

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

Graeme Blackwood, Adrià de Gispert, and William Byrne

Machine Intelligence Laboratory

Cambridge University Engineering Department



24th of August 2010

Introduction and Motivation

- ▶ Translation quality is often described¹ in terms of **fluency** and **adequacy**
 - ▶ Adequacy should be more important than fluency
 - ▶ Humans rate less fluent translations as less adequate
- ▶ Therefore not enough to focus only on adequacy
- ▶ We propose a novel, robust framework for improving SMT fluency:
 1. Segment lattice using posterior-based confidence measure
 2. Apply fluency constraints in regions of low confidence
 3. Perform lattice MBR search over the refined hypothesis space
 4. Leads to improved fluency as judged by native speakers

¹C. Callison-Burch, P. Koehn, C. Monz, and J. Schroeder. *Findings of the 2009 Workshop on Statistical Machine Translation*. WMT 2009

Fluency in Maximum Likelihood Decoding

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

- ▶ Main fluency issues for maximum likelihood decoder:
 1. LM can only encourage production of fluent hypotheses
 2. Difficult to enforce constraints or introduce new hypotheses
- ▶ $P(E)$ is language model probability of hypothesis E
 - ▶ Closest thing to ‘fluency component’ in most SMT systems
- ▶ $P(E)$ only affects hypotheses likely under the translation model
 - ▶ SMT systems often assign $P(F|\bar{E}) = 0$ to reference \bar{E} of F
- ▶ Hierarchical and syntax-based SMT sometimes also lack fluency
- ▶ Fluent output requires tightly constraining target language grammar
 - ▶ This is at odds with broad coverage parser needed for robust translation

Conflict Between Robustness and Fluency

- ▶ Two main issues for maximum likelihood decoding fluency:
 1. SMT may fail to generate fluent hypotheses
 - ▶ No simple way to introduce them into the search
 2. SMT produces many translations that are not fluent
 - ▶ Enforcing fluency constraints can hurt robustness

Minimum Bayes-Risk Decoding Framework

- ▶ Propose novel framework to improve fluency of any SMT system
- ▶ Minimum Bayes-Risk search² over space of fluent hypotheses \mathcal{H} :

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

- ▶ We distinguish between the evidence space and hypothesis space:
 1. MBR evidence space \mathcal{E} produced by baseline SMT system
 2. MBR search for translations in collection of fluent sentences \mathcal{H}
- ▶ Choose translation closest to top baseline SMT hypotheses
 - ▶ As determined by the loss function $L(E, E')$ (e.g. $1 - \text{BLEU}$)

²S. Kumar and W. Byrne. *Minimum Bayes-risk decoding for statistical machine translation*. NAACL 2004.

MBR Decoding and Fluency

- ▶ Decoupling \mathcal{H} from first-pass translation offers great flexibility
 - ▶ Lexical, syntactic, and semantic constraints are easily applied
- ▶ \mathcal{H} can also be augmented with entirely new hypotheses
 - ▶ Produced by a stochastic Natural Language Generation system³
- ▶ Robust since constraints on \mathcal{H} do not affect evidence space \mathcal{E}
- ▶ In this work, we search out fluent hypotheses from the vast number of translations produced by the baseline decoder
- ▶ We use high confidence subsequences of \hat{E} to guide this process
 1. Trusted subsequences retained, alternatives considered elsewhere
 2. Fluency of sentence fragments can be refined in context

³J. Oberlander and C. Brew. *Stochastic text generation*. Philosophical Transactions of the Royal Society, 2000.

Lattice Minimum Bayes-Risk Decoding

- ▶ Linearised lattice MBR⁴ maximises conditional expected gain

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

- ▶ $p(u|\mathcal{E})$ is the “path posterior” probability of n -gram u

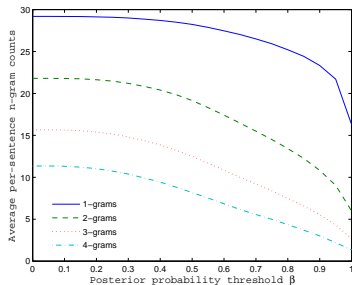
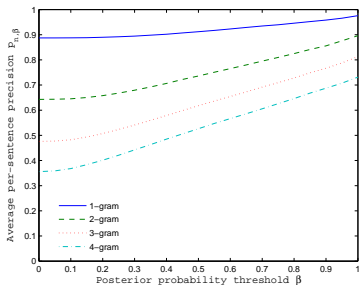
$$p(u|\mathcal{E}) = \sum_{E \in \mathcal{E}_u} P(E|F)$$

- ▶ Summation over subset of lattice hypotheses containing u

⁴R. Tromble, S. Kumar, F. Och, and W. Macherey. *Lattice Minimum Bayes-Risk decoding for statistical machine translation*. EMNLP 2008.

