

Improving Chinese-English Machine Translation through Better Source-side Linguistic Processing



Pi-Chuan Chang

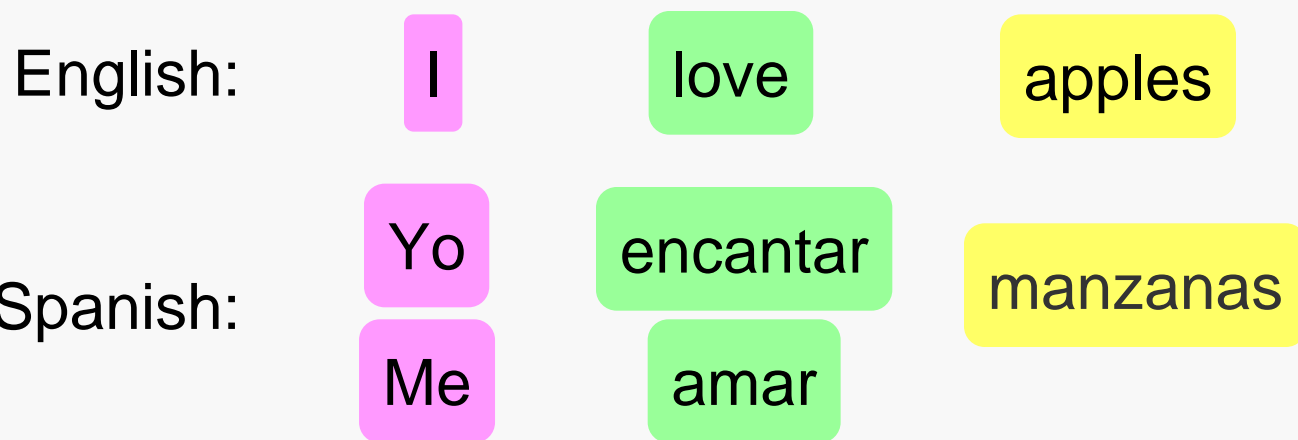


Translation

- Definition from Wikipedia:
 - **Translation** is the interpreting of the meaning of a text and the subsequent production of an equivalent text that communicates the same message in another language.

How to translate?

- How about dictionary lookup?



Correct translation: Me encantan las manzanas

reflexive
pronoun
(myself)

verb conjugation

feminine; plural



Why is translation hard

- Words translate differently in context
 - Meaning
 - Morphology (e.g., verb tenses, plural forms, gender, etc)
- Word mapping is not trivial
 - Not one-to-one mapping
- Word orders differ among languages
 - e.g., SVO languages: Chinese and English
 - SOV languages: Japanese
- More than words:
 - non-literal translation; idioms; cultural references



Research on Statistical Machine Translation

Focus on general inter-lingual phenomenon:

- Word Alignment
- Translation models
- Reordering models
- Decoding
- etc

Mostly language independent



Difficult language pair: Chinese-English

- Chinese-English is a difficult language pair
- From NIST MT evaluation 2008
 - Arabic-English, Chinese-English
- On constrained track
 - Evaluated in BLEU
 - Arabic (top 5 groups): 42 ~ 46
 - Chinese (top 5 groups): 24 ~ 31
- Arabic-English MT: close to readable quality
- Chinese-English MT: still much worse



Example Arabic-English MT output

- Russian President Vladimir Putin participates in the work of the European summit starts today in the capital of the current presidency of the EU, with the participation of leaders at a working dinner on Sunday evening.
 - Some grammatical errors
 - Readable and understandable
 - Relationships between important elements are clear



Example Chinese-English MT output

- The increases in the price of beef noodle soup, a whopping 20 percent; statistical data, this is the beginning of the 1980s, Lanzhou, for the first time since the rise in the price of beef noodle soup, after a short, one of the biggest increase.
 - Pieces of information don't connect together
 - Hard to understand



Why is Chinese-English MT hard?

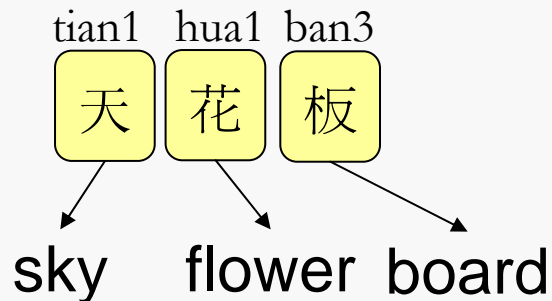
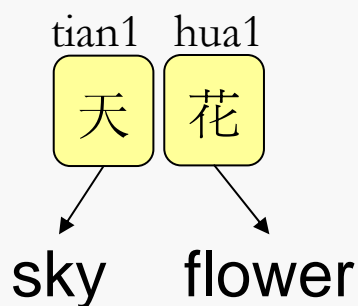
Segmentation

Segmentation

国际天花生物安全高峰会议在日内瓦举行
这次为期两天的会议将于10月21日至22日举行, 重点是讨论针对故意使用天花病毒的应对措施。

Geneva to Host International Smallpox Biosecurity Summit
The two-day meeting, which focuses on measures to counteract the deliberate use of the smallpox virus, will take place on October 21-22 .

- Word-for-word... no word boundaries (spaces) in Chinese
- Can we use characters as minimum units?

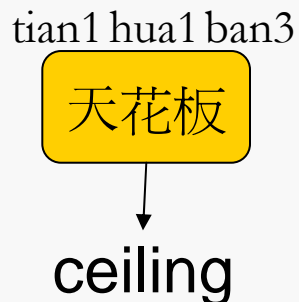
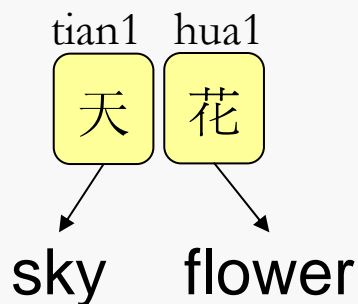


Segmentation

国际天花生物安全高峰会议在日内瓦举行
这次为期两天的会议将于10月21日至22日举行, 重点是讨论针对故意使用天花病毒的应对措施。

Geneva to Host International Smallpox Biosecurity Summit
The two-day meeting, which focuses on measures to counteract the deliberate use of the smallpox virus, will take place on October 21-22 .

- Word-for-word... no word boundaries (spaces) in Chinese
- Can we use characters as minimum units?





Segmentation

Segmented text

国际 天花 生物 安全 高峰 会议 在 日内瓦 举行
这次 为期 两天 的 会议 将 于 10月 21日 至 22日 举行 ， 重点 是 讨论 针对
故意 使用 天花 病毒 的 应对 措施 。

- introduce errors in the first step
 - 同样 / 的 / 玩具 / 在 / 沃尔 / 玛不到 / 40 / 美元
Wal-mart
- Unclear what is the best segmentation for translation
 - E.g., Chinese Person name: 胡锦涛 or 胡 / 锦涛
Hu Jintao

胡锦涛 和 胡海峰
Hu Jintao and Hu Haifeng



Why is Chinese-English MT hard?

