A Discriminative Latent Variable Model for Statistical Machine Translation

Phil Blunsom, Trevor Cohn and **Miles Osborne**
School of Informatics, University of Edinburgh
2 Buccleuch Place, Edinburgh, EH8 9LW, UK
{pblunsom,tcohn,miles}@inf.ed.ac.uk

Abstract

Large-scale discriminative machine translation promises to further the state-of-the-art, but has failed to deliver convincing gains over current heuristic frequency count systems. We argue that a principle reason for this failure is not dealing with multiple, equivalent translations. We present a translation model which models derivations as a latent variable, in both training and decoding, and is fully discriminative and globally optimised. Results show that accounting for multiple derivations does indeed improve performance. Additionally, we show that regularisation is essential for maximum conditional likelihood models in order to avoid degenerate solutions.

1 Introduction

Statistical machine translation (SMT) has seen a resurgence in popularity in recent years, with progress being driven by a move to phrase-based and syntax-inspired approaches. Progress within these approaches however has been less dramatic. We believe this is because these frequency count based[1] models cannot easily incorporate non-independent and overlapping features, which are extremely useful in describing the translation process. Discriminative models of translation can include such features without making assumptions of independence or explicitly modelling their interdependence. However, while discriminative models promise much, they have not been shown to deliver significant gains over their simpler cousins. We argue that this is due to a number of inherent problems that discriminative models for SMT must address, in particular the problems of spurious ambiguity and degenerate solutions. These occur when there are many ways to translate a source sentence to the same target sentence by applying a sequence of steps (a *derivation*) of either phrase translations or synchronous grammar rules, depending on the type of system. Existing discriminative models require a reference derivation to optimise against, however no parallel corpora annotated for derivations exist. Ideally, a model would account for this ambiguity by marginalising out the derivations, thus predicting the best *translation* rather than the best *derivation*. However, doing so exactly is NP-complete. For this reason, to our knowledge, all discriminative models proposed to date either side-step the problem by choosing simple model and feature structures, such that spurious ambiguity is lessened or removed entirely (Ittycheriah and Roukos, 2007; Watanabe et al., 2007), or else ignore the problem and treat derivations as translations (Liang et al., 2006; Tillmann and Zhang, 2007).

In this paper we directly address the problem of spurious ambiguity in discriminative models. We use a synchronous context free grammar (SCFG) translation system (Chiang, 2007), a model which has yielded state-of-the-art results on many translation tasks. We present two main contributions. First, we develop a log-linear model of translation which is globally trained on a significant number of parallel sentences. This model maximises the conditional likelihood of the data, $p(\mathbf{e}|\mathbf{f})$, where $\mathbf{e}$ and $\mathbf{f}$ are the English and foreign sentences, respectively. Our estimation method is theoretically sound, avoiding the biases of the heuristic relative frequency estimates

[1] We class approaches using minimum error rate training (Och, 2003) *frequency count based* as these systems re-scale a handful of generative features estimated from frequency counts and do not support large sets of non-independent features.

Proceedings of ACL-08: HLT, pages 200–208,
Columbus, Ohio, USA, June 2008. ©2008 Association for Computational Linguistics

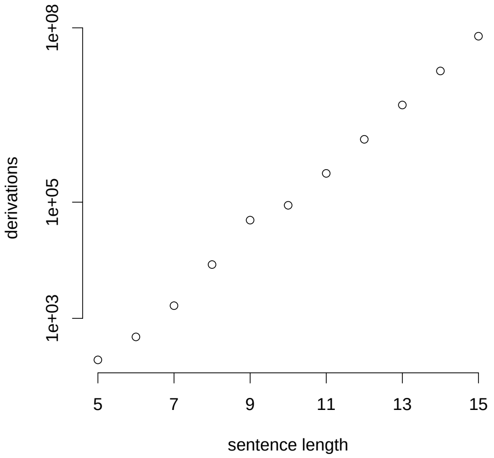

Figure 1. Exponential relationship between sentence length and the average number of derivations (on a log scale) for each reference sentence in our training corpus.

(Koehn et al., 2003). Second, within this framework, we model the derivation, $\mathbf{d}$, as a latent variable, $p(\mathbf{e}, \mathbf{d}|\mathbf{f})$, which is marginalised out in training and decoding. We show empirically that this treatment results in significant improvements over a maximum-derivation model.

The paper is structured as follows. In Section 2 we list the challenges that discriminative SMT must face above and beyond the current systems. We situate our work, and previous work, on discriminative systems in this context. We present our model in Section 3, including our means of training and decoding. Section 4 reports our experimental setup and results, and finally we conclude in Section 5.

