



Maximum Entropy Based Phrase Reordering Model for Statistical Machine Translation

Deyi Xiong, Qun Liu and Shouxun Lin

Multilingual Interaction Technology & Evaluation Lab

Institute of Computing Technology

Chinese Academy of Sciences

{dyxiong, liuqun, sxlin} at ict dot ac dot cn

Homepage: <http://mtgroup.ict.ac.cn/~devi/>



Outline

- Previous work
- Maximum entropy based phrase reordering
- System overview
- Experiments
- Conclusions

Previous Work

- Content-independent reordering models
 - E.g. distance-based or flat reordering models
 - Learn nothing for reordering from real-world bitexts
- Content-dependent reordering models
 - Lexicalized reordering models [Tillmann, 04; Och et. al. 04; ...]
 - Totally dependent on bilingual phrases
 - With a large number of parameters to estimate
 - Without generalization capabilities



Can We Build **Such** a Reordering Model?

- ❑ Content-dependent but not restricted by phrases
- ❑ Without introduction of too large number of parameters but still powerful
- ❑ With generalization capabilities



Build It Conditioned on **Features**, not **Phrases**

- Features can be:
 - Some special words or their classes in phrases
 - Syntactic properties of phrases
 - Surface attributes like distance of swapping
- **Advantages** of feature-based reordering:
 - Flexibility
 - Less parameters
 - Generalization capabilities

Feature-based Reordering

- Discriminative Reordering Model proposed by Zens & Ney in NAACL 2006 Workshop on SMT
 - We have become aware of this work when we prepare the talk
 - Very close to our work
 - But still different:
 - Implemented under the IBM constraints
 - Using different feature selection mechanism



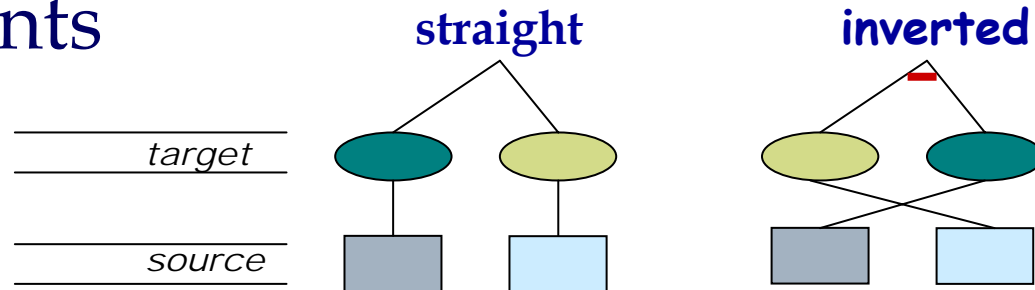
Outline

- Previous work
- Maximum entropy based phrase reordering
- System overview
- Experiments
- Conclusions

Reordering as Classification

- Regard reordering as a problem of classification
 - Two-class problem under the ITG constraints
 - {straight, inverted}
 - Multi-class problem if positions are considered under the IBM constraints

- Our work focused on the reordering under the ITG constraints



MaxEnt-based Reordering Model (MERM)

- ❑ A reordering framework for BTG

$$\Omega = f(o, A^1, A^2), o \in \{straight, inverted\}$$

- ❑ MERM under this framework

$$\Omega = p_{\theta}(o | A^1, A^2) = \frac{\exp(\sum_i \theta_i h_i(o, A^1, A^2))}{\sum_o \exp(\sum_i \theta_i h_i(o, A^1, A^2))}$$

- ❑ Feature function

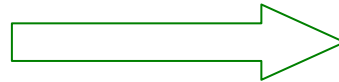
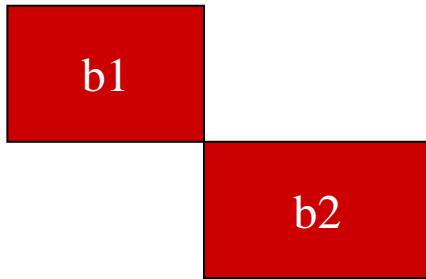
$$h_i(o, A^1, A^2) = \begin{cases} 1 & \text{if } f(A^1, A^2) = T, o = O \\ 0 & \text{otherwise} \end{cases}$$

