

# Edinburgh System Description for 2005 IWSLT Speech Translation Evaluation

Philipp Koehn, Amittai Axelrod, Alexandra Birch Mayne,  
Chris Callison-Burch, Miles Osborne, David Talbot

School of Informatics  
University of Edinburgh



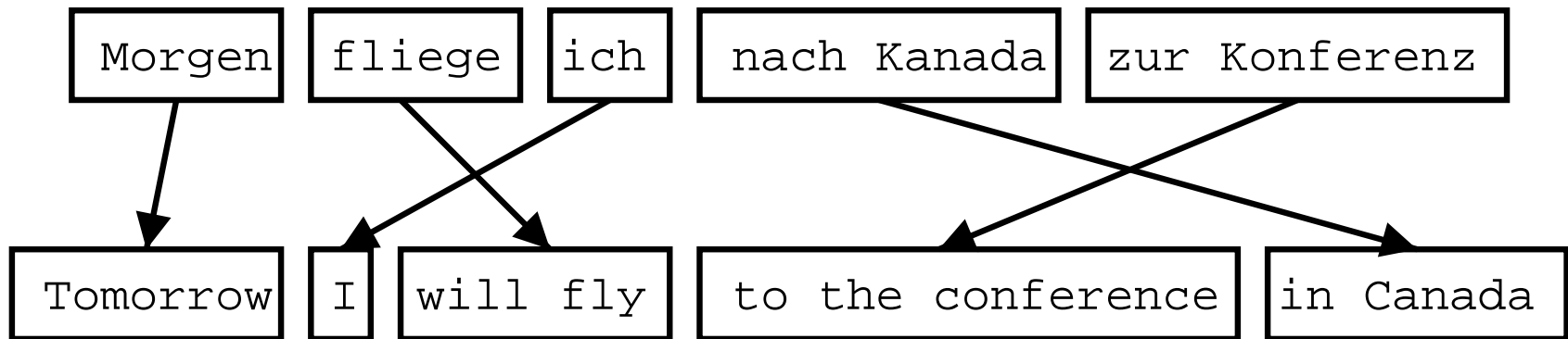
# Outline

- SMT at the University of Edinburgh
- Adaptations to the IWSLT task
  - optimizing word alignment
  - optimizing lexicalized reordering
  - optimizing reordering limit
- Results

# SMT at the University of Edinburgh

- Currently involved: 2 faculty, 6 graduate students
- Systems for
  - DARPA/NIST (GALE): Arabic-English, Chinese-English
  - Euromatrix: European Union languages
  - IWSLT: first attempt at speech domain
- Available software
  - Pharaoh (incl. training code)
  - New open source decoder (with Linear-B)

## Phrase-Based Translation



- Foreign input is segmented in phrases
  - any sequence of words, not necessarily linguistically motivated
- Each phrase is translated into English
- Phrases are reordered

