ACS Statistical Machine Translation

Lecture 10: Minimum Error Training, Rescoring and System Combination

Department of Engineering
University of Cambridge

Bill Byrne
bill.byrne@cam.ac.uk

Lent 2013

Statistical Machine Translation (SMT)

Fundamental idea:

> "*Any* word sequence of a target language is a possible translation of a given input sentence of a source language"

- Each target word sequence is a translation hypothesis, and it has a certain probability
- This probability is defined according to a **statistical model** describing the translation process
- The model parameters are **estimated automatically** from big parallel texts

Translation $\equiv$ **finding** the highest-probability hypothesis

$$\hat{t} = \text{argmax}_t \, P(t|s) = \text{argmax}_t \quad \underbrace{P(s|t)}_{\substack{\text{Translation} \\ \text{Model}}} \quad \underbrace{P(t)}_{\substack{\text{Language} \\ \text{Model}}}$$

SMT Depends on Data

Statistical translation models try to *learn* from the parallel corpus of translation 'examples' in order to *generalize* to unseen input data.

The corpus is crucial:

- Parallel texts *and* monolingual texts
  - size ← *the bigger, the better*
  - language pairs, direction ← *which was the original?*
  - domain, literalness, domain scalability
  - quality, noise, complete errors ← *automatically generated?*
  - linguistic annotation ← *Part-Of-Speech, lemmas, dependencies, ...*
- Consistent pre-processing
  - tokenization, normalization, length-ratio filtering
- Evaluation metrics
  - number of golden reference translations
  - quality of golden reference translations ← *similar to training material?*

SMT General Concepts

Model definition and estimation:

- Translation unit: defines the model and search algorithms
- Fully automatic, inferred
- Data sparsity: parallel corpus quality and task definition
- Complexity: exact vs approximated estimation

Search:

- Complexity: monotonic vs reordered
- Completeness: exact vs pruned

Minimum Error Training (MERT)

In practice, the translation process is formulated as a log-linear combination of features:

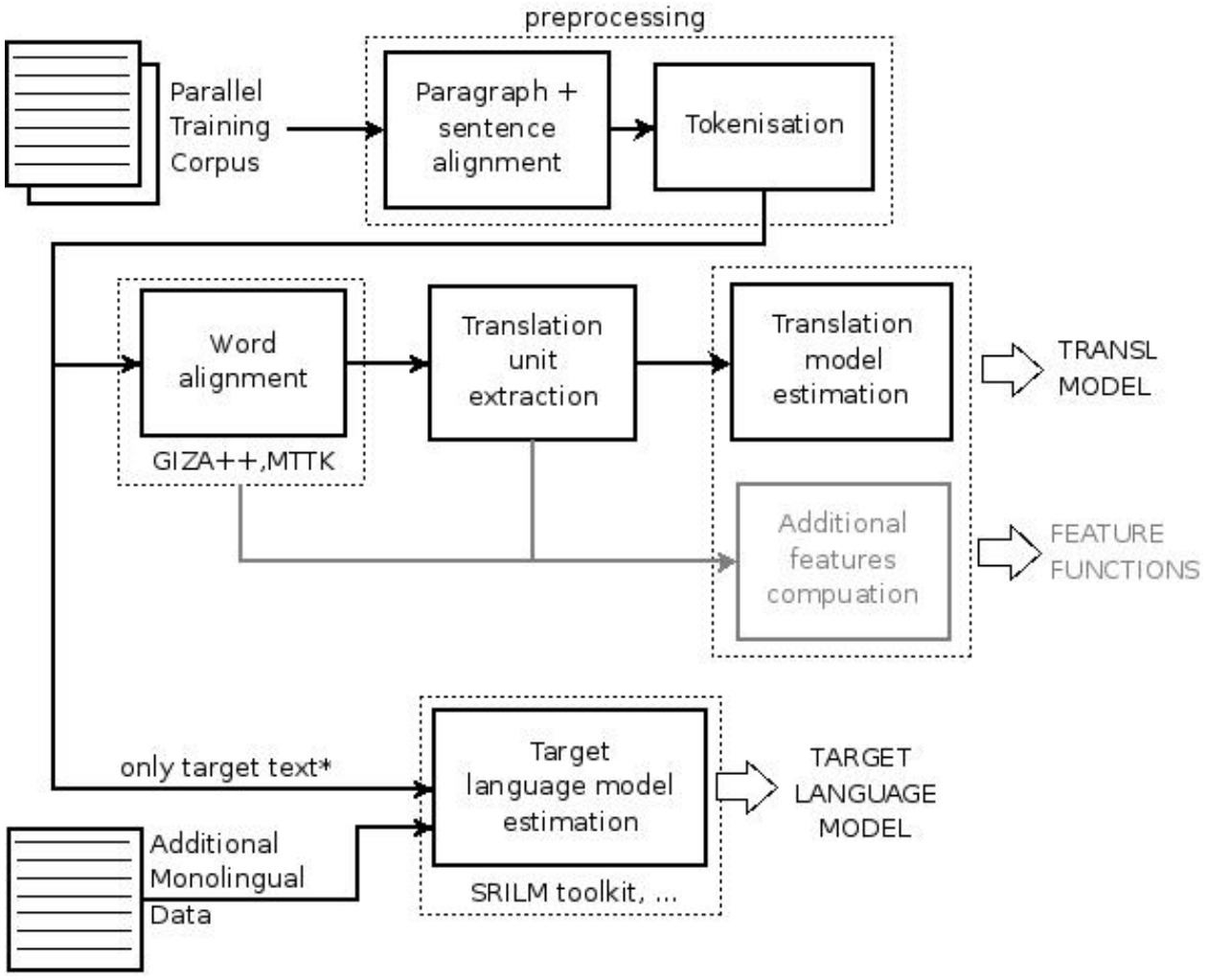
$$\hat{t}_1^I = \underset{t_1^I}{\operatorname{argmax}} \left\{ \sum_{m=1}^{M} \lambda_m h_m(s_1^J, t_1^I) \right\}$$

