

# Minimum Bayes-Risk Lattice Rescoring Methods for Statistical Machine Translation

Graeme Blackwood

Machine Intelligence Laboratory

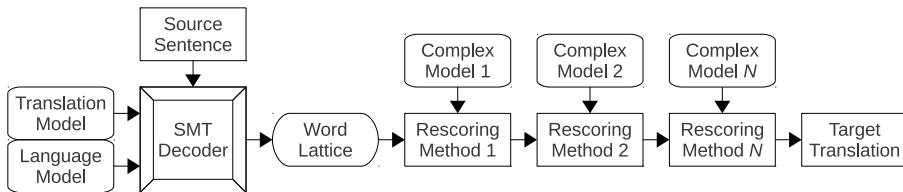
Cambridge University Engineering Department



20th of May, 2011

# Statistical Machine Translation Pipeline

- ▶ Decoding in SMT is often factored as a cascaded series of modules:



- ▶ Preserve uncertainty by passing as much information as possible to subsequent modules, usually in the form of a lattice or  $k$ -best list
- ▶ Lattices allow models difficult or impossible to use in first-pass decoding to be applied to very large subset of likely hypotheses

# Talk Outline

- ▶ Three related large-scale SMT lattice rescoring procedures based on minimum Bayes-risk decoding and realized in terms of WFSTs<sup>1</sup>:
  1. Efficient Lattice Minimum Bayes-Risk Decoding
    - ▶ Use weighted path counting transducers to efficiently compute the statistics required for MBR decoding over large SMT lattices
  2. Lattice Minimum Bayes-Risk System Combination
    - ▶ Generalize the efficient lattice MBR decoder to the task of combining multiple lattices produced by different SMT systems
  3. Confidence-Based Lattice Segmentation and MBR Fluency Constraints
    - ▶ Segment translation lattices under  $n$ -gram confidence measure in order to apply models that target specific deficiencies in SMT hypotheses

---

<sup>1</sup>Mehryar Mohri. *Finite-State Transducers in Language and Speech Processing*. 1997.

## Part I

# Efficient Path Counting Transducers for Minimum Bayes-Risk Decoding of SMT Lattices

# Statistical Machine Translation

- ▶ Data-driven approach to natural language translation
- ▶ Define distribution over possible target language translations
- ▶ Learn distribution parameters from large aligned parallel corpus
- ▶ SMT decoding: choose target language sentence  $E$  that maximises conditional probability  $P(E|F)$  for given source language sentence  $F$
- ▶ Source-channel model of SMT (Brown et. al, (1993)):

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

- ▶ Log-linear direct model of translation (Och and Ney, (2002)):

$$\hat{E} = \operatorname{argmax}_E P(E|F) = \operatorname{argmax}_E \left\{ \sum_{i=1}^M \lambda_i h_i(E, F) \right\}$$

- ▶ Feature weights  $\lambda_1^M$  optimised on held-out development set

## Lattice Minimum Bayes-Risk Decoding

- ▶ Minimum Bayes-risk decoding can be applied to any MT system that defines a posterior distribution over translation hypotheses:

$$P(E|F) = \frac{\exp(\alpha \sum_{i=1}^M \lambda_i h_i(E, F))}{\sum_{E'} \exp(\alpha \sum_{i=1}^M \lambda_i h_i(E', F))} \quad (1)$$

- ▶ MBR decoding under loss function  $L(E, E')$  has the general form<sup>2</sup>

$$\hat{E} = \operatorname{argmin}_{E' \in \mathcal{E}} \underbrace{\sum_{E \in \mathcal{E}} L(E, E') P(E|F)}_{\text{Bayes-risk of hypothesis } E'} \quad (2)$$

- ▶ MBR decoding over lattices can be implemented efficiently using path counting transducers to compute the required statistics
- ▶ This enables efficient decoding even for large SMT lattices

---

<sup>2</sup>Shankar Kumar and William Byrne. *Minimum Bayes-risk decoding for statistical machine translation*. NAACL 2004.

# Linearized Lattice Minimum Bayes-Risk Decoding

- ▶ Linearized lattice MBR<sup>3</sup> maximizes conditional expected gain:

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

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

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

- ▶ Note that this is not the same as a conditional expected count:

$$c(u|\mathcal{E}) = \sum_{E \in \mathcal{E}} \#_u(E) P(E|F) \quad (5)$$

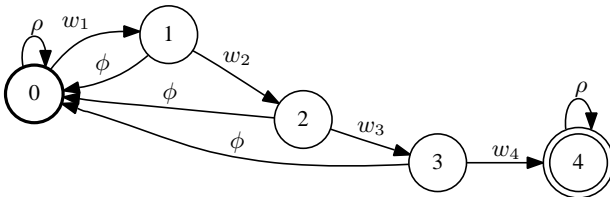
- ▶ **AIM:** Efficient and exact implementation of Equation (3)

---

<sup>3</sup>Roy Tromble, Shankar Kumar, Franz Och, and Wolfgang Macherey. *Lattice Minimum Bayes-Risk decoding for statistical machine translation*. EMNLP 2008.

