

# Fluency Constraints for Minimum Bayes-Risk Decoding of Statistical Machine Translation Lattices

Graeme Blackwood and Adrià de Gispert and William Byrne

Cambridge University Engineering Department, Trumpington Street, CB2 1PZ, U.K.

{gwb24 | ad465 | wjb31}@cam.ac.uk

## Abstract

A novel and robust approach to improving statistical machine translation fluency is developed within a minimum Bayes-risk decoding framework. By segmenting translation lattices according to confidence measures over the maximum likelihood translation hypothesis we are able to focus on regions with potential translation errors. Hypothesis space constraints based on monolingual coverage are applied to the low confidence regions to improve overall translation fluency.

## 1 Introduction and Motivation

Translation quality is often described in terms of *fluency* and *adequacy*. Fluency reflects the ‘nativeness’ of the translation while adequacy indicates how well a translation captures the meaning of the original text (Ma and Cieri, 2006).

From a purely utilitarian view, adequacy should be more important than fluency. But fluency and adequacy are subjective and not easy to tease apart (Callison-Burch et al., 2009; Villar et al., 2007). There is a human tendency to rate less fluent translations as less adequate. One explanation for this is that errors in grammar cause readers to be more critical. A related phenomenon is that the nature of translation errors changes as fluency improves so that whatever errors are present in fluent translations must necessarily be relatively subtle. It is therefore not enough to focus solely on adequacy. SMT systems must also be fluent if they are to be accepted and trusted. The reliance on automatic metrics may have led SMT researchers to pay insufficient attention to fluency. Automatic metrics like BLEU (Papineni et al., 2002), TER (Snover et al., 2006), and METEOR (Lavie and Denkowski,

2009) show broad correlation with human rankings of MT quality, but are not capable of making fine distinctions between fluency and adequacy.

There is concern that the fluency of current SMT is inadequate (Knight, 2007b). SMT is robust, in that a translation is nearly always produced. But unlike translators who should be skilled in at least one of the languages, SMT systems are limited in both source and target language competence. Fluency and accuracy therefore tend to suffer together as translation quality degrades. This should not be the case. Ideally, an SMT system should never be any less fluent than the best *stochastic text generation* system available in the target language (Oberlander and Brew, 2000). What is needed is a good way to enhance the fluency of SMT hypotheses.

The maximum likelihood (ML) formulation (Brown et al., 1990) of translation of source language sentence  $F$  to target language sentence  $\hat{E}$

$$\hat{E} = \operatorname{argmax}_E P(F|E)P(E) \quad (1)$$

makes it clear why improving SMT fluency is a difficult modelling problem. The language model  $P(E)$ , the closest thing to a ‘fluency component’ in most systems, only affects candidates likely under the translation model  $P(F|E)$ . Given the weakness of current translation models this is a severe limitation. For example, it often happens that SMT systems assign  $P(F|\bar{E}) = 0$  to a correct reference translation  $\bar{E}$  of  $F$ . The problem is that in ML decoding the language model can only encourage the production of fluent translations; it cannot easily enforce constraints on fluency or introduce new hypotheses.

In Hiero (Chiang, 2007) and syntax-based SMT (Knight and Graehl, 2005; Knight, 2007a), the primary role of syntax is to drive the translation process. Translations produced by these

systems respect the syntax of their translation models, but this does not force them to be grammatical in the way that a typical human sentence is grammatical; they produce many translations which are not fluent. The problem is robustness. Generating fluent translations demands a tightly constraining target language grammar but such a grammar is at odds with broad-coverage parsing needed for robust translation.

We have described two problems in translation fluency: (1) SMT may fail to generate fluent hypotheses and there is no simple way to introduce them into the search; (2) SMT produces many translations which are not fluent but enforcing constraints to improve fluency can hurt robustness. Both problems are rooted in the ML decoding framework in which robustness and fluency are conflicting objectives.

We propose a novel framework to improve the fluency of any SMT system, whether syntactic or phrase-based. We will perform Minimum Bayes-risk search (Kumar and Byrne, 2004) over a space of fluent hypotheses  $\mathcal{H}$ :

$$\hat{E}_{\text{MBR}} = \operatorname{argmin}_{E' \in \mathcal{H}} \sum_{E \in \mathcal{E}} L(E, E') P(E|F) \quad (2)$$

In this approach the MBR evidence space  $\mathcal{E}$  is generated by an SMT system as a  $k$ -best list or lattice. The system runs in its best possible configuration, ensuring both translation robustness and good baselines. Rather than decoding in the output of the baseline SMT system, translations will be sought among a collection of fluent sentences that are close to the top SMT hypotheses as determined by the loss function  $L(E, E')$ .

Decoupling the MBR hypothesis space from first-pass translation offers great flexibility. Hypotheses in  $\mathcal{H}$  may be arbitrarily constrained according to lexical, syntactic, semantic, or other considerations, with no effect on translation robustness. This is because constraints on fluency do not affect the production of the evidence space by the baseline system. Robustness and fluency are no longer conflicting objectives. This framework also allows the MBR hypothesis space to be augmented with hypotheses produced by an NLG system, although this is beyond the scope of the present paper.

