, the number of larger circles which imply greater ambiguity in target translations is greater than that of larger rectangles; indicating MMS performs better in these ambiguous clusters, implying that the target translations also contain useful information about phrase movements.

Comparing the two discriminative models, the cluster improvement of DPR over ME is smaller than that over LR, represented by the reduced number of circles and the increased number of rectangles. However, DPR with MMS still achieves a stable improvement over ME. This is especially true when the training samples are not adequate (represented by more circles in Figure 13 (c) than in Figure 13 (d)), where the ME model is more likely to over-fit while the DPR with MMS still performs well.

Finally we illustrate three examples on Figure 14, where we observed a great improvement of DPR over LR. The first (top) example demonstrates the benefit from the target translations as by translating the Chinese source phrase "you guan" into different English words (i.e., "relating", "relates" or "relevant"), the phrase pairs usually have different but regular movements. The second

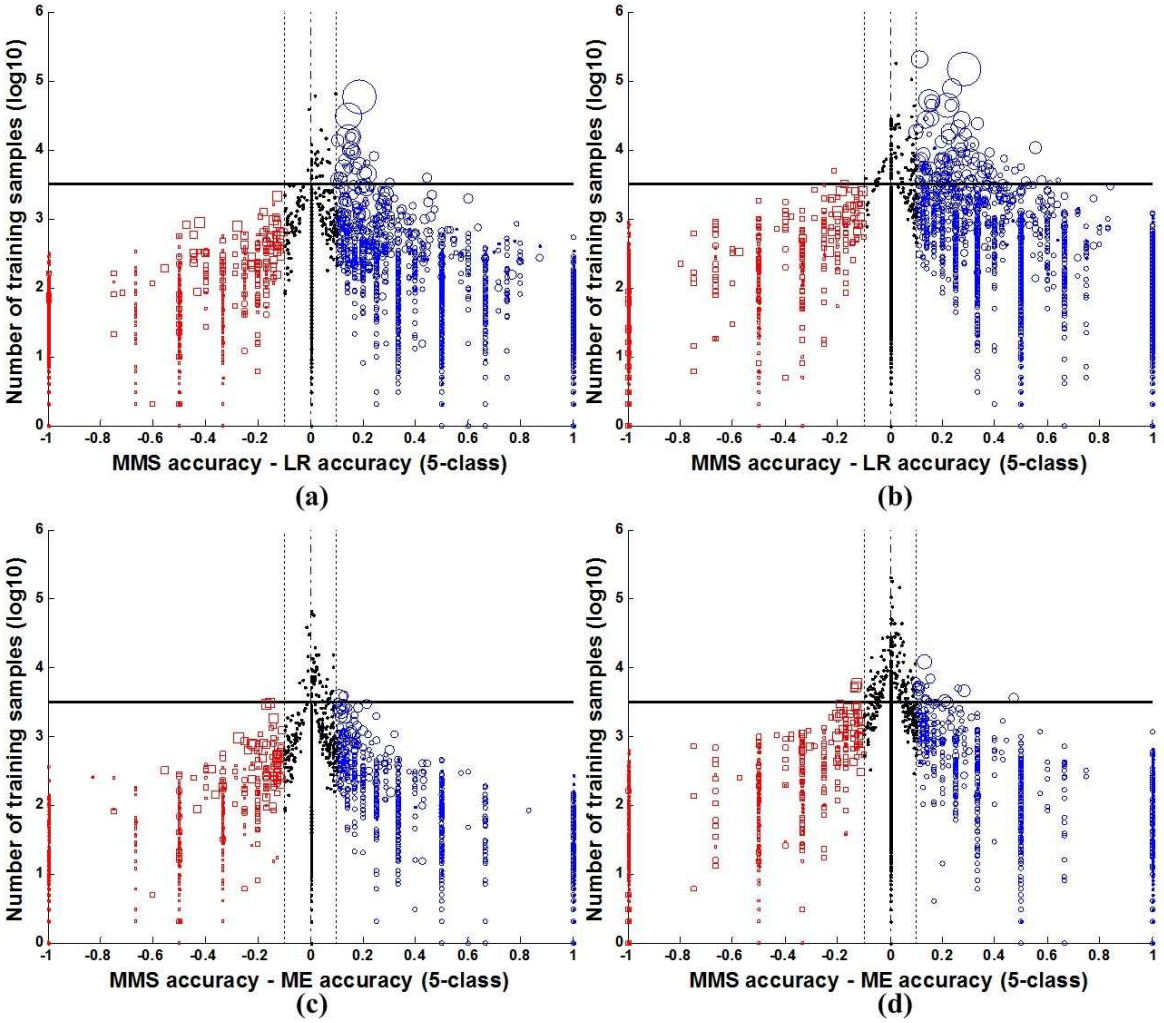


Figure 13: Scatter-plots comparing the cluster accuracies of DPR with MMS with the LR (top) and ME (bottom) models on 50K-sentence task (left) and 150K-sentence task (right). Each circle/rectangle/point represents a cluster that contains all phrase pairs with a unique source phrase (e.g., Figure 8 (a)). Those clusters for which the performance difference (x-axes) is greater than 0.1 are shown as rectangles and circles, the areas of which are proportional to the number of target translations in them. The y-axes show the number of training samples (in log 10 scale) for each cluster.

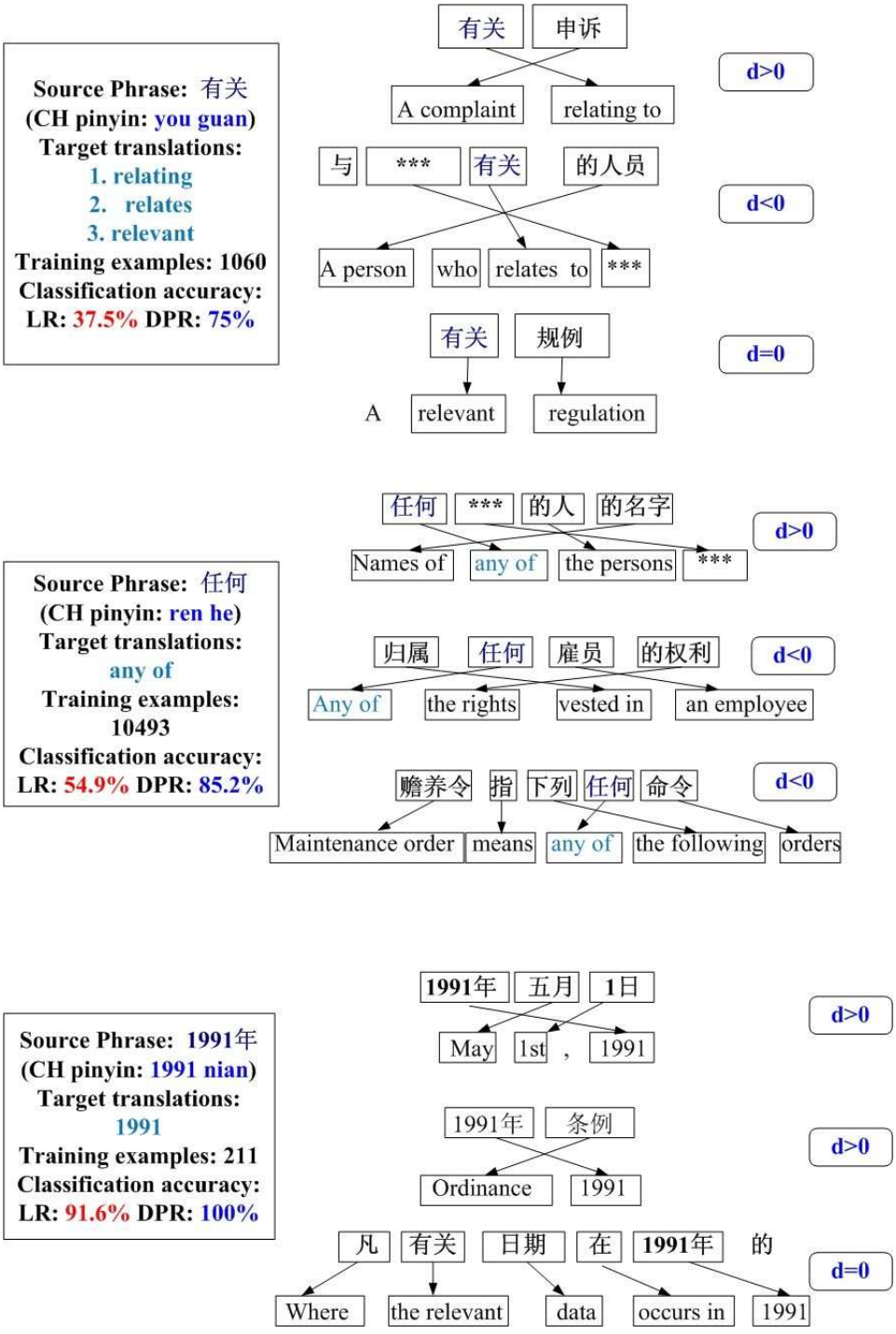


Figure 14: Phrase movements captured by the DPR model with MMS on the 50*K*-sentence task.