# Posterior Probability Computation using Acceptors

- ▶ Path posteriors  $p(u|\mathcal{E})$  can be computed using FSAs<sup>4</sup>
  - ▶ Intersect acceptor for  $\Sigma^*u\Sigma^*$  with  $\mathcal{E}$  to obtain  $\mathcal{E}_u$
  - ▶ Then sum path weight  $P(E|F)$  for each  $E \in \mathcal{E}_u$
  - ▶ Repeated one-by-one in sequence for each  $n$ -gram  $u \in \mathcal{N}$
  - ▶ This “sequential” method can be very slow for large  $|\mathcal{N}|$
  
- ▶ Simplified path matching acceptor for the  $n$ -gram  $u = w_1w_2w_3w_4$

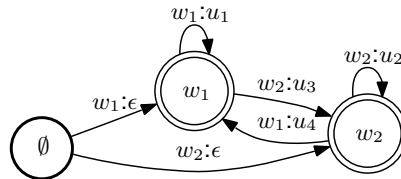


<sup>4</sup>Roy Tromble, Shankar Kumar, Franz Och, and Wolfgang Macherey. *Lattice Minimum Bayes-Risk decoding for statistical machine translation*. EMNLP 2008.

# Posterior Probability Computation using Transducers

- ▶ We show that exact  $p(u|\mathcal{E})$  can be computed simultaneously<sup>5</sup>
  - ▶ Using a single counting transducer for each order  $n = 1 \dots 4$
- ▶ We simplify counting by transducing lattice  $\mathcal{E}$  to lattice of  $n$ -grams  $\mathcal{E}_n$ 
  - ▶ Easier to count “unigrams” in  $\mathcal{E}_n$  than to count  $n$ -grams in  $\mathcal{E}$
- ▶ Example mapping transducer for any sequence of bigrams in  $\{u_1, u_2, u_3, u_4\}^*$  formed from the unigram alphabet  $\Sigma_1 = \{w_1, w_2\}$

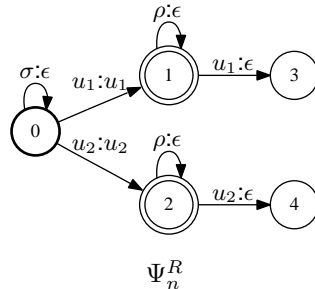
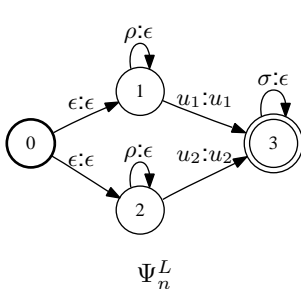
words	bigram
$w_1 w_1$	$u_1$
$w_2 w_2$	$u_2$
$w_1 w_2$	$u_3$
$w_2 w_1$	$u_4$



<sup>5</sup>Graeme Blackwood, Adrià de Gispert, and William Byrne. *Efficient path counting transducers for minimum Bayes-risk decoding of SMT lattices*. ACL 2010.

# Efficient Path Counting 1

- ▶ Transducer  $\Psi_n$  computes simultaneously  $p(u|\mathcal{E})$  for all  $u \in \mathcal{N}_n$
- ▶ Example:  $\Psi_n^L$  and  $\Psi_n^R$  for  $u_{1,2} \in \mathcal{N}_n$  for some order  $n$ :



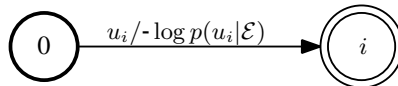
- ▶  $\Psi_n^L$  counts **first** (left-most) occurrence of each  $n$ -gram  $u \in \mathcal{N}_n$  on path<sup>6</sup>
- ▶  $\Psi_n^R$  counts **last** (right-most) occurrence of each  $n$ -gram  $u \in \mathcal{N}_n$  on path<sup>7</sup>

<sup>6</sup>Cyril Allauzen, Shankar Kumar, Wolfgang Macherey, Mehryar Mohri, and Michael Riley. *Expected sequence similarity maximization*. NAACL 2010.

<sup>7</sup>Graeme Blackwood. *Lattice rescoring methods for statistical machine translation*. Ph.D. Thesis. Cambridge University Engineering Department and Clare College. 2010.