This paper focuses on searching out fluent strings amongst the vast number of hypotheses encoded in SMT lattices. Oracle BLEU scores computed over  $k$ -best lists (Och et al., 2004) show that many high quality hypotheses are produced by first-pass SMT decoding. We propose reducing the difficulty of enhancing the fluency of complete hypotheses by first identifying regions of high-confidence in the ML translations and using these to guide the fluency refinement process. This has two advantages: (1) we keep portions of the baseline hypotheses that we trust and search for alternatives elsewhere, and (2) the task is made much easier since the fluency of sentence fragments can be refined in context.

In what follows, we use posterior probabilities computed from SMT lattices to identify useful subsequences in the ML translations (Section 3). These subsequences drive the segmentation and transformation of lattices into smaller subproblems (Sections 4 & 5). These subproblems are mined for fluent strings (Section 6), resulting in improved translation fluency (Sections 7 & 8).

## 2 Lattice MBR Decoding

The formulation of the MBR decoder in Equation (2) separates the hypothesis space from the evidence space. We can apply the linearised form of the lattice MBR decision rule (Tromble et al., 2008) to give

$$\hat{E}_{\text{LMBR}} = \operatorname{argmax}_{E' \in \mathcal{H}} \left\{ \theta_0 |E'| + \sum_{u \in \mathcal{N}} \theta_u \#_u(E') p(u|\mathcal{E}) \right\}, \quad (3)$$

where  $\mathcal{H}$  is the hypothesis space,  $\mathcal{E}$  is the evidence space,  $\mathcal{N}$  is the set of all  $n$ -grams in  $\mathcal{H}$  (typically,  $n = 1 \dots 4$ ), and  $\theta$  are constants estimated on held-out data. The quantity  $p(u|\mathcal{E})$  is the path posterior probability of  $n$ -gram  $u$

$$p(u|\mathcal{E}) = \sum_{E \in \mathcal{E}_u} P(E|F), \quad (4)$$

where  $\mathcal{E}_u = \{E \in \mathcal{E} : \#_u(E) > 0\}$  is the subset of paths containing  $n$ -gram  $u$  at least once. The path posterior probabilities  $p(u|\mathcal{E})$  of Equation (4) can be easily calculated using general purpose WFST operations (Mohri et al., 2002).

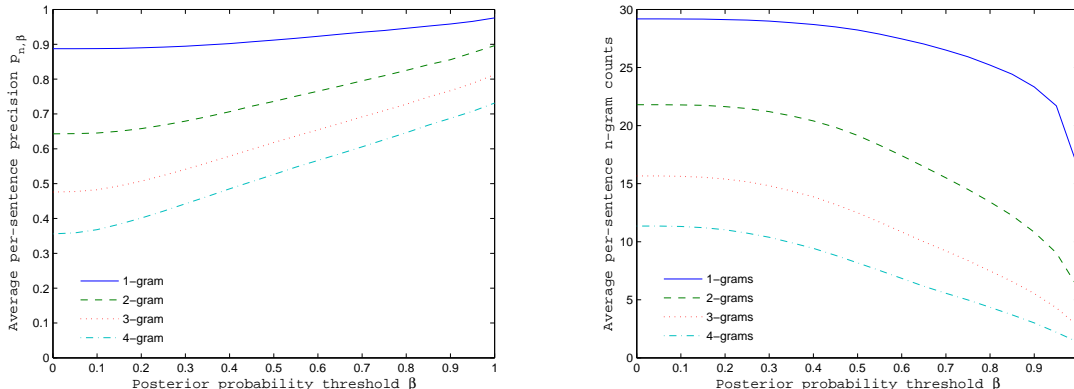


Figure 1: Average  $n$ -gram precisions (left) and counts (right) for 2075 sentences of NIST Arabic→English ML translations at a range of posterior probability thresholds  $0 \leq \beta \leq 1$ . The left plot shows at  $\beta = 0$  the  $n$ -gram precisions used in the BLEU score of the ML baseline system.

### 3 Posterior Probability Confidence Measures

In the formulation of Equations (3) and (4) the path posterior  $n$ -gram probabilities play a crucial role. MBR decoding under the linear approximation to BLEU is driven mainly by the presence of high posterior  $n$ -grams in the lattice; the low posterior  $n$ -grams contribute relatively little to the MBR decision criterion. Here we investigate the predictive power of these statistics. We will show that the  $n$ -gram posterior is a good predictor as to whether or not an  $n$ -gram is to be found in a set of reference translations.

Let  $\mathcal{N}_n$  denote the set of  $n$ -grams of order  $n$  in the ML hypothesis  $\hat{E}$ , and let  $\mathcal{R}_n$  denote the set of  $n$ -grams of order  $n$  in the union of the references. For confidence threshold  $\beta$ , let  $\mathcal{N}_{n,\beta} = \{u \in \mathcal{N}_n : p(u|\mathcal{E}) \geq \beta\}$  denote the  $n$ -grams in  $\mathcal{N}_n$  with posterior probability greater than or equal to  $\beta$ , where  $p(u|\mathcal{E})$  is computed using Equation (4). This is equivalent to identifying all substrings of length  $n$  in the translation hypotheses for which the system assigns a posterior probability of  $\beta$  or higher. The precision at order  $n$  for threshold  $\beta$  is the proportion of  $n$ -grams in  $\mathcal{N}_{n,\beta}$  also present in the references:

$$\mathcal{P}_{n,\beta} = \frac{|\mathcal{R}_n \cap \mathcal{N}_{n,\beta}|}{|\mathcal{N}_{n,\beta}|} \quad (5)$$