(middle) example shows a grammatical structure captured by the DPR model: in English the phrase "any of" usually stays in front of the subjects (or objects) it modifies. In general, when given enough training samples a discriminative model such as DPR is able to capture various grammatical structures (modelled by phrase movements) better than a generative model. The final (bottom) example depicts one type of phrase movements caused by the constant expressions in different languages (e.g., date expression). Although such expressions can be covered manually with a rule-based MT system, they can easily be captured by a DPR model as well. Hence, we conclude that the frequent phrase movements, whether caused by different grammatical structures or rule-based expressions, can be captured and the movement information is then passed on to an MT decoder to organise the target sentence structures.

4.1.3 A COMPARISON OF THE TRAINING TIME

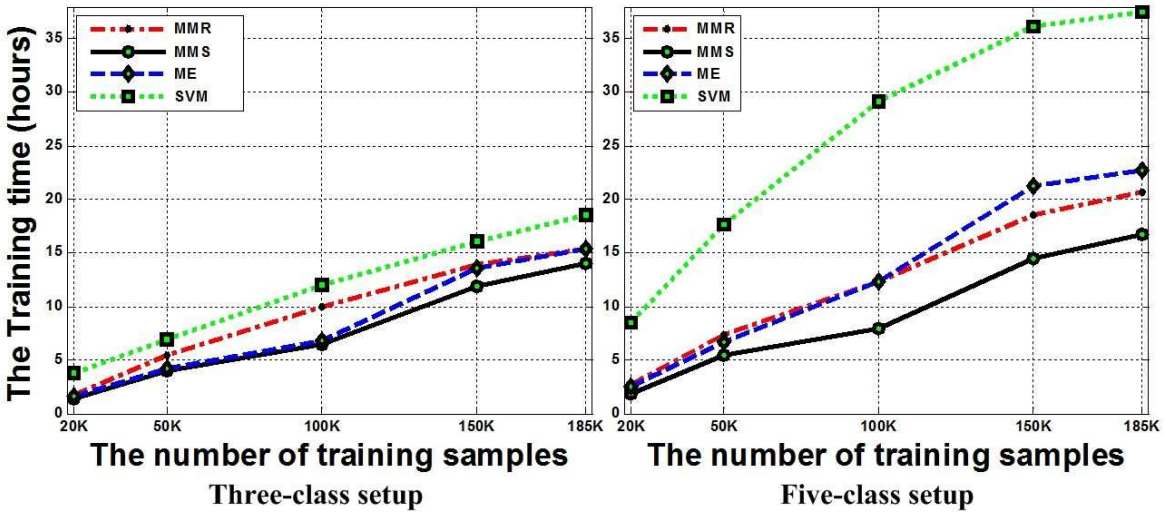


Figure 15: The training time of MMR, ME, MMS (coded in Python) and SVM (coded in C++) to reach the same training error tolerance.

As a comparison, we plot on Figure 15 the training time of MMS, MMR, ME and SVM to reach the same training error tolerance.[9] For the DPR model, MMS is the fastest as expected where in contrast the SVM technique is the slowest. Moreover, training a DPR model with MMS is faster than training an ME model, especially when the number of classes increase. This is because the *generalised iterative scaling* (GIS) algorithm for an ME model requires going through all samples twice at each round: one is for updating the conditional distributions $p(o|\overline{f}_j,\overline{e}_i)$ and the other is for updating $\{\mathbf{w}_o\}_{o\in O}$. Alternatively, the PSL algorithm only goes through all examples once at each round, making it faster and more applicable for larger data sets.

9. The MMS, MMR and ME models are coded in Python while SVM-multiclass is coded in C++.

4.2 Machine Translation Experiments

We now test the effectiveness of the DPR model in an MT system, using a state-of-the-art SMT system—MOSES (Koehn et al., 2005) that models phrase movements with the LR models as a baseline system. To keep the comparison fair, our MT system just replaces MOSES's LR models with DPR while sharing all other components (i.e., a phrase translation probability model, a 4-gram language model (Stolcke, 2002) and the beam search decoder). In addition, we also compared the DPR model with the ME model in Zens and Ney (2006) on the 50K-sentence MT task, where the results confirmed that DPR can lead to improvement performance.