## Efficient Path Counting 2

- ▶ We form weighted path counts acceptor  $\mathcal{X}_n = \mathcal{E}_n \circ \Psi_n$
- ▶ Project output, map to log semiring,  $\epsilon$ -removal, determinize, minimize
- ▶  $\mathcal{X}_n$  has one arc from the start state for each  $u \in \mathcal{N}_n$ :



- ▶  $\mathcal{E}_n \circ \Psi_n$  can have many states and arcs for large  $|\mathcal{N}_n|$ 
  - ▶ Slow log semiring  $\epsilon$ -removal and determinization operations
- ▶ If  $\Psi_n^R$  is used instead of  $\Psi_n^L$ , then
  1. Each path in  $\mathcal{E}_n \circ \Psi_n$  has a single non- $\epsilon$  output label  $u$
  2. All paths leading to the same final state share the same output label  $u$
- ▶ This allows a lattice traversal procedure to be used to compute  $p(u | \mathcal{E})$ 
  - ▶ Simply requires propagating symbols as well as probabilities<sup>8</sup>

---

<sup>8</sup>Graeme Blackwood, Adrià de Gispert, and William Byrne. *Efficient path counting transducers for minimum Bayes-risk decoding of SMT lattices*. ACL 2010.

## Efficient LMBR Decoder Implementation

- ▶ We use exact values of  $p(u|\mathcal{E})$  at all orders to compute

$$\hat{E} = \operatorname{argmax}_{E' \in \mathcal{E}} \left\{ \theta_0 |E'| + \sum_{n=1}^4 g_n(E') \right\} \quad (6)$$

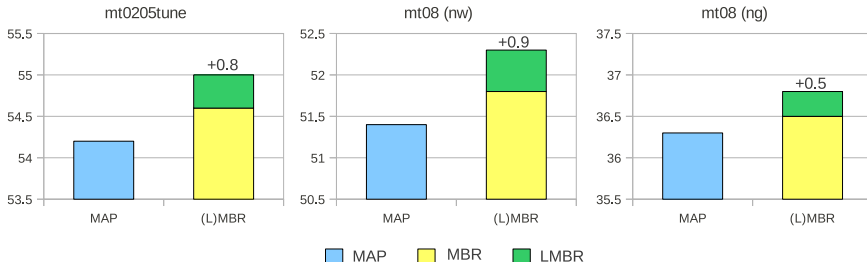
where  $g_n(E')$  is partial gain of  $E'$  associated with  $n$ -grams of order  $n$

- ▶ Construct acceptor  $\Omega_n$  to apply  $g_n(E')$  to all paths in the lattice  $\mathcal{E}$
- ▶ Form hypothesis space  $\mathcal{E}_0$  as  $\mathcal{E}$  with weight  $\theta_0$  on all arcs
- ▶  $\hat{E}$  is maximum weight string in LMBR decoder automaton:

$$\mathcal{E}_0 \circ \Omega_1 \circ \Omega_2 \circ \Omega_3 \circ \Omega_4 \quad (7)$$

## Lattice MBR Decoding Performance

- ▶ NIST MT 2008 Arabic→English translation task (constrained)
- ▶ HiFST: a hierarchical phrase-based lattice decoder<sup>9</sup>
- ▶ First-pass lattices rescored with large zero-cutoff 5-gram LM (6B words)
- ▶ IBM BLEU scores for 1-best, 1000-best MBR, and lattice MBR:

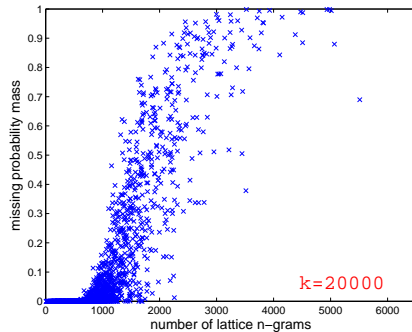
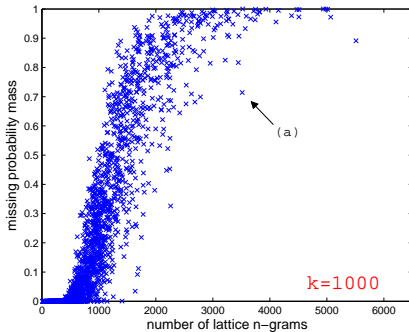


- ▶ BLEU gains over 1-best: +0.8 (tune), +0.9 (test nw), +0.5 (test ng)

<sup>9</sup>Adrià de Gispert, Gonzalo Iglesias, Graeme Blackwood, Eduardo R. Banga, and William Byrne. *Hierarchical phrase-based translation with weighted finite state transducers and shallow- $n$  grammars*. Computational Linguistics 2010.