- Each feature contributes differently according to a weight λ_m
- **Minimum Error Training (MERT)** used to optimize weights according to a given development set and evaluation metric (BLEU)
- Significant gains over uniform weights

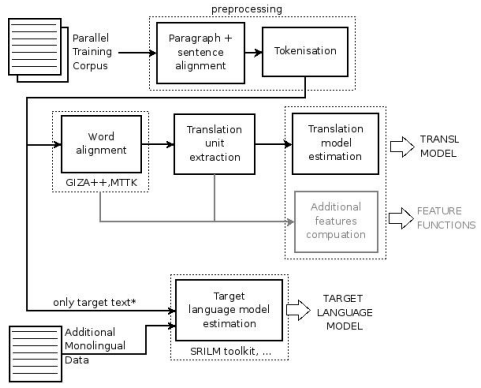
Typical feature functions:

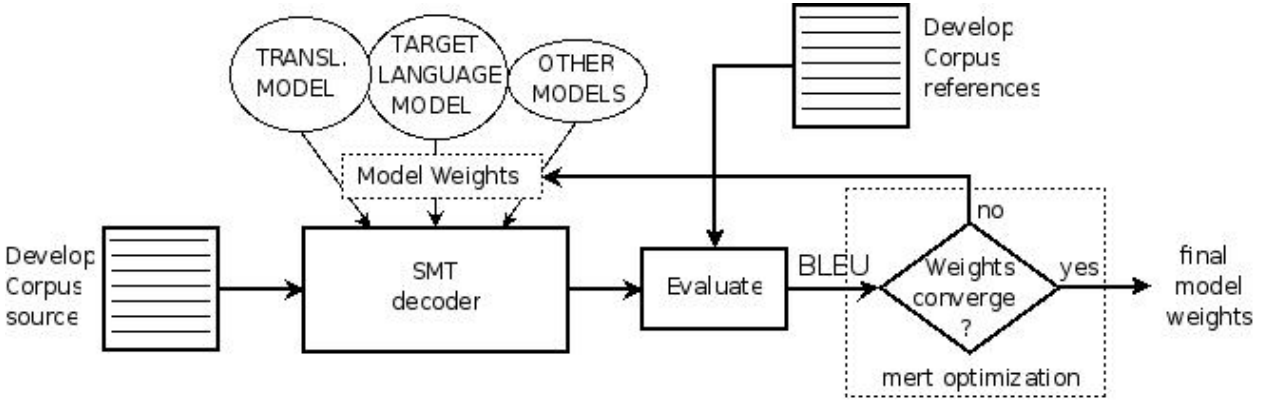
- **translation model** (included in both directions)
- **target language model**
- word and phrase penalties
- lexical features based on IBM model 1 word-to-word probabilities (both directions)
- additional phrase counters

