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

Graeme Blackwood and **Adrià de Gispert** and **William Byrne**
Machine Intelligence Laboratory, Department of Engineering, Cambridge University
{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; Vilar et al., 2007). There is a human tendency to rate less fluent translations as less adequate. One explanation 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 any errors in fluent translations must 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. It is possible that the reliance on automatic metrics may have led SMT researchers to pay insufficient attention to fluency: 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 incapable of 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 $\hat{E}$

$$\hat{E} = \underset{E}{\operatorname{argmax}}\, P(F|E)P(E) \qquad (1)$$

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 the original formulation, only affects candidates likely under the translation model $P(F|E)$. Given the weakness of current translation models this is a severe limitation. It often happens that SMT systems assign $P(F|\bar{E}) = 0$ to a correct reference translation $\bar{E}$ of F (see the discussion in Section 9). 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

Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010), pages 71–79,
Beijing, August 2010

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 $\mathcal{H}$:

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

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 over SMT lattices to identify useful subsequences in the ML translations (Sections 2 & 3). These subsequences drive the segmentation and transformation of lattices into smaller subproblems (Sections 4 & 5). Subproblems are mined for fluent strings (Section 6), resulting in improved translation fluency (Sections 7 & 8). Our results show that, when guided by the careful selection of subproblems, fluency can be improved with no real degradation of the BLEU score.

2 Lattice MBR Decoding

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

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

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, $n = 1 \ldots 4$), and θ are constants estimated on held-out data. The quantity $p(u|\mathcal{E})$ is the path posterior probability of n-gram u

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

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 efficiently calculated (Blackwood et al., 2010) using general purpose WFST operations (Mohri et al., 2002).

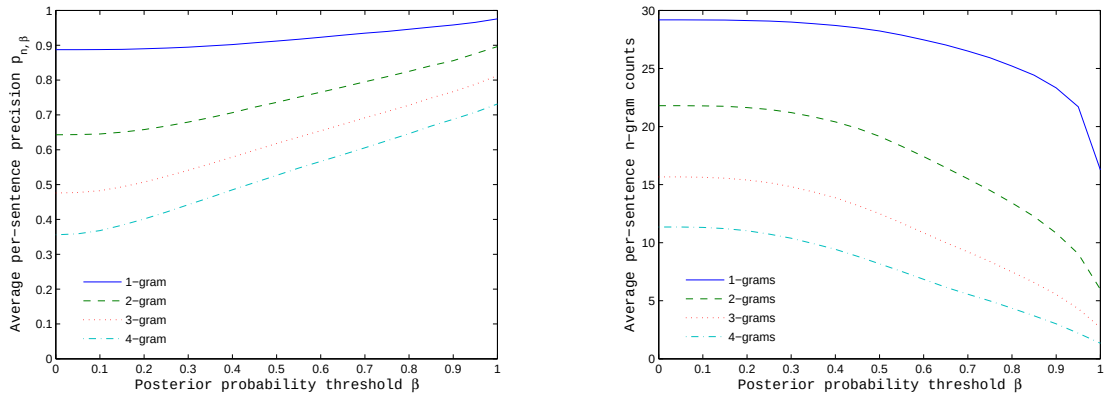


Figure 1: Average n-gram precisions (left) and counts (right) for 2075 sentences of NIST Arabic→English ML translations at a range of posterior probability thresholds $0 \leq \beta \leq 1$. The left plot shows at $\beta = 0$ the n-gram precisions used in the BLEU score of the ML baseline system.

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 β, 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 β, 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 β or higher. The precision at order n for threshold β is the proportion of n-grams in $\mathcal{N}_{n,\beta}$ also present in the references:

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

The left plot in Figure 1 shows average per-sentence n-gram precisions $\mathcal{P}_{n,\beta}$ at orders $1 \ldots 4$ 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 unigram precisions. We note that precision at all orders improves as the threshold β 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 for the same range of β. We see that for high β, there are 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 on average three 4-grams per sentence with posterior probabilities that exceed β. Even at this very high confidence level, high posterior n-grams occur frequently enough that we can expect them to be useful.