The left plot in Figure 1 shows average per-sentence  $n$ -gram precisions  $\mathcal{P}_{n,\beta}$  at orders 1...4

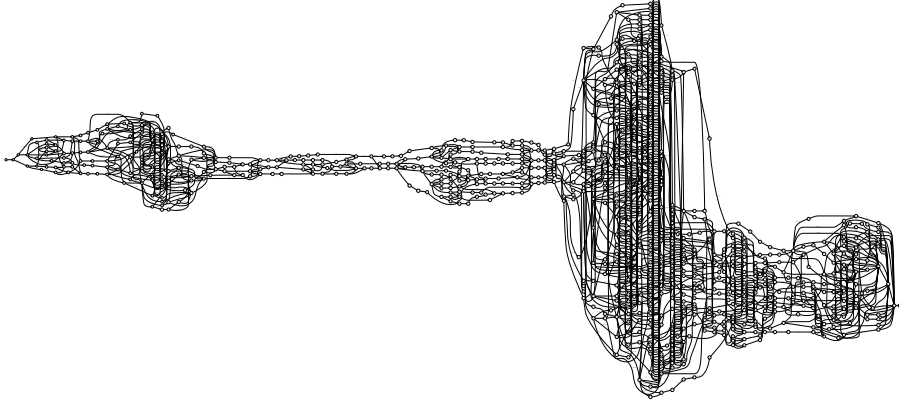
for an Arabic→English translation task at a range of thresholds  $0 \leq \beta \leq 1$ . Sentence start and end tokens are ignored when computing uni-gram precisions. We note that precision at all orders improves as the threshold  $\beta$  increases. This confirms that these intrinsic measures of translation confidence have strong predictive power.

The right-hand side of the figure shows the average number of  $n$ -grams per sentence at each order for the same range of  $\beta$ . We see that for high  $\beta$ , there are relatively few  $n$ -grams with  $p(u|\mathcal{E}) \geq \beta$ ; this is as expected. However, even at a high threshold of  $\beta = 0.9$  there are still an average of three 4-grams per sentence with posterior probabilities that exceed that threshold. Even at this very high confidence level, high posterior  $n$ -grams occur frequently enough that we can expect them to be useful.

These precision results motivate our use of path posterior  $n$ -gram probabilities as a confidence measure. We assign confidence  $p(\hat{E}_i^j|\mathcal{E})$  to subsequences  $\hat{E}_i \dots \hat{E}_j$  of the ML hypothesis.

Prior work focuses on word-level confidence measures extracted from  $n$ -best lists and word graphs (Ueffing and Ney, 2007; Koehn, 2010), while Zens and Ney (2006) rescore  $n$ -best lists with  $n$ -gram posterior probabilities. Similar experiments using different statistics and with a different motivation are reported by DeNero et al. (2009) who show that expected counts of  $n$ -grams in a lattice can be used to predict which  $n$ -grams appear in the references.

the newspaper “ constitution ” quoted brigadier abdullah krishan , the chief of police in karak governorate ( 521 km south @-@ west of amman ) as saying that the seizure took place after police received information that there were attempts by the group to sell for more than \$ 100 thousand dollars , the police rushed to the arrest in possession .



$\mathcal{H}_1$	$\mathcal{H}_2$	$\mathcal{H}_3$	$\mathcal{H}_4$	$\mathcal{H}_5$	$\mathcal{H}_6$	$\mathcal{H}_7$	$\mathcal{H}_8$	$\mathcal{H}_9$
433	1	4	1	6	1	6860	1	76

Figure 2: ML translation  $\hat{E}$ , word lattice  $\mathcal{E}$ , and decomposition as a sequence of four string and five sublattice regions  $\mathcal{H}_1 \dots \mathcal{H}_9$  using  $n$ -gram posterior probability threshold  $p(u|\mathcal{E}) \geq 0.8$ .

#### 4 Lattice Segmentation

The previous section shows that current SMT systems, although flawed, can identify with confidence partial hypotheses that can be trusted. We wish to constrain MBR decoding to include these trusted partial hypotheses but allow decoding to consider alternatives in regions of low confidence. In this way we aim to improve the best possible output of the best available systems.

We use the path posterior  $n$ -gram probabilities of Equation (4) to segment the lattice  $\mathcal{E}$  into regions of high and low confidence. An example of this segmentation is shown in Figure 2. As that example shows, lattice segmentation is performed relative to the ML hypothesis  $\hat{E}$ , i.e. relative to the best path through  $\mathcal{E}$ .

For confidence threshold  $\beta$ , we find all 4-grams  $u = \hat{E}_i, \dots, \hat{E}_{i+3}$  in the ML translation hypothesis for which  $p(u|\mathcal{E}) > \beta$ . We then segment  $\hat{E}$  into regions of high and low confidence where the high confidence regions are identified

by consecutive, overlapping high confidence 4-grams. The high confidence regions are contiguous strings of words for which there is consensus amongst the translations in the lattice. If we trust the path posterior  $n$ -gram probabilities, any hypothesised translation should include these high confidence substrings.

