NAACL HLT – Rochester, NY – April 2007

Chunk-level Reordering of Source Language Sentences with Automatically Learned Rules for Statistical Machine Translation

Yuqi Zhang, Richard Zens and Hermann Ney

Human Language Technology and Pattern Recognition
Lehrstuhl für Informatik 6
Computer Science Department
RWTH Aachen University, Germany

Overview

- Introduction
- Baseline system
- Chunk parsing
- Rules extraction
- Reordering lattice generation
- Results
- Conclusions and outlook

Introduction

goal:

improve MT utilizing syntactic knowledge

idea:

reordering at the chunk level

approach:

1. **chunk source sentence**
2. **reorder chunks**
3. **represent alternative reorderings in a lattice**
4. **translate lattice**

Phrase-based SMT

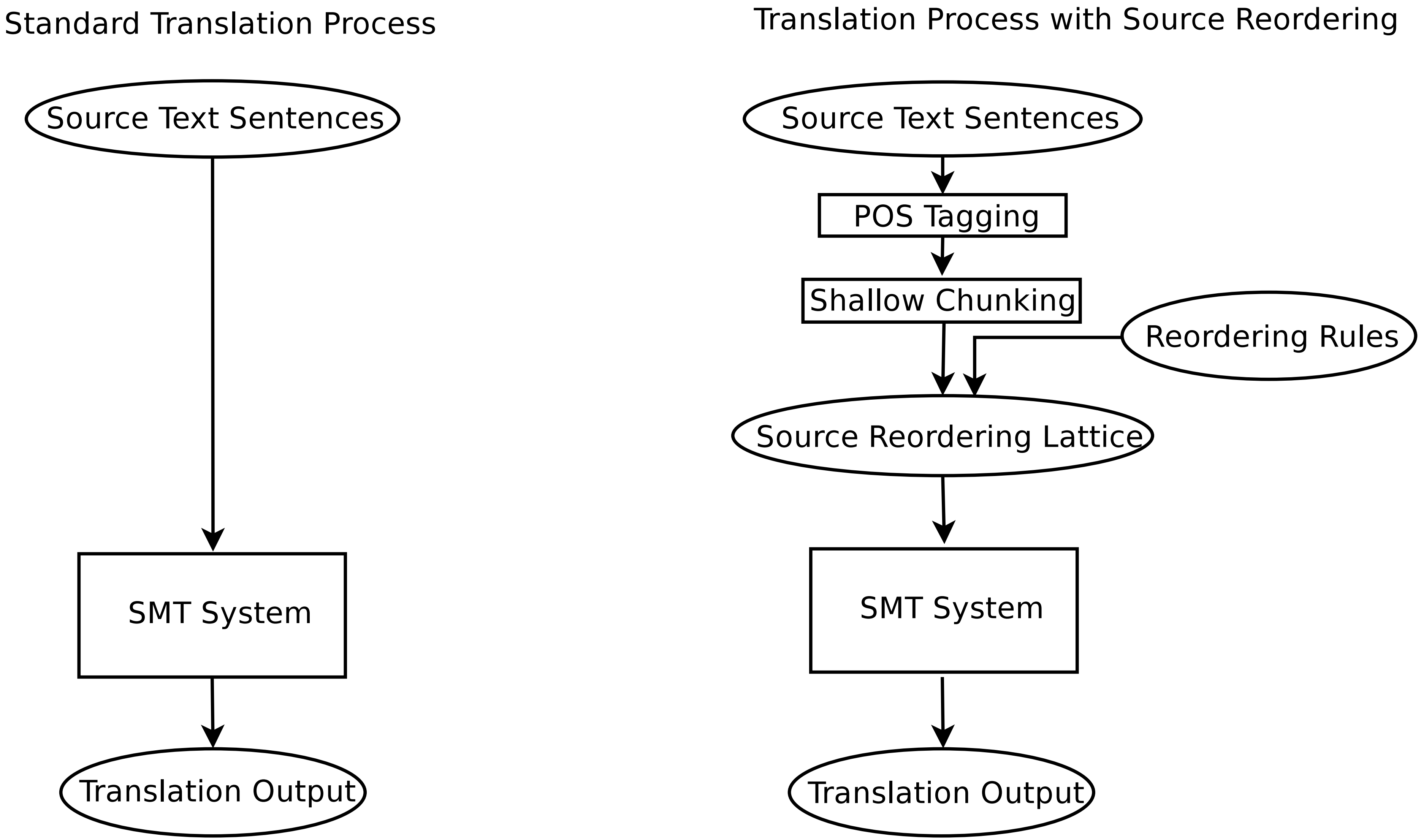
log-linear combination of several model:

$$Pr(e_1^I|f_1^J) = \frac{\exp\left(\sum_{m=1}^M \lambda_m h_m(e_1^I, f_1^J)\right)}{\sum\limits_{I', e'^{I'}_1} \exp\left(\sum_{m=1}^M \lambda_m h_m(e'^{I'}_1, f_1^J)\right)}$$

models:

- **phrase translation model**
- **phrase count features**
- **word-based translation model**
- **word and phrase penalty**
- **target language model (6-gram)**
- **distortion penalty model**

System Architecture



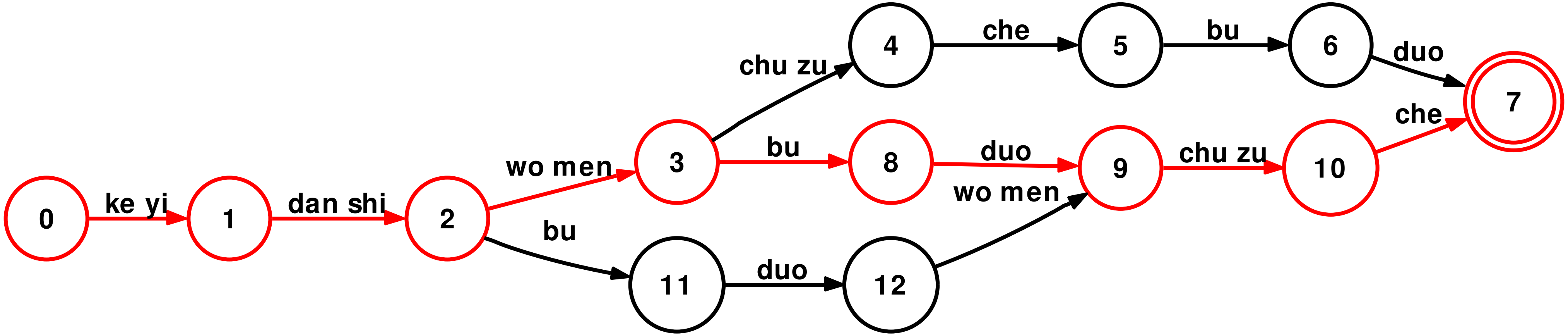
Example

source	ke yi	dan shi	wo men	chu zu	che	bu	duo
POS	v	c	r	v	n	d	m
chunks	v	c	r	NP		VP	
English gloss	yes	but	we	taxi		not many	

reordering rules
NP VP → VP NP
r NP VP → r VP NP
r NP VP → VP r NP

Example (cont'd)