Posterior Probability Confidence Measures

- ▶ We use $p(u|\mathcal{E})$ of n -grams in \hat{E} to predict whether u is in the references
- ▶ Precisions and counts by order for confidence threshold $0 \leq \beta \leq 1$:



- ▶ High posterior n -grams good predictor; occur often enough to be useful

Lattice Segmentation

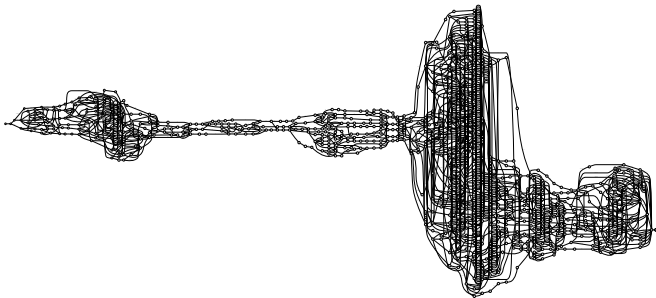
- ▶ We have shown we can identify “trusted” subsequences of \hat{E}
- ▶ We constrain MBR decoding to include these trusted subsequences and search for more fluent alternative translations elsewhere
- ▶ High posterior n -grams can be used to segment the first-pass lattice:
 1. Find all 4-grams u with $p(u|\mathcal{E}) \geq \beta$ for some confidence β
 2. Segment lattice \mathcal{E} into sequence of high and low confidence regions
- ▶ High confidence regions contain consecutive high confidence 4-grams
- ▶ Low confidence regions contain no high confidence 4-grams
- ▶ Confidence threshold β can be used to tighten or relax constraints on \mathcal{H}

Lattice Segmentation

- ▶ First-pass Arabic \rightarrow English ML 1-best translation hypothesis \hat{E} :










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 .

- ▶ Example translation lattice \mathcal{E} produced during first-pass decoding:



Lattice Segmentation

- ▶ Segment \hat{E} into R alternating high & low confidence subsequences
- ▶ Each subsequence is associated with an unweighted subspace \mathcal{H}_r
- ▶ Each low confidence subspace \mathcal{H}_r has a high confidence left context \hat{e}_{r-1} and a high confidence right context \hat{e}_{r+1} (both possibly empty)
- ▶ We build a transducer to implement $\text{regex } / \cdot * \hat{e}_{r-1} (\cdot *) \hat{e}_{r+1} \cdot * / \setminus 1 /$
- ▶ Low confidence regions obtained through standard FST composition

\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

Lattice segmentation example ($\beta = 0.8$)

Hypothesis Space Construction

- ▶ We now refine the quality of hypotheses in low confidence regions using the segmentation of the lattice to guide this process
- ▶ We specify a general transformation function Ψ that operates on each low confidence region and its neighbouring high confidence contexts:

$$\mathcal{H}_r \leftarrow \Psi(\mathcal{H}_{r-1}, \mathcal{H}_r, \mathcal{H}_{r+1})$$

- ▶ The final lattice MBR hypothesis space \mathcal{H} is then assembled by concatenation of the trusted string and refined sublattice regions:

$$\mathcal{H} = \bigotimes_{r=1}^R \mathcal{H}_r$$

Monolingual Coverage Constraints

- ▶ We propose one implementation of Ψ for improving SMT fluency
- ▶ Where possible, we constrain each low confidence region based on its coverage of high-order n -grams in a large monolingual corpus
- ▶ We build coverage acceptor \mathcal{C}_n to identify fluent partial hypotheses
 - ▶ \mathcal{C}_n has a similar form to WFSA backoff n -gram language model⁵
 - ▶ \mathcal{C}_n assigns a cost proportional to the number of times backed-off
- ▶ We use \mathcal{C}_n to constrain low confidence regions \mathcal{H}_r to contain only partial hypotheses completely covered by high-order n -grams
- ▶ \mathcal{C}_n can also be viewed as a very simplistic NLG system
 - ▶ Generates strings by concatenation of n -grams observed in our corpus

⁵C. Allauzen, M. Mohri, and B. Roark. *Generalized algorithms for constructing statistical language models*. ACL 2003.