We chose MMS as the learning agent for the DPR model in consideration of its prominent classification performance. In detail, all consistent phrase pairs (up to length 8) were extracted from the training sentence pairs and form the sample pool. The DPR model was then trained by the PSL algorithm and the function $h(z) = \exp(z)$ was applied to Equation (3) to transform the prediction scores.

To make use of the phrase reordering probabilities, two strategies were applied: one is to use the probabilities directly as the reordering cost (dotted line in Figure 1), which is also used in Xiong et al. (2006); Zens and Ney (2006); the other is to use them to adjust the word distance-based reordering cost (solid line in Figure 1), where the reordering cost of a sentence is computed as

$$h_d(\overline{\mathbf{f}}^I, \overline{\mathbf{e}}^I) = - \sum_{(\overline{f}_{j_m}, \overline{e}_{i_m}) \in (\overline{\mathbf{f}}^I, \overline{\mathbf{e}}^I)} \frac{d_m}{\beta p(o | \overline{f}_{j_m}, \overline{e}_{i_m})} \tag{7}$$

with tuning parameter β. Intuitively, if the DPR model has a large orientation set (i.e., the phrase movements are modelled in a precise way) and the orientation predictions are good enough, it is reasonable to use the reordering probabilities directly. However, as we experienced in Section 4.1, the DPR predictions with five-class setup still need improvement, especially for the switching orientations (i.e., $d \neq 0$). On the other hand, if the DPR model only uses a small orientation set (e.g., three-class setup), it is able to provide very good orientation predictions. But all long distance phrase movements will have the same reordering probabilities, which may mislead the SMT decoder and spoil the translations. In this case, the distance-sensitive expression (7) is able to fill the deficiency of a small-class setup of DPR by penalising long distance phrase movements. Hence in the MT experiments, we used the five-class phrase reordering probabilities directly while the three-class probabilities were used to adjust the word distance-based reordering cost.

For parameter tuning, minimum-error-rating training (MERT) (Och, 2003) is used to tune the parameters. Note that there are seven parameters which need tuning in MOSES's LR models, while there is only one for DPR. The translation performance is evaluated by four standard MT measurements, namely *word error rate* (WER) (Tillmann et al., 1997), *BLEU* (Papineni et al., 2002), *NIST* (Doddington, 2002) and *METEOR* (Banerjee and Lavie, 2005). In effect, WER and NIST weight more on word/phrase translation accuracy; BLEU biases towards translation fluency; and METEOR emphasises translation adequacy (i.e., word/phrase translation recall). The reader is referred to Callison-Burch et al. (2007) and Ni (2010) for detailed discussions on these measurements.

We first demonstrate on Table 7 a comparison of the DPR model with the LR and the ME models on the 50K-sentence Chinese-to-English MT task. The improvements on most evaluations over LR and ME are consistent with what are observed on the reordering classification experiments. However, the MT results show no difference between three-class setup and five-class setup, possibly due to the low classification accuracy of DPR with five-class setup (especially on the switching

Measure	MOSES	DPR		ME	
		3-class	5-class	3-class	5-class
WER [%]	**24.3** $\pm$ 0.6	24.6 $\pm$ 1.5	24.7 $\pm$ 1.1	25.3 $\pm$ 1.7	26.0 $\pm$ 2.1
BLEU [%]	44.5 $\pm$ 1.2	47.1 $\pm$ 1.3	**47.5** $\pm$ 1.2	46.17 $\pm$ 1.7	45.0 $\pm$ 2.5
NIST	8.73 $\pm$ 0.11	**9.04** $\pm$ 0.26	9.03 $\pm$ 0.32	8.72 $\pm$ 0.26	8.49 $\pm$ 0.49
METEOR [%]	66.1 $\pm$ 0.8	**66.4** $\pm$ 1.1	66.1 $\pm$ 1.1	65.0 $\pm$ 1.7	63.9 $\pm$ 2.6

Table 7: The comparison of the DPR model with the LR and the ME models on the 50K-sentence MT task.

classes). How to improve the classification accuracy of DPR with a large class setup is hence a main challenge for our future work.

Since the three-class DPR achieves the same translation quality but it is faster, for the other MT tasks we only used DPR with three-class setup as the phrase reordering model. Figure 16 illustrates the comparison of the DPR MT system with the baseline MOSES according to the four MT evaluations, where we observed consistent improvements on most evaluations. Furthermore, the larger the sample size is, the better results DPR will achieve. This again shows the learning ability of the DPR model when given enough samples. In particular, both systems produce similar predictions in sentence content (represented by similar WERs), but our MT system does better at phrase reordering and produces more fluent translations (represented by better BLEUs).