- **reordering lattice:**



- **translation result:**

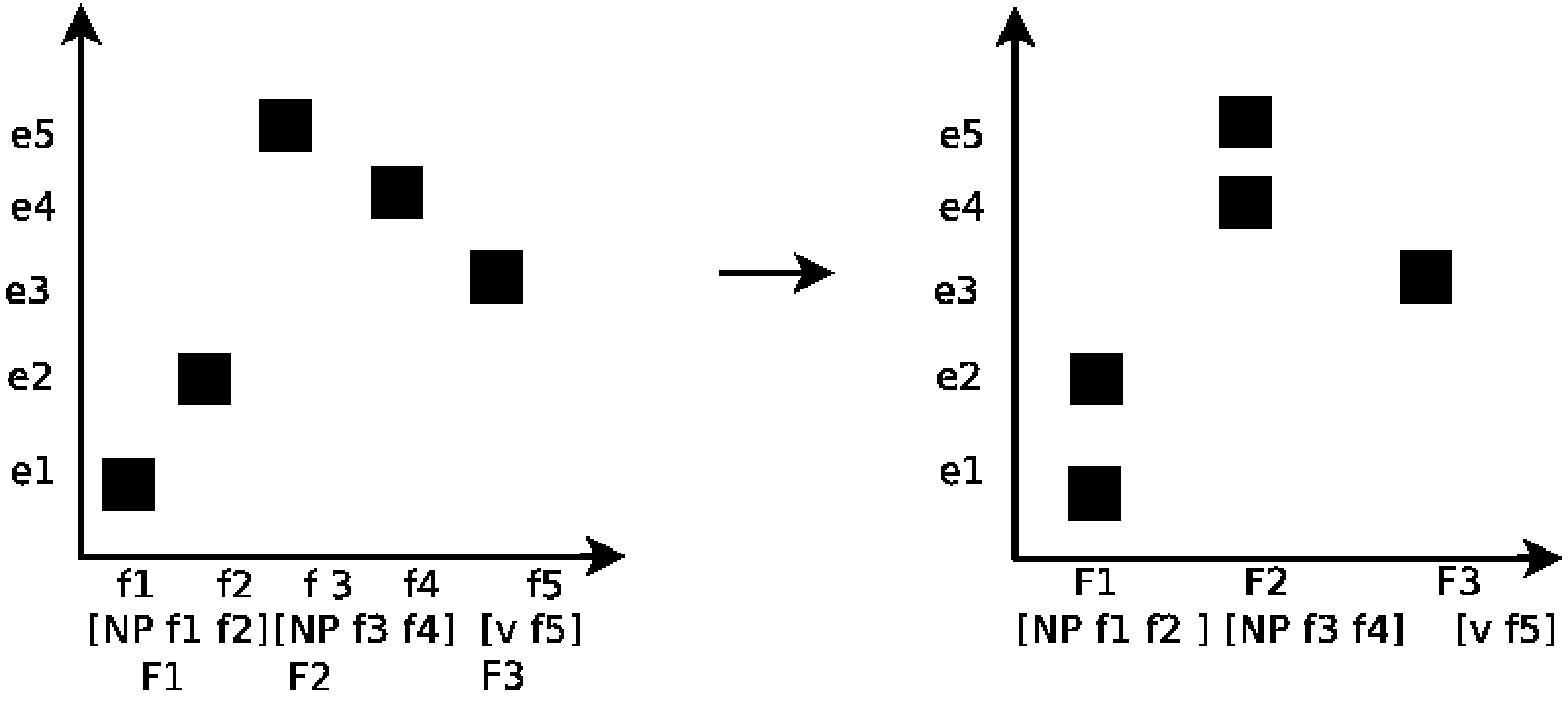
reference	yes, but there are not many rental cars here
baseline	yes , but we do rent car is not
chunk-reordering	yes , but we do not have much rental car

Chunk Parsing

- POS tagging + word segmentation with ICTCLAS tool
  Institute of Computing Technology, Chinese Academy of Sciences
- training data for chunker: Chinese Treebank (LDC2005T01)
- 24 chunk types
- MaxEnt tagger
  - input features: word + POS tag
  - output: chunk types + chunk boundary