General Pipeline. Model estimation

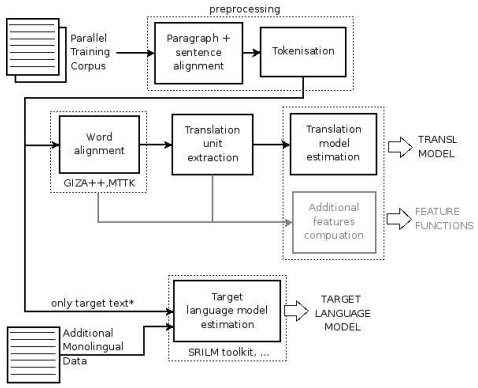


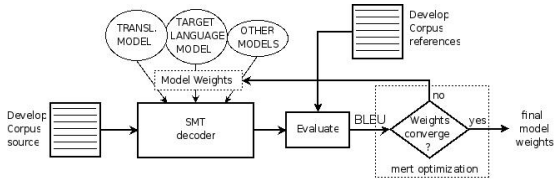
General Pipeline. Discriminative training

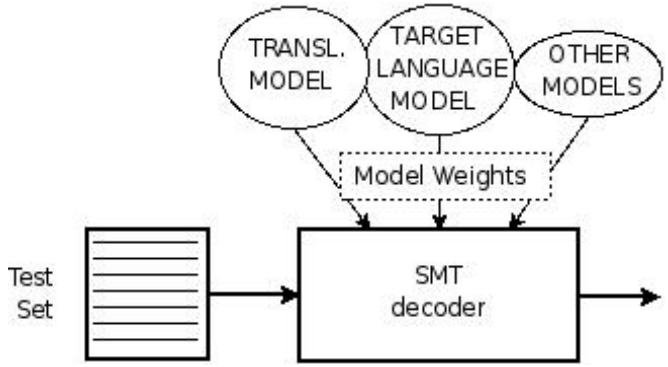




General Pipeline. Decoding







Rescoring or hypothesis re-ranking

- Not all features can be straightforwardly integrated in decoding
  - Complexity
  - Unavailable contextual information
- Additional complex features can be integrated via Rescoring
- Over N-best lists or lattices
- Rescoring models
  - Large Language Models: higher orders, extra material, ...
  - Phrase Segmentation Models
  - Minimum Bayes Risk Decoding
  - Lattice-to-String word-to-word Model 1 alignment, Syntactic features ...

Large Language Model Rescoring

Stupid backoff zero cut-off 5-gram language model [1]

Directly build sentence-specific LMs:

- Counts are extracted beforehand from all monolingual English data
- 5-grams are extracted from first-pass lattices
- All observed n-grams are kept and backoff weight α is fixed for all orders:

$$S(s_i|s_{i-n+1}^{i-1}) = \begin{cases} \frac{\#(s_{i-n+1}^{i})}{\#(s_{i-n+1}^{i-1})} & \text{if } \#(s_{i-n+1}^{i}) > 0 \\ \alpha\, S(s_i|s_{i-n+2}^{i-1}) & \text{otherwise} \end{cases}$$