## Evidence Space Size Analysis

- ▶ LMBR beats  $k$ -best MBR by exploiting a much larger evidence space<sup>10</sup>
- ▶ Let  $\phi(\mathcal{E}) = \sum_{E \in \mathcal{E}} P(E|F)$ . The lattice probability mass missing from the  $k$ -best list of the most likely hypotheses is  $1 - \phi(\mathcal{E}_K)/\phi(\mathcal{E}_L)$
- ▶  $\phi(\mathcal{E}_K)$  and  $\phi(\mathcal{E}_L)$  can be computed exactly by pushing weights



- ▶ Even 20,000-best lists discard much of the first-pass evidence

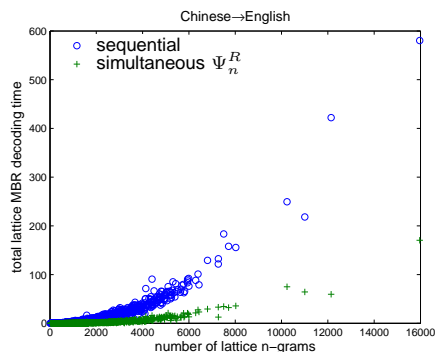
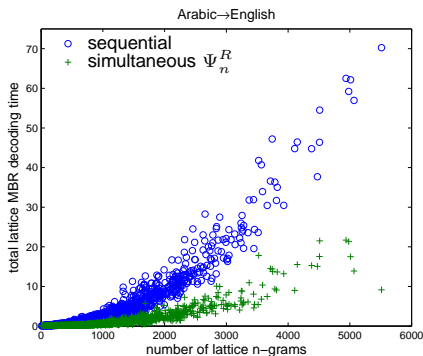
<sup>10</sup>Roy Tromble, Shankar Kumar, Franz Och, and Wolfgang Macherey. *Lattice Minimum Bayes-Risk decoding for statistical machine translation*. EMNLP 2008.

# Lattice MBR Decoding Efficiency

- ▶ Average posterior probability computation time (seconds/sentence):

Method	mt0205tune	mt0205test
sequential	1.52	1.62
$\Psi_n^L$	3.32	3.62
$\Psi_n^R$	0.84	0.87

- ▶ Total lattice MBR decoder time as a function of number of  $n$ -grams:



## Part II

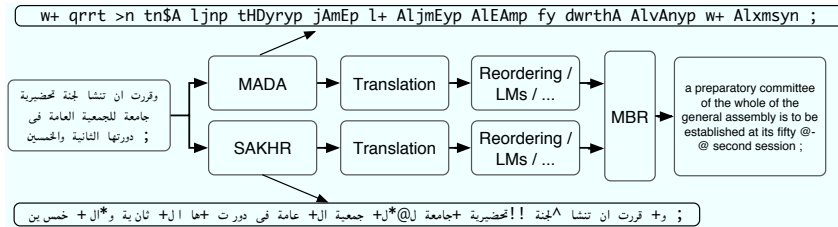
# Multiple-Lattice Minimum Bayes-Risk System Combination

# Statistical Machine Translation System Combination

- ▶ System combination is an effective technique for improving MT quality
  - ▶ Exploits differences in the nature of errors made by individual systems
- ▶ **Multi-Input Translation**
  - ▶ Multiple representations of the source sentence are available
  - ▶ E.g. alternative morphological analyses, word segmentations, automatic speech recognition transcriptions, source sentence paraphrases...
- ▶ **Multi-Source Translation**
  - ▶ The source sentence is available in multiple languages
  - ▶ E.g. the source sentence has already been translated into some other language(s) for which automatic MT system(s) are available

# Lattice MBR System Combination Pipeline

- ▶ Lattice MBR provides simple framework for combining lattices
- ▶ E.g. multi-input translation of alternative morphological analyses<sup>11</sup>



- ▶ Efficient multiple-lattice MBR decoder using path counting transducers
- ▶ Exploits full evidence space of exponentially many diverse translations
- ▶ Possible to obtain large gains in BLEU score over individual systems

<sup>11</sup>Adrià de Gispert, Sami Virpioja, Mikko Kurimo, and William Byrne. *MBR Combination of Translation Hypotheses from Alternative Morphological Decompositions*. NAACL 2009.