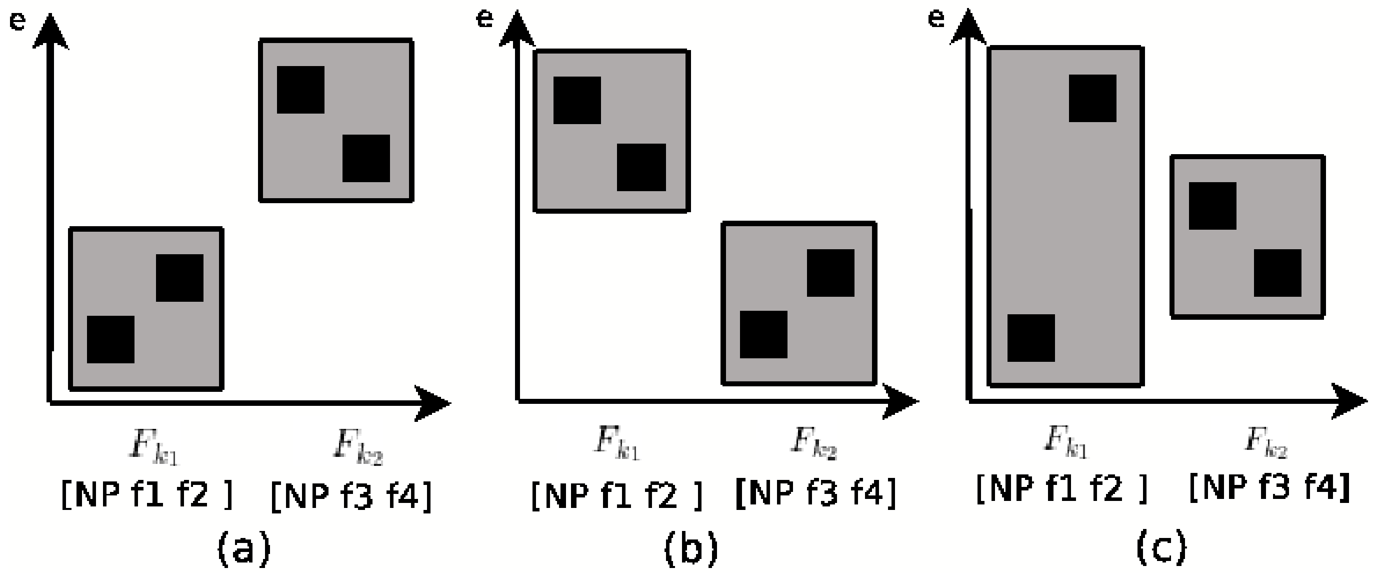
Reordering Rules Extraction

- **convert word-to-word alignment to chunk-to-word alignment**



- **run standard phrase extraction on chunk-to-word alignment**

Reordering Rules Extraction (cont'd)

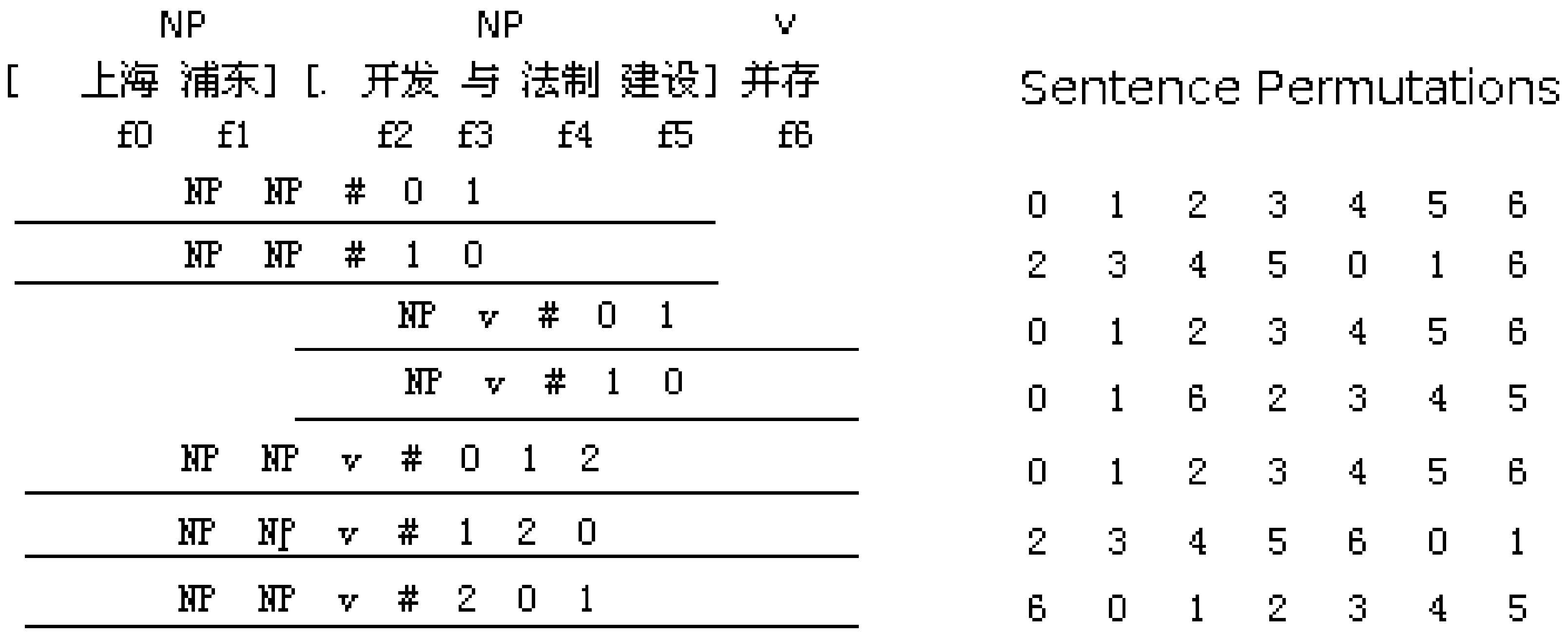


(a) monotone phrase, (b) reordering phrase, (c) cross phrase

- extract rules from monotone phrases and reordering phrases
  - e.g. $NP_0 NP_1 \# \; NP_0 \; NP_1 \quad NP_0 NP_1 \# \; NP_1 \; NP_0$
  - within a subsentence, not across punctuations