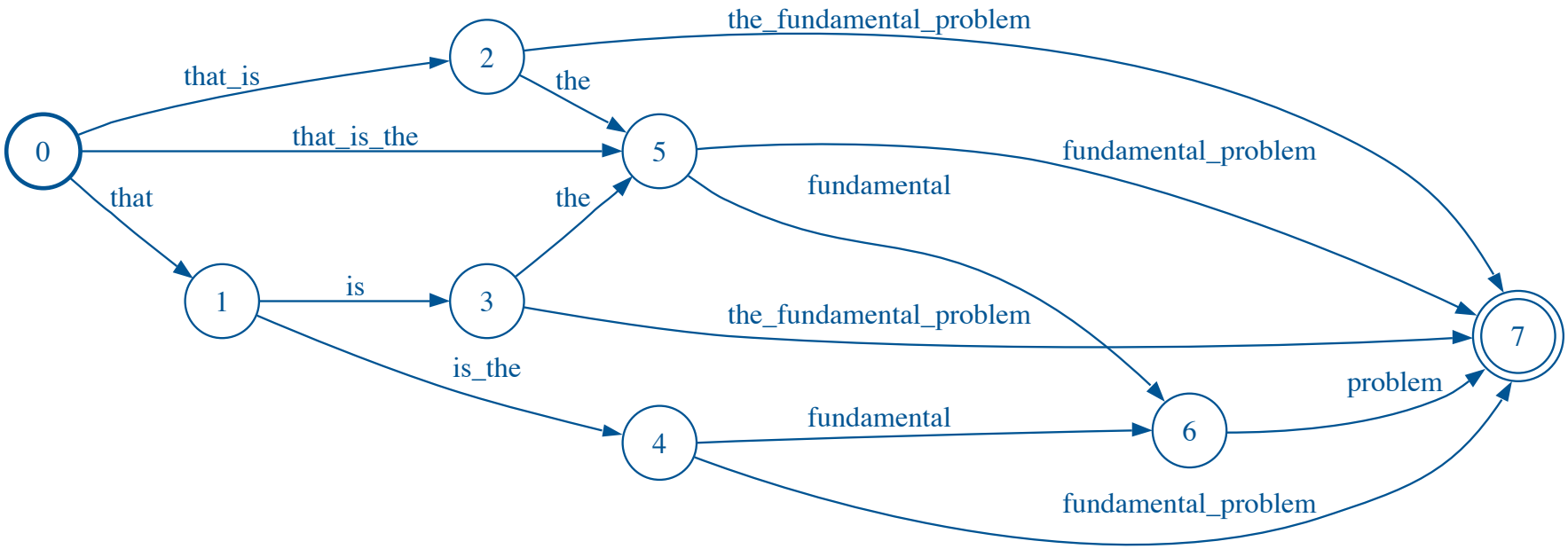
- equal weight interpolation with first-pass 4-gram LM
- exact search with OpenFST libraries in second-pass lattice rescoring

[1] T. Brants et al. 2007. Large Language Models in Machine Translation. EMNLP

Phrasal Segmentation Model Rescoring

Assign probability to sequences of English phrases

- phrases are translatable word sequences
- complement word-based N-grams



Phrase segmentation transducer can assign 'bigram' probabilities to phrases:

$$P(u_1^K|s_1^I) = \prod_k P(u_k \mid u_{k-1}\,,\, s_1^K)$$

Minimum Bayes Risk Decoding [3]

Taking the goal as BLEU maximization

- A baseline translation model to give the probabilities over translations: $P(\mathbf{S}|\mathbf{T})$
- A set $\mathcal{N}$ of N-Best Translations of T
- A Loss function $L(\mathbf{S}', \mathbf{S})$ that measures the the quality of $\mathbf{S}'$ relative to $\mathbf{S}$

MBR Decoder

$$\widehat{\mathbf{S}} = \underset{\mathbf{S} \in \mathcal{N}}{\operatorname{argmin}} \sum_{\mathbf{S}' \in \mathcal{N}} -\text{BLEU}(\mathbf{S}', \mathbf{S}) P(\mathbf{S}'|\mathbf{T})$$

$\widehat{\mathbf{S}}$ is sometimes called the 'consensus hypothesis'

- picks from the middle of the similar, relatively likely translation hypotheses
- typically over an N-Best list, but also lattices [2]

Rationale is to balance estimation criteria (e.g. MLE) with translation criteria (e.g. BLEU)

[2] Tromble, R. et al. 2008. Lattice Minimum Bayes-Risk Decoding for Statistical Machine Translation. EMNLP.