## System Combination Minimum Bayes-Risk Decoder

- ▶  $M$  distinct lattices  $\mathcal{E}^{(i)}$ ,  $i = 1 \dots M$  for each source sentence
- ▶ MBR search space formed as the union of individual lattices:

$$\mathcal{E} = \bigoplus_{i=1}^M \mathcal{E}^{(i)} \quad (8)$$

- ▶ Let  $\mathcal{N}$  denote the set of all  $n$ -grams in the unioned lattices:

$$\mathcal{N} = \bigcup_{i=1}^M \mathcal{N}^{(i)} \quad (9)$$

- ▶ With these definitions, the multiple-lattice MBR decoder has the same form as the single-lattice decoder:

$$\hat{E} = \operatorname{argmax}_{E' \in \mathcal{E}} \left\{ \theta_0 |E'| + \sum_{u \in \mathcal{N}} \theta_u \#_u(E') p(u|\mathcal{E}) \right\} \quad (10)$$

- ▶ The difference is in computation of  $n$ -gram posterior probabilities  $p(u|\mathcal{E})$

## System Combination Posterior Probabilities

- ▶ Posterior probability of  $n$ -gram  $u$  computed as a linear interpolation of individual lattice posteriors:

$$p(u|\mathcal{E}) = \sum_{i=1}^M \lambda_i p_i(u|\mathcal{E}^{(i)}) \quad (11)$$

- ▶  $0 \leq \lambda_i \leq 1$  are system-specific interpolation weights with  $\sum_{i=1}^M \lambda_i = 1$
- ▶ Lattice posteriors required for the interpolation are computed as

$$p_i(u|\mathcal{E}^{(i)}) = \sum_{E \in \mathcal{E}_u^{(i)}} P_i(E|F) \quad (12)$$

where  $\mathcal{E}_u^{(i)} = \{E \in \mathcal{E}^{(i)} : \#_u(E) > 0\}$  is paths in  $\mathcal{E}^{(i)}$  containing  $u$

- ▶ Posteriors computed efficiently using path counting transducers  $\Psi_n$

# Multi-Input Translation Experiments

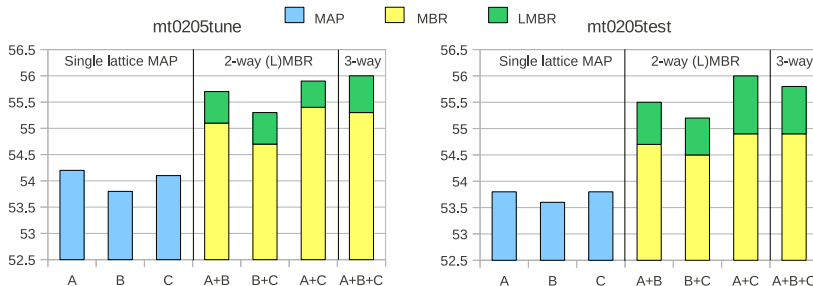
- ▶ Arabic→English NIST MT 2008 constrained data track
- ▶ Arabic data preprocessed with one of three morphological analysers:
  - ▶ Two different analyses using MADA (Habash and Rambow, 2005)
  - ▶ One analysis using the Arabic Morphological Tagger of Sakhr Software<sup>12</sup>
- ▶ Chinese→English GALE P4 evaluation framework
- ▶ Chinese data preprocessed with two different word segmentations:
  - ▶ AGILE segmentation distributed by BBN for AGILE P4
  - ▶ Oxford Chinese word segmentation (Zhang and Clark, 2007)
- ▶ Same MT pipeline applied to each Ar→En and Zh→En input variant:
  - ▶ Alignment, rule extraction, MERT optimization, 5-gram rescoring (around 6B words), then extract  $k$ -best lists or lattices for system combination

---

<sup>12</sup><http://www.sakhr.com/default.aspx>

# Multi-Input Translation Results (Arabic→English)

- Multi-input combination of alternative morphological analyses<sup>13</sup>

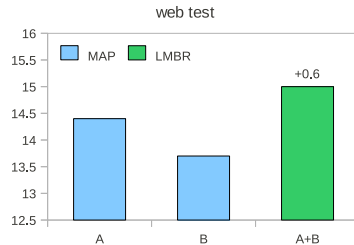
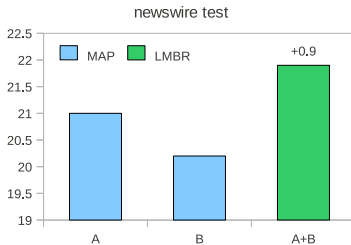


- 2-way gains over best single system between +1.1 and +2.2 BLEU
- 3-way LMBR offers no improvement over 2-way – insufficient diversity?
- Lattice MBR improves over  $k$ -best MBR by between 0.5 and 1.0 BLEU

<sup>13</sup>A = Mada1, B = Mada2, C = Sakhr

## Multi-Input Translation Results (Chinese→English)

- ▶ Multi-input combination of alternative Chinese word segmentations<sup>14</sup>



- ▶ Test BLEU gains over best individual system: +0.9 (nw) and +0.6 (web)
- ▶ Smaller gains than Arabic→English – lower quality translation lattices