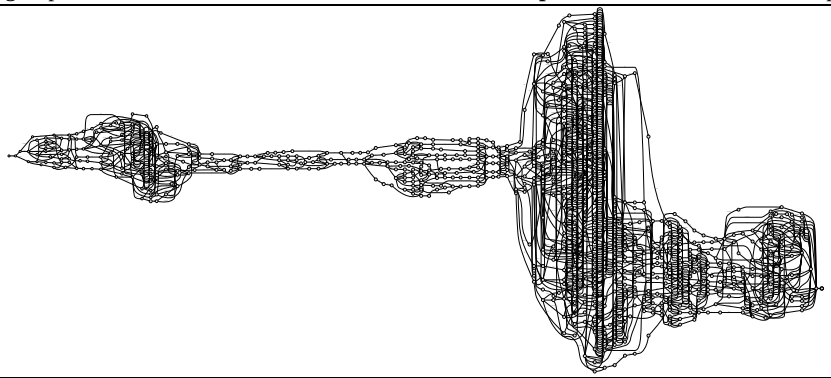
These precision results motivate our use of path posterior n-gram probabilities as a confidence measure. We assign confidence $p(\hat{E}_i^j|\mathcal{E})$ to subsequences $\hat{E}_i \ldots \hat{E}_j$ of the ML hypothesis.

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

4 Lattice Segmentation

We have shown that current SMT systems, although flawed, can identify with confidence par-

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 .



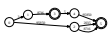
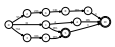
$\mathcal{H}_1$	$\mathcal{H}_2$	$\mathcal{H}_3$	$\mathcal{H}_4$	$\mathcal{H}_5$	$\mathcal{H}_6$	$\mathcal{H}_7$	$\mathcal{H}_8$	$\mathcal{H}_9$
								
433	1	4	1	6	1	6860	1	76

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

tial 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 lattice $\mathcal{E}$ into regions of high and low confidence. As shown in the example of Figure 2, the lattice segmentation process is performed relative to the ML hypothesis $\hat{E}$, i.e. relative to the best path through $\mathcal{E}$.

For confidence threshold β, we find all 4-grams $u = \hat{E}_i, \ldots, \hat{E}_{i+3}$ in the ML translation 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. This approach differs from simple posterior-based pruning in that we discard paths, rather than words or n-grams, which are not consistent with high-confidence regions of the ML hypothesis.

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 $/.*\hat{e}_{r-1}(.*)\hat{e}_{r+1}.*/\backslash 1/$.[1] 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 $\mathcal{H}_1, \ldots, \mathcal{H}_R$ defines the lattice segmentation.

5 Hypothesis Space Construction

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

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 need only 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 Ψ:

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

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

$$\mathcal{H} = \mathcal{E} \circ \bigotimes_{1 \le r \le R} \mathcal{H}_r \tag{7}$$

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.

Size of the Hypothesis Space If no new hypotheses are introduced by the operations Ψ, the size of the hypothesis space $\mathcal{H}$ is determined by the posterior probability threshold β. 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 are formed from subsequences that occur on every path in $\mathcal{E}$.

We can therefore use β 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 β 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

This section describes one implementation of the transformation function Ψ 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 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

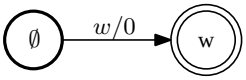
[1] In this notation parentheses indicate string matches so that $/.*y(a*)w.*/\backslash 1/$ applied to $xyaaawzz$ yields aaa.

