Local Phrase Reordering Models for Statistical Machine Translation

Shankar Kumar* and Bill Byrne

Center for Language and Speech Processing, Johns Hopkins University
Machine Intelligence Laboratory, Cambridge University Engineering Department
* Now at Google

October 6, 2005

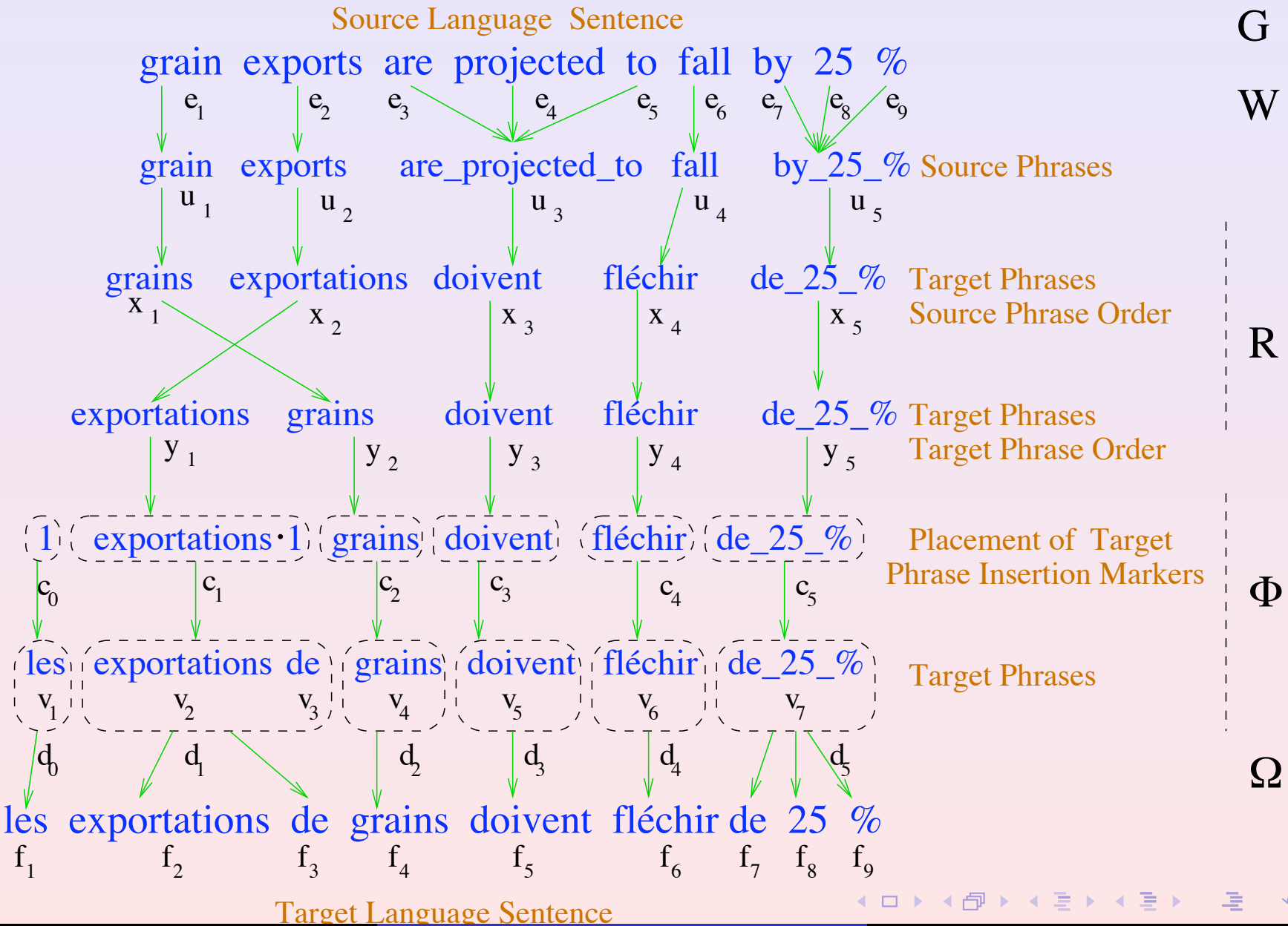
Introduction

- Word & Phrase Reordering is Important for Statistical MT
- Reordering in MT is expensive, sometimes NP-complete (Knight '99)
- Translation Scheme should balance
  - Complexity of the Reordering Model
  - Ability to realize the model without approximation.
- Goal Formulate models of local phrase movement embedded inside a generative, phrase-based translation model.
- We will present
  - Models of Local Phrase Reordering
  - Parameter Estimation with Expectation Maximization Techniques
  - Translation Experiments