The hypothesis string  $\hat{E}$  is in this way segmented into  $R$  alternating subsequences of high and low confidence. The segment boundaries are  $i_r$  and  $j_r$  so that  $\hat{E}_{i_r}^{j_r}$  is either a high confidence or a low confidence subsequence. Each subsequence is associated with an unweighted subspace  $\mathcal{H}_r$ ; this subspace has the form of a string for high confidence regions and the form of a lattice for low confidence regions.

If the  $r^{th}$  segment is a high confidence region then  $\mathcal{H}_r$  accepts only the string  $\hat{E}_{i_r}^{j_r}$ . If the  $r^{th}$  segment is a region of low confidence, then  $\mathcal{H}_r$  is built to accept relevant substrings from  $\mathcal{E}$ . It is constructed as follows. The  $r^{th}$  low confidence

region  $\hat{E}_{i_r}^{j_r}$  has a high confidence left context  $\hat{e}_{r-1}$  and a high confidence right context  $\hat{e}_{r+1}$  formed from subsequences of the ML translation hypothesis  $\hat{E}$  as

$$\hat{e}_{r-1} = \hat{E}_{i_{r-1}}^{j_{r-1}}, \quad \hat{e}_{r+1} = \hat{E}_{i_{r+1}}^{j_{r+1}}$$

Note that when  $r = 1$  the left context  $\hat{e}_{r-1}$  is the empty string and when  $r = R$  the right context  $\hat{e}_{r+1}$  is the empty string. We build a transducer  $\mathcal{T}_r$  for the regular expression  $/. * \hat{e}_{r-1} (.*) \hat{e}_{r+1} . * \wedge 1 /$ .<sup>1</sup> Composition with  $\mathcal{E}$  yields  $\mathcal{H}_r = \mathcal{E} \circ \mathcal{T}_r$ , so that  $\mathcal{H}_r$  contains all the reasonable alternatives to  $\hat{E}_{i_r}^{j_r}$  in  $\mathcal{E}$  consistent with the high confidence left and right contexts  $\hat{e}_{r-1}$  and  $\hat{e}_{r+1}$ . If  $\mathcal{H}_r$  is aligned to a high confidence subsequence of  $\hat{E}$ , we call it a *string region* since it contains a single path; if it is aligned to a low confidence region it is a lattice and we call it a *sublattice region*. The series of high and low confidence subspace regions  $\mathcal{H}_1, \dots, \mathcal{H}_R$  defines the segmentation of the lattice.

## 5 Hypothesis Space Construction

We now describe a general framework for improving the fluency of the hypothesis space in lattice MBR decoding.

The segmentation of the lattice described in Section 4 considerably simplifies the problem of improving the fluency of its hypotheses since each region of low confidence may be considered independently. The low confidence regions can be transformed one-by-one and then reassembled to form a new MBR hypothesis space.

In order to transform the hypothesis region  $\mathcal{H}_r$  it is important to know the context in which it occurs, i.e. the sequences of words that form its prefix and suffix. Some transformations might only need a short context; others may need a sentence-level context, i.e. the full sequence of ML words  $\hat{E}_1^{j_{r-1}}$  and  $\hat{E}_{i_{r+1}}^N$  to the left and right of the region  $\mathcal{H}_r$  that is to be transformed.

To put this formally, each low confidence sublattice region is transformed by the application of some function  $\Psi$ :

$$\mathcal{H}_r \leftarrow \Psi(\hat{E}_1^{j_{r-1}}, \mathcal{H}_r, \hat{E}_{i_{r+1}}^N) \quad (6)$$

<sup>1</sup>In this notation parentheses indicate string matches, e.g.  $/. * y(a*)w . * \wedge 1 /$  applied to  $xyaaawzz$  yields  $aaa$ .

The hypothesis space is then constructed from the concatenation of high confidence string and transformed low confidence sublattice regions

$$\mathcal{H} = \mathcal{E} \circ \bigotimes_{1 \leq r \leq R} \mathcal{H}_r \quad (7)$$

The composition with the original lattice  $\mathcal{E}$  discards any new hypotheses that might be created via the unconstrained concatenation of strings from the  $\mathcal{H}_r$ . It may be that in some circumstances the introduction of new paths is good, but in what follows we test the ability to improve fluency by searching among existing hypotheses, and this ensures that nothing new is introduced.

### 5.1 Size of the Hypothesis Space

If no new hypotheses are introduced by the operations  $\Psi$ , the size of the hypothesis space  $\mathcal{H}$  is determined by the posterior probability threshold  $\beta$ . Only the ML hypothesis remains at  $\beta = 0$ , since all its subsequences are of high confidence, i.e. can be covered by  $n$ -grams with non-zero path posterior probability. At the other extreme, for  $\beta = 1$ , it follows that  $\mathcal{H} = \mathcal{E}$  and no paths are removed, since any string regions created will contain subsequences from every path in  $\mathcal{E}$ .

We can use  $\beta$  to tighten or relax constraints on the LMBR hypothesis space. At  $\beta = 0$ , LMBR returns only the ML hypothesis; at  $\beta = 1$ , LMBR is done over the full translation lattice. This is shown in Table 1, where the BLEU score approaches the BLEU score of unconstrained LMBR as  $\beta$  increases.