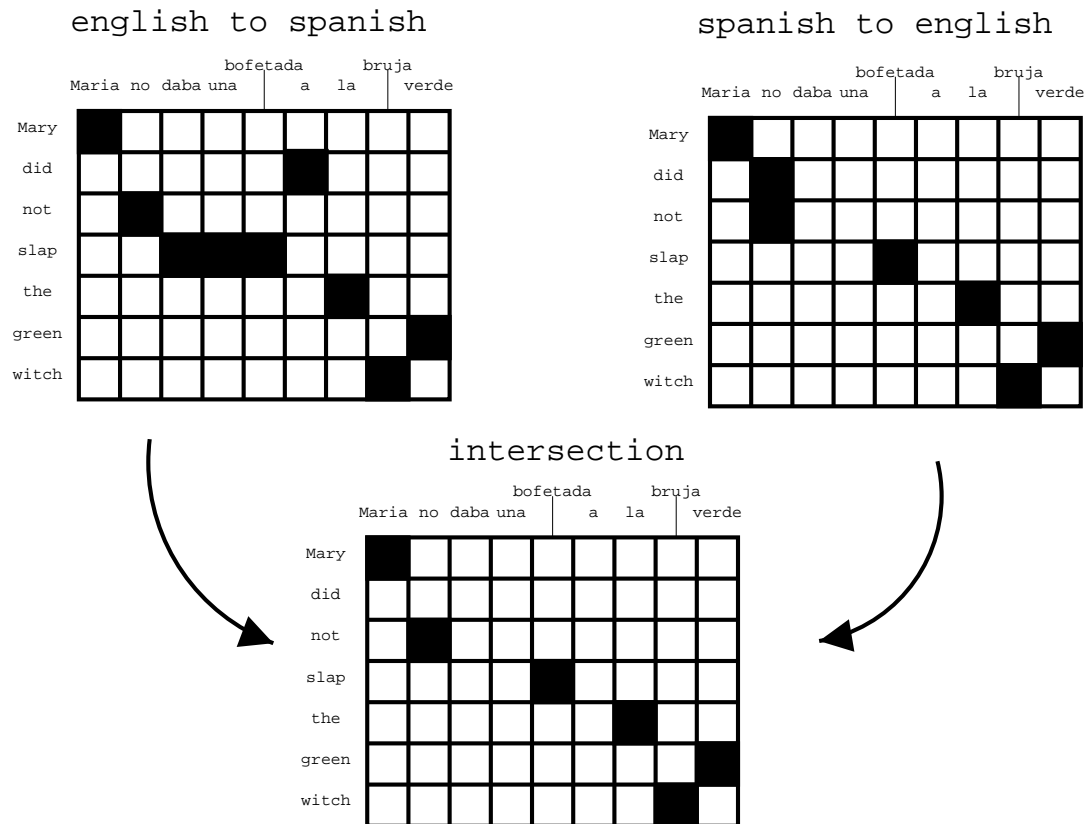
## Log-Linear Model

- Several model components combined in log-linear model

$$\begin{aligned}\hat{\mathbf{e}} &= \arg \max_e p(\mathbf{e}|\mathbf{f}) \\ &= \arg \max_e \sum_{m=1}^M \lambda_m h_m(\mathbf{e}, \mathbf{f})\end{aligned}$$

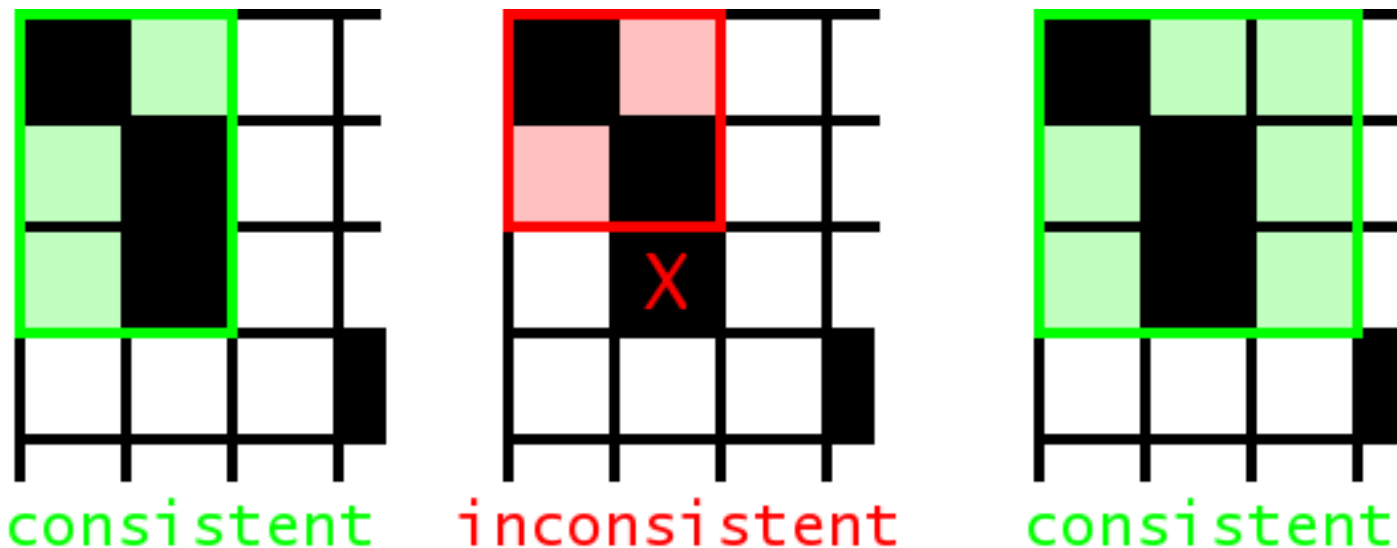
- Tuning with minimum error rate training [Och, 2003]

# Word Alignment based on IBM Models



- Intersection of GIZA++ bidirectional alignments
- Growing additional alignment points

# Extract Phrase Pairs



- Consistent with the word alignment :=  
phrase alignment has to contain all alignment points for all covered words

$$(\bar{e}, \bar{f}) \in BP \Leftrightarrow \begin{array}{l} \forall e_i \in \bar{e} : (e_i, f_j) \in A \rightarrow f_j \in \bar{f} \\ \text{AND } \forall f_j \in \bar{f} : (e_i, f_j) \in A \rightarrow e_i \in \bar{e} \end{array}$$

## Probability Distribution over Phrase Pairs

- We need a probability distribution  $\phi(\bar{f}|\bar{e})$  over the collected phrase pairs