<sup>14</sup>A = AGILE segmentation, B = Oxford segmentation

## Lattice MBR System Combination Example

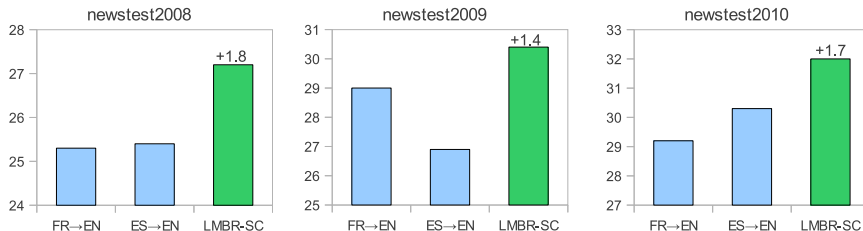
- ▶ Arabic→English mt0205tune system combination example:

Source	Tokenized translation string	BLEU <sub>S</sub>
$R_1$	over the next 13 years , peking invested in the construction of 7 new plants .	-
$R_2$	peking invested in the construction of 7 new plants over the next 13 years .	-
$R_3$	beijing has invested in building 7 new plants over the following 13 years .	-
$R_4$	peking has invested in the construction of 7 new plants in the next 13 years .	-
$\hat{E}_A$	the beijing invested in the construction of 7 new factor in the next 13 years .	0.6865
$\hat{E}_B$	beijing has invested in building new plants in the next 13 years .	0.7882
$\hat{E}_+$	beijing has invested in the construction of 7 new plants in the next 13 years .	1.0000

- ▶  $\hat{E}_A$  includes all of the required information but has poor fluency
- ▶  $\hat{E}_B$  is fluent but omits important content: the number of new plants
- ▶  $\hat{E}_+$  is both fluent and contains all of the required information content

# Multi-Source Translation Experiments

- ▶ LMBR combination of French→English and Spanish→English lattices
- ▶ WMT 2010 evaluation framework and baseline HiFST decoder<sup>15</sup>



- ▶ Good testset BLEU gains over best individual system LMBR
- ▶ Large gains for BLEU computed with respect to a single reference

<sup>15</sup>Juan Pino, Gonzalo Iglesias, Adrià de Gispert, Graeme Blackwood, Jamie Brunning, and William Byrne. *The CUED HiFST system for the WMT10 translation shared task*. WMT 2010.

## Part III

# Confidence-Based Lattice Segmentation and Monolingual Fluency Constraints for Statistical Machine Translation

# Machine Translation Fluency and Adequacy

- ▶ MT quality is often described in terms of **fluency** and **adequacy**<sup>16</sup>
  - ▶ Translation adequacy ought to be more important than fluency
  - ▶ Humans tend to rate less fluent translations as less adequate
- ▶ Therefore not enough to focus solely on adequacy
  - ▶ Fluency required to achieve widespread acceptance of SMT
- ▶ Novel and 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
- ▶ Leads to improved fluency as judged by native speakers

---

<sup>16</sup>C. 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

- ▶ Main fluency issues for maximum likelihood decoder:

$$\hat{E}_{\text{ML}} = \underset{E}{\operatorname{argmax}} P(F|E)P(E) \quad (13)$$

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 translation can also lack fluency
    - ▶ Fluent output requires tightly constraining target language grammar which 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 compromise robustness

# Minimum Bayes-Risk Decoding Framework

- ▶ Propose novel framework to improve fluency of any SMT system<sup>17</sup>
- ▶ Minimum Bayes-Risk search over space of fluent hypotheses  $\mathcal{H}$ :

$$\hat{E}_{\text{MBR}} = \underset{E' \in \mathcal{H}}{\operatorname{argmin}} \sum_{E \in \mathcal{E}} L(E, E') P(E|F) \quad (14)$$

- ▶ The hypothesis space is decoupled from the evidence 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 MBR loss function  $L(E, E')$  (e.g. 1 – BLEU)

---

<sup>17</sup>Graeme Blackwood, Adrià de Gispert, and William Byrne. Fluency constraints for minimum Bayes-risk decoding of statistical machine translation lattices. COLING 2010.