Note also that the size of the resulting hypothesis space is the product of the number of sequences in the sublattice regions. For Figure 2 at  $\beta = 0.8$ , this product is  $\sim 5.4$  billion hypotheses. Even for fairly aggressive constraints on the hypothesis space, many hypotheses remain.

## 6 Monolingual Coverage Constraints for Translation Fluency

This section describes one implementation of the transformation  $\Psi$  that we will show leads to improved fluency of machine translation output. This transformation is based on  $n$ -gram coverage in a large target language text collection: where

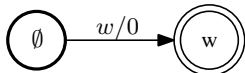
LM	Translation hypothesis $\bar{E}$	Score
4g	<s> <sub>1</sub> the <sub>2</sub> reactor <sub>3</sub> produces <sub>3</sub> plutonium <sub>2</sub> <i>needed<sub>2</sub></i> to <sub>3</sub> <i>manufacture<sub>4</sub></i> atomic <sub>3</sub> bomb <sub>2</sub> . <sub>3</sub> </s> <sub>4</sub>	-22.59
	<s> <sub>1</sub> the <sub>2</sub> reactor <sub>3</sub> produces <sub>3</sub> plutonium <sub>2</sub> <i>needed<sub>2</sub></i> to <sub>3</sub> <i>manufacture<sub>4</sub></i> the <sub>4</sub> atomic <sub>2</sub> bomb <sub>3</sub> . <sub>4</sub> </s> <sub>4</sub>	-23.61
5g	<s> <sub>1</sub> the <sub>2</sub> reactor <sub>3</sub> produces <sub>4</sub> plutonium <sub>5</sub> <i>needed<sub>3</sub></i> to <sub>3</sub> <i>manufacture<sub>4</sub></i> atomic <sub>5</sub> bomb <sub>2</sub> . <sub>3</sub> </s> <sub>4</sub>	-16.04
	<s> <sub>1</sub> the <sub>2</sub> reactor <sub>3</sub> produces <sub>4</sub> plutonium <sub>5</sub> <i>needed<sub>3</sub></i> to <sub>3</sub> <i>manufacture<sub>4</sub></i> the <sub>4</sub> atomic <sub>4</sub> bomb <sub>5</sub> . <sub>4</sub> </s> <sub>5</sub>	-17.96

Figure 3: Scores and  $n$ -gram orders for hypotheses using 4-gram Kneser-Ney and 5-gram stupid-backoff (estimated from 1.1B and 6.6B tokens, resp.) LMs. Low confidence regions are in italics.

possible, we filter the sublattice regions so that they contain only long-span  $n$ -grams observed in the text. Our motivation is that large monolingual text collections are good guides to fluency. If a hypothesis is composed entirely of previously seen high order  $n$ -grams, it is likely to be fluent and should be favoured.

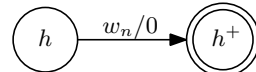
Initial attempts to identify fluent hypotheses in sublattice regions by ranking according to  $n$ -gram LM scores were ineffective. Figure 3 shows the difficulties. We see that both the 4-gram Kneser-Ney and 5-gram stupid-backoff language models favour the shorter but disfluent hypothesis; normalising by length was not effective. However, the stupid-backoff LM has far better coverage and we see that the backing-off behaviour is a clue to the presence of disfluency. The shorter hypothesis backs off to a bigram for “atomic bomb”, whereas the longer hypothesis covers the same words with 4-grams. We therefore disregard the language model scores and focus on  $n$ -gram coverage. This is an example where robustness and fluency are at odds. The  $n$ -gram models are robust, but often favour shorter less fluent hypotheses.

Let  $\mathcal{S}$  denote the set of all  $n$ -grams in the monolingual training data. To identify partial hypotheses in sublattice regions that have complete monolingual coverage at the maximum order  $n$ , we build a coverage acceptor  $\mathcal{C}_n$  with a similar form to the WFST representation of an  $n$ -gram backoff language model (Allauzen et al., 2003).  $\mathcal{C}_n$  assigns a penalty to every  $n$ -gram not found in  $\mathcal{S}$ . In  $\mathcal{C}_n$  word arcs have no cost and backoff arcs are assigned a fixed cost of 1. Firstly, arcs from the start state are added for each unigram  $w \in \mathcal{N}_1$ :

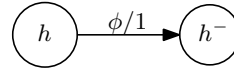


Then for  $n$ -grams  $u \in \mathcal{S} \cap \{\cup_{i=2}^n \mathcal{N}_i\}$ , where

$u = w_1^n$  consisting of history  $h = w_1^{n-1}$  and target word  $w_n$ , arcs are added



where  $h^+ = w_2^{n-1}$  if  $u$  has order  $n$  and  $h^+ = w_1^n$  if  $u$  has order less than  $n$ . Backoff arcs are added for each  $u$  as



where  $h^- = w_2^{n-1}$  if  $u$  has order  $> 2$ , and bigrams backoff to the null history start state  $\emptyset$ .