However, if the sample size is small (e.g., the 20K-sentence task), DPR is unable to collect adequate phrase reordering information. In this case the application of DPR to an MT system may involve a risk of a reduction in translation quality (represented by the low qualities on WER and METEOR).

5. Conclusions and Future Work

We have proposed a distance phrase reordering (DPR) model using a classification scheme and trained and evaluated it in a structured learning framework. The phrase reordering classification tasks have shown that DPR is better at capturing phrase movements over the LR and ME models. The MMS learning agent in particular, achieves outstanding performance in terms of classification accuracy and computational efficiency. An analysis of performance confirms that the proposed MMS method is shown to perform particularly well when there is a large amount of training data, and on translation examples with large ambiguity in the target language domain.

Machine translation experiments carried out on the Chinese-English corpus show that DPR gives more fluent translation results, which confirms its effectiveness. On the other hand, when training data is sparse, the process may involve a risk of a reduction in translation quality.

For future work, we aim to improve the prediction accuracy of the five-class setup before applying it to an MT system, as DPR can be more powerful if it is able to provide more precise phrase positions for the decoder. We also aim to formulate the phrase reordering problem as an ordinal regression problem rather than a classification one proposed in this paper. Furthermore, we will refine the learning framework of DPR by carefully designing or automatically learning the distance matrix $\triangle(o_n, o')$. A richer feature set to better characterise the grammar reorderings is also a direc-

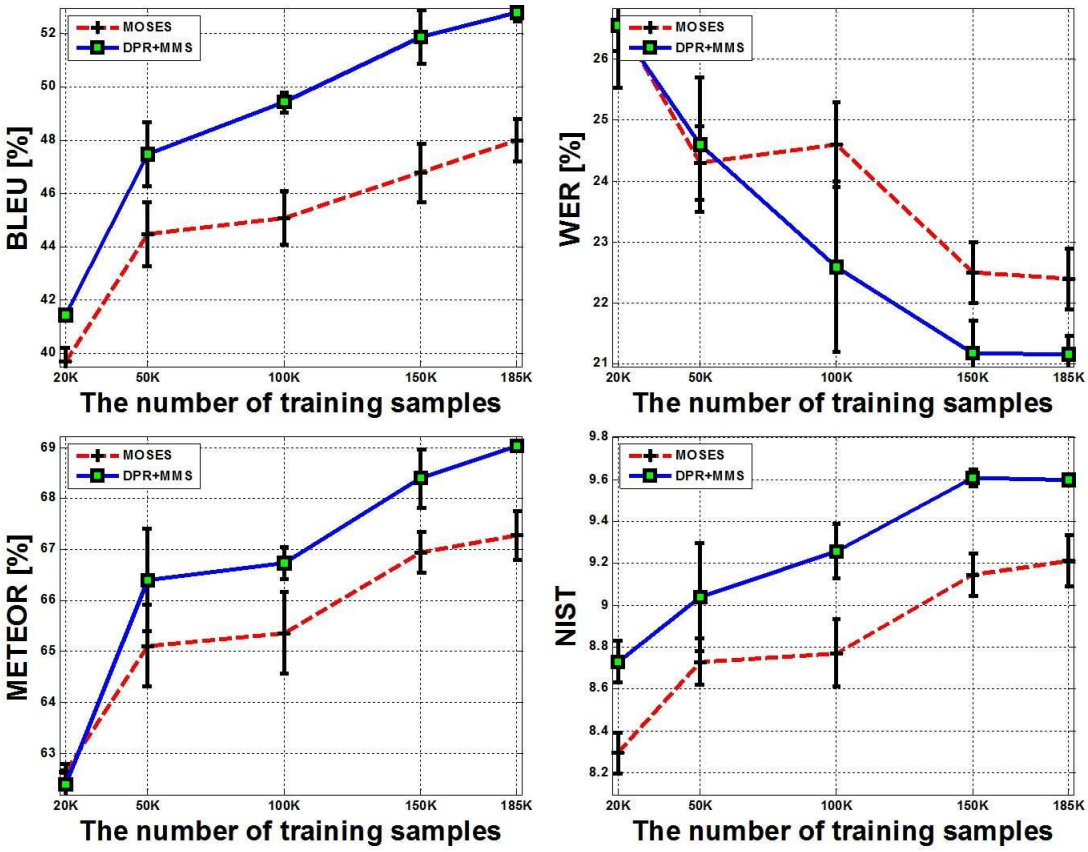


Figure 16: The translation evaluations.

tion of our current investigations. Finally we will try the DPR model on larger corpora (e.g., the NIST Chinese-English corpus), with the purpose of verifying its ability in scaling up to large data collections.