# MBR Decoding and Fluency

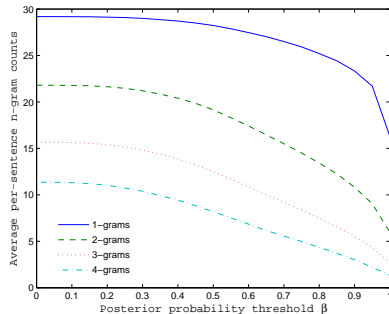
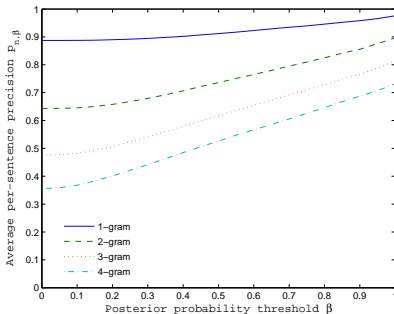
- ▶ Decoupling  $\mathcal{H}$  from first-pass translation offers great flexibility
  - ▶ Lexical, syntactic, and semantic constraints are easily applied
- ▶  $\mathcal{H}$  can also be augmented with entirely new hypotheses
  - ▶ Produced by a stochastic natural language generation system<sup>18</sup>
- ▶ Robust since constraints on  $\mathcal{H}$  do not affect evidence space  $\mathcal{E}$
- ▶ In initial experiments, fluent hypotheses are sought amongst the vast number of alternative translations produced by the baseline decoder
- ▶ Use high confidence subsequences of  $\hat{E}_{ML}$  to guide this process
  1. Trusted subsequences retained, alternatives considered elsewhere
  2. Allows fluency of sentence fragments to be refined in context

---

<sup>18</sup>J. Oberlander and C. Brew. *Stochastic text generation*. Philosophical Transactions of the Royal Society, 2000.

# Posterior Probability Confidence Measures

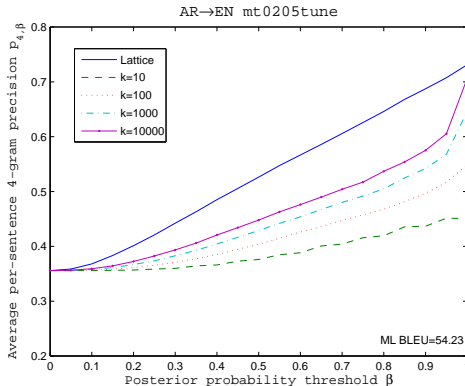
- ▶ Compute  $n$ -gram posteriors from lattice:  $p(u|\mathcal{E}) = \sum_{E \in \mathcal{E}_u} P(E|F)$
- ▶ Use  $p(u|\mathcal{E})$  of  $n$ -grams in  $\hat{E}_{ML}$  to predict whether  $u$  is in the references<sup>19</sup>
- ▶ Reference precisions and counts by order for confidence  $0 \leq \beta \leq 1$ :



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

## Evidence Space Size and Reference Precisions

- ▶ 4-gram reference precisions of  $n$ -gram posteriors computed from  $k$ -best lists of size 10, 100, 1000, 10000 and the full lattice evidence space<sup>20</sup>

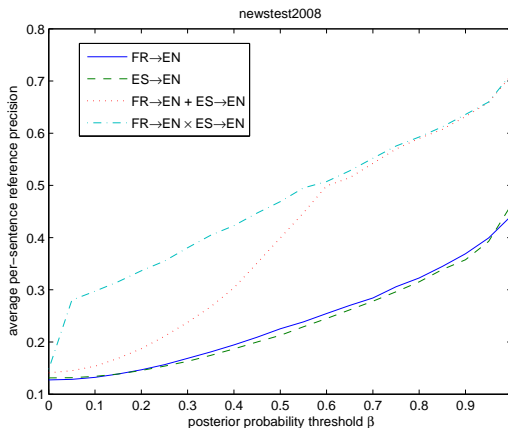


- ▶ Using the full lattice evidence space improves confidence estimates

<sup>20</sup>Adrià de Gispert, Gonzalo Iglesias, Graeme Blackwood, Eduardo R. Banga, and William Byrne. *Hierarchical phrase-based translation with weighted finite state transducers and shallow- $n$  grammars*. Computational Linguistics 2010.

## System Combination Reference Precisions

- 4-gram reference precisions for WMT 2010 French→English and Spanish→English single system and multi-source MBR 1-best:



- Multiple evidence spaces significantly improve confidence estimates

# Lattice Segmentation

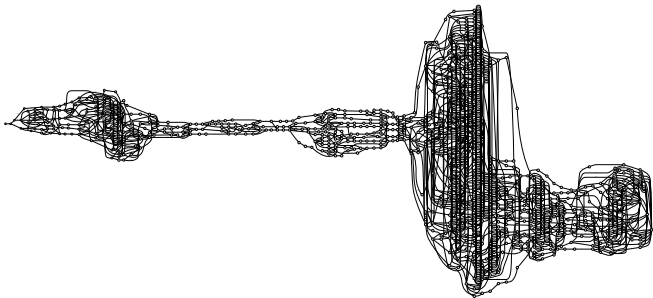
- ▶ Precision plots show we can identify “trusted” subsequences of  $\hat{E}_{ML}$
- ▶ We constrain MBR decoding to include these trusted subsequences and search for more fluent alternative translations elsewhere
- ▶ High posterior  $n$ -grams can be used to segment the first-pass lattice:
  1. Find 4-grams  $u$  in  $\hat{E}_{ML}$  with  $p(u|\mathcal{E}) \geq \beta$  for some confidence threshold  $\beta$
  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  $\beta$  acts to tighten or relax constraints on  $\mathcal{H}$