For each sublattice region  $\mathcal{H}_r$ , we wish to penalise each path proportionally to the number of its  $n$ -grams not found in the monolingual text collection  $\mathcal{S}$ . We wish to do this in context, so that we include the effect of the neighbouring high confidence regions  $\mathcal{H}_{r-1}$  and  $\mathcal{H}_{r+1}$ . Given that we are counting  $n$ -grams at order  $n$  we form the left context machine  $\mathcal{L}_r$  which accepts the last  $n - 1$  words in  $\mathcal{H}_{r-1}$ ; similarly,  $\mathcal{R}_r$  accepts the first  $n - 1$  words of  $\mathcal{H}_{r+1}$ . The concatenation  $\mathcal{X}_r = \mathcal{L}_r \otimes \mathcal{H}_r \otimes \mathcal{R}_r$  represents the partial translation hypotheses in  $\mathcal{H}_r$  padded with  $n - 1$  words of left and right context from the neighbouring high confidence regions. Composing  $\mathcal{X}_r \circ \mathcal{C}_n$  assigns each partial hypothesis a cost equal to the number of times it was necessary to back off to lower order  $n$ -grams while reading the string. Partial hypotheses with cost 0 did not back off at all and contain only maximum order  $n$ -grams.

In the experiments that follow, we look at each  $\mathcal{X}_r \circ \mathcal{C}_n$  and if there are paths with cost 0, only these are kept and all others are discarded. We introduce this as a constraint on the hypothesis space which we will evaluate for improvement on fluency. Here the transformation function  $\Psi$  returns  $\mathcal{H}_r$  as  $\mathcal{X}_r \circ \mathcal{C}_n$  after pruning. If  $\mathcal{X}_r \circ \mathcal{C}_n$  has no zero cost paths, the transformation function  $\Psi$  returns  $\mathcal{H}_r$  as we find it, since there is not enough coverage in the monolingual text to

guide the selection of fluent hypotheses.

After applying monolingual coverage constraints to each region, the modified hypothesis space used for MBR search is formed by concatenation using Equation (7).

We note that  $\mathcal{C}_n$  is a simplistic NLG system. It generates strings by concatenating  $n$ -grams found in  $\mathcal{S}$ . We do not allow it to run ‘open loop’ in these experiments, but instead use it to find the strings in  $\mathcal{X}_r$  with good  $n$ -gram coverage.

## 7 LMBR Over Segmented Lattices

The effect of fluency constraints on LMBR decoding is evaluated in the context of the NIST Arabic→English MT task. The set *tune* consists of the odd numbered sentences of the MT02–MT05 testsets; the even sentences form *test*. MT08 performance on *nw08* (newswire) and *ng08* (newsgroup) data is also reported.

First-pass translation is performed using HiFST (Iglesias et al., 2009b), a hierarchical phrase-based decoder. The first-pass LM is a modified Kneser-Ney (Kneser and Ney, 1995) 4-gram estimated over the English side of the parallel text and an 881M word subset of the English GigaWord 3rd Edition (Graff et al., 2007). Prior to LMBR, the first-pass lattices are rescored with zero-cutoff stupid-backoff 5-gram language models (Brants et al., 2007) estimated over more than 6B words of English text. The factors  $\theta_0, \dots, \theta_4$  of the LMBR decoder are set as in Tromble et al. (2008) using unigram precision  $p = 0.85$  and average recall ratio  $r = 0.74$ .

The effect of performing LMBR over the segmented hypothesis space is shown in Table 1. The hypothesis spaces  $\mathcal{H}_r$  are constructed at various confidence thresholds as described in Section 4 with  $\mathcal{H}$  formed via Equation (7); no coverage constraints are applied yet. Constraining the search space using  $\beta = 0.6$  leads to little degradation in LMBR performance under BLEU. This shows lattice segmentation works as intended.

We next investigate the effect on BLEU of applying monolingual coverage constraints. We build the acceptors  $\mathcal{C}_n$  as described in Section 6 with  $\mathcal{S}$  consisting of all 5-grams in the full English GigaWord collection of approximately

		<i>tune</i>	<i>test</i>	<i>nw08</i>	<i>ng08</i>
ML		54.2	53.8	51.3	36.3
$\beta$	0.0	54.2	53.8	51.3	36.3
	0.2	54.3	53.8	51.3	36.3
	0.4	54.6	54.2	51.6	36.7
	0.6	54.9	54.4	52.1	36.6
	0.8	54.9	54.4	52.1	36.6
	1.0	54.9	54.4	52.2	36.7
LMBR		54.9	54.4	52.2	36.8

Table 1: BLEU scores for ML hypotheses and LMBR decoding in  $\mathcal{H}$  over  $0 \leq \beta \leq 1$ .

4.4B words. At  $\beta = 0.6$  we found 181 sentences with sublattices  $\mathcal{H}_r$  that could be spanned by 5-grams from  $\mathcal{S}$ , i.e. for which  $\mathcal{X}_r \circ \mathcal{C}_n$  have paths with cost 0; these are filtered as described. LMBR over these coverage-constrained sublattices is denoted LMBR+CC. On *nw08* the BLEU score for LMBR+CC is 52.0 which is +0.7 over the ML decoder and only -0.2 BLEU below the unconstrained LMBR decoding. Done in this way, constraining hypotheses to have 5-grams from the GigaWord has little impact on BLEU.