2 Challenges for Discriminative SMT

Discriminative models allow for the use of expressive features, in the order of thousands or millions, which can reference arbitrary aspects of the source sentence. Given most successful SMT models have a highly lexicalised grammar (or grammar equivalent), these features can be used to smuggle in linguistic information, such as syntax and document context. With this undoubted advantage come four major challenges when compared to standard frequency count SMT models:

1. There is no one reference derivation. Often there are thousands of ways of translating a source sentence into the reference translation. Figure 1 illustrates the exponential relationship between sentence length and the number of derivations. Training is difficult without a clear target, and predicting only one derivation at test time is fraught with danger.

2. Parallel translation data is often very noisy, with such problems as non-literal translations, poor sentence- and word-alignments. A model which exactly translates the training data will inevitably perform poorly on held-out data. This problem of over-fitting is exacerbated in discriminative models with large, expressive, feature sets. Regularisation is essential for models with more than a handful of features.

3. Learning with a large feature set requires many training examples and typically many iterations of a solver during training. While current models focus solely on efficient decoding, discriminative models must also allow for efficient training.

Past work on discriminative SMT only address some of these problems. To our knowledge no systems directly address Problem 1, instead choosing to ignore the problem by using one or a small handful of reference derivations in an n-best list (Liang et al., 2006; Watanabe et al., 2007), or else making local independence assumptions which side-step the issue (Ittycheriah and Roukos, 2007; Tillmann and Zhang, 2007; Wellington et al., 2006). These systems all include regularisation, thereby addressing Problem 2. An interesting counterpoint is the work of DeNero et al. (2006), who show that their unregularised model finds degenerate solutions. Some of these discriminative systems have been trained on large training sets (Problem 3); these systems are the local models, for which training is much simpler. Both the global models (Liang et al., 2006; Watanabe et al., 2007) use fairly small training sets, and there is no evidence that their techniques will scale to larger data sets.

Our model addresses all three of the above problems within a global model, without resorting to n-best lists or local independence assumptions. Furthermore, our model explicitly accounts for spurious ambiguity without altering the model structure or arbitrarily selecting one derivation. Instead we model the translation distribution with a latent variable for the derivation, which we marginalise out in training and decoding.

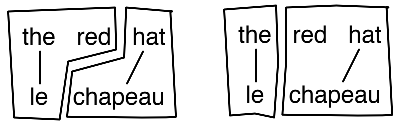

Figure 2. The dropping of an adjective in this example means that there is no one segmentation that we could choose that would allow a system to learn $le \rightarrow the$ and $chapeau \rightarrow hat$.

$$\begin{aligned}
&\langle S \rangle \rightarrow \langle S_{\boxed{1}}\, X_{\boxed{2}},\ S_{\boxed{1}}\, X_{\boxed{2}} \rangle \\
&\langle S \rangle \rightarrow \langle X_{\boxed{1}},\ X_{\boxed{1}} \rangle \\
&\langle X \rangle \rightarrow \langle ne\ X_{\boxed{1}}\, pas,\ does\ not\ X_{\boxed{1}} \rangle \\
&\langle X \rangle \rightarrow \langle va,\ go \rangle \\
&\langle X \rangle \rightarrow \langle il,\ he \rangle
\end{aligned}$$

Figure 3. A simple SCFG, with non-terminal symbols S and X, which performs the transduction: $il\ ne\ vas\ pas \Rightarrow he\ does\ not\ go$

This itself provides robustness to noisy data, in addition to the explicit regularisation from a prior over the model parameters. For example, in many cases there is no one perfect derivation, but rather many imperfect ones which each include some good translation fragments. The model can learn from many of these derivations and thereby learn from all these translation fragments. This situation is illustrated in Figure 2 where the non-translated adjective *red* means neither segmentation is 'correct', although both together present positive evidence for the two lexical translations.

We present efficient methods for training and prediction, demonstrating their scaling properties by training on more than a hundred thousand training sentences. Finally, we stress that our main findings are general ones. These results could – and should – be applied to other models, discriminative and generative, phrase- and syntax-based, to further progress the state-of-the-art in machine translation.

3 Discriminative Synchronous Transduction

A synchronous context free grammar (SCFG) consists of paired CFG rules with co-indexed non-terminals (Lewis II and Stearns, 1968). By assigning the source and target languages to the respective sides of a SCFG it is possible to describe translation as the process of parsing the source sentence using a CFG, while generating the target translation from the other (Chiang, 2007). All the models we present use the grammar extraction technique described in Chiang (2007), and are bench-marked against our own implementation of this hierarchical model (Hiero). Figure 3 shows a simple instance of a hierarchical grammar with two non-terminals. Note that our approach is general and could be used with other synchronous grammar transducers (e.g., Galley et al. (2006)).

3.1 A global log-linear model

Our log-linear translation model defines a conditional probability distribution over the target translations of a given source sentence. A particular sequence of SCFG rule applications which produces a translation from a source sentence is referred to as a *derivation*, and each translation may be produced by many different derivations. As the training data only provides source and target sentences, the derivations are modelled as a latent variable.