Acknowledgments

This work was funded from an European Community project SMART (FP7-033917). Y. Ni was funded by a scholarship from the School of Electronics and Computer Science, University of Southampton. Integration of the work as part of MOSES was funded by the PASCAL Network of Excellence. Particular thanks go to Prof Philip Koehn and Mr Hieu Hoang, University of Edinburgh, for help with implementation details of MOSES.

Appendix A.

In this appendix we infer the optimisation problem (6) from (4).

To consider adding a regularisation term, we upper bound the norm of each $\mathbf{w}_o$ by $\|\mathbf{w}_o\| \leq R$. Then minimising (5) with respect to $\{\mathbf{w}_o\}_{o\in O}$ is equivalent to solving the following optimisation problem

$$\min_{\|\mathbf{w}_o\|\leq R} L\big(\{\mathbf{w}_o\}_{o\in O}\big), \tag{8}$$

where the cumulative loss $L\big(\{\mathbf{w}_o\}_{o\in O}\big) = \sum_n \rho\big(o_n, \bar{f}^n, \bar{e}^n, \{\mathbf{w}_o\}_{o\in O}\big)$.

We can then express the sub maximisation problem $\max_{o\neq o_n}[\triangle(o_n,o) + \mathbf{w}_o^T\phi(\bar{f}^n,\bar{e}^n)] - \mathbf{w}_{o_n}^T\phi(\bar{f}^n,\bar{e}^n)$ in Equation (5) as a *linear programming* problem

$$\begin{array}{ll} \max\limits_{\mathbf{z}_n} & \sum\limits_{o\in O} z_n^o[\triangle(o_n,o) + \mathbf{w}_o^T\phi(\bar{f}^n,\bar{e}^n)] \\ s.t. & \sum\limits_c z_n^o = 0 \\ & z_n^{o_n} = -1 \\ & z_n^o \geq 0,\ o \in \{O\backslash o_n\} \end{array} . \tag{9}$$

Let $\mathcal{Z}_n$ denote the closed set of $\mathbf{z}_n$, $\mathbf{z} = \{\mathbf{z}_1,\ldots,\mathbf{z}_N\}$ and $\mathcal{Z} = \mathcal{Z}_1 \times \ldots \times \mathcal{Z}_N$, substituting (9) into (8) yields a natural saddle-point form

$$\min_{\|\mathbf{w}_o\|\leq R} \max_{\mathbf{z}\in\mathcal{Z}} L(\{\mathbf{w}_o\}_{o\in O}, \{\mathbf{z}_o\}_{o\in O})$$

with

$$\begin{array}{rl} L(\{\mathbf{w}_o\}_{o\in O}, \{\mathbf{z}_o\}_{o\in O}) = & \sum\limits_{n=1}^{N} \max\Big\{0, \sum\limits_{o\in O} z_o^n\big(\triangle(o_n,o) + \mathbf{w}_o^T\phi(\bar{f}^n,\bar{e}^n)\big)\Big\} \\ s.t. & \begin{cases} z_o^n = -1 & o = o_n \\ z_o^n \geq 0 & o \neq o_n \\ \sum\limits_{o\in O} z_o^n = 0 \end{cases} \quad n = 1,\ldots,N \end{array} ,$$

which is the optimisation problem (6).

Appendix B.

In this appendix, we describe more details about maximum margin regression (MMR).

To illuminate the background of the MMR, let the constraints

$$\begin{aligned} \text{s.t.} \quad & \sum_{o\in O} \varphi(o_n,o)\mathbf{w}_o^T\phi(\bar{f}^n,\bar{e}^n) \geq 1-\xi(\bar{f}^n,\bar{e}^n) \\ & \xi(\bar{f}^n,\bar{e}^n) \geq 0,\ \forall(\bar{f}^n,\bar{e}^n)\in\Upsilon \end{aligned}$$

be transformed into an inner product based form

$$\begin{aligned} & \sum_{o\in O} \varphi(o_n,o)\mathbf{w}_o^T\phi(\bar{f}^n,\bar{e}^n) \geq 1-\xi(\bar{f}^n,\bar{e}^n) \\ & \Rightarrow \langle\varphi(o_n),\mathbf{W}\phi(\bar{f}^n,\bar{e}^n)\rangle \geq 1-\xi(\bar{f}^n,\bar{e}^n) \end{aligned} \quad , \tag{10}$$

where the following short hand notations are applied: $\mathbf{W}=\{\mathbf{w}_o^T\}\in\mathbb{R}^{|O|\times dim(\phi)}$ is a matrix in which the rows correspond to the row vectors $\mathbf{w}_o^T$, and $\varphi(o_n)=(\varphi(o_n,o)),\ o\in O$ is a vector of the indicator values.