$$O \in \{straight, inverted\}$$

Training for MERM

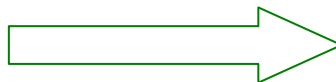
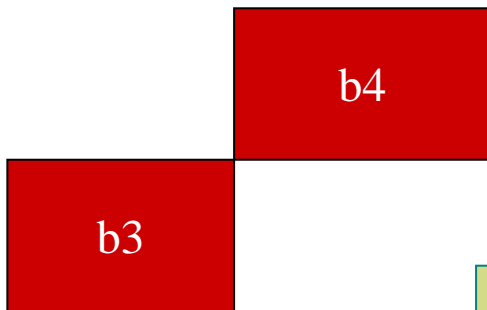
- Training procedures: 3 steps
 - Learning reordering examples
 - Generating features (h_i) from reordering examples
 - Parameter (θ_i) estimation using off-the-shelf MaxEnt toolkits

Reordering Example



$\langle b1; b2 \rangle \rightarrow \text{STRAIGHT}$

E.g. $\langle \text{今天有棒球比赛} | \text{Are there any baseball games today; 吗? } | ? \rangle \rightarrow \text{STRAIGHT}$



$\langle b3; b4 \rangle \rightarrow \text{INVERT}$

E.g. $\langle \text{澳门政府} | \text{the Macao government; 有关部门} | \text{related departments of} \rangle \rightarrow \text{INVERT}$

Features

- Lexical feature: source or target boundary words
- Collocation feature: combinations of boundary words

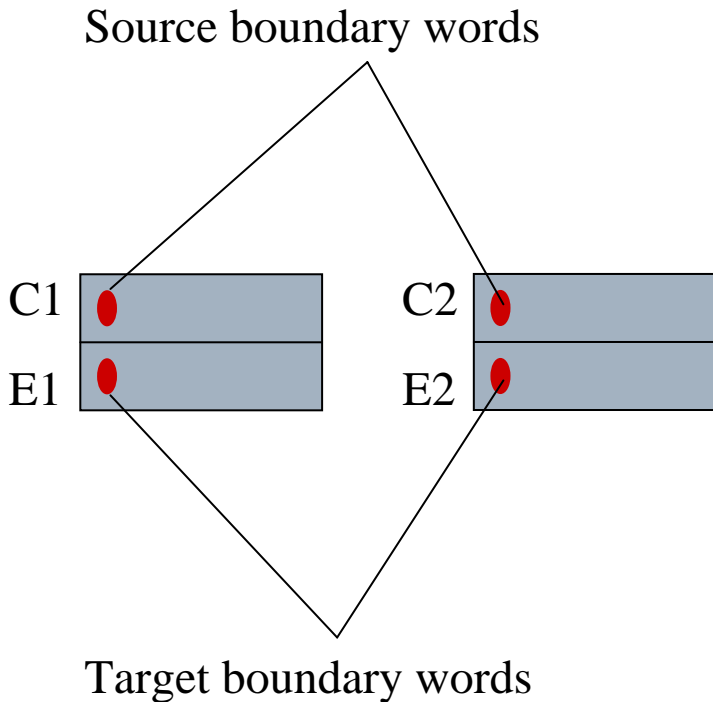
<与他们|with them; 保持联系|keep contact> → INVERT

Feature selection

$$h_{lex}(o, A^1, A^2) = \begin{cases} 1 & \text{if } A^2.t_1 = \textit{keep}, o = \textit{invert} \\ 0 & \text{otherwise} \end{cases}$$

$$h_{col}(o, A^1, A^2) = \begin{cases} 1 & \text{if } A^1.t_1 = \textit{with}, A^2.t_1 = \textit{keep}, o = \textit{invert} \\ 0 & \text{otherwise} \end{cases}$$

Why Do We Use Boundary Words as Features: Information Gain Ratio



feature	IGR
Phrases	.02655
C1C2E1E2	.0263687
E1E2	.0239286
C1C2	.023363
C2E2	.0192932
C1E1	.0153117
C2	.011371
E2	.00994372
E1	.00899752
C1	.00758598



Outline

- Previous work
- Maximum entropy based phrase reordering
- System overview (Bruin)
- Experiments
- Conclusions

Translation Model

- Built upon BTG

$$A \xrightarrow{\square} (A^1, A^2)$$

$$A \xrightarrow{\diamond} (A^1, A^2)$$

$$A \rightarrow (x, y)$$

- The whole model is built in the log-linear form
 - The score to apply lexical rules is calculated using features similar to many state-of-the-art systems
 - The score to apply merging rules is divided into two parts
 - The reordering model score
 - The increment of language model score