The conditional probability of a derivation, $\mathbf{d}$, for a target translation, $\mathbf{e}$, conditioned on the source, $\mathbf{f}$, is given by:

$$p_\Lambda(\mathbf{d}, \mathbf{e}|\mathbf{f}) = \frac{\exp \sum_k \lambda_k H_k(\mathbf{d}, \mathbf{e}, \mathbf{f})}{Z_\Lambda(\mathbf{f})} \qquad (1)$$

$$\text{where} \quad H_k(\mathbf{d}, \mathbf{e}, \mathbf{f}) = \sum_{r \in \mathbf{d}} h_k(\mathbf{f}, r) \qquad (2)$$

Here k ranges over the model's features, and $\Lambda = \{\lambda_k\}$ are the model parameters (weights for their corresponding features). The feature functions H_k are predefined real-valued functions over the source and target sentences, and can include overlapping and non-independent features of the data. The features must decompose with the derivation, as shown in (2). The features can reference the entire source sentence coupled with each rule, r, in a derivation. The distribution is globally normalised by the partition function, $Z_\Lambda(\mathbf{f})$, which sums out the numerator in (1) for every derivation (and therefore every translation) of $\mathbf{f}$:

$$Z_\Lambda(\mathbf{f}) = \sum_{\mathbf{e}} \sum_{\mathbf{d} \in \Delta(\mathbf{e}, \mathbf{f})} \exp \sum_k \lambda_k H_k(\mathbf{d}, \mathbf{e}, \mathbf{f})$$

Given (1), the conditional probability of a target translation given the source is the sum over all of its derivations:

$$p_\Lambda(\mathbf{e}|\mathbf{f}) = \sum_{\mathbf{d} \in \Delta(\mathbf{e}, \mathbf{f})} p_\Lambda(\mathbf{d}, \mathbf{e}|\mathbf{f}) \qquad (3)$$

where $\Delta(\mathbf{e}, \mathbf{f})$ is the set of all derivations of the target sentence **e** from the source **f**.

Most prior work in SMT, both generative and discriminative, has approximated the sum over derivations by choosing a single 'best' derivation using a Viterbi or beam search algorithm. In this work we show that it is both tractable and desirable to directly account for derivational ambiguity. Our findings echo those observed for latent variable log-linear models successfully used in monolingual parsing (Clark and Curran, 2007; Petrov et al., 2007). These models marginalise over derivations leading to a dependency structure and splits of non-terminal categories in a PCFG, respectively.

3.2 Training

The parameters of our model are estimated from our training sample using a maximum *a posteriori* (MAP) estimator. This maximises the likelihood of the parallel training sentences, $\mathcal{D} = \{(\mathbf{e}, \mathbf{f})\}$, penalised using a prior, i.e., $\Lambda^{MAP} = \arg\max_\Lambda p_\Lambda(\mathcal{D})p(\Lambda)$. We use a zero-mean Gaussian prior with the probability density function $p_0(\lambda_k) \propto \exp\left(-\lambda_k^2/2\sigma^2\right)$.[2] This results in the following log-likelihood objective and corresponding gradient:

$$\mathcal{L} = \sum_{(\mathbf{e},\mathbf{f})\in\mathcal{D}} \log p_\Lambda(\mathbf{e}|\mathbf{f}) + \sum_k \log p_0(\lambda_k) \tag{4}$$

$$\frac{\partial \mathcal{L}}{\partial \lambda_k} = E_{p_\Lambda(\mathbf{d}|\mathbf{e},\mathbf{f})}[h_k] - E_{p_\Lambda(\mathbf{e}|\mathbf{f})}[h_k] - \frac{\lambda_k}{\sigma^2} \tag{5}$$

In order to train the model, we maximise equation (4) using L-BFGS (Malouf, 2002; Sha and Pereira, 2003). This method has been demonstrated to be effective for (non-convex) log-linear models with latent variables (Clark and Curran, 2004; Petrov et al., 2007). Each L-BFGS iteration requires the objective value and its gradient with respect to the model parameters. These are calculated using inside-outside inference over the feature forest defined by the SCFG parse chart of **f** yielding the partition function, $Z_\Lambda(\mathbf{f})$, required for the log-likelihood, and the marginals, required for its derivatives.

Efficiently calculating the objective and its gradient requires two separate packed charts, each representing a derivation forest. The first one is the full chart over the space of possible derivations given the source sentence. The inside-outside algorithm over this chart gives the marginal probabilities for each chart cell, from which we can find the feature expectations. The second chart contains the space of derivations which produce the reference translation from the source. The derivations in this chart are a subset of those in the full derivation chart. Again, we use the inside-outside algorithm to find the 'reference' feature expectations from this chart. These expectations are analogous to the empirical observation of maximum entropy classifiers.