Lack of Word level Information



Chinese: morphology-poor

- Morphology:
 - the structure and content of word forms
- Inflectional morphology:
 - Verb tenses: eat (present tense)
ate (past tense)
 - Chinese: 吃
 - Word-for-word... 吃 (past tense) is implicit
 - Noun morphology:
 - English: “-s”
 - Chinese: plural marking is not required
- Derivational morphology:
 - In Chinese, nouns and verbs usually have the same form
 - English: develop (verb); *development* (noun)
 - Chinese: 发展 (verb; noun)



Lack of word information

- Chinese:
 - a source language where a lots of information is hidden
 - Poor morphology: verbs and nouns
 - Determiners for nouns not required
 - etc



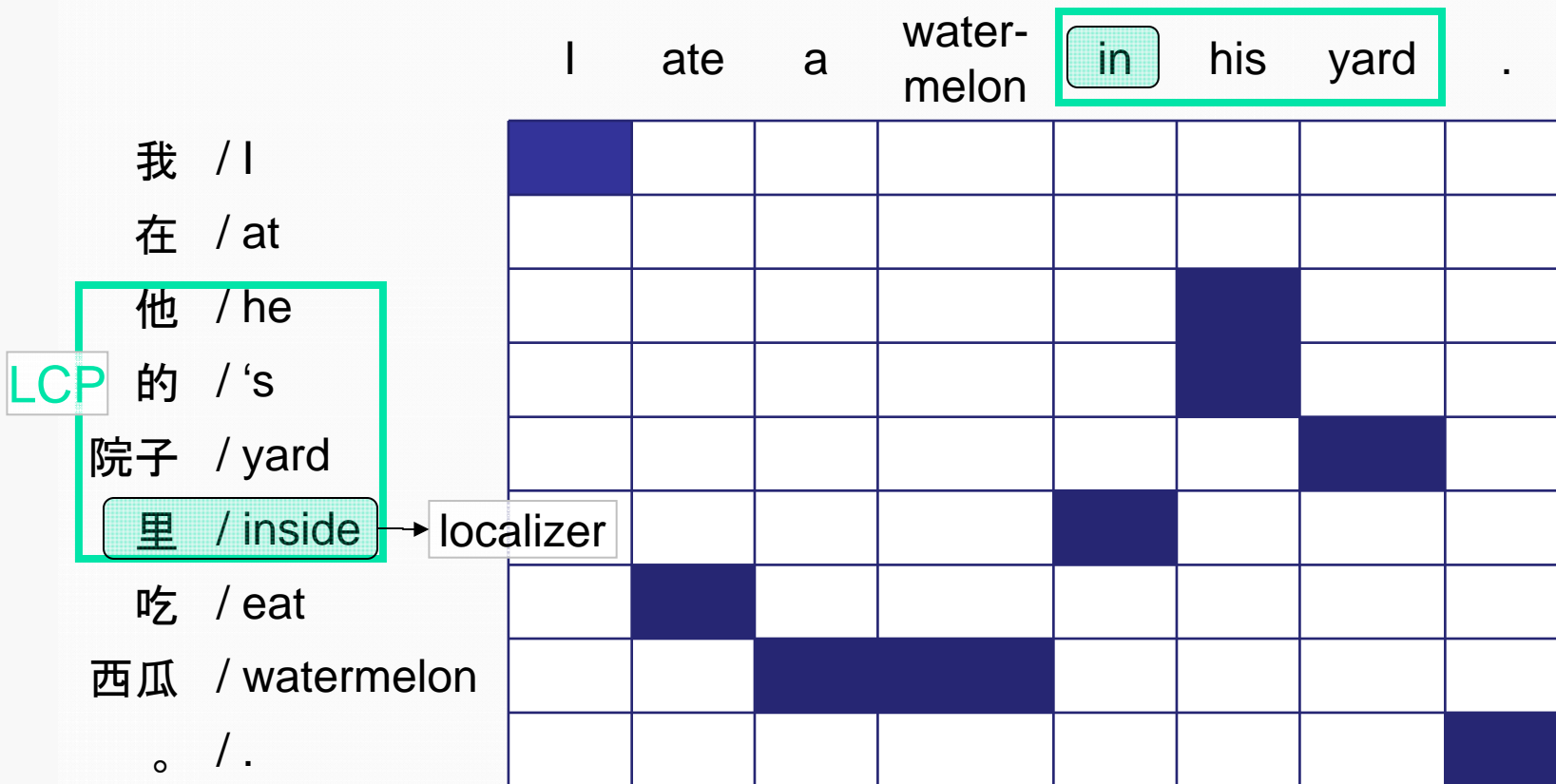


Why is Chinese-English MT hard?

Syntactic Structural Differences

Lots of different word orders – Localizers

- Localizers
 - like a post-phrasal preposition
 - often used with temporal or locative phrases



Lots of different word orders – The notorious 的 (DE)

- 的 (DE):
 - Ch [Complementizer Phrase that pre-modify an NP]
→ En [relative clause]

	the	companies	that	closed	last	year
CP						
NP						

去年 / last year

CP 关闭 / close

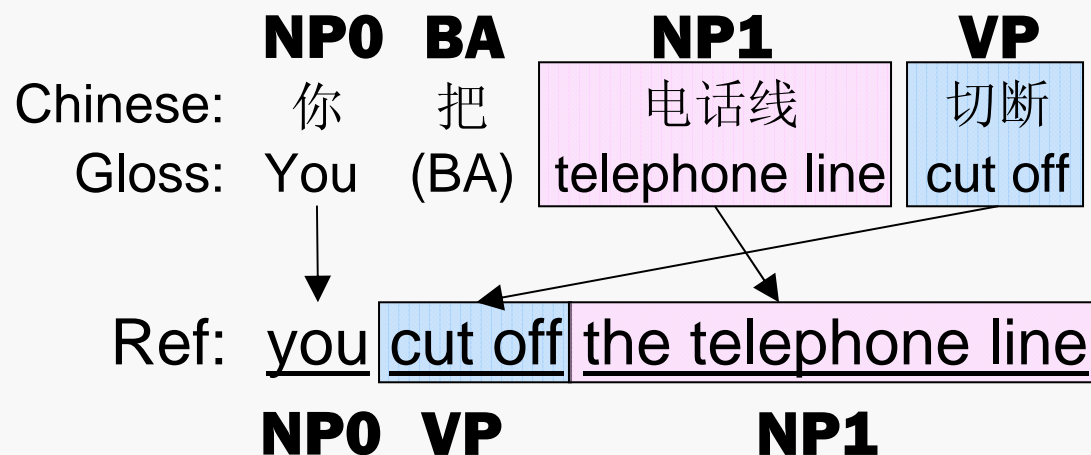
的 / "DE"

NP 公司 / company

Lots of different word orders – special Chinese constructions

- 把 (BA) construction
- 被 (BEI) construction
- Ch [SOV] → En [SVO]

Example:





Contribution of my thesis

- Segmentation
 - optimized for MT
- Chinese Grammatical Relations
 - Bi-lexical relations
 - Use them to improve reordering models in MT
- Analyze common function words
 - Disambiguating 的 (DE) is useful for MT
- MT system Error analysis