Different Reordering Models Are Embedded in the Whole Translation Model

- ❑ The reordering framework

$$\Omega = f(o, A^1, A^2), o \in \{straight, inverted\}$$

- ❑ MaxEnt-based reordering model

$$\Omega = p_{\theta}(o | A^1, A^2) = \frac{\exp(\sum_i \theta_i h_i(o, A^1, A^2))}{\sum_o \exp(\sum_i \theta_i h_i(o, A^1, A^2))}$$

- ❑ Distance-based reordering model

$$\Omega = \begin{cases} \exp(0) & o = straight \\ \exp(|A^1| + |A^2|) & o = inverted \end{cases}$$

- ❑ Flat reordering model

$$\Omega = \begin{cases} p_m & o = straight \\ 1 - p_m & o = inverted \end{cases}$$



CKY-style Decoder

- Core algorithm
 - Borrowed from CKY parsing algorithm
- Edge pruning
 - Histogram pruning
 - Thresholding pruning
- Language model incorporation
 - Record the leftmost & rightmost n words for each edge



Outline

- Previous work
- Maximum entropy based phrase reordering
- System overview
- Experiments**
- Conclusions

Experiment Design

- To test MERM against various reordering models, we carried out experiments on:
 - Bruin with MERM
 - Bruin with monotone search
 - Bruin with distance-based reordering model
 - Bruin with flat reordering model
 - Pharaoh, a distance-based state-of-the-art system (Koehn 2004)

Systems Settings

Systems	Phrase table	Phrase table pruning	Stack pruning	Reordering
Pharaoh	Same	$b = 100$ (default 20)	$n = 100$; $\beta = 10^{-5}$	Limited distortions to 4 (default 0)
Bruin	Same	$b = 100$	$n = 40$; $\beta = 0.5$	Monotone/flat /distance/ME RM

Small Scale Experiments

■ NIST MT 05

- Training data: FBIS (7.06M + 9.15M)
- Language model: 3-gram trained on 81M English words (most from UN corpus) using SRILM toolkit
- Development set: 580 sentences length of at most 50 Chinese characters from NIST MT 02

■ IWSLT 04

- Small data track
- 20k sentences for training of TM and LM
- 506 sentences as the development set

MERM Training

	Item	Num
NIST MT 05	Straight Reordering Examples	2.7M
	Inverted Reordering Examples	367K
	Lexical Features	112K
	Collocation Features	1.7M
IWSLT 04	Straight Reordering Examples	79.5k
	Inverted Reordering Examples	9.3k
	Lexical Features	16.9K
	Collocation Features	89.6K

Results on Small Scale Data

Systems	NIST MT 05	IWSLT 04
Bruin with monotone search	20.1	37.8
Bruin with distance-based reordering	20.9	38.8
Bruin with flat reordering	20.5	38.7
Pharaoh	20.8	38.9
Bruin with MERM (lex)	22.0	42.4
Bruin with MERM (lex+col)	22.2	42.8

Scaling to Large Bitexts

- Just used lexical features for MaxEnt reordering model
- Training data
 - 2.4M sentence pairs (68.1M Chinese words and 73.8M English words)
- Two 3-gram language models
 - One was trained on the English side
 - The other was trained on the Xinhua portion of the Gigaword corpus with 181.1M words
- Used simple rules to translate number, time expressions and Chinese person names
- BLEU score: 0.22 → 0.29

Results Comparison

他将参加世界领袖六日在印尼首都雅加达举行的会议



Bruin (MaxEnt): he will attend the meeting held in the Indonesian capital Jakarta on world leaders



Pharaoh: he will participate in the leader of the world on 6 Indonesian capital of Jakarta at the meeting



Bruin (Distortion): he will join the world 's leaders of the Indonesian capital of Jakarta meeting held on 6



Bruin (monotone): he will participate in the world leaders on 6 the Indonesian capital of Jakarta at the meeting of the

Ref: he will attend the meeting of world leaders to be held on the 6th in the Indonesian capital of Jakarta



Outline

- Previous work
- Maximum entropy based phrase reordering
- System overview
- Experiments
- Conclusions**

Comparisons

	distance/flat	lexicalized	MaxEnt
Content-dependent	No	Yes	Yes
Generalized	/	No	Yes
Parameter number	/	Large	Small
Parameter estimation	/	MLE	Discriminative

Conclusions

- MaxEnt-based reordering model is
 - Feature-based
 - Content-dependent
 - Capable of generalization
 - Trained discriminatively
 - Easy to be integrated into systems under the IBM constraints

Future Work

- More features
 - Syntactic features
 - Global features of the whole sentences
 - ...
- Other language pairs
 - English-Arabic
 - Chinese-Mongolian



Thank you!