⇒ Different scoring methods

- relative frequency of collected phrases:  $\phi(\bar{f}|\bar{e}) = \frac{\text{count}(\bar{f},\bar{e})}{\sum_{\bar{f}} \text{count}(\bar{f},\bar{e})}$
- conversely  $\phi(\bar{e}|\bar{f})$
- use lexical translation probabilities, also both directions
- word penalty
- phrase penalty

## IWSLT Task vs. DARPA/NIST Task

- Participated in supplied track for all language pairs
- Much less training data
  - 20,000 sentences vs. millions
- Different text type
  - shorter sentences
  - questions and answers
  - travel domain vs. news
- Faster Training
  - 15 minutes vs. many days

# Adaptations for IWSLT Task

- Optimizing word alignment
  - small corpus → high alignment error rate
  - high precision alignments may be better than high recall
- Optimizing lexicalized reordering
  - new model similar to block-orientation model [Tillmann, 2004]
  - tuning of model variants
- Optimizing reordering limit
  - so far, used reordering limit of maximum movement of 4 words
  - better reordering model may allow more reordering

# Optimizing Word Alignment

- Heuristics with increasing number of alignment points:
  - intersection of bidirectional GIZA++ alignments (intersection)
  - add union alignment points, if directly (block) neighboring (grow)
  - above + also allow diagonal neighborhood (grow-diag)
  - above + also add points that connect two unaligned words (final-and)
  - above + also add points that connect one unaligned words (final)
- Effect on phrase table size (Japanese–English):

	final	final-and	grow-diag	grow	intersect
English words	187,843	187,843	187,843	187,843	187,843
Alignment points	282,110	234,027	220,318	185,714	79,200
Distinct phrase pairs	61,168	270,654	447,550	854,680	2,561,715

# Optimizing Word Alignment

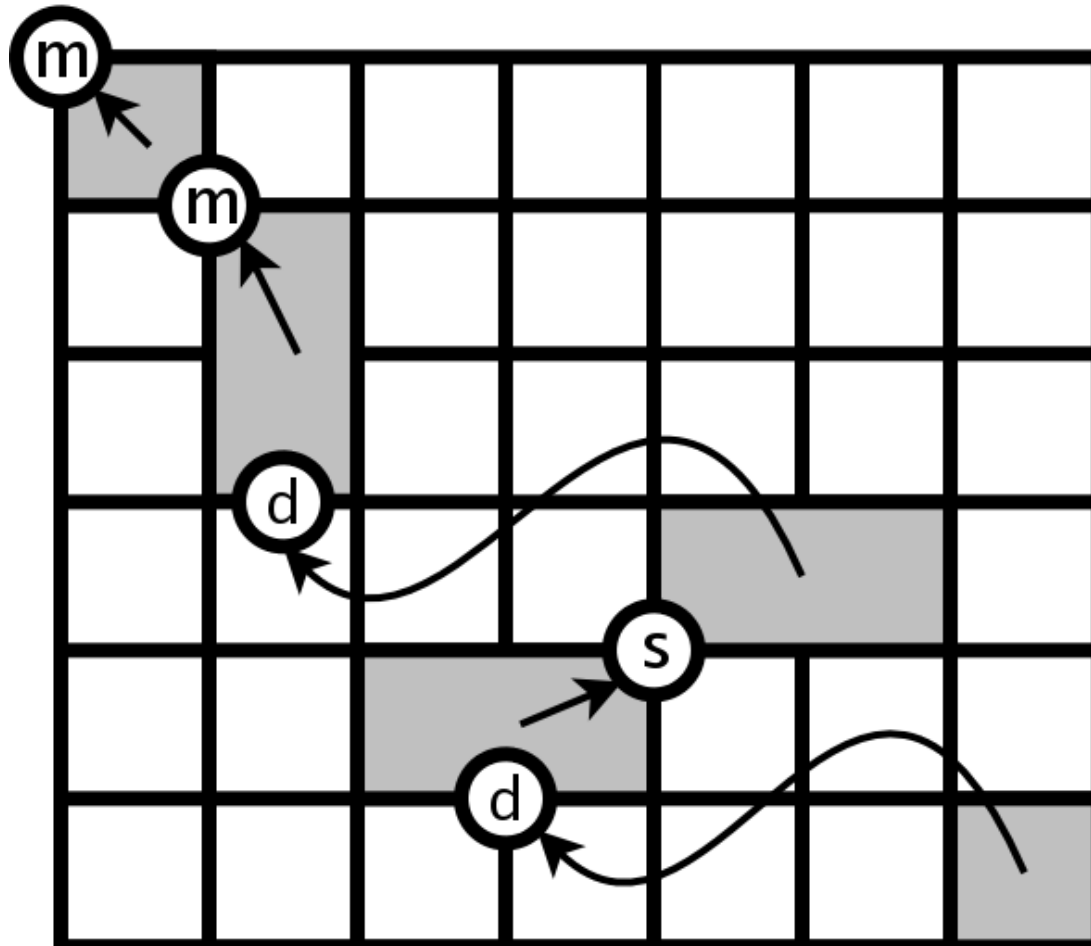
- Fewer alignment points, larger phrase table often better
  - intersection: +5% BLEU for Japanese, Chinese