At this value of  $\beta$ , 116 of the 813 *nw08* sentences have a low confidence region (1) completely covered by 5-grams, and (2) within which the ML hypothesis and the LMBR+CC hypothesis differ. It is these regions which we will inspect for improved fluency.

## 8 Human Fluency Evaluation

We asked 17 native speakers to judge the fluency of sentence fragments from *nw08*. We compared hypotheses from the maximum likelihood decoder (ML) and the LMBR+CC decoder. Each fragment consisted of the partial translation hypothesis from a low confidence region together with its left and right high confidence contexts (examples given in Figure 4). For each sample, judges were asked: ‘‘Could this fragment occur in a fluent sentence?’’

The results are shown in Table 2. Most of the time, the ML and LMBR+CC sentence fragments were both judged to be fluent; it often happened that they differed by only a single noun or verb substitution which didn’t affect fluency. In a small number of cases, both ML and LMBR+CC were judged to be disfluent. We are most interested in the ‘off-diagonal’ cases. In cases when

ML	... view , especially with <i>the open chinese economy</i> to the world and ...
+LMBR	... view , especially with <i>the open chinese economy</i> to the world and ...
+LMBR+CC	... view , especially with <i>the opening of the chinese economy</i> to the world and ...
ML	... revision of the constitution <i>of the japanese public</i> , which dates back ...
+LMBR	... revision of the constitution <i>of the japanese public</i> , which dates back ...
+LMBR+CC	... revision of the constitution <i>of japan</i> , which dates back ...

Figure 4: Improved fluency through the application of monolingual coverage constraints to the hypothesis space in MBR decoding of NIST MT 08 Arabic→English newswire lattices.

		LMBR+CC	
		Fluent	Not Fluent
ML	Fluent	1175 (59.6%)	192 (9.7%)
	Not Fluent	530 (26.9%)	75 (3.8%)

Table 2: Partial hypothesis fluency judgements.

one system was judged to be fluent and the other was not, LMBR+CC was preferred about twice as often as the ML baseline (26.9% to 9.7%). Put another way, the monolingual fluency constraints were judged to have improved the fluency of the low confidence region more than twice as often as a fluent hypothesis was made disfluent.

Some examples of improved fluency are shown in Figure 4. Although both the ML and unconstrained LMBR hypotheses might satisfy adequacy, they lack the fluency of the LMBR+CC hypotheses generated using monolingual fluency constraints.

## 9 Summary and Conclusions

We have described a general framework for improving SMT fluency. Decoupling the hypothesis space from the evidence space allows for much greater flexibility in lattice MBR search.

We have shown that high path posterior probability  $n$ -grams in the ML translation can be used to guide the segmentation of a lattice into regions of high and low confidence. Segmenting the lattice simplifies the process of refining the hypothesis space since low confidence regions can be refined in the context of their high confidence context. This can be done independently before reassembling the refined regions. Lattice segmentation facilitates the application of post-processing techniques targeted to address particular deficiencies in ML decoding.

The techniques we presented are related to consensus decoding and system combination for SMT (Matusov et al., 2006; Sim et al., 2007),

Testset	Reachability
tune	15%
test	14%
nw08	11%
ng08	9%

Table 3: MT08 reference reachability.

and to segmental MBR for automatic speech recognition (Goel et al., 2004).

Hypothesis space constraints based on  $n$ -gram coverage in monolingual text were applied to low confidence regions. An evaluation of the constrained regions by native speakers showed improved fluency with only a minor degradation in BLEU score relative to unconstrained LMBR decoding. This approach is limited by the coverage of sublattices using monolingual text. We expect this to improve with larger text collections or in tightly focused scenarios where in-domain text is less diverse.

However, fluency will be best improved by integrating more sophisticated NLG systems. NLG systems capable of generating sentence fragments in context can be incorporated directly into this framework. If the MBR hypothesis space  $\mathcal{H}$  contains a generated hypothesis  $\bar{E}$  for which  $P(F|\bar{E}) = 0$ ,  $\bar{E}$  could still be produced as a translation, since it can be ‘voted for’ by nearby hypotheses produced by the underlying system. Table 3 shows the proportion of NIST MT08 testset sentences that can be aligned to one of the reference translations using our baseline hierarchical decoder. The low level of reachability of the reference translations suggests that generation may be required to achieve high levels of translation quality and fluency.

We note that our approach could also be used to improve the fluency of ASR, OCR and other language processing tasks where the goal is to produce fluent natural language output.