[3] S. Kumar W. Byrne. 2004. Minimum Bayes-risk decoding for statistical machine translation. HLT-NAACL

Rescoring Results. Task

NIST 2008 Arabic-English MT task [4]

- development set (*mt02-05-tune*): odd-numbered sentences of the NIST MT02 through MT05 evaluation sets
- validation set (*mt02-05-test*): even-numbered sentences of the NIST MT02 through MT05 evaluation sets
- test sets: MT06 (with newswire and newsgroup subsets) and MT08

TTM system trained on all available Arabic-English data for NIST MT08

- $\sim$ 6M sentences, $\sim$ 150M words
- word-aligned using MTTK

English Language Model data includes:

- first-pass 4-gram: parallel corpus + subset from English GigaWord Third Edition ($\sim$ 965M words)
- second-pass zero-cutoff 5-gram: $\sim$ 4.7B words (newswire text)
- phrasal segmentation model: $\sim$ 1.8B word subset of the above text

[4] nist.gov/speech/tests/mt/2006 nist.gov/speech/tests/mt/2008

Rescoring Results. TTM

TTM. Phrase-based system implemented with WFST:

- Reordering model is MJ1 (maximum one phrase swap)

Lowercase BLEU/TER scores over five test sets from 2002 through 2008:

Method	mt02-05-tune	mt02-05-test	mt06-nist-nw	mt06-nist-ng	mt08-nist
TTM+MET	50.9 / 42.8	50.3 / 43.3	48.1 / 44.3	37.5 / 53.5	43.1 / 49.5
+5g	53.5 / 41.8	52.4 / 42.4	49.6 / 43.9	39.0 / 54.0	43.7 / 49.3
+PSM	53.9 / 42.1	53.3 / 42.7	50.1 / 44.3	39.0 / 54.7	44.3 / 49.3
+MBR	54.0 / 41.7	53.7 / 42.2	51.0 / 43.9	39.4 / 54.1	45.0 / 48.9

- Important gains from lattice rescoring (improved fluency)
- Phrasal segmentation model complements 5-gram rescoring with further (yet smaller) gains
- Minimum Bayes Risk on the 1000-best list produces consistent gains
- Ranks among the group of top single-system entries in NIST 2008 official results

Rescoring Results. HiFST

HiFST. Hierarchical Phrase-based system implemented with WFST:

- Reordering model is explicit in rules
- CYK parsing

Method	*mt02-05-tune*		*mt02-05-test*		*mt08*	
	BLEU	TER	BLEU	TER	BLEU	TER
HiFST+MET	52.2	41.5	51.6	42.1	42.4	48.7
+5gram	53.3	40.6	52.7	41.3	43.7	48.1
+MBR	53.7	40.4	53.3	40.9	44.0	48.0

For **Chinese-to-English**:

	tune-nw		*test-nw*		*mt08*	
	BLEU	TER	BLEU	TER	BLEU	TER
HiFST+MET	32.0	60.1	32.2	60.0	27.1	60.5
+5gram	32.7	58.3	33.1	58.4	28.1	59.1
+MBR	32.9	58.4	33.4	58.5	28.9	58.9

- Very important contribution of large monolingual data
- MBR decoding always helps

System Combination

Given a set of alternative MT outputs

Different SMT approaches:

- Phrase-based systems
- Hierarchical phrase-based systems
- Syntax-based systems
- ...

Different configurations:

- Features set
- Search parameters, feature weights
- ...