Source-side linguistic processing in MT

1. Efficient

- Source sentence is fully observed
- Linguistic processing is more efficient

2. Brings more clues for making the prediction

- Disambiguate the confusing parts
- Better match with target language

Segmentation (optimized for MT)

[Chang et al 2008; Tseng et al 2005]



Word segmentation and MT

- The Chinese segmentation task has been defined as optimizing for a given segmentation standard
 - Performance measured by F1 score
- Better segmentation performance
≠ Better MT performance



Segmentation performance \neq MT performance

- Example:
A segmenter that performs *much better* on segmentation can be *worse* for MT than a simple max-match segmenter

	Segmentation Performance SIGHAN 2006 F1	MT Performance MT05 BLEU (Moses)
CRF-basic	87.7	
Lexicon max-match	82.8	



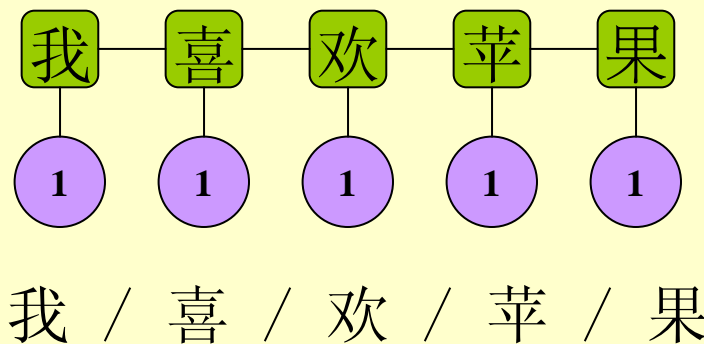
Segmentation performance

- Example:
A segmenter that performs
segmentation can be compared to
max-match segmentation

CRF-basic
Lexicon max-match

CRF (conditional random fields)

- binary sequence labeler
 - 1 – separated from previous
 - 0 – continuous





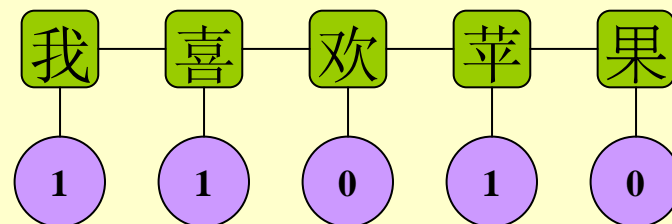
Segmentation performance

- Example:
A segmenter that performs segmentation can outperform a max-match segmenter

CRF-basic
Lexicon max-match

CRF (conditional random fields)

- binary sequence labeler
 - 1 – separated from previous
 - 0 – continuous



我 / 喜欢 / 苹果

- advantages
 1. easy to add new features
 2. good performance
 3. can recognize unseen words



Segmentation performance

- Example:
A segmenter that uses
segmentation can
max-match segmenter

CRF-basic
Lexicon max-match

Lexicon Max-match

- start with a lexicon
- left-to-right greedy match
 - always matches longest word possible
- very efficient
- cannot recognize OOV words
- lower segmentation performance
 - longer match isn't always correct



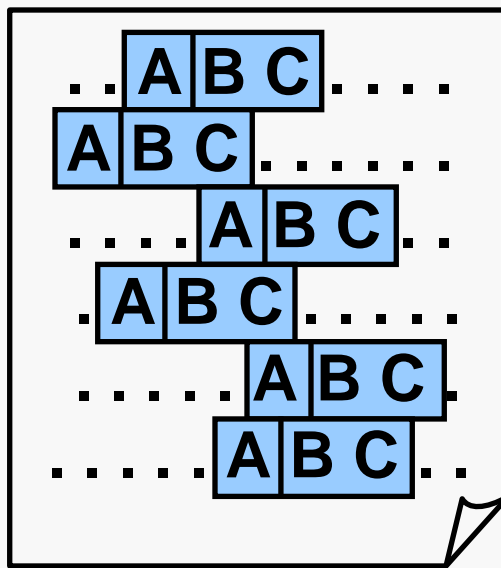
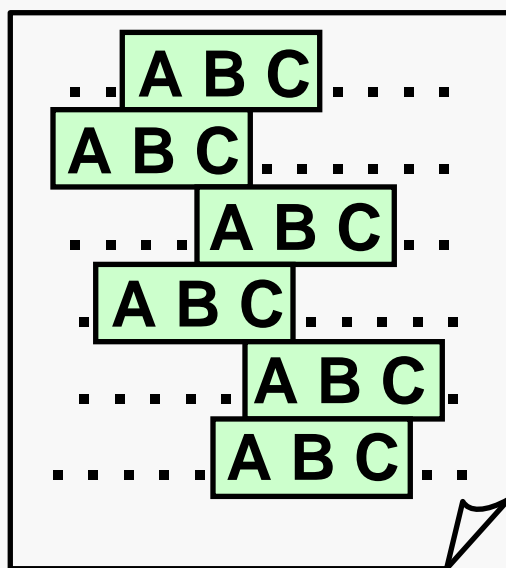
Segmentation performance \neq MT performance

- Example:
A segmenter that performs *much better* on segmentation can be *worse* for MT than a simple max-match segmenter

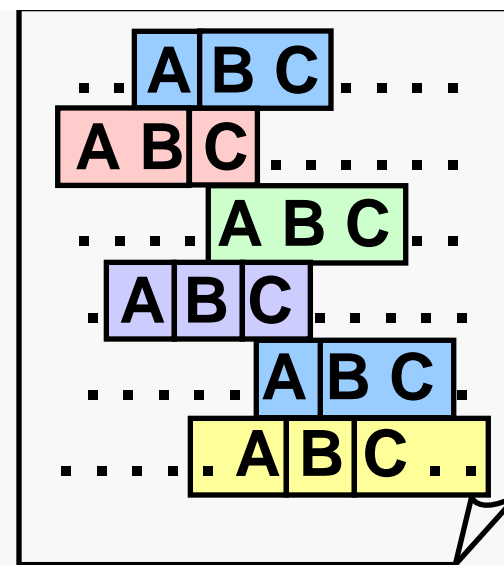
	Segmentation Performance SIGHAN 2006 F1	MT Performance MT05 BLEU (Moses)
CRF-basic	87.7	30.4
Lexicon max-match	82.8	30.7

Important characteristics of good segmentation for MT

1. Segmentation consistency



Inconsistent segmentation



**If the segmenter segments all the occurrences as “A” “BC”, it is still consistent.
(but this will be penalized in the Segmentation F1 score)**