LM	Translation hypothesis E and n-gram orders used by the LM to score each word	Score
4g	<s>$_1$ the$_2$ reactor$_3$ produces$_3$ plutonium$_2$ *needed*$_2$ *to*$_3$ *manufacture*$_4$ atomic$_3$ bomb$_2$.$_3$ </s>$_4$	-22.59
	<s>$_1$ the$_2$ reactor$_3$ produces$_3$ plutonium$_2$ *needed*$_2$ *to*$_3$ *manufacture*$_4$ *the*$_4$ atomic$_2$ bomb$_3$.$_4$ </s>$_4$	-23.61
5g	<s>$_1$ the$_2$ reactor$_3$ produces$_4$ plutonium$_5$ *needed*$_3$ *to*$_3$ *manufacture*$_4$ atomic$_5$ bomb$_2$.$_3$ </s>$_4$	-16.04
	<s>$_1$ the$_2$ reactor$_3$ produces$_4$ plutonium$_5$ *needed*$_3$ *to*$_3$ *manufacture*$_4$ *the*$_4$ atomic$_4$ bomb$_5$.$_4$ </s>$_5$	-17.96

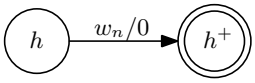
Figure 3: Scores and n-gram orders for hypotheses using 4-gram Kneser-Ney and 5-gram stupid-backoff (estimated from 1.1B and 6.6B tokens, resp.) LMs. Low confidence regions are in italics.

favour the shorter but disfluent hypothesis; normalising by length was not effective. However, the stupid-backoff LM has better coverage and the backing-off behaviour is a clue to the presence of disfluency. Similar cues have been observed in ASR analysis (Chase, 1997). The shorter hypothesis backs off to a bigram for "atomic bomb", whereas the longer hypothesis covers the same words with 4-grams and 5-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 less fluent hypotheses.

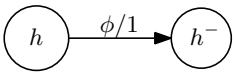
Let $\mathcal{S}$ denote the set of all n-grams in the monolingual training data. To identify partial hypotheses in sublattice regions that have complete monolingual coverage at the maximum order n, we build a coverage acceptor $\mathcal{C}_n$ with a similar form to the WFST representation of an n-gram 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 following experiments, we look at each $\mathcal{X}_n \circ \mathcal{C}_n$ and if there are paths with cost 0, only these are kept and all others discarded. We introduce this as a constraint on the hypothesis space which we will evaluate for improvement on fluency. Here the transformation function Ψ 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 Ψ returns $\mathcal{H}_r$ as we find it, since there is not enough monolingual coverage 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

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

of the odd numbered sentences of the MT02–MT05 testsets; the even numbered 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., 2009), 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. 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 LMBR factors $\theta_0, \ldots, \theta_4$ are set as in Tromble et al. (2008) using unigram precision $p = 0.85$ and recall ratio $r = 0.74$.

The effect of performing LMBR over the segmented hypothesis space is shown in Table 1. The hypothesis subspaces $\mathcal{H}_r$ 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 of monolingual coverage constraints on BLEU. We build acceptors $\mathcal{C}_n$ as described in Section 6 with $\mathcal{S}$ consisting of all n-grams in the English GigaWord. At $\beta = 0.6$ we found 181 sentences with sublattices $\mathcal{H}_r$ spanned by maximum order n-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 unconstrained LMBR decoding. Done in this way, constraining hypotheses to have 5-grams from the GigaWord has little impact on BLEU.

		tune	*test*	*nw08*	*ng08*
ML		54.2	53.8	51.3	36.3
β	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 $0 \leq \beta \leq 1$.

At this value of β, 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 ML and the LMBR+CC decoders. 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 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%). In other words, the monolingual fluency constraints were judged

		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.

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 Discussion

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 neighbours. This can be done independently before reassembling the refined regions. Lattice segmentation facilitates the application of post-processing and rescoring 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), and to segmental MBR for automatic speech recognition (Goel et al., 2004). Mohit et al. (2009) describe an alternative approach to improving specific portions of translation hypotheses. They use an SVM classifier to identify a single phrase in each source language sentence that is "difficult to translate"; such phrases are then translated using an adapted language model estimated from parallel data. In contrast to their approach, our approach is able to exploit large collections of monolingual data to refine multiple low confidence regions using posterior probabilities obtained from a high-quality evidence space of first-pass translations.

Testset	Sentences	Reachability
tune	2075	15%
test	2040	14%
nw08	813	11%
ng08	547	9%

Table 3: Arabic→English reference reachability.