Use of Neighbouring Context

- ▶ For coverage at order n , we construct $\mathcal{X}_r = \mathcal{L}_r \otimes \mathcal{H}_r \otimes \mathcal{R}_r$
 - ▶ \mathcal{L}_r accepts the last $n - 1$ words of \mathcal{H}_{r-1}
 - ▶ \mathcal{R}_r accepts the first $n - 1$ words of \mathcal{H}_{r+1}
- ▶ \mathcal{X}_r contains all of the partial hypotheses in the low confidence subspace \mathcal{H}_r padded with $n - 1$ words of high confidence context
- ▶ Composing $\mathcal{X}_r \circ \mathcal{C}_n$ assigns cost proportional to use of backoff
- ▶ If $\mathcal{X}_r \circ \mathcal{C}_n$ contains zero cost paths then all others are discarded
 - ▶ At least one path is completely covered by maximum order n -grams
- ▶ If $\mathcal{X}_r \circ \mathcal{C}_n$ contains no zero cost paths then \mathcal{H}_r is left unchanged
 - ▶ Insufficient monolingual coverage to guide selection of fluent hypotheses
- ▶ After applying these constraints to each low confidence region, we perform LMBR over the concatenation of string and sublattice regions

MBR Over Segmented Lattices

- ▶ NIST MT Arabic→English translation task (constrained)
- ▶ HiFST: a hierarchical phrase-based lattice decoder⁶
- ▶ Effect of confidence threshold β on LMBR BLEU score:

		nw08	ng08
ML		51.3	36.3
β	0.0	51.3	36.3
	0.2	51.3	36.3
	0.4	51.6	36.7
	0.6	52.1	36.6
	0.8	52.1	36.6
	1.0	52.2	36.7
LMBR		52.2	36.8

- ▶ Constraining \mathcal{H} with high posterior u from \hat{E} leads to little degradation

⁶Gonzalo Iglesias, Adrià de Gispert, Eduardo R. Banga, and William Byrne. *Hierarchical phrase based translation with weighted finite state transducers*. NAACL 2009.

Effect of Monolingual Coverage Constraints

- ▶ Build coverage acceptors \mathcal{C}_n using 5-grams from English GigaWord
- ▶ \mathcal{H}_r with zero cost paths in $\mathcal{X}_r \circ \mathcal{C}_n$ are constrained using $\beta = 0.6$
- ▶ 181 sentences of `nw08` have \mathcal{H}_r completely covered by 5-grams
- ▶ BLEU score for LMBR over coverage constrained lattices is 52.0
 - ▶ +0.7 over ML 1-best; only -0.2 below LMBR in unconstrained \mathcal{H}
- ▶ Constraining \mathcal{H}_r using GigaWord 5-grams has little impact on BLEU

Human Fluency Evaluation

- ▶ We analyse effect of coverage constraints where \hat{E}_{ML} and $\hat{E}_{LMBR-CC}$ differ
- ▶ 17 native speakers judged fluency of ML and LMBR+CC fragments
- ▶ Each fragment shown with high confidence left and right context
- ▶ Judges asked “Could this fragment occur in a fluent sentence?”
 - ▶ Both fluent: 1175 (59.6%)
 - ▶ Both not fluent: 75 (3.8%)
 - ▶ ML fluent, LMBR+CC not fluent: 192 (9.7%)
 - ▶ ML not fluent, LMBR+CC fluent: 530 (26.9%)
- ▶ Example: improved fluency through monolingual coverage constraints

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 ...

Natural Language Generation

- ▶ Our long-term goal is to integrate stochastic natural language generation in SMT without sacrificing robustness
- ▶ Our flexible framework already supports direct integration of NLG
 - ▶ We just need to be able to generate sentence fragments in context
- ▶ If generated \mathcal{H} contains generated \bar{E} for which $P(F|\bar{E}) = 0$
 - ▶ It can still be output by LMBR if voted for by similar ML hypotheses
- ▶ Reference reachability with large-scale hierarchical translation grammar used in CUED submission to NIST MT09:

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

Summary and Discussion

- ▶ We proposed a general MBR framework for improving SMT fluency
- ▶ Related to ideas in consensus decoding (Matusov et al., 2006; Sim et al., 2007), segmental MBR for ASR (Goel et al., 2004), and LM adaptation (Mohit et al., 2009)
- ▶ Showed how high posterior n -grams in \hat{E} guide lattice segmentation
 1. Simplifies hypothesis space refinement process
 2. Low confidence regions refined in high confidence context
- ▶ Monolingual coverage constraints are one way of improving fluency
 - ▶ Constrained low confidence regions to retain partial hypotheses with consecutive, overlapping n -grams from a fluent target language corpus
- ▶ Led to improved fluency with no real degradation in BLEU score
- ▶ Our decoding framework allows robust integration of stochastic natural language generation of sentence fragments in SMT

Thanks!

Questions?