Important characteristics of good segmentation for MT

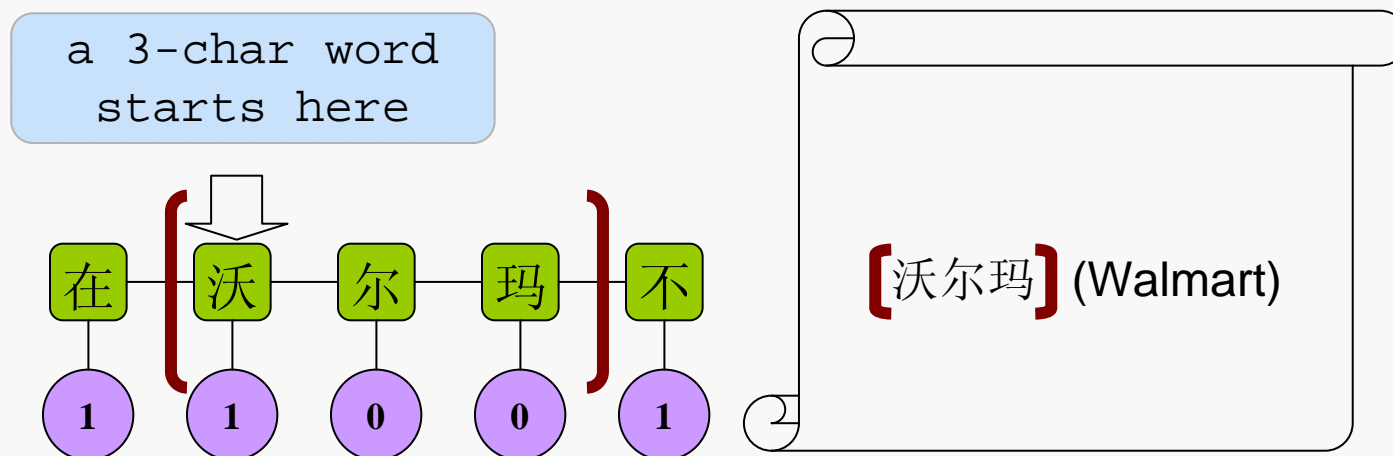
1. Segmentation consistency
2. Word granularity better matched with English

胡锦涛 (“Hu Jintao”) (CTB standard segmentation)
胡 / 锦涛 might be a better segmentation

內政部 (“Department of Internal Affairs”)
內政 = internal affairs
部 = department
內政 / 部 might be a better segmentation

Improving Segmentation Consistency

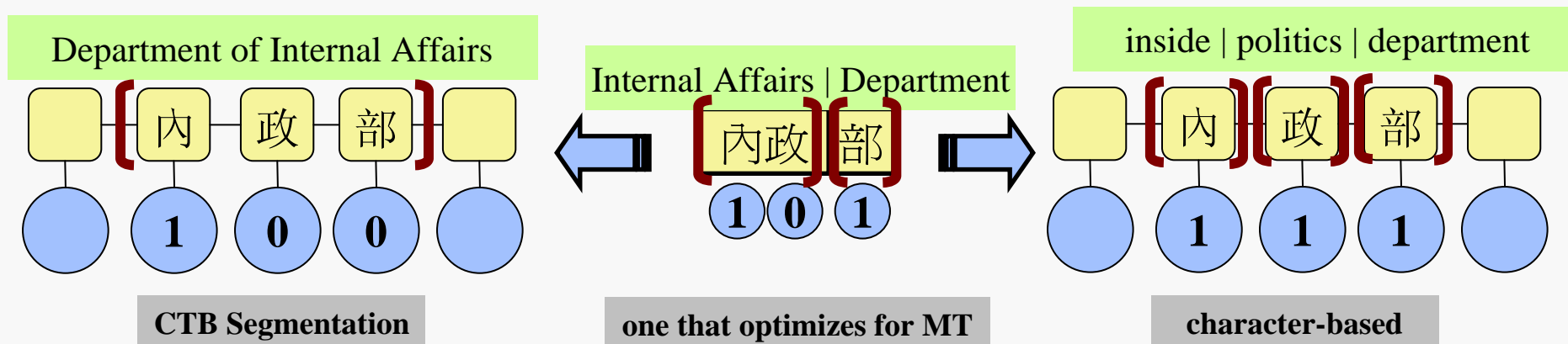
- CRF segmenter + lexicon-based features
- Use external lexicons, and check if each character is a prefix / suffix / infix of existing words



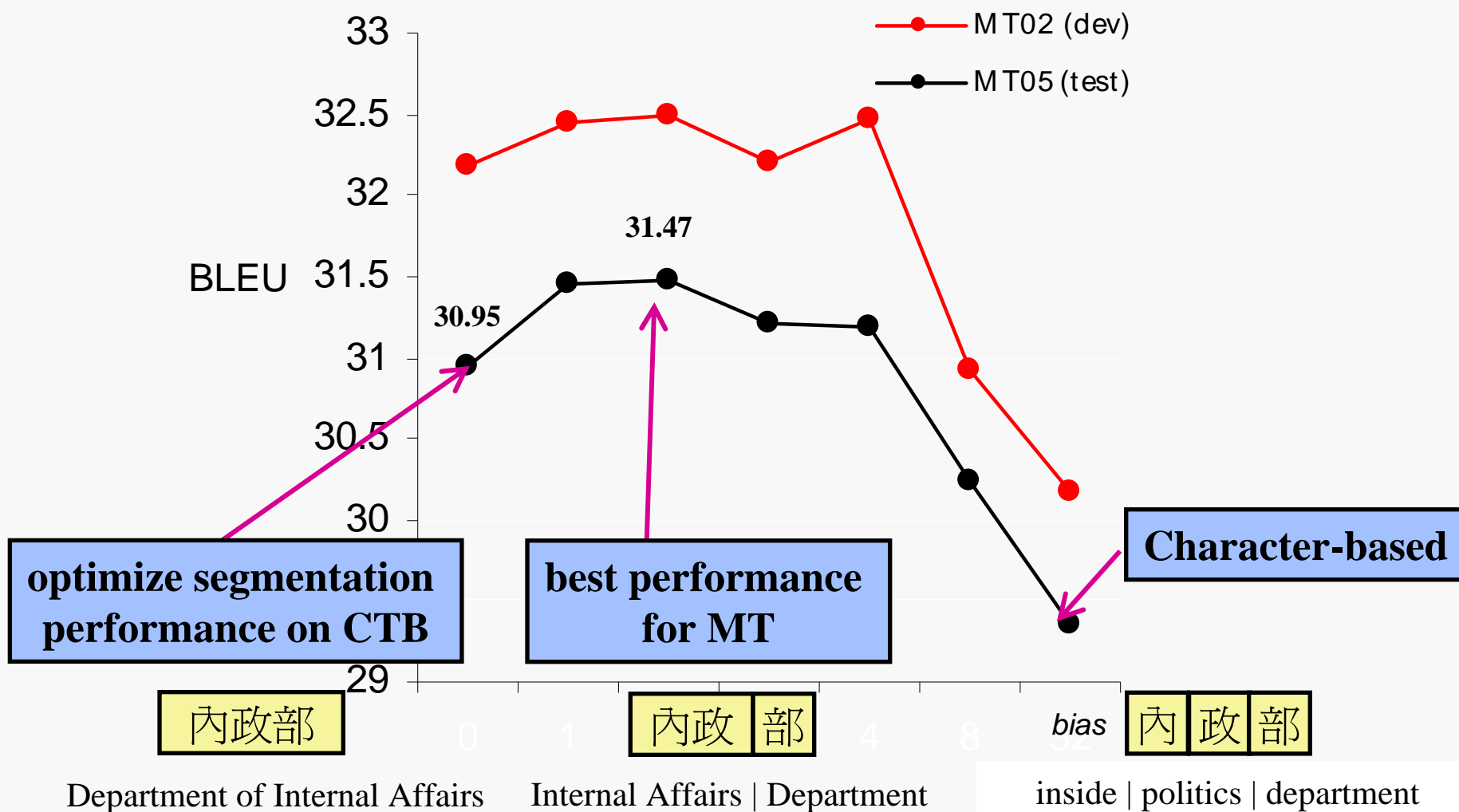
- The lexicon features help CRF segment more consistently