Training

- Run GIZA++ in both directions
- Use grow-diag-final refinement rules
- Maximum phrase length: 7 words on the Chinese side
- Length ratio: $\max(|s|, |t|) / \min(|s|, |t|) \leq 3$

Language Model Incorporation (further)

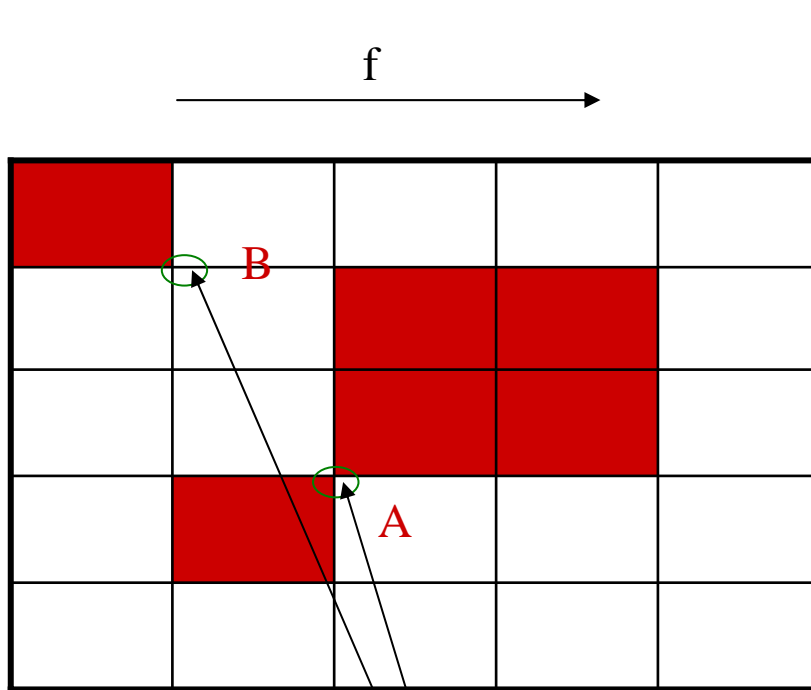
- The edge spans only part of the source sentence
 - The history of LM is not available
 - the language model score has to be approximated by computing the score for the generated target words alone
- Combination of two neighbor edges
 - only need to compute the increment of the LM score:

$$\Delta_{LM} = LM(s_1^r s_2^l) - LM(s_1^r) - LM(s_2^l)$$

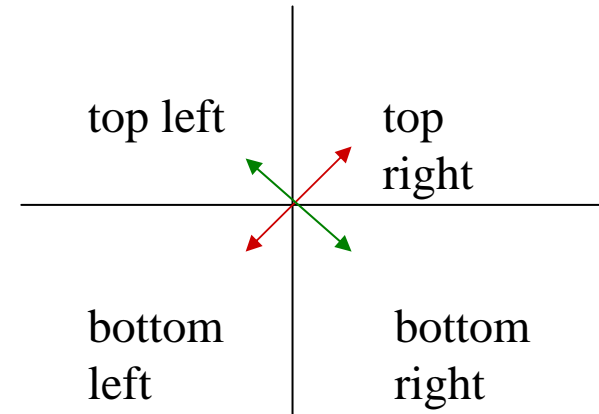
Tree

```
<tree>
(MONO (Ukraine|乌克兰)
      (MONO (MONO (MONO (because of|因)
                    (INVERT (INVERT (the chaos|的混乱)
                                   (caused by|引发))
                                   (the presidential election|总统选举)))
                    (in the|进入))
            (third week of|第三周)))
</tree>
```

Related Definitions



Every crossing point is a corner



STRAIGHT link and *INVERT link* can not co-exist



The Algorithm of Extracting Reordering Examples



- For each sentence pair do
 - Extract bilingual phrases
 - Update the links for the four corners of each extracted phrases
- For each corner do
 - If it has a STRAIGHT link with phrase a and b
 - Extract the pattern: $\langle a; b \rangle \rightarrow \text{STRAIGHT}$
 - If it has a INVERT link with phrase a and b
 - Extract the pattern: $\langle a; b \rangle \rightarrow \text{INVERT}$