Language Pair	final (default)	final-and	grow-diag	grow	intersect
Arabic-English	48.8	48.5	<b>49.9</b>	39.9	47.5
Japanese-English	40.4	39.9	39.0	39.1	<b>45.1</b>
Korean-English	33.9	<b>35.7</b>	27.7	13.5	35.4
Chinese-English	28.9	32.4	31.7	32.8	<b>34.6</b>
English-Chinese	<b>15.4</b>	9.6	8.1	<b>15.4</b>	15.2

## New Lexicalized Reordering Model

- Traditionally, only a distance-based reordering model
- Lexicalized reordering: conditioned on translated phrase
- Orientation types: monotone, swap, discontinuous
- Extension of work at IBM [Tillmann, 2004]

# Orientation Types



monotone (m), swap (s), discontinuous (d)

## Lexicalized Reordering Model Variants

- Distinguish between monotone, swap, and discontinuous  
... or just check monotonicity?
- Condition on foreign phrase  
... or on both foreign and English phrase?
- Model reordering in respect to previously translated phrase  
... or also in respect to following phrase?

## Different Methods for Language Pairs

- Optimized alignment heuristic and lexicalized reordering model variant
  - baseline: only distance-based reordering model
  - often intersection gives best results
  - mixed picture for lexicalized reordering model

Language Pair	Reordering	Word Alignment	Baseline	Improved
Arabic-English	orientation-bi-fe	final-and	49.9	50.9
Japanese-English	orientation-fe	intersect	45.1	47.6
Korean-English	orientation-fe	intersect	35.7	42.3
Chinese-English	monotonicity-fe	intersect	34.6	38.6
English-Chinese	monotonicity-bi-fe	grow-diag	15.2	16.6

# Optimizing Reordering Limit

- Traditionally, we used a reordering limit 4 words
  - no reordering worse and faster
  - unlimited reordering worse and slower
- Lexicalized reordering model allows more reordering

Reordering Limit	3	4	5	6	7	8
Arabic-English	50.3	50.4	50.1	<b>50.6</b>	50.0	50.1
Japanese-English	46.4	48.3	48.8	49.1	49.0	<b>49.9</b>
Korean-English	37.8	41.8	42.0	44.1	44.1	<b>45.2</b>
Chinese-English	36.8	36.7	37.2	<b>37.5</b>	36.9	37.2
English-Chinese	16.6	16.8	16.0	16.4	<b>17.2</b>	17.1

## Also Tried, but no Success

- Optimizing GIZA++ parameters
  - no consistent gains
  - not given up yet — this needs to be revisited
- Manual reordering rules for Japanese
  - move verb in front of sentence
  - move markers in front of nouns
  - ... but we have no Japanese POS tagger, parser

# Results

Language Pair	BLEU	NIST	Rank
Arabic-English	0.5105 (0.93)	7.6382 (0.70)	5th of 8
Japanese-English	0.3778 (0.81)	4.0784 (0.41)	4th of 7
Korean-English	0.3672 (0.88)	5.6172 (0.60)	1st of 4
Chinese-English	0.4650 (0.90)	6.4922 (0.62)	3rd of 10
English-Chinese	0.2127 (0.94)	5.1807 (0.98)	1st of 2

- Comparably good results
- Why so low NIST scores?
  - tuned for shortest reference length
  - relatively short output (length penalty in parenthesis)

## Tuning for *Average Reference Length*

Language Pair	BLEU	NIST
Arabic-English	0.5180 (0.98)	9.7749 (0.94)
Japanese-English	0.3941 (0.95)	8.1209 (0.91)
Korean-English	0.3859 (1.00)	8.4455 (0.99)
Chinese-English	0.4364 (1.00)	9.0834 (0.99)
English-Chinese	0.2230 (0.91)	5.2391 (0.97)

- Much higher NIST scores
  - for Japanese-English: 4.07 → 8.12
- BLEU not changed much
  - 4 out of 5 higher, but not by much

## Conclusion

- Successful participation, competitive results
- System could be easily applied
- Lessons from adaptation
  - high precision word alignment heuristics better
  - lexicalized reordering model validated
  - longer reordering windows