Different data:

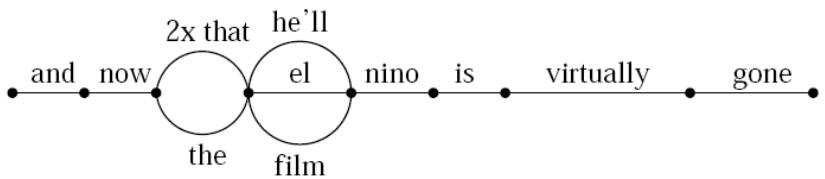
- Corpora, subcorpora weights
- Pre-processing, Morphological segmentations
- ...

ROVER for Combining Speech Recognisers

ROVER [5] is a method of system combination based on consensus networks:

- Originally devised for combining ASR output

1. Take one output hypothesis as reference
2. Time-align all other hyps to the reference to build confusion network / word graph
3. Re-decode using consensus scores and language models

reference:	...	And now that El Nino is virtually gone ...
recognizer 1 (1 error):	...	and now that he'll nino is virtually gone ...
recognizer 2 (no error):	...	and now that el nino is virtually gone ...
recognizer 3 (2 errors):	...	and now the film nino is virtually gone ...
word transition network:		
ROVER arbitrary ties (1 error) :	...	and now that he'll nino is virtually gone ...
ROVER with LM (no errors) :	...	and now that el nino is virtually gone ...

From Schwenk,H. and Gauvain, J-L. ICSLP 2000

[5] Fiscus, J.G. 1997. A post-processing system to yield reduced error word rates: Recognizer output voting error reduction (ROVER). Proc. ASRU

ROVER for MT System Combination

ROVER can be applied successfully to the translation task [6]:

- Select Reference Hypothesis: least cost, MBR hypothesis, ...
- Align other hypothesis to reference ← allow shifts?
- Re-Decode including extra features

System/ Combination	2003		2004	
	TER	BLEU	TER	BLEU
BBN Phrase	41.56	53.32	41.71	45.40
BBN Hiero	42.36	52.03	44.26	42.67
Edinburgh	42.05	52.5	44.20	**47.76**
ISI Hiero	**40.53**	**54.54**	**42.21**	46.49
ISI Phrase	41.94	52.35	43.09	45.21
ISI Syntax	42.96	52.36	45.00	44.11
MBR-BLEU	39.71	56.16	41.29	48.37
Confusion	39.37	55.67	41.21	46.45
ConMBR-BLEU	**39.02**	**56.64**	**40.23**	**48.93**

- Alignment strategy is crucial for good performance [7]

[6] Sim, C-K. et al. 2007. Consensus network decoding for statistical machine translation system combination. In ICASSP.

[7] Rosti, A-V. et al. 2007. Combining Outputs from Multiple Machine Translation Systems. Proc. NAACL.

MBR-based System Combination

Minimum Bayes Risk decoding over N-best list:

- A baseline translation model to give the probabilities over translations: $P(\mathbf{S}|\mathbf{T})$
- A set $\mathcal{N}$ of N-Best Translations of T
- A Loss function $L(\mathbf{S}', \mathbf{S})$ that measures the the quality of $\mathbf{S}'$ relative to $\mathbf{S}$

MBR Decoder

$$\widehat{\mathbf{S}} = \underset{\mathbf{S} \in \mathcal{N}}{\operatorname{argmin}} \sum_{\mathbf{S}' \in \mathcal{N}} -\text{BLEU}(\mathbf{S}', \mathbf{S}) P(\mathbf{S}'|\mathbf{T})$$

Steps:

- Take N-best hypotheses **from each system**
- Apply scaling factor to probs. of each system $\rightarrow$ **tune scale factor on dev data**
- Decode using MBR as usual

MBR-based System Combination. Results (1)

Alternative translation systems

Phrase-based vs Hierarchical Phrase-based

Chinese-to-English large-data task: $\sim$ 11M sentences

newswire system		Tune.text IBLEU	Tune.text TER	SysCombTune.text IBLEU	SysCombTune.text TER	Test.text IBLEU	Test.text TER
TTM	MET	30.0	61.07	30.6	60.26	20.5	63.77
	+5grams	30.7	59.92	31.4	59.58	20.7	63.25
	+PSM (A)	30.8	60.80	31.6	60.18	21.1	63.59
	+MBR	31.0	60.80	31.6	60.36	21.4	63.37
HiFST	MET	32.1	60.37	32.6	59.72	22.3	63.21
	+6grams (B)	33.1	58.61	34.0	57.69	22.5	61.83
	+MBR	33.3	58.70	34.1	58.03	22.9	61.65
MBRCOMB	A+B	**33.5**	58.61	**34.3**	57.96	**23.2**	61.50