Given these two charts we can calculate the log-likelihood of the reference translation as the inside-score from the sentence spanning cell of the reference chart, normalised by the inside-score of the spanning cell from the full chart. The gradient is calculated as the difference of the feature expectations of the two charts. Clark and Curran (2004) provides a more complete discussion of parsing with a log-linear model and latent variables.

The full derivation chart is produced using a CYK parser in the same manner as Chiang (2005), and has complexity $O(|\mathbf{e}|^3)$. We produce the reference chart by synchronously parsing the source and reference sentences using a variant of CYK algorithm over two dimensions, with a time complexity of $O(|\mathbf{e}|^3|\mathbf{f}|^3)$. This is an instance of the ITG alignment algorithm (Wu, 1997). This step requires the reference translation for each training instance to be contained in the model's hypothesis space. Achieving full coverage implies inducing a grammar which generates all observed source-target pairs, which is difficult in practise. Instead we discard the unreachable portion of the training sample (24% in our experiments). The proportion of discarded sentences is a function of the grammar used. Extraction heuristics other than the method used herein (Chiang, 2007) could allow complete coverage (e.g., Galley et al. (2004)).

3.3 Decoding

Accounting for all derivations of a given translation should benefit not only training, but also decoding. Unfortunately marginalising over derivations in decoding is NP-complete. The standard solution is to approximate the maximum probability translation using a single derivation (Koehn et al., 2003).

Here we approximate the sum over derivations directly using a beam search in which we produce a beam of high probability translation sub-strings for each cell in the parse chart. This algorithm is sim-

[2] In general, any conjugate prior could be used instead of a simple Gaussian.

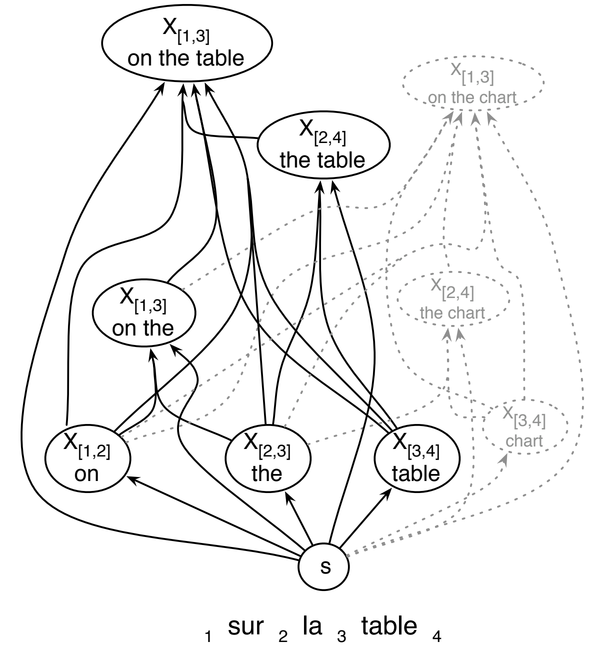

Figure 4. Hypergraph representation of max translation decoding. Each chart cell must store the entire target string generated.

ilar to the methods for decoding with a SCFG intersected with an n-gram language model, which require language model contexts to be stored in each chart cell. However, while Chiang (2005) stores an abbreviated context composed of the $n-1$ target words on the left and right edge of the target substring, here we store the entire target string. Additionally, instead of maximising scores in each beam cell, we sum the inside scores for each derivation that produces a given string for that cell. When the beam search is complete we have a list of translations in the top beam cell spanning the entire source sentence along with their approximated inside derivation scores. Thus we can assign each translation string a probability by normalising its inside score by the sum of the inside scores of all the translations spanning the entire sentence.

Figure 4 illustrates the search process for the simple grammar from Table 2. Each graph node represents a hypothesis translation substring covering a sub-span of the source string. The space of translation sub-strings is exponential in each cell's span, and our algorithm can only sum over a small fraction of the possible strings. Therefore the resulting probabilities are only estimates. However, as demonstrated in Section 4, this algorithm is considerably more effective than maximum derivation (Viterbi) decoding.

4 Evaluation

Our model evaluation was motivated by the following questions: (1) the effect of maximising translations rather than derivations in training and decoding; (2) whether a regularised model performs better than a maximum likelihood model; (3) how the performance of our model compares with a frequency count based hierarchical system; and (4) how translation performance scales with the number of training examples.

We performed all of our experiments on the Europarl V2 French-English parallel corpus.[3] The training data was created by filtering the full corpus for all the French sentences between five and fifteen words in length, resulting in 170K sentence pairs. These limits were chosen as a compromise between experiment turnaround time and leaving a large enough corpus to obtain indicative results. The development and test data was taken from the 2006 NAACL and 2007 ACL workshops on machine translation, also filtered for sentence length.[4] Tuning of the regularisation parameter and MERT training of the benchmark models was performed on *dev2006*, while the test set was the concatenation of *devtest2006*, *test2006* and *test2007*, amounting to 315 development and 1164 test sentences.