Reordering Lattice Generation I

- apply reordering rules to chunked source sentence
- represent alternative reorderings as a lattice
- example:



NP NP v

[上海 浦东] [开发 与 法制 建设] 并存

f0 f1 f2 f3 f4 f5 f6

NP NP # 0 1

NP NP # 1 0

NP v # 0 1

NP v # 1 0

NP NP v # 0 1 2

NP NP v # 1 2 0

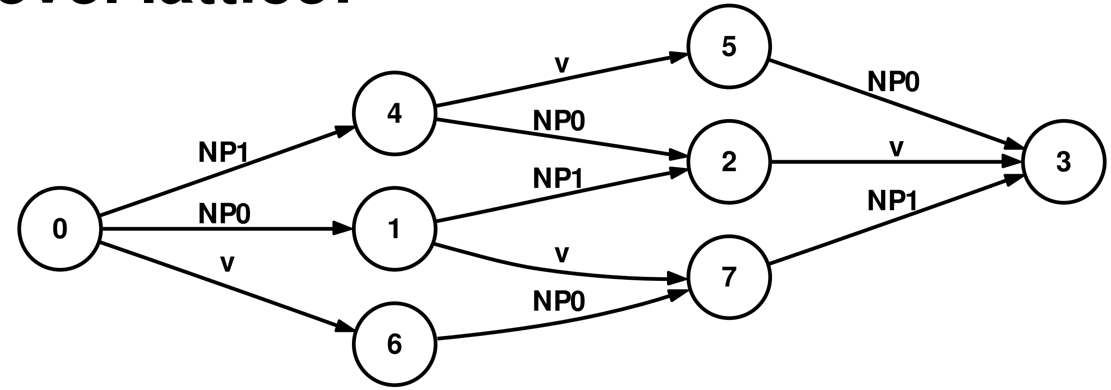
NP NP v # 2 0 1

Sentence Permutations

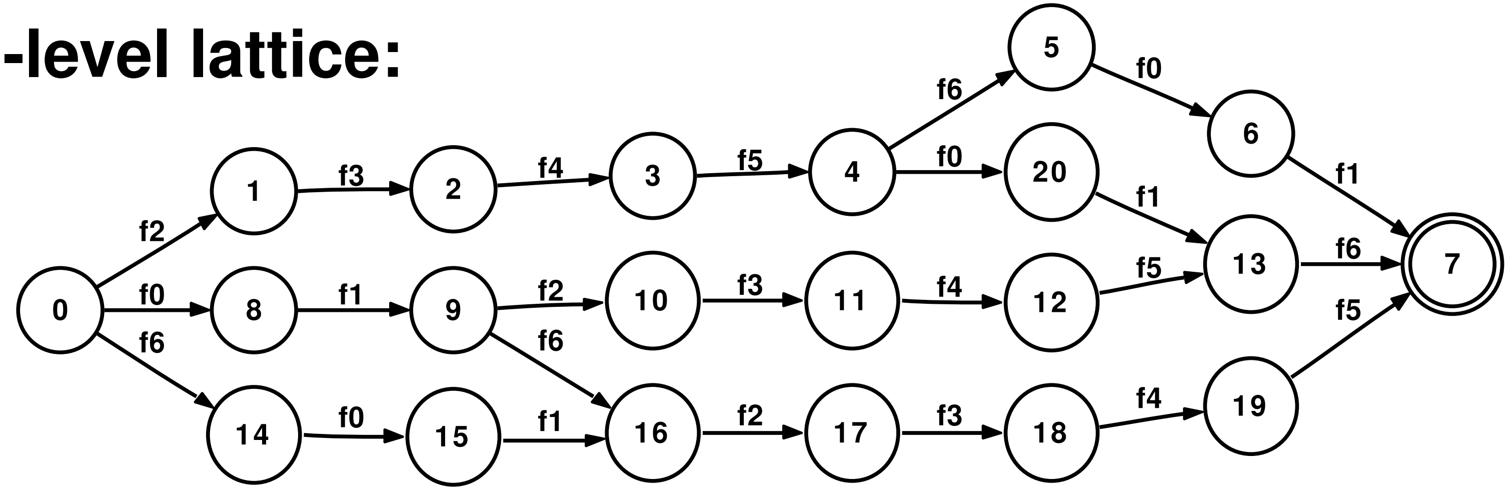
0	1	2	3	4	5	6
2	3	4	5	0	1	6
0	1	2	3	4	5	6
0	1	6	2	3	4	5
0	1	2	3	4	5	6
2	3	4	5	6	0	1
6	0	1	2	3	4	5