Finding better granularity – tuning word length in CRF

- **Our CRF model makes a binary prediction. 1 means separated from previous character; 0 means continuing.**
 - Trained on Chinese Treebank 6 data
- **We add a bias feature that fires only when the prediction is 1**
 - Extreme case: Very large positive bias cause the segmenter to predict only 1 (character-based)
 - We find optimal middle ground



Continuum between word- and character-based segmentation





Summary for Segmentation (optimized for MT)

- Segmentation is useful for MT
 - Characters are not the smallest units of meaning
 - Even a simple segmenter is better than nothing
- Important characteristics for good segmentation
 - Consistency – improved by lexicon-based features in CRF
 - Word granularity – can be tuned with one extra feature

Syntactic Structural differences:
Grammatical Relations
help MT reordering models

[Chang et al 2009]

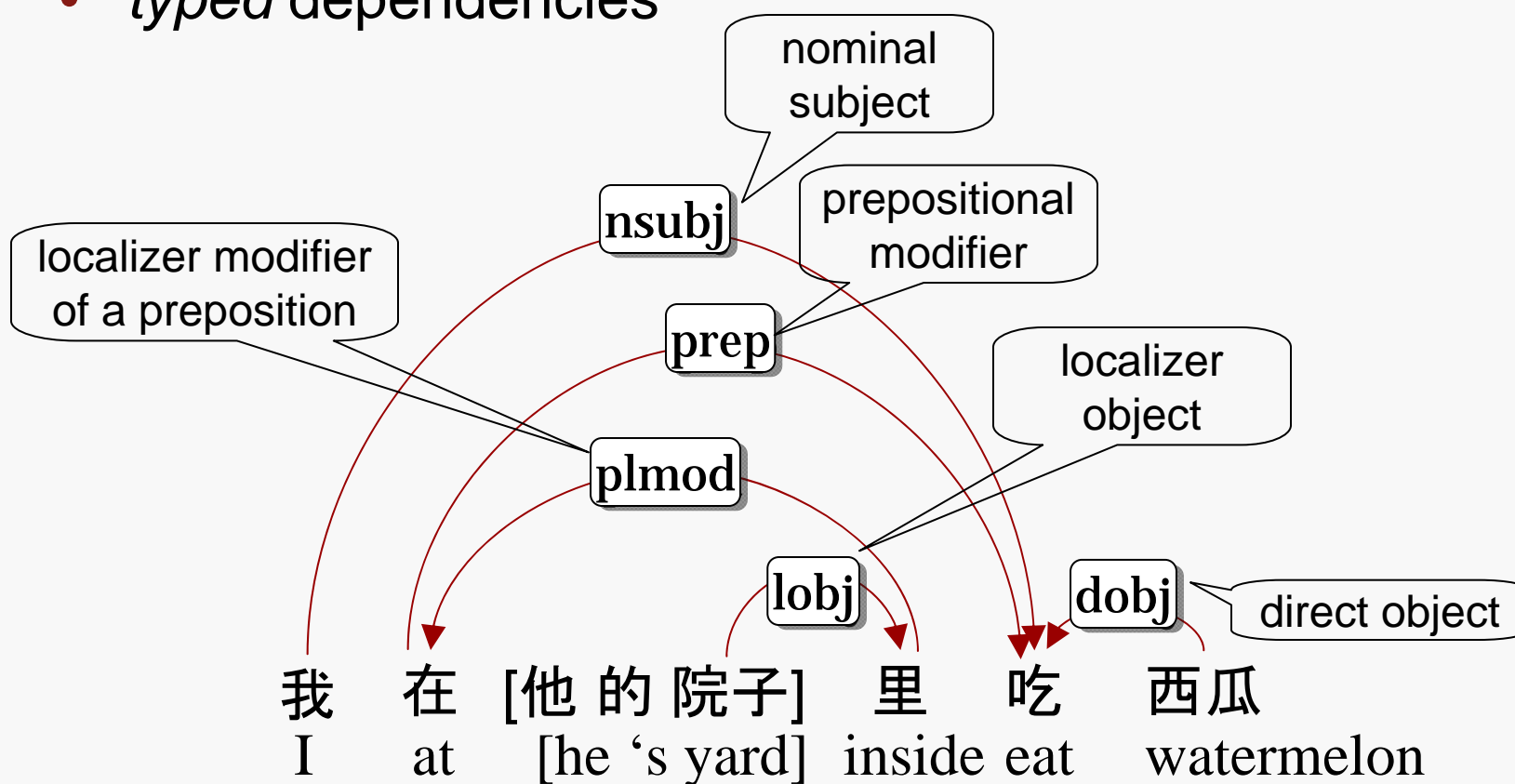


Reordering (Chinese → English)