Here we focus on evaluating our model's basic ability to learn a conditional distribution from simple binary features, directly comparable to those currently employed in frequency count models. As such, our base model includes a single binary identity feature per-rule, equivalent to the $p(e|f)$ parameters defined on each rule in standard models.

As previously noted, our model must be able to derive the reference sentence from the source for it to be included in training. For both our discriminative and benchmark (Hiero) we extracted our grammar on the 170K sentence corpus using the approach described in Chiang (2007), resulting in 7.8 million rules. The discriminative model was then trained on the training partition, however only 130K of the sentences were used as the model could not produce a derivation of the reference for the remaining sentences. There were many grammar rules that the discriminative model did not observe in a reference derivation, and thus could not assign their feature a positive weight. While the benchmark model has a

[3] http://www.statmt.org/europarl/

[4] http://www.statmt.org/wmt0{6,7}

	Decoding	
Training	derivation	translation
All Derivations	28.71	31.23
Single Derivation	26.70	27.32
ML ($\sigma^2 = \infty$)	25.57	25.97

Table 1. A comparison on the impact of accounting for all derivations in training and decoding (development set).

positive count for every rule (7.8M), the discriminative model only observes 1.7M rules in actual reference derivations. Figure 1 illustrates the massive ambiguity present in the training data, with fifteen word sentences averaging over 70M reference derivations.

Performance is evaluated using cased BLEU4 score on the test set. Although there is no direct relationship between BLEU and likelihood, it provides a rough measure for comparing performance.

Derivational ambiguity Table 1 shows the impact of accounting for derivational ambiguity in training and decoding.[5] There are two options for training, we could use our latent variable model and optimise the probability of all derivations of the reference translation, or choose a single derivation that yields the reference and optimise its probability alone. The second option raises the difficult question of which one, of the thousands available, we should choose? We use the derivation which contains the most rules. The intuition is that small rules are likely to appear more frequently, and thus generalise better to a test set. In decoding we can search for the maximum probability derivation, which is the standard practice in SMT, or for the maximum probability translation which is what we actually want from our model, i.e. the best translation.

The results clearly indicate the value in optimising translations, rather than derivations. Max-translation decoding for the model trained on single derivations has only a small positive effect, while for the latent variable model the impact is much larger.[6]

For example, our max-derivation model trained on the Europarl data translates *carte sur la table* as *on the table card*. This error in the reordering of *card* (which is an acceptable translation of *carte*) is due to the rule $\langle X \rangle \rightarrow \langle carte\ X_{\boxed{1}},\ X_{\boxed{1}}\ card \rangle$ being the highest scoring rule for *carte*. This is reasonable, as *carte* is a noun, which in the training data, is often observed with a trailing adjective which needs to be reordered when translating into English. In the example there is no adjective, but the simple hierarchical grammar cannot detect this. The max-translation model finds a good translation *card on the table*. This is due to the many rules that enforce monotone ordering around *sur la*, $\langle X \rangle \rightarrow \langle X_{\boxed{1}}\ sur,\ X_{\boxed{1}}\ in \rangle$ $\langle X \rangle \rightarrow \langle X_{\boxed{1}}\ sur\ la\ X_{\boxed{2}},\ X_{\boxed{1}}\ in\ the\ X_{\boxed{2}} \rangle$ etc. The scores of these many monotone rules sum to be greater than the reordering rule, thus allowing the model to use the weight of evidence to settle on the correct ordering.

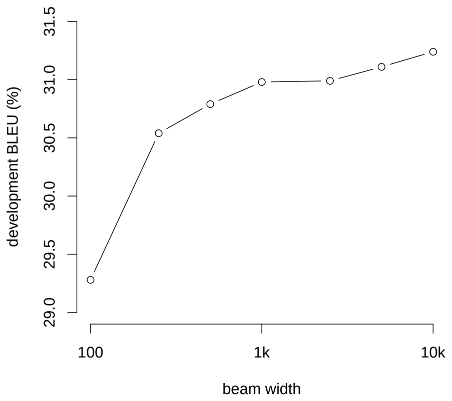

Figure 5. The effect of the beam width (log-scale) on max-translation decoding (development set).

Having established that the search for the best translation is effective, the question remains as to how the beam width over partial translations affects performance. Figure 5 shows the relationship between beam width and development BLEU. Even with a very tight beam of 100, max-translation decoding outperforms maximum-derivation decoding, and performance is increasing even at a width of 10k. In subsequent experiments we use a beam of 5k which provides a good trade-off between performance and speed.

Regularisation Table 1 shows that the performance of an unregularised maximum likelihood model lags well behind the regularised max-translation model. From this we can conclude that the maximum likelihood model is overfitting the training set. We suggest that is a result of the degenerate solutions of the conditional maximum likelihood estimate, as described in DeNero et al. (2006). Here we assert that our regularised *maximum a pos-*

[5] When not explicitly stated, both here and in subsequent results, the regularisation parameter was set to one, $\sigma^2 = 1$.