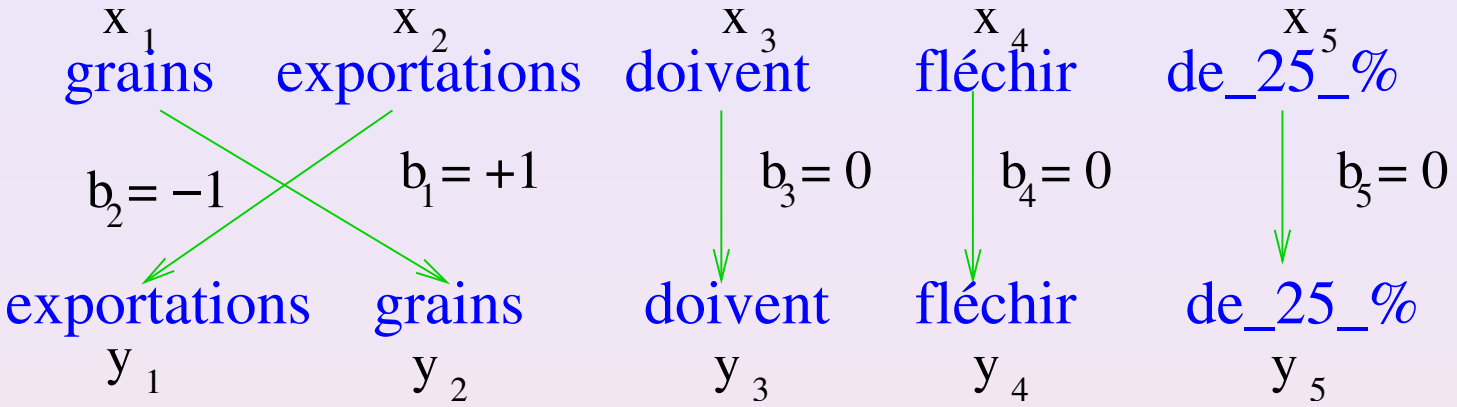
Translation Template Model

- Translation Template Model (TTM) (Kumar et. al. '05)
  - A generative, phrase-based translation model
    - Stochastic process generates English sentences which are then transformed into French.
    - TTM relies on a phrase-pair inventory
    - TTM component distributions realized as Weighted Finite State Transducers (WFSTs)
    - Phrase Alignment and Translation can be performed with standard WFST operations involving the component transducers
    - Monotone Phrase Order in Translation
- Local Phrase Reordering Models inside the generative process

Generative Process : Noisy Channel Model



Local Phrase Reordering Model



- Associate a jump sequence b_1^K with each sequence y_1^K
- $x_k \longrightarrow y_{k+b_k}$
  - $P(y_1^K|x_1^K, u_1^K) = P(b_1^K|x_1^K, u_1^K)$
- Jump sequence ensures that y_1^K is a permutation of x_1^K
  - Construct Non-deficient Models

Model Structure

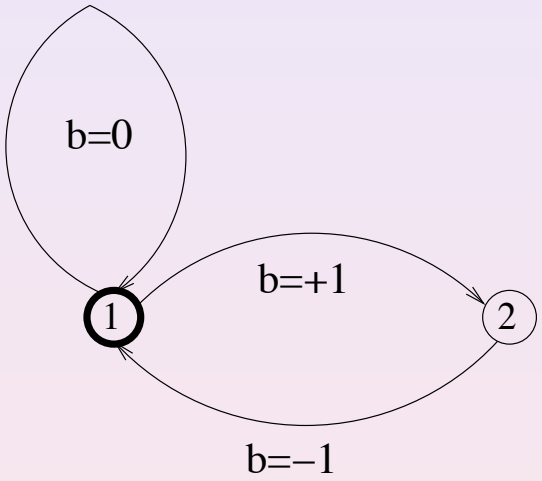
- Make Model parameters phrase-pair specific
  - Inspired by the block orientation model (Tillmann '04)
- b_1^K is a FSM

$$P(b_1^K | x_1^K, u_1^K) = \prod_{k=1}^{K} P(b_k | x_k, u_k, \underbrace{\phi(b_1^{K-1})}_{\text{FSM state } b_1^{K-1}})$$