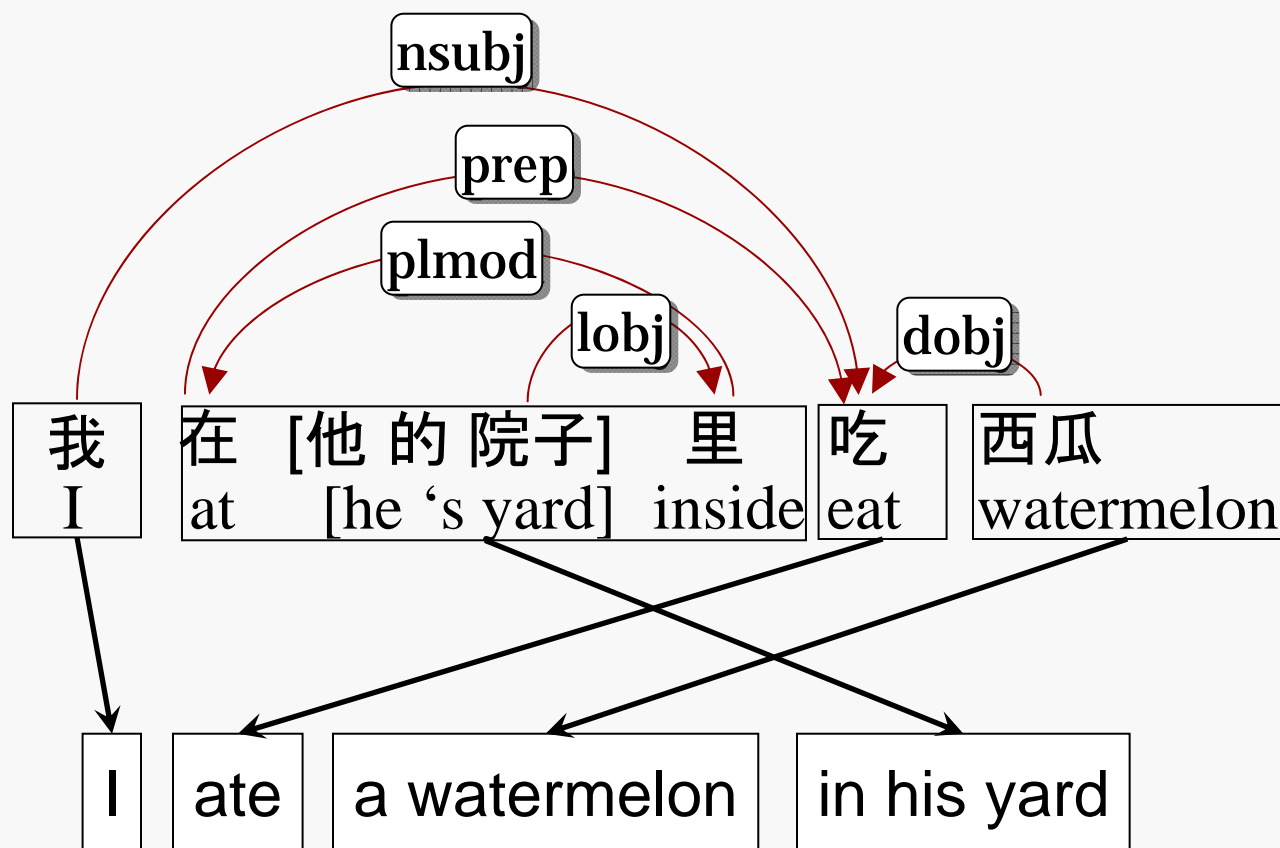
- Certain Chinese syntactic structures are more likely to be reordered in English
 - Ch [PP VP] → En [VP PP]
 - Localizers (post-phrasal prepositions)
 - DE constructions (in English: relative clause, PP, etc)
 - BA, BEI constructions (Ch [SOV] → En [SVO])
- Chinese syntactic structures bring useful signals for deciding word orders

Chinese Grammatical Relations

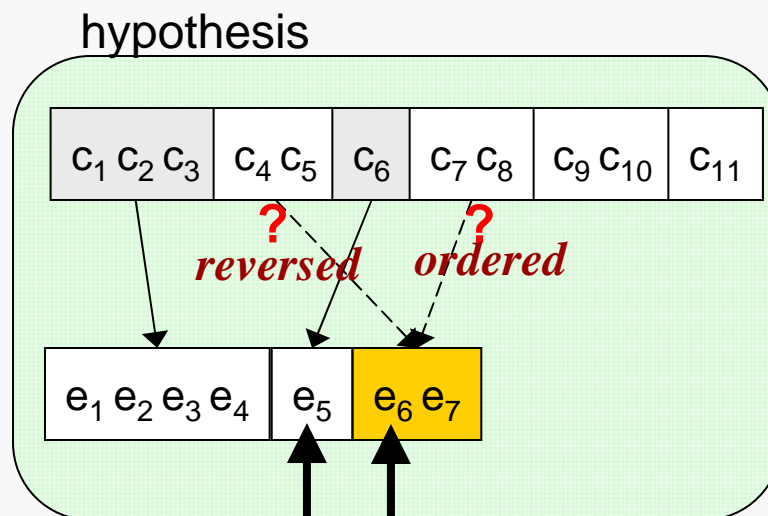
- Binary relations between words
- *typed* dependencies



In MT decoding: Which Chinese words to translate next?



Discriminative Reordering Model (Zens and Ney 2006)

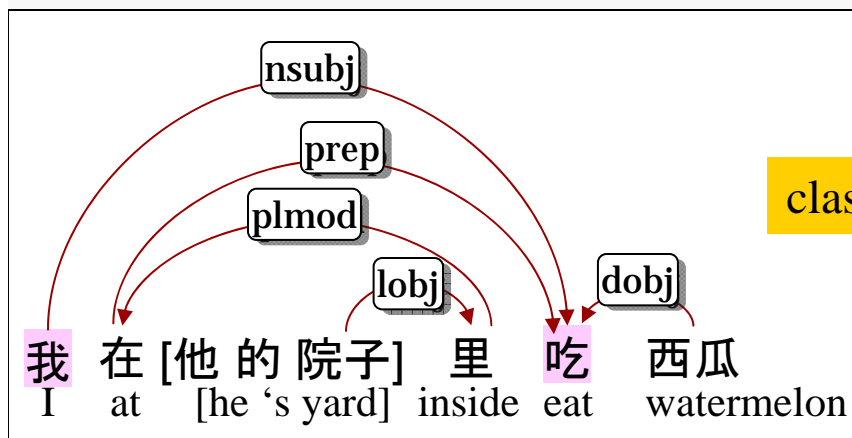


Decide the order of source Chinese words to translate: {reversed, ordered}

- log-linear binary classifier
 - Basic feature: lexical features
- We improved the model by adding syntactic information



PATH features in the reordering model



class = *ordered*

I ate a water-melon in his yard .
 0 1 2 3 4 5 6 7

nsubj

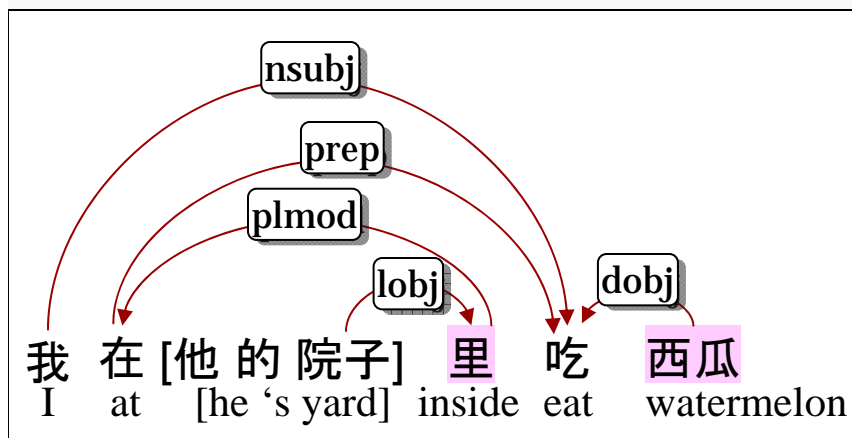
我 / I
 在 / at
 他 / he
 的 / 's
 院子 / yard
 里 / inside
 吃 / eat
 西瓜 / watermelon
 。 / .