MBR-based System Combination. Results (2)

Identical translation system

Alternative data segmentations

Arabic-to-English HiFST system trained on multiple data segmentations [8]

Arabic	wqrrt An tn$A ljnp tHDyryp jAmEp lljmEyp AlEAmp fY dwrthA AlvAnyp wAlxmsyn
MADA D2	w+ qrrt >n tn$A ljnp tHDyryp jAmEp l+ AljmEyp AlEAmp fy dwrthA AlvAnyp w+ Alxmsyn
SAKHR	w+ qrrt An tn$A ljnp tHDyryp jAmEp l*l+ jmEyp Al+ EAmp fY dwrt +hA Al+ vAnyp w*Al+ xmsyn
English	a preparatory committee of the whole of the general assembly is to be established at its fifty-second session

Large gains:

	mt02-05- -tune	-test	mt08
MADA-based	53.3	52.7	43.7
+MBR	53.7	53.3	44.0
SAKHR-based	52.7	52.8	43.3
+MBR	53.2	53.2	43.8
MBR-combined	**54.6**	**54.6**	**45.6**

[8] de Gispert, A. et al. 2009. Minimum Bayes Risk Combination of Translation Hypotheses from Alternative Morphological Decompositions. Proc. of HLT-NAACL: short papers.

MBR-based System Combination. Results (3)

Identical translation system

Alternative data segmentations

Finnish-to-English HiFST system trained on multiple data segmentations [9]

Finnish	vaarallisten aineiden kuljetusten turvallisuusneuvonantaja
Morfessor Linguistic	$\text{vaara}_{\text{STM}}$ $\text{llisten}_{\text{STM}}$ aine_{STM} iden_{SUF} $\text{kuljetus}_{\text{PRE}}$ ten_{STM} $\text{turvallisuus}_{\text{PRE}}$ $\text{neuvo}_{\text{STM}}$ n_{SUF} $\text{antaja}_{\text{STM}}$ vaara llis t en aine i den kuljet us t en turva llis uus neuvo n anta ja
English	safety adviser for the transport of dangerous goods

Large gains:

	devel	test
Word-based	30.2	27.9
Morph-based	29.4	27.4
MBR-combined	**30.5**	**28.9**

[9] de Gispert, A. et al. 2009. Minimum Bayes Risk Combination of Translation Hypotheses from Alternative Morphological Decompositions. Proc. of HLT-NAACL: short papers.

Lattice Minimum Bayes-Risk Decoding

- Lattice-based MBR uses a linear approximation to BLEU for efficiency
- Linearized lattice MBR[10] maximizes conditional expected gain:

$$\hat{E} = \underset{E' \in \mathcal{E}}{\operatorname{argmax}} \left\{ \theta_0 |E'| + \sum_{u \in \mathcal{N}} \theta_u \#_u(E') p(u|\mathcal{E}) \right\} \quad (1)$$

- $p(u|\mathcal{E})$ is "path posterior probability" of n-gram u

$$p(u|\mathcal{E}) = \sum_{E \in \mathcal{E}_u} P(E|F) \quad (2)$$

- Efficient and exact implementation of Equation (1) can be achieved with WFST operations[11]
- LMBR outperforms N-best MBR, both for single system and system combination
- due to considering more options when carrying out consensus decisions (better n-gram posterior estimates)