Reordering Lattice Generation II

- **chunk-level lattice:**



- **word-level lattice:**



Reordering Model

- use language model to weigh lattice
- training:
  - chunk source training data
  - generate chunk-to-word alignment
  - reorder source chunks to monotonize alignment
  - train LM on reordered source training data
- word-level LM

Chunking Result

- corpus statistics (Chinese Treebank LDC2005T01):

	train	test
sentences	17 785	1 000
words	486 468	21 851
chunks	105 773	4 680

- tagging results:

word-level	chunk-level		
accuracy [%]	precision [%]	recall [%]	F-measure [%]
74.51	65.2	61.5	63.3

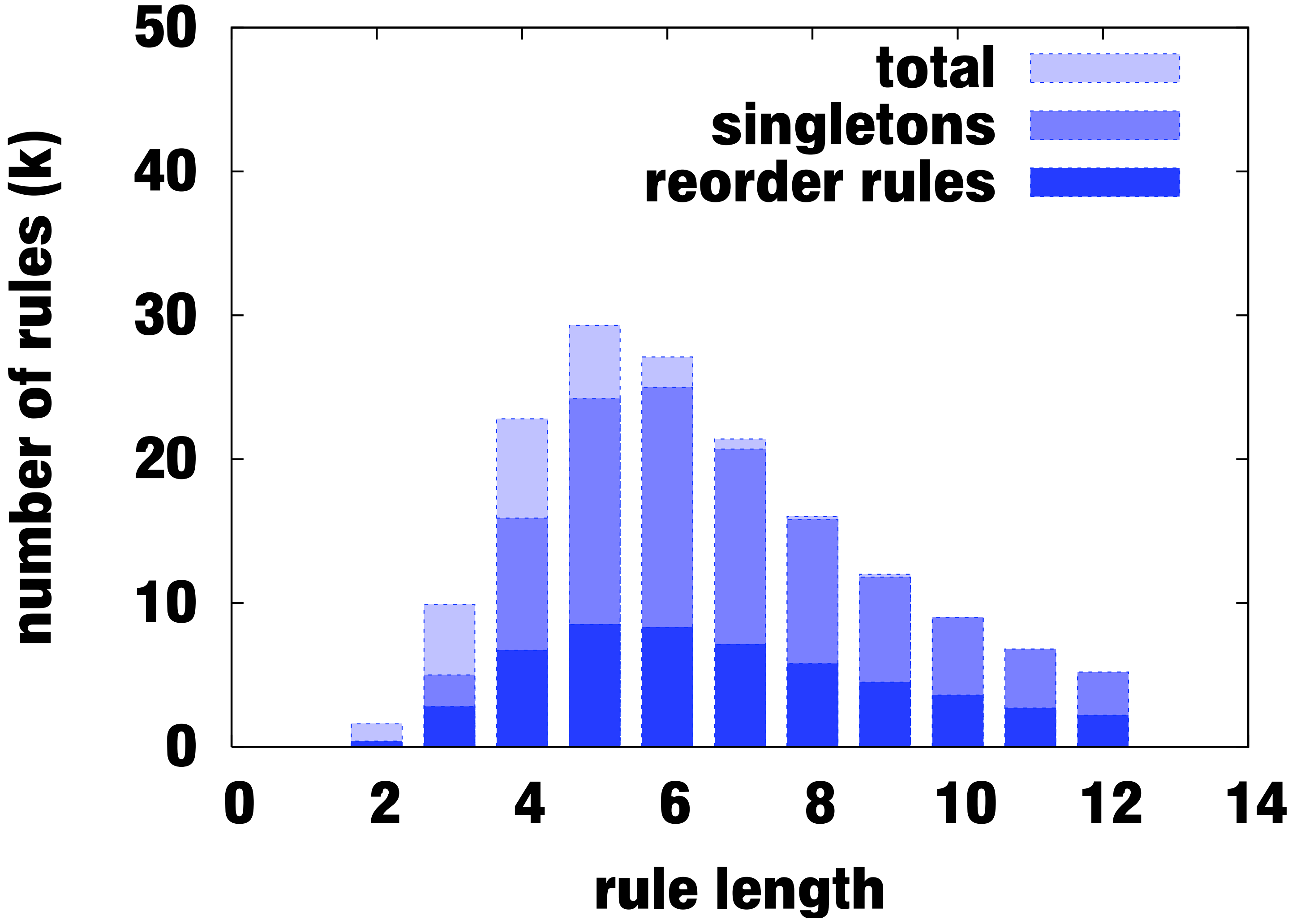
- number of chunk types: 24

- both chunk type and boundary have to be correct

Corpus Statistics

		Chinese	English
Train	Sentences	40k	
	Words	308k	377k
Dev(dev4)	Sentences	489	
	Words	5 478	6 008
Test	Sentences	500	
IWSLT04	Words	3 866	3 581
Test	Sentences	506	
IWSLT05	Words	3 652	3 579
Test	Sentences	500	
IWSLT06	Words	5 846	–

Statistics of Reordering Rules



total: 184k, singletons: 88%, reorder rules: 34%

Translation Results

		WER [%]	PER [%]	NIST	BLEU [%]
IWSLT04	baseline	47.3	38.2	7.78	39.1
	chunk reordering	46.3	37.2	7.70	40.9
IWSLT05	baseline	45.0	37.3	7.40	41.8
	chunk reordering	44.6	36.8	7.51	42.3
IWSLT06	baseline	67.4	50.0	6.65	22.4
	chunk reordering	65.6	50.4	6.46	23.3

- evaluation without punctuation marks and in lower case
- baseline: RWTH IWSLT 2006 system without rescoring

Chunk-level vs. POS-level

Translation performance (IWSLT 2004):

	WER [%]	PER [%]	NIST	BLEU [%]
Baseline	47.3	38.2	7.78	39.1
POS	46.9	37.5	7.38	39.7
Chunk	46.3	37.2	7.70	40.9

Lattice statistics:

	avg. density per sent	used rules	translation time [min:sec]
Baseline	-	-	1:22
POS	15.7	6 868	7:08
Chunk	8.2	3 685	3:47

Translation Examples (IWSLT04)

reference	about twenty-five seconds