To give a possible interpretation to the inner product based constraints consider the well known cosine rule connecting the inner product and the distance via norms in a L_2 norm space, namely for every pair of vectors $\mathbf{a},\mathbf{b}$ of this space the following holds $\|\mathbf{a}-\mathbf{b}\|_2^2=\|\mathbf{a}\|_2^2+\|\mathbf{b}\|_2^2-2\langle\mathbf{a},\mathbf{b}\rangle$. Exploiting this equality the inner product based constraints (10) can be transformed into an equivalent, norm based one

$$\|\varphi(o_n)\|_2^2+\|\mathbf{W}\phi(\bar{f}^n,\bar{e}^n)\|_2^2-2+2\xi(\bar{f}^n,\bar{e}^n)\geq\|\varphi(o_n)-\mathbf{W}\phi(\bar{f}^n,\bar{e}^n)\|_2^2 \ . \tag{11}$$

Constraints (11) state that the squared distance between the vector valued representation of the outputs and the image of the input vectors with respect to the linear operator matrix $\mathbf{W}$ is bounded above by summing of the square norm of outputs, the norm of the input image and a tolerance given by the margin. Therefore if the Frobenius norm of $\mathbf{W}$ is minimised then the constraints force the distance between the outputs and the image of the inputs to be small. If the norm of all outputs are the same this minimisation works uniformly, otherwise larger distance error is allowed for outputs with greater norm.

The optimisation problem of MMR allows to use implicit representation not only for the inputs but also for the outputs. To see that one can introduce Lagrangian multipliers, $\alpha(\bar{f}^n,\bar{e}^n)$, to each constraints and write up the dual problem for MMR

$$\begin{aligned} \min \quad & \tfrac{1}{2}\sum_{(\bar{f}^n,\bar{e}^n)}\sum_{(\bar{f}^m,\bar{e}^m)}\alpha(\bar{f}^n,\bar{e}^n)\alpha(\bar{f}^m,\bar{e}^m)\overbrace{\langle\varphi(o_n),\varphi(o_m)\rangle}^{\kappa^{\varphi}_{nm}}\overbrace{\langle\phi(\bar{f}^n,\bar{e}^n),\phi(\bar{f}^m,\bar{e}^m)\rangle}^{\kappa^{\phi}_{nm}} \\ & -\sum_{(\bar{f}^n,\bar{e}^n)}\alpha(\bar{f}^n,\bar{e}^n) \\ \text{w.r.t} \quad & \alpha(\bar{f}^n,\bar{e}^n),\ \forall(\bar{f}^n,\bar{e}^n)\in\Upsilon \\ \text{s.t.} \quad & 0\leq\alpha(\bar{f}^n,\bar{e}^n)\leq C,\ \forall(\bar{f}^n,\bar{e}^n)\in\Upsilon \end{aligned} \quad ,$$

where κ^{ϕ}_{nm} and κ^{φ}_{nm} stand for the inner products between the input and the output pairs respectively. Therefore, to solve the dual problem requires only the knowledge of the values of inner products of the output pairs and not their concrete feature representation. See further motivations behind the formulation of MMR in Szedmak and Hussain (2009).

References

S. Banerjee and A. Lavie. Meteor: an automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and Summarisation*, pages 65–72, Ann Arbor, Michigan, June 2005.

W. S. Bennett and J. Slocum. The lrc machine translation system. *Computational Linguistics*, 11 (2-3):111–121, 1985.

A. Berger, S. D. Pietra, and V. D. Pietra. A maximum entropy approach to natural language processing. *Computational Linguistics*, 22(1):39–72, March 1996.

C. M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

C. Callison-Burch, C. Fordyce, P. Koehn, C. Monz, and J. Schroeder. (meta-) evaluation of machine translation. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 136–158, Prague, Czech Republic, June 2007.

M. Collins. Discriminative training methods for hidden markov models: theory and experiments with percepton algorithms. In *C. Sammut, A. G. Hoffmann (Eds.) Proceedings of the 19th International Conference on Machine Learning (ICML 2002)*, 2002.

N. Cristianini and J. Shawe-Taylor. *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods*. Cambridge University Press, 2000.