# Lattice Segmentation Example

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










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 Procedure

- ▶ Segment  $\hat{E}_{ML}$  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  $\text{regex } /. * \hat{e}_{r-1} (.*) \hat{e}_{r+1} . * /\backslash 1/$
- ▶ Low confidence regions extracted by composing with this transducer

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

Lattice segmentation example ( $\beta = 0.8$ )

# Hypothesis Space Construction

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

$$\mathcal{H}_r \leftarrow \Psi(\mathcal{H}_{r-1}, \mathcal{H}_r, \mathcal{H}_{r+1}) \quad (15)$$

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

$$\mathcal{H} = \bigotimes_{r=1}^R \mathcal{H}_r \quad (16)$$

# Monolingual Coverage Constraints

- ▶ One possible implementation of  $\Psi$  for improving SMT fluency:
  - ▶ Where possible, each low confidence region is constrained in accordance with its high-order  $n$ -gram coverage in a large, fluent monolingual corpus
- ▶ Coverage acceptor  $\mathcal{C}_n$  identifies fluent partial hypotheses:
  - ▶  $\mathcal{C}_n$  has a similar form to WFSA backoff  $n$ -gram language model<sup>21</sup>
  - ▶  $\mathcal{C}_n$  assigns a cost proportional to the number of times backed-off
- ▶  $\mathcal{C}_n$  is used 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 that generates strings by concatenation of  $n$ -grams observed in the training corpus

---

<sup>21</sup>C. 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

- ▶ Effect of confidence threshold  $\beta$  on LMBR BLEU score:

		mt08nw	mt08ng
ML		51.3	36.3
$\beta$	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
LMBR		52.2	36.8

- ▶ Confidence-based lattice segmentation procedure works as intended
- ▶ Constraining  $\mathcal{H}$  with high posterior  $u$  from  $\hat{E}_{ML}$  leads to little degradation

# 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 mt08nw 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 ...

# Integrating Natural Language Generation

- ▶ Long-term goal is to integrate stochastic natural language generation in SMT without sacrificing robustness
- ▶ MBR decoding framework already supports direct integration of NLG
  - ▶ Just need to be able to generate sentence fragments in context
- ▶ If hypothesis space  $\mathcal{H}$  contains generated  $\bar{E}$  for which  $P(F|\bar{E}) = 0$ , it can still be selected by LMBR if voted for by similar hypotheses

## Summary and Discussion

- ▶ Described lattice MBR framework for improving SMT fluency
  - ▶ Related to ideas in confusion network decoding (Matusov et al., 2006; Sim et al., 2007; Rosti et al., 2007), segmental MBR for ASR (Goel et al., 2004), and LM adaptation for “difficult” phrases (Mohit et al., 2009)
- ▶ Showed how high confidence  $n$ -grams in  $\hat{E}_{ML}$  guide segmentation
  1. Considerably 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
- ▶ Lattice segmentation and efficient MBR decoder potentially allow for robust integration of stochastic natural language generation of sentence fragments in SMT

# Questions

## Hybrid Decision Rule Accuracy

- ▶ Hybrid decision rule for linearised lattice MBR<sup>22</sup> can be written as:

$$\hat{E} = \operatorname{argmax}_{E' \in \mathcal{E}} \left\{ \theta_0 |E'| + \sum_{u \in \mathcal{N}: 1 \leq |u| \leq k} \theta_{|u|} \#_u(E') p(u|\mathcal{E}) + \sum_{u \in \mathcal{N}: k < |u| \leq N} \theta_{|u|} \#_u(E') c(u|\mathcal{E}) \right\} \quad (17)$$

- ▶ Range of  $n$ -gram orders using  $p(u|\mathcal{E})$  and  $c(u|\mathcal{E})$  is determined by  $k$
- ▶ Arabic→English MAP 1-best and hybrid LMBR BLEU for  $k = 0 \dots 4$

		mt0205tune	mt0205test
MAP		54.2	53.8
$k$	0	52.6	52.3
	1	54.8	54.4
	2	54.9	54.5
	3	54.9	54.5
	4	55.0	54.6

<sup>22</sup>Cyril Allauzen, Shankar Kumar, Wolfgang Macherey, Mehryar Mohri, and Michael Riley. *Expected sequence similarity maximization*. NAACL 2010.