baseline	seconds about twenty-five
chunk reorder	about twenty five seconds
reference	could n’t you make it a little cheaper
baseline	could not you some better
chunk reorder	can’t you make it a little cheaper ones
reference	how much is admission
baseline	admission fees how much is it
chunk reorder	how much is the admission
reference	may i have that gift wrapped please
baseline	wrap can i have a gift
chunk reorder	can i have a gift wrapped please

Summary

- idea:

  1. chunk input sentence
  2. reorder chunks
  3. represent alternative reorderings as lattice
  4. translate lattice

- nice improvements on IWSLT task
- chunk-level reordering better than POS-level reordering

Outlook

- large data task (e.g. NIST)
- other language pairs
- improve chunk parsing
- better reordering model
- analyze what kind of rules work well

THANK YOU FOR YOUR ATTENTION!

ICTCLAS POS Tag Set

n	noun	r	pron
nr	person name	rg	pron morpheme
ns	location name	m	number
ng	noun morpheme	q	quantity
t	time	d	adverb
s	location	p	prep
f	position word	c	conjunction
v	verb	u	auxiliary
vd	verb adv	e	interjection
vn	noun verb	y	modal particle
vg	verb morpheme	o	onomatopoeia
a	adj	h	prefix
ad	adv adj	k	suffix
an	adj noun	w	punctuation
ag	adj morpheme	b	determiner

Syntactic Tag Set of Chunks

ADJP	adjective phrase
ADVP	adverbial phrase headed by AD (adverb)
CLP	classifier phrase
CP	clause headed by C (complementizer)
DNP	phrase formed by $'XP+DEG'$
DP	determiner phrase
DVP	phrase formed by $'XP+DEV'$
FRAG	fragment
IP	simple clause headed by I (INFL)
LCP	phrase formed by $'XP+LC'$
LST	list marker
NP	noun phrase
PP	preposition phrase
PRN	parenthetical
QP	quantifier phrase
UCP	unidentical coordination phrase
VP	verb phrase

Details are in "The bracketing gudelines for penn chinese treebank 3.0", Technical Report 00-08, University of Pennsylvania(2000) IRCS Report.