G. Doddington. 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)*, 2002.

C. Gentile. A new approximate maximal margin classification algorithm. *Journal of Machine Learning Research*, 2:213–242, 2001.

J. Giménez and L. Màrquez. Context–aware discriminative phrase selection for statistical machine translation. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 159–166, Prague, June 2007.

J. Kivinen, A. J. Smola, and R. C. Williamson. Online learning with kernels. *IEEE Transactions on Signal Processing*, 52(8):2165–2176, 2004.

K. Knight. Decoding complexity in word replacement translation models. *Computational Linguistics*, 25(2):607–615, 1999.

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

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

Y. Ni. *Beyond Multi-class–Structured Learning for Machine Translation*. PhD thesis, University of Southampton, Southampton, U.K., 2010.

Y. Ni, M. Niranjan, C. Saunders, and S. Szedmak. Distance phrase reordering for moses - user manual and code guide. Technical report, School of Electronics and Computer Science, University of Southampton, Southampton, UK, 2010a.

Y. Ni, C. Saunders, S. Szedmak, and M. Niranjan. The application of structured learning in natural language processing. *Machine Translation Journal*, 24(2):71–85, 2010b.

F. J. Och. Minimum error rate training in statistical machine translation. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics (ACL 2003)*, Japan, September 2003.

F. J. Och, C. Tillmann, and H. Ney. Improved alignment models for statistical machine translation. In *Proceedings of the 1999 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, pages 20–28, University of Maryland, College Park, MD, June 1999.

K. Papineni, S. Roukos, T. Ward, and W. Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL 2002)*, 2002.

A. Stolcke. Srilm - an extensible language modeling toolkit. In *J. H. L. Hansen and B. Pellom (Eds.) Proceedings of the 7th International Conference on Spoken Language Processing (ICSLP 2002)*, Denver, Colorado, USA, September 2002.

S. Szedmak and Z. Hussain. A universal machine learning optimization framework for arbitrary outputs. Technical report, PASCAL, Southampton, UK, 2009.

S. Szedmak, J. Shawe-Taylor, and E. Parado-Hernandez. Learning via linear operators: maximum margin regression; multiclass and multiview learning at one-class complexity. Technical report, PASCAL, Southampton, UK, 2006.

B. Taskar, C. Guestrin, and D. Koller. Max–margin markov networks. In *S. Thrun, L. K. Saul, B. Schölkopf (Eds.) Proceedings of 7th Annual Conference on Neural Information Processing Systems (NIPS 2003)*, Vancouver, Canada, 2003.

B. Taskar, S. Lacoste-Julien, and M. I. Jordan. Structured prediction, dual extragradient and bregman projections. *Journal of Machine Learning Research, "Special Topic on Machine Learning and Optimization"*, pages 1627–1653, 2006.

C. Tillmann. A block orientation model for statistical machine translation. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL 2004)*, Boston, MA, USA, 2004.

C. Tillmann and T. Zhang. A localized prediction model for statistical machine translation. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL 2005)*, pages 557–564, Ann Arbor, MI, June 2005.

C. Tillmann, S. Vogel, H. Ney, H. Sawaf, and A. Zubiaga. Accelerated dp based search for statistical translation. In *Proceedings of the 5th European Conference on Speech Communication and Technology*, 1997.

I. Tsochantaridis, T. Hofmann, T. Joachims, and Y. Altun. Support vector machine learning for interdependent and structured output spaces. In *Russ Greiner, Dale Schuurmans (Eds.) Proceedings of the 21st International Machine Learning Conference (ICML 2004)*. ACM Press, 2004.

D. Vickrey, L. Biewald, M. Teyssier, and D. Koller. Word–sense disambiguation for machine translation. In *Proceedings of the Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing (HLT–EMNLP 2005)*, pages 771–778, 2005.

W. Weaver. Translation. In W. N. Locke and A. D. Booth, editors, *Machine Translation of Languages: Fourteen Essays*, pages 15–23, 1949.

D. Xiong, Q. Liu, and S. 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 Association for Computational Linguistics*, pages 521–528, Sydney, July 2006.

R. Zens and H. Ney. Discriminative reordering models for statistical machine translation. In *Proceedings of the Workshop on Statistical Machine Translation*, pages 55–63, New York City, June 2006.

R. Zens, O. Bender, S. Hasan, S. Khadivi, E. Matusov, J. Xu, Y. Zhang, and H. Ney. The rwth phrase–based statistical machine translation system. In *Proceedings of the International Workshop on Spoken Language Translation (IWSLT 2005)*, pages 155–162, Pittsburgh, PA, October 2005.