- Investigate two models by restricting the maximum allowable jump
  - MJ-1: $b_k \in \{0, +1, -1\}$
  - MJ-2: $b_k \in \{0, +1, -1, +2, -2\}$

Phrase Reordering Process for MJ-1

Jump Sequence b_1^K respects the MJ-1 process



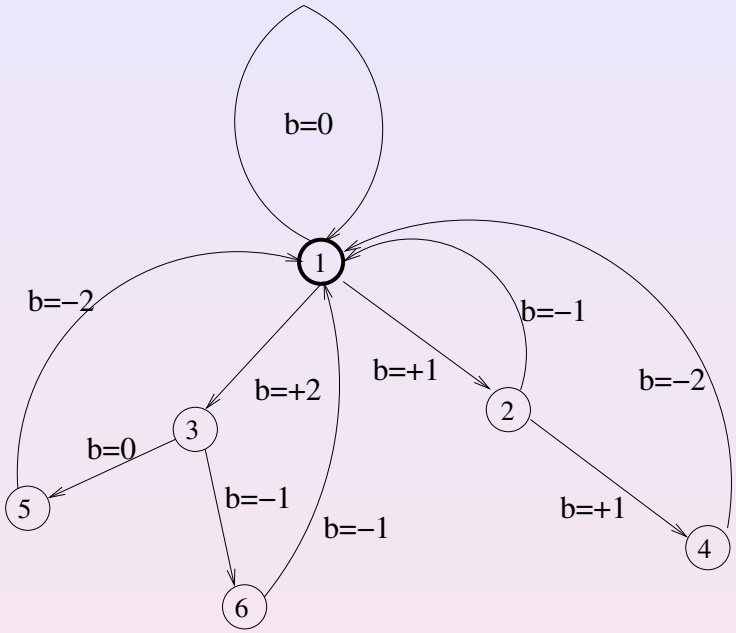
For FSM state $\phi(b_1^{K-1}) = 1$

$$P(b_k|x_k, u_k) = \begin{cases} \beta_1(x_k, u_k) & b_k = +1 \\ 1 - \beta_1(x_k, u_k) & b_k = 0 \end{cases}$$

Single Parameter $\beta_1(x, u)$ for each phrase-pair (x, u)

Phrase Reordering Process for MJ-2

Jump Sequence respects the MJ-2 process



For FSM State $\phi(b_1^{K-1}) = 1$

$$P(b_k|x_k, u_k) = \begin{cases} \beta_1(x_k, u_k) & b_k = +1 \\ \beta_2(x_k, u_k) & b_k = +2 \\ 1 - \beta_1(x_k, u_k) - \beta_2(x_k, u_k) & b_k = 0 \end{cases}$$

Two parameters $\beta_1(x, u)$ and $\beta_2(x, u)$ for each phrase-pair (x, u)

Reordering Model WFST

Translating from French $\rightarrow$ English

- Input : Lattice of French phrase sequences
- Output: English phrase sequences in English phrase order
- WFST R is constructed using reordering model & phrase translation probabilities (details in paper)

- Input Phrase Sequence
  exportations grains doivent flechir de_25_%
- Possible MJ-1 Output : b_1^5
  grain exports are_projected_to fall by_25_% : +1,-1,0,0,0
- Possible MJ-2 Output : b_1^5
  are_projected_to grain exports fall by_25_% : +2,0,-2,0,0

- Translation of a French sentence (F) is obtained by
  - $\mathcal{T} = G \circ W \circ R \circ \Phi \circ \Omega \circ F \rightarrow$ best-path in $\mathcal{T}$
- Phrase-Alignment of a Sentence pair (E, F) is obtained by
  - $\mathcal{B} = E \circ W \circ R \circ \Phi \circ \Omega \circ F \rightarrow$ best-path in $\mathcal{B}$