0								
1								
2								
3								
4								
5								
6								
7								
8								

PATH feature = nsubj

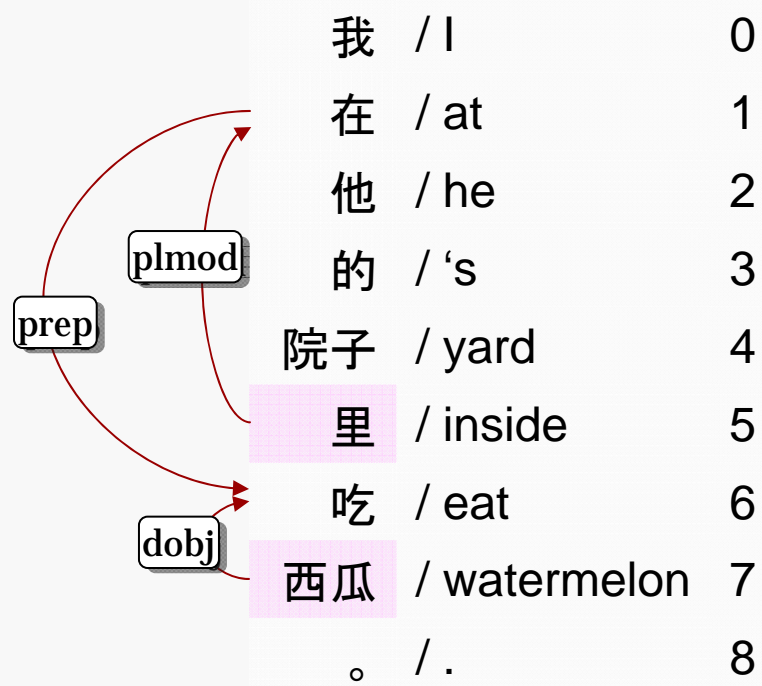


PATH features in the reordering model



class = *reversed*

I ate a water-melon in his yard .
 0 1 2 3 4 5 6 7



0	█							
1								
2								
3								
4								█
5					█			
6		█						
7			█	█				
8								█

PATH feature = plmod-prep-dobjR

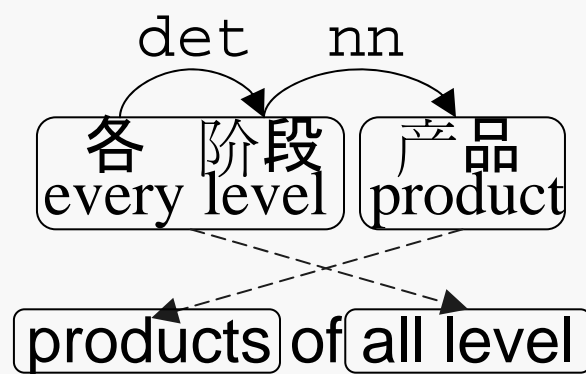


Analysis: features for “reversed”

- In the log-linear classifier, features highly weighted for being “reversed”:
 - Many of them match linguistic intuitions
- 1. Chinese [PP VP] → English [VP PP]
 - *prep-dobjR*
- 2. Chinese [CP NP] → English [NP relative clause]
 - *rcmod* (relative clause modifier)

Analysis: features for “reversed”

- Another example for a highly “reversed” path:
 - *det-nn* :



- A longer nominal modifier (with a DT) in Chinese is more likely to become a prepositional modifier in English

MT experiments

- Experiments:
 - Exp1. Moses (+ lexicalized reordering)
 - Exp2. (Exp1) + DiscrimReorderModel (no PATH feature)
 - Exp3. (Exp1) + DiscrimReorderModel with PATH features

	MT06(tune)	MT02	MT03	MT05
Exp1	32.6	32.6	32.7	31.9
Exp2	32.7(+0.1)	32.6(+0.0)	33.0(+0.3)	31.8 (-0.1)
Exp3	33.0(+0.4)	33.2(+0.6)	33.7(+1.0)	32.7(+0.8)



Summary:

Grammatical Relations help reordering in MT

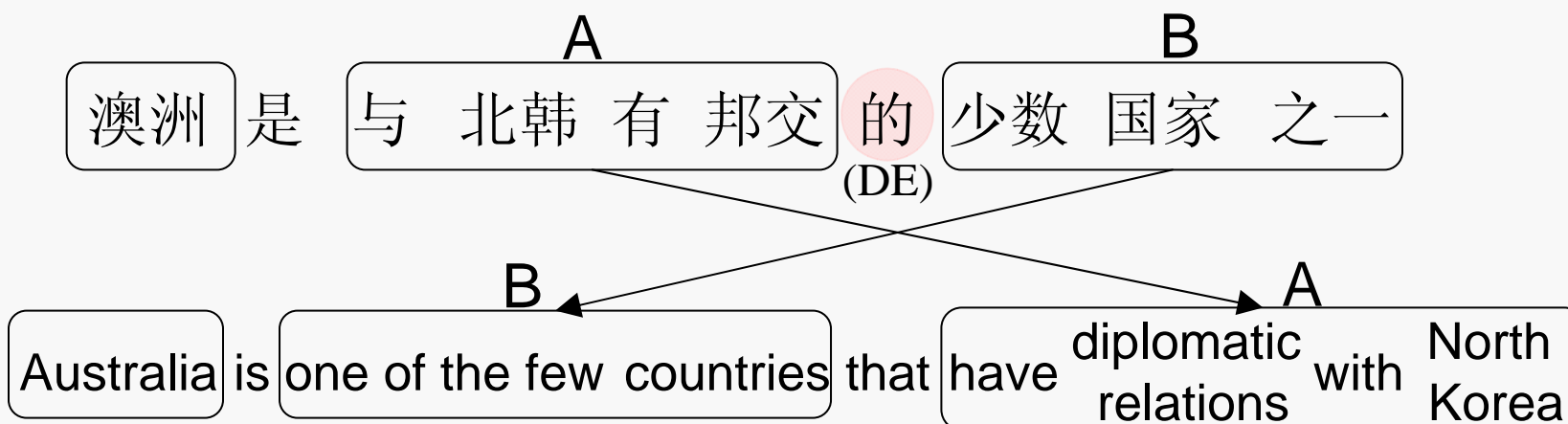
- We designed a set of Chinese grammatical relations
 - Useful for describing relations between words
- MT reordering models are improved by source-side syntactic information
 - We improved a discriminative reordering model by adding features from Chinese grammatical relations

的 (DE) Disambiguation

[Chang et al 2009]

的 (DE)

- 的 (DE) is the most common word in Chinese.