[6] We also experimented with using max-translation decoding for standard MER trained translation models, finding that it had a small negative impact on BLEU score.

Grammar Rules	ML ($\sigma^2 = \infty$)	MAP ($\sigma^2 = 1$)
$\langle X \rangle \rightarrow \langle carte,\ map \rangle$	1.0	0.5
$\langle X \rangle \rightarrow \langle carte,\ notice \rangle$	0.0	0.5
$\langle X \rangle \rightarrow \langle sur,\ on \rangle$	1.0	1.0
$\langle X \rangle \rightarrow \langle la,\ the \rangle$	1.0	1.0
$\langle X \rangle \rightarrow \langle table,\ table \rangle$	1.0	0.5
$\langle X \rangle \rightarrow \langle table,\ chart \rangle$	0.0	0.5
$\langle X \rangle \rightarrow \langle carte\ sur,\ notice\ on \rangle$	1.0	0.5
$\langle X \rangle \rightarrow \langle carte\ sur,\ map\ on \rangle$	0.0	0.5
$\langle X \rangle \rightarrow \langle sur\ la,\ on\ the \rangle$	1.0	1.0
$\langle X \rangle \rightarrow \langle la\ table,\ the\ table \rangle$	0.0	0.5
$\langle X \rangle \rightarrow \langle la\ table,\ the\ chart \rangle$	1.0	0.5
Training data: carte sur la table ↔ map on the table carte sur la table ↔ notice on the chart		

Table 2. Comparison of the susceptibility to degenerate solutions for a ML and MAP optimised model, using a simple grammar with one parameter per rule and a monotone glue rule: $\langle X \rangle \rightarrow \langle X_{\boxed{1}} X_{\boxed{2}},\ X_{\boxed{1}} X_{\boxed{2}} \rangle$

teriori model avoids such solutions.

This is illustrated in Table 2, which shows the conditional probabilities for rules, obtained by locally normalising the rule feature weights for a simple grammar extracted from the ambiguous pair of sentences presented in DeNero et al. (2006). The first column of conditional probabilities corresponds to a maximum likelihood estimate, i.e., without regularisation. As expected, the model finds a degenerate solution in which overlapping rules are exploited in order to minimise the entropy of the rule translation distributions. The second column shows the solution found by our model when regularised by a Gaussian prior with unit variance. Here we see that the model finds the desired solution in which the true ambiguity of the translation rules is preserved. The intuition is that in order to find a degenerate solution, dispreferred rules must be given large negative weights. However the prior penalises large weights, and therefore the best strategy for the regularised model is to evenly distribute probability mass.

Translation comparison Having demonstrated that accounting for derivational ambiguity leads to improvements for our discriminative model, we now place the performance of our system in the context of the standard approach to hierarchical translation. To do this we use our own implementation of Hiero (Chiang, 2007), with the same grammar but with the traditional generative feature set trained in a linear model with minimum BLEU training. The feature set includes: a trigram language model (*lm*) trained on the English side of the unfiltered Europarl corpus; direct and reverse translation scores estimated as relative frequencies (p_d, p_r); lexical translation scores (p_d^{lex}, p_r^{lex}), a binary flag for the glue rule which allows the model to (dis)favour monotone translation (*gr*); and rule and target word counts (*rc, wc*).

System	Test (BLEU)
Discriminative max-derivation	25.78
Hiero (p_d, gr, rc, wc)	26.48
Discriminative max-translation	27.72
Hiero ($p_d, p_r, p_d^{lex}, p_r^{lex}, gr, rc, wc$)	28.14
Hiero ($p_d, p_r, p_d^{lex}, p_r^{lex}, gr, rc, wc, lm$)	32.00

Table 3. Test set performance compared with a standard Hiero system

Table 3 shows the results of our system on the test set. Firstly we show the relative scores of our model against Hiero without using reverse translation or lexical features.[7] This allows us to directly study the differences between the two translation models without the added complication of the other features. As well as both modelling the same distribution, when our model is trained with a single parameter per-rule these systems have the same parameter space, differing only in the manner of estimation.

Additionally we show the scores achieved by MERT training the full set of features for Hiero, with and without a language model.[8] We provide these results for reference. To compare our model directly with these systems we would need to incorporate additional features and a language model, work which we have left for a later date.

The relative scores confirm that our model, with its minimalist feature set, achieves comparable performance to the standard feature set without the language model. This is encouraging as our model was trained to optimise likelihood rather than BLEU, yet it is still competitive on that metric. As expected, the language model makes a significant difference to BLEU, however we believe that this effect is orthogonal to the choice of base translation model, thus we would expect a similar gain when integrating a language model into the discriminative system.