Estimation of Reordering Parameters $\beta_1(x, u)$

- Estimate parameters over the training bitext
  - Could use a large phrase-pair Inventory to cover all bitext
  - Instead, we restrict the inventory to the test set
- ML estimate
  - $C_{x,u}(\phi, b)$ = Expected # of times (x, u) is aligned with a jump of b phrases when the jump history is in State ϕ.
  - $\hat{\beta}_1(x, u) = \frac{C_{x,u}(0,+1)}{C_{x,u}(0,+1)+C_{x,u}(0,0)}$
- Viterbi procedure $\longrightarrow$ counts over the best alignments
- If a phrase-pair never seen in alignments, back-off : $\beta_1 = 0.05$.
- Training Scheme
  - Start with a flat model : $\beta_1(x, u) = 0.05$ for all phrase-pairs
  - WFST operations to obtain Viterbi phrase alignments of the bitext
  - Get Counts $C_{x,u}(\phi, b)$ over the phrase alignments
  - Obtain ML estimates of $\beta_1(x, u)$
- MJ-2 model is estimated via a similar procedure

Experiments on Medium-Sized Bitexts

- C-E: FBIS/11.7M Eng words/8.9M Chn words/674K chunk pairs
- A-E: LDC News/3.8M Eng words/3.2M Arb words/137K chunk pairs
- Performance measured using uncased BLEU score (Papineni et. al.)

Reordering	BLEU (%)	
	A-E	C-E
None	37.8	25.0
MJ-1 flat	40.7	26.2
MJ-1 VT	41.6	26.5
MJ-2 flat	41.1	26.7
MJ-2 VT	42.0	26.8

- Gains higher on A-E

- MJ-1 VT outperforms flat MJ-1: Useful to estimate reordering parameters from bitext
- flat MJ-2 is better than flat MJ-1: Larger search space allows LM to select a better hypothesis
- MJ-2 VT marginally better than MJ-1 VT

Reordering over Variable Span Language Models

Reordering	BLEU (%)					
	A-E			C-E		
	2g	3g	4g	2g	3g	4g
None	21.0	36.8	37.8	16.1	24.8	25.0
MJ-1 VT	23.4	40.4	41.6	16.2	25.9	26.5
MJ-2 VT	23.5	40.6	42.0	16.0	26.1	26.8

- Reordering gives bigger gains under higher order LMs
- AE: BLEU Score Improvements from 3g LM $\rightarrow$ 4g LM
  - 1.0 (No reordering), 1.2 (MJ-1), 1.4 (MJ-2)

Reordering across Test Set Genres (NIST '04)

News, Editorials, Speeches

Reordering	BLEU (%)					
	A-E			C-E		
	News	Eds	Spcs	News	Eds	Spcs
None	41.1	30.8	33.3	23.6	25.9	30.8
MJ-1 VT	45.6	32.6	35.7	24.8	27.8	33.3
MJ-2 VT	46.2	32.7	35.5	24.8	27.8	33.7

- A-E: Larger Improvements on News
- C-E: Larger Gains on Speeches/Editorials
- Hypothesis
  - Different degrees of movement across different genres

Phrase Reordering over Large Bitexts

- JHUCU NIST 2005 evaluation systems.
- MJ-1 models trained over all bitext (AE), non-UN bitext (CE)

	BLEU (%)			
	Arabic-English		Chinese-English	
Reordering	02	03	02	03
None	40.2	42.3	28.9	27.4
MJ-1 VT	43.1	45.0	30.2	28.2

- MJ-1 VT gives improvements relative to monotone phrase order

Conclusions

- Non-deficient phrase reordering inside a generative translation model
  - EM-style embedded estimation : parameters estimated from statistics of the complete model
  - Estimation of reordering parameters from phrase alignments is novel!
- Local Phrase Movement Models can be integrated inside WFST translation to perform phrase alignment and translation
  - No pruning or approximation in construction
  - No permutation acceptors!
- Improvements in Chinese-English and Arabic-English MT
- Reordering Model can be trained over large bitexts (in paper)
- TTM toolkit available for research : contact wjb31@cam.ac.uk

Thank you!