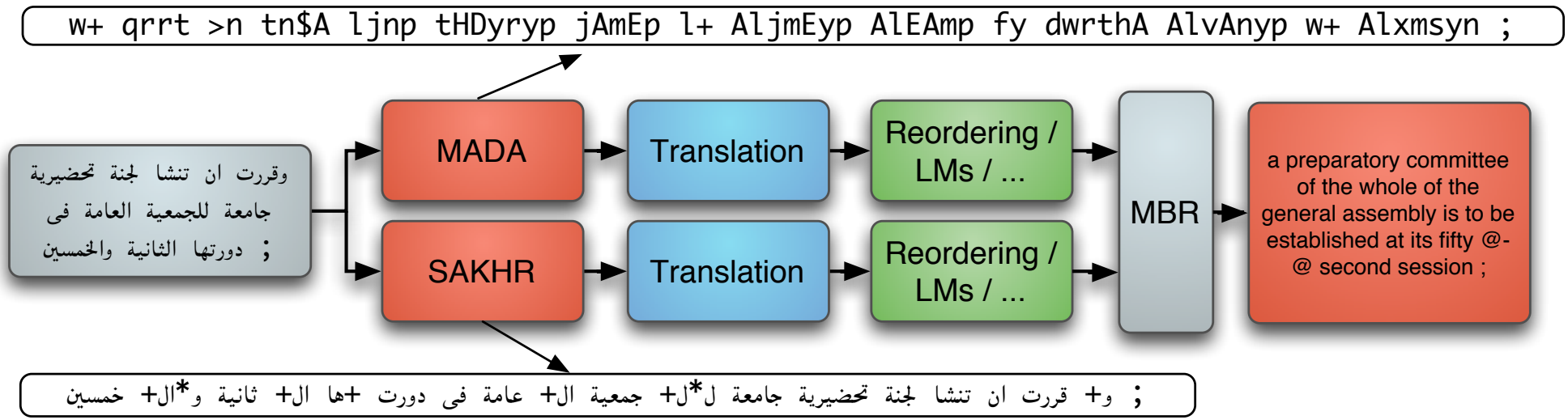
[10] Roy Tromble, Shankar Kumar, Franz Och, and Wolfgang Macherey. *Lattice Minimum Bayes-Risk decoding for statistical machine translation*. EMNLP 2008.

[11] Graeme Blackwood, Adrià de Gispert and Bill Byrne. *Efficient Path Counting Transducers for Minimum Bayes-Risk decoding of Statistical Machine Translation Lattices*. COLING 2010.

Morphological Analysis in Arabic-English MT

Multiple analysis schemes are available. Early or late fusion ?

- Early fusion: combine analyses into a lattice and perform lattice translation[12]
- Late fusion: hypothesis (lattice) combination at the output



[12] C. Dyer, S. Muresan, and P. Resnik. Generalizing Word Lattice Translation. In Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL), Columbus, Ohio, July 2008.

MBR – Easy Integration of Multiple Morphological Analyses

Arabic-to-English results when using alternative Arabic decompositions[13],[14]

- Contrast N-best-based and lattice-based Minimum Bayes Risk rescoring
- Nice gains from LMBR
  - Further advantages in direct generation of lattices
  - Very robust for noisy text – see NIST MT09 Web results

configuration		*mt02-05-tune*	*mt02-05-test*	*mt08*
a	HiFST+5g	54.2	53.8	44.9
b	HiFST+5g	53.8	53.6	45.0
c	HiFST+5g	54.1	53.8	44.7
a+b	+MBR (N=1000)	55.1	54.7	46.1
	+LMBR	55.7	55.4	46.7
a+c	+MBR (N=500x2)	55.4	54.9	46.5
	+LMBR	56.0	55.9	46.9
a+b+c	+MBR (N=$\frac{1000}{3}$x3)	55.3	54.9	46.5
	+LMBR	56.0	55.7	47.3

[13] A. de Gispert, S. Virpioja, M. Kurimo, and W. Byrne. Minimum Bayes risk combination of translation hypotheses from alternative morphological decompositions. Procedings of NAACL-HLT, 2009.

[14] M. Kurimo, S. Virpioja, V. T. Turunen, G. W. Blackwood, W. Byrne. Overview and results of Morpho Challenge 2009. 10th Workshop of the Cross-Language Evaluation Forum - CLEF 2009