An informal comparison of the outputs on the development set, presented in Table 4, suggests that the

[7] Although the most direct comparison for the discriminative model would be with p_d model alone, omitting the gr, rc and wc features and MERT training produces poor translations.

[8] Hiero ($p_d, p_r, p_d^{lex}, p_r^{lex}, gr, rc, wc, lm$) represents state-of-the-art performance on this training/testing set.

S:	C'est pourquoi nous souhaitons que l'affaire nous soit renvoyée.
R:	We therefore want the matter re-referred to ourselves.
D:	That is why we want the that matters we to be referred back.
T:	That is why we would like the matter to be referred back.
H:	That is why we wish that the matter we be referred back.
S:	Par contre, la transposition dans les États membres reste trop lente.
R:	But implementation by the Member States has still been too slow.
D:	However, it is implemented in the Member States is still too slow.
T:	However, the implementation measures in Member States remains too slow.
H:	In against, transposition in the Member States remains too slow.
S:	Aussi, je considère qu'il reste énormément à faire dans ce domaine.
R:	I therefore consider that there is an incredible amount still to do in this area.
D:	So I think remains a lot to be done in this field.
T:	So I think there is still much to be done in this area.
H:	Therefore, I think it remains a vast amount to do in this area.

Table 4. Example output produced by the max-derivation (D), max-translation (T) decoding algorithms and Hiero($p_d, p_r, p_d^{lex}, p_r^{lex}, gr, rc, wc$) (H) models, relative to the source (S) and reference (R).

translation optimising discriminative model more often produces quite fluent translations, yet not in ways that would lead to an increase in BLEU score.[9] This could be considered a side-effect of optimising likelihood rather than BLEU.

Scaling In Figure 6 we plot the scaling characteristics of our models. The systems shown in the graph use the full grammar extracted on the 170k sentence corpus. The number of sentences upon which the iterative training algorithm is used to estimate the parameters is varied from 10k to the maximum 130K for which our model can reproduce the reference translation. As expected, the more data used to train the system, the better the performance. However, as the performance is still increasing significantly when all the parseable sentences are used, it is clear that the system's performance is suffering from the large number (40k) of sentences that are discarded before training.

5 Discussion and Further Work

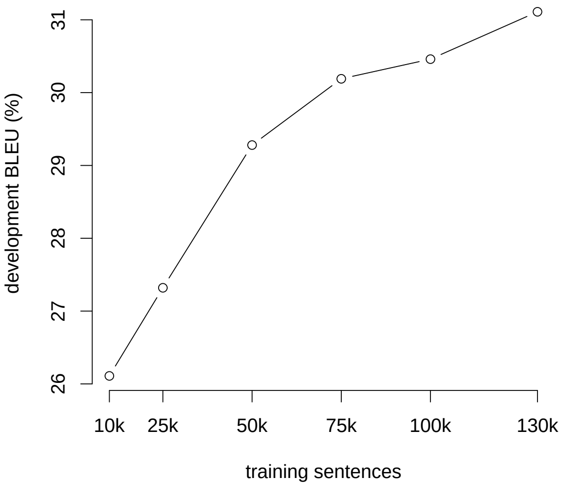

Figure 6. Learning curve showing that the model continues to improve as we increase the number of training sentences (development set)

We have shown that explicitly accounting for competing derivations yields translation improvements. Our model avoids the estimation biases associated with heuristic frequency count approaches and uses standard regularisation techniques to avoid degenerate maximum likelihood solutions.

Having demonstrated the efficacy of our model with very simple features, the logical next step is to investigate more expressive features. Promising features might include those over source side reordering rules (Wang et al., 2007) or source context features (Carpuat and Wu, 2007). Rule frequency features extracted from large training corpora would help the model to overcome the issue of unreachable reference sentences. Such approaches have been shown to be effective in log-linear word-alignment models where only a small supervised corpus is available (Blunsom and Cohn, 2006).

Finally, while in this paper we have focussed on the science of discriminative machine translation, we believe that with suitable engineering this model will advance the state-of-the-art. To do so would require integrating a language model feature into the max-translation decoding algorithm. The use of richer, more linguistic grammars (e.g., Galley et al. (2004)) may also improve the system.

Acknowledgements

The authors acknowledge the support of the EPSRC (Blunsom & Osborne, grant EP/D074959/1; Cohn, grant GR/T04557/01).

[9] Hiero was MERT trained on this set and has a 2% higher BLEU score compared to the discriminative model.

References

Phil Blunsom and Trevor Cohn. 2006. Discriminative word alignment with conditional random fields. In *Proc. of the 44th Annual Meeting of the ACL and 21st International Conference on Computational Linguistics (COLING/ACL-2006)*, pages 65–72, Sydney, Australia, July.

Marine Carpuat and Dekai Wu. 2007. Improving statistical machine translation using word sense disambiguation. In *Proc. of the 2007 Conference on Empirical Methods in Natural Language Processing (EMNLP-2007)*, pages 61–72, Prague, Czech Republic.