We applied hypothesis space constraints based on monolingual coverage to low confidence regions resulting in improved fluency with no real 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 natural language generation. 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 testset sentences that can be aligned to any of the reference translations using our high quality baseline hierarchical decoder with a powerful grammar. The low level of reachability suggests that NLG may be required to achieve high levels of translation quality and fluency. Other rescoring approaches (Kumar et al., 2009; Li et al., 2009) may also benefit from NLG when the baseline is incapable of generating the reference.

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.

Acknowledgments

We would like to thank Matt Gibson and the human judges who participated in the evaluation. This work was supported in part under the GALE program of the Defense Advanced Research Projects Agency, Contract No. HR0011-06-C-0022 and the European Union Seventh Framework Programme (FP7-ICT-2009-4) under Grant Agreement No. 247762.

References

Allauzen, Cyril, Mehryar Mohri, and Brian Roark. 2003. Generalized algorithms for constructing statistical language models. In *Proceedings of ACL 2003*.

Blackwood, Graeme, Adrià de Gispert, and William Byrne. 2010. Efficient path counting transducers for minimum Bayes-risk decoding of statistical machine translation lattices. In *Proceedings of ACL 2010*.

Brants, Thorsten, Ashok C. Popat, Peng Xu, Franz J. Och, and Jeffrey Dean. 2007. Large language models in machine translation. In *Proceedings of the EMNLP 2007*.

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 *WMT 2009*.

Chase, Lin Lawrance. 1997. Error-responsive feedback mechanisms for speech recognizers, Ph.D. Thesis, Carnegie Mellon University.

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 ACL-IJCNLP 2009*.

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.

Iglesias, Gonzalo, Adrià de Gispert, Eduardo R. Banga, and William Byrne. 2009. Hierarchical phrase-based translation with weighted finite state transducers. In *Proceedings of the 2009 Annual Conference of the NAACL*.

Kneser, R. and H. Ney. 1995. Improved backing-off for m-gram language modeling. In *Acoustics, Speech, and Signal Processing*.

Knight, K and J Graehl. 2005. An overview of probabilistic tree transducers for natural language processing. In *Proceedings of CICLING 2005*.

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: Keynote Address*.

Kumar, Shankar and William Byrne. 2004. Minimum Bayes-risk decoding for statistical machine translation. In *NAACL 2004*.

Kumar, Shankar, Wolfgang Macherey, Chris Dyer, and Franz Och. 2009. Efficient minimum error rate training and minimum bayes-risk decoding for translation hypergraphs and lattices. In *Proceedings of ACL-IJCNLP 2009*.

Lavie, Alon and Michael J. Denkowski. 2009. The METEOR metric for automatic evaluation of machine translation. *Machine Translation Journal*.

Li, Zhifei, Jason Eisner, and Sanjeev Khudanpur. 2009. Variational decoding for statistical machine translation. In *Proceedings of ACL-IJCNLP 2009*.

Ma, Xiaoyi and Christopher Cieri. 2006. Corpus support for machine translation at LDC. In *LREC 2006*.

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 EACL*.

Mohit, B., F. Liberato, and R. Hwa. 2009. Language model adaptation for difficult-to-translate phrases. In *Proceedings of the 13th Annual Conference of the EAMT*.

Mohri, Mehryar, Fernando Pereira, and Michael Riley. 2002. Weighted finite-state transducers in speech recognition. In *CSL*, volume 16, pages 69–88.

Oberlander, Jon and Chris Brew. 2000. Stochastic text generation. In *Philosophical Transactions of the Royal Society*.

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 HLT Conference of the NAACL*.

Papineni, Kishore, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of ACL 2002*.

Sim, K.-C., W. Byrne, M. Gales, H. Sahbi, and P.C. Woodland. 2007. Consensus network decoding for statistical machine translation system combination. In *ICASSP 2007*.

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 AMTA*.

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 EMNLP*.

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 WMT 2007*.

Zens, Richard and Hermann Ney. 2006. N -gram posterior probabilities for statistical machine translation. In *Proceedings of WMT 2006*.