## References

- Allauzen, Cyril, Mehryar Mohri, and Brian Roark. 2003. Generalized algorithms for constructing statistical language models. In *Proceedings of the 41st Meeting of the Association for Computational Linguistics*, pages 557–564.
- Allauzen, Cyril, Michael Riley, Johan Schalkwyk, Wojciech Skut, and Mehryar Mohri. 2007. OpenFst: a general and efficient weighted finite-state transducer library. In *Proceedings of the 9th International Conference on Implementation and Application of Automata*, pages 11–23. Springer.
- Brants, Thorsten, Ashok C. Popat, Peng Xu, Franz J. Och, and Jeffrey Dean. 2007. Large language models in machine translation. In *Proceedings of the 2007 Joint Conference on EMNLP and CoNLL*, pages 858–867.
- Brown, Peter F., John Cocke, Stephen A. Della Pietra, Vincent J. Della Pietra, Fredrick Jelinek, John D. Lafferty, Robert L. Mercer, and Paul S. Roossin. 1990. A statistical approach to machine translation. *Computational Linguistics*, 16(2):79–85.
- Callison-Burch, Chris, Philipp Koehn, Christof Monz, and Josh Schroeder. 2009. Findings of the 2009 Workshop on Statistical Machine Translation. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, pages 1–28, Athens, Greece, March. Association for Computational Linguistics.
- Chiang, David. 2007. Hierarchical phrase-based translation. *Computational Linguistics*, 33(2):201–228.
- DeNero, John, David Chiang, and Kevin Knight. 2009. Fast consensus decoding over translation forests. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 567–575, Suntec, Singapore, August. Association for Computational Linguistics.
- Deng, Yonggang and William Byrne. 2008. HMM word and phrase alignment for statistical machine translation. *IEEE Transactions on Audio, Speech, and Language Processing*, 16(3):494–507.
- Goel, V., S. Kumar, and W. Byrne. 2004. Segmental minimum Bayes-risk decoding for automatic speech recognition. *IEEE Transactions on Speech and Audio Processing*, 12:234–249, May.
- Graff, David, Junbo Kong, Ke Chen, and Kazuaki Maeda. 2007. English Gigaword Third Edition.
- Iglesias, Gonzalo, Adrià de Gispert, Eduardo R. Banga, and William Byrne. 2009a. Rule filtering by pattern for efficient hierarchical translation. In *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009)*, pages 380–388, Athens, Greece, March. Association for Computational Linguistics.
- Iglesias, Gonzalo, Adrià de Gispert, Eduardo R. Banga, and William Byrne. 2009b. Hierarchical phrase-based translation with weighted finite state transducers. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 433–441, Boulder, Colorado, June. Association for Computational Linguistics.
- Kneser, R. and H. Ney. 1995. Improved backing-off for m-gram language modeling. In *Acoustics, Speech, and Signal Processing*, pages 181–184.
- Knight, K and J Graehl. 2005. An overview of probabilistic tree transducers for natural language processing. In *Proceedings of the Conference on Intelligent Text Processing and Computational Linguistics (CICLING)*.
- Knight, K. 2007a. Capturing practical natural language transformations. *Machine Translation*, 21(2).
- Knight, Kevin. 2007b. Automatic language translation generation help needs badly. In *MT Summit XI Workshop on Using Corpora for NLG: Language Generation and Machine Translation, Keynote Address*.
- Koehn, Phillip. 2010. *Statistical Machine Translation*. Cambridge University Press.
- Kumar, Shankar and William Byrne. 2004. Minimum Bayes-risk decoding for statistical machine translation. In *Proceedings of Human Language Technologies: The 2004 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 169–176.
- Lavie, Alon and Michael J. Denkowski. 2009. The METEOR metric for automatic evaluation of machine translation. *Machine Translation Journal*.
- Ma, Xiaoyi and Christopher Cieri. 2006. Corpus support for machine translation at LDC. In *Proc of LREC 2006: Fifth International Conference on Language Resources and Evaluation*.
- Matusov, Evgeny, Nicola Ueffing, and Hermann Ney. 2006. Computing consensus translation from multiple machine translation systems using enhanced hypotheses alignment. In *11th Conference of the*

- European Chapter of the Association for Computational Linguistics*, pages 33–40.
- Mohri, Mehryar, Fernando Pereira, and Michael Riley. 2002. Weighted finite-state transducers in speech recognition. In *Computer Speech and Language*, volume 16, pages 69–88.
- Oberlander, Jon and Chris Brew. 2000. Stochastic text generation. In *Philosophical Transactions of the Royal Society*, volume 358, pages 1373–1387.
- Och, F., D. Gildea, S. Khudanpur, A. Sarkar, K. Yamada, A. Fraser, S. Kumar, L. Shen, D. Smith, K. Eng, V. Jain, Z. Jin, and D. Radev. 2004. A smorgasbord of features 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*.
- Papineni, Kishore, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Meeting of the Association for Computational Linguistics*, pages 311–318, Morristown, NJ, USA.
- Sim, K.-C., W. Byrne, M. Gales, H. Sahbi, and P.C. Woodland. 2007. Consensus network decoding for statistical machine translation system combination. In *IEEE Conference on Acoustics, Speech and Signal Processing*.
- Snover, Matthew, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, , and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of Association for Machine Translation in the Americas*, pages 223–231, August.
- Tromble, Roy, Shankar Kumar, Franz Och, and Wolfgang Macherey. 2008. Lattice minimum Bayes-risk decoding for statistical machine translation. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 620–629, Honolulu, Hawaii, October. Association for Computational Linguistics.
- Ueffing, Nicola and Hermann Ney. 2007. Word-level confidence estimation for machine translation. *Computational Linguistics*, 33(1):9–40.
- Vilar, D, G Leusch, H Ney, and R Banchs. 2007. Human evaluation of machine translation through binary system comparisons. In *Proceedings of the Second Workshop on Statistical Machine Translation*.
- Zens, Richard and Hermann Ney. 2006. N -gram posterior probabilities for statistical machine translation. In *Technology Conference of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL): Proc. of the Workshop on Statistical Machine Translation*.