David Chiang. 2005. A hierarchical phrase-based model for statistical machine translation. In *Proc. of the 43rd Annual Meeting of the ACL (ACL-2005)*, pages 263–270, Ann Arbor, Michigan, June.

David Chiang. 2007. Hierarchical phrase-based translation. *Computational Linguistics*, 33(2):201–228.

Stephen Clark and James R. Curran. 2004. Parsing the WSJ using CCG and log-linear models. In *Proc. of the 42nd Annual Meeting of the ACL (ACL-2004)*, pages 103–110, Barcelona, Spain.

Stephen Clark and James R. Curran. 2007. Wide-coverage efficient statistical parsing with CCG and log-linear models. *Computational Linguistics*, 33(4).

John DeNero, Dan Gillick, James Zhang, and Dan Klein. 2006. Why generative phrase models underperform surface heuristics. In *Proc. of the HLT-NAACL 2006 Workshop on Statistical Machine Translation*, pages 31–38, New York City, June.

Michel Galley, Mark Hopkins, Kevin Knight, and Daniel Marcu. 2004. What's in a translation rule? In *Proc. of the 4th International Conference on Human Language Technology Research and 5th Annual Meeting of the NAACL (HLT-NAACL 2004)*, Boston, USA, May.

Michel Galley, Jonathan Graehl, Kevin Knight, Daniel Marcu, Steve DeNeefe, Wei Wang, and Ignacio Thayer. 2006. Scalable inference and training of context-rich syntactic translation models. In *Proc. of the 44th Annual Meeting of the ACL and 21st International Conference on Computational Linguistics (COLING/ACL-2006)*, pages 961–968, Sydney, Australia, July.

Abraham Ittycheriah and Salim Roukos. 2007. Direct translation model 2. In *Proc. of the 7th International Conference on Human Language Technology Research and 8th Annual Meeting of the NAACL (HLT-NAACL 2007)*, pages 57–64, Rochester, USA.

Philipp Koehn, Franz Josef Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *Proc. of the 3rd International Conference on Human Language Technology Research and 4th Annual Meeting of the NAACL (HLT-NAACL 2003)*, pages 81–88, Edmonton, Canada, May.

Philip M. Lewis II and Richard E. Stearns. 1968. Syntax-directed transduction. *J. ACM*, 15(3):465–488.

Percy Liang, Alexandre Bouchard-Côté, Dan Klein, and Ben Taskar. 2006. An end-to-end discriminative approach to machine translation. In *Proc. of the 44th Annual Meeting of the ACL and 21st International Conference on Computational Linguistics (COLING/ACL-2006)*, pages 761–768, Sydney, Australia, July.

Robert Malouf. 2002. A comparison of algorithms for maximum entropy parameter estimation. In *Proc. of the 6th Conference on Natural Language Learning (CoNLL-2002)*, pages 49–55, Taipei, Taiwan, August.

Franz Josef Och. 2003. Minimum error rate training in statistical machine translation. In *Proc. of the 41st Annual Meeting of the ACL (ACL-2003)*, pages 160–167, Sapporo, Japan.

Slav Petrov, Adam Pauls, and Dan Klein. 2007. Discriminative log-linear grammars with latent variables. In *Advances in Neural Information Processing Systems 20 (NIPS)*, Vancouver, Canada.

Fei Sha and Fernando Pereira. 2003. Shallow parsing with conditional random fields. In *Proc. of the 3rd International Conference on Human Language Technology Research and 4th Annual Meeting of the NAACL (HLT-NAACL 2003)*, pages 134–141, Edmonton, Canada.

Christoph Tillmann and Tong Zhang. 2007. A block bigram prediction model for statistical machine translation. *ACM Transactions Speech Language Processing*, 4(3):6.

Chao Wang, Michael Collins, and Philipp Koehn. 2007. Chinese syntactic reordering for statistical machine translation. In *Proc. of the 2007 Conference on Empirical Methods in Natural Language Processing (EMNLP-2007)*, pages 737–745, Prague, Czech Republic.

Taro Watanabe, Jun Suzuki, Hajime Tsukada, and Hideki Isozaki. 2007. Online large-margin training for statistical machine translation. In *Proc. of the 2007 Conference on Empirical Methods in Natural Language Processing (EMNLP-2007)*, pages 764–773, Prague, Czech Republic.

Benjamin Wellington, Joseph Turian, Chris Pike, and I. Dan Melamed. 2006. Scalable purely-discriminative training for word and tree transducers. In *Proc. of the 7th Biennial Conference of the Association for Machine Translation in the Americas (AMTA)*, Boston, USA.

Dekai Wu. 1997. Stochastic inversion transduction grammars and bilingual parsing of parallel corpora. *Computational Linguistics*, 23(3):377–403.