1. “的” can cause longer distance reordering
 - In Chinese, A *pre-modifies* B
 - In English, A is a relative clause *following* B
2. “的” translates ambiguously
 - It doesn't always translate to a relative clause.
 - The order doesn't always reverse



Why 的 DE construction is hard to translate

X 的 Y →

- X Y 同样的玩具 the same toy
- X's Y 墨西哥的 2000万人 Mexico's 20 million people
- Y X 门后的那条小河 that small river behind my house
- Y of X 河的两头 both sides of the river
- Y that X 传染给对方的天花病 smallpox that infected the opponents
- Y in X 周围的玩伴 playmates in the neighborhood
- Y with X 名声不佳的中学 school with a bad reputation



5 “DE” classes

- We labeled 3412 DEs from 1-325 in Chinese Treebank
A 的 B →

1. A B	693	(24.05%)
2. A 's B	91	(3.16%)
3. A <i>prep.</i> B	48	(1.66%)
4. <i>Relative clause</i>	669	(23.21%)
5. B <i>prep.</i> A	1381	(47.92%)

Reorder A and B in English
- 530 DEs labeled as “other”
 - Not included for training the DE classifier



Features for DE classification

1. Semantic Class features:

- Look up semantic classes in a Chinese thesaurus “CILIN”

	韩国	的	投资	对象
gloss:	Korea	(DE)	investment	target
class:	location	(DE)	activity	abstraction

2. Discourse features: topicality

- Whether a noun is mentioned before, can affect the decision between of-genitive and s-genitive

Of-genitive: the mother of the boy

S-genitive: the boy's mother

3. Syntactic features:

- from the Chinese parse trees

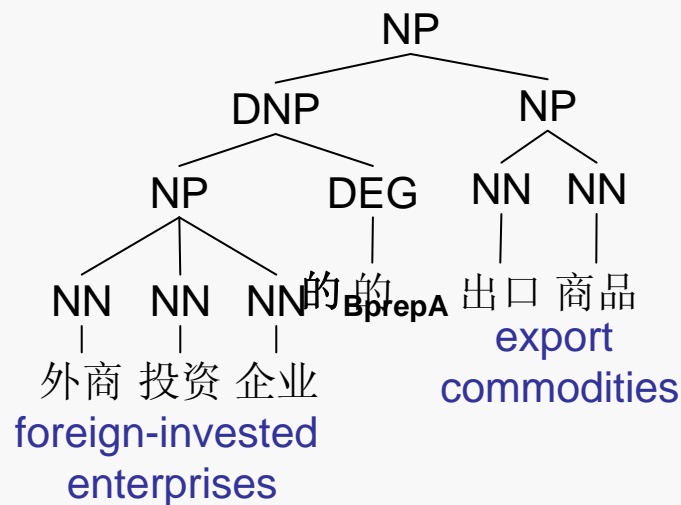
4. Lexical features

Use DE-annotated text for MT

1. Use the labeled DEs to train a 5-class classifier
2. Parse the Chinese sentences in MT training data
3. Use the DE classifier to annotate DEs.
4. Reorder 的_{BprepA} and 的_{relc}

ChineseInput =

外商 投资 企业 的 出口 商品



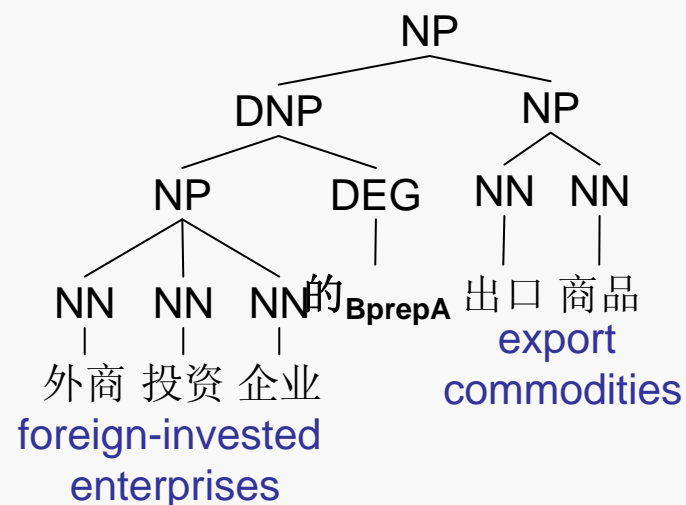


Use DE-annotated text for MT

1. Use the labeled DEs to train a 5-class classifier
2. Parse the Chinese sentences in MT training data
3. Use the DE classifier to annotate DEs.
4. Reorder 的_{BprepA} and 的_{relc}

Chinese =
外商 投资 企业 的 出口 商品

Processed Chinese =
出口 商品 的_{BprepA} 外商 投资 企业





Annotated DE helps in MT systems

ChineseInput =

遭 工会 强烈 反对 的_{relc} 就业 改革 方案

Gloss: suffer labor union strongly oppose (DE) employment reform plan

Reference: employment reform plan that was strongly opposed by the labor union



Annotated DE helps in MT systems

ChineseInput =

遭 工会 强烈 反对 的_{relc} 就业 改革 方案

Gloss: suffer labor union strongly oppose (DE) employment reform plan

Reference: employment reform plan that was strongly opposed by the labor union

- The classifier labeled the 的 as 的_{relc}.
- Reorder the Chinese words

Processed ChineseInput =

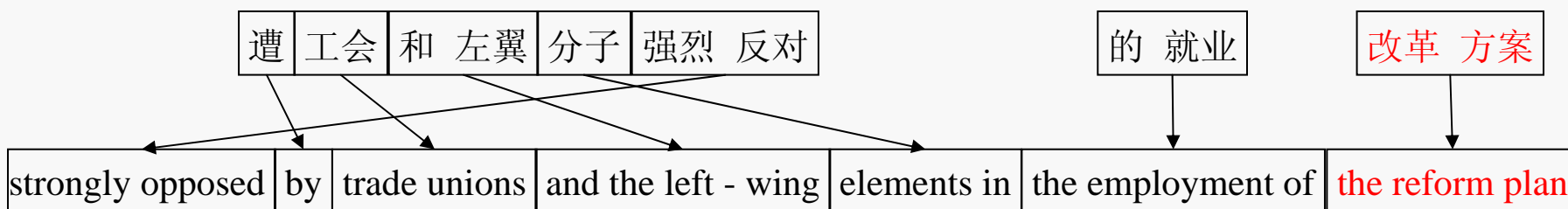
就业 改革 方案 的_{relc} 遭 工会 强烈 反对

Gloss: employment reform plan (DE) suffer labor union strongly oppose

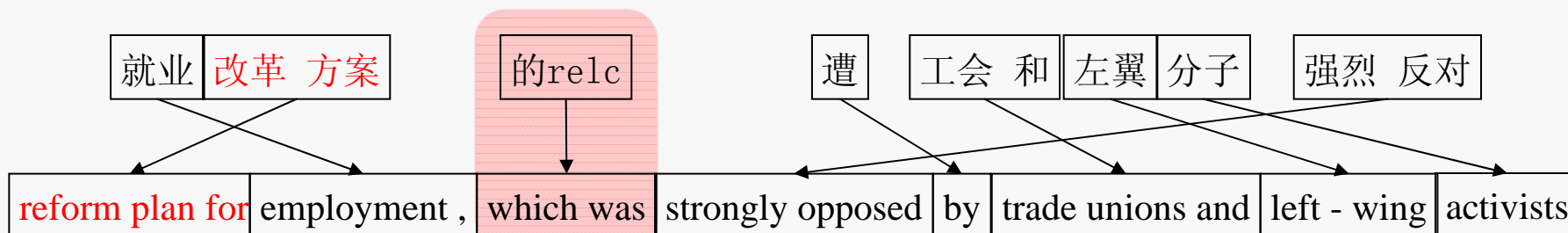
Reference: employment reform plan that was strongly opposed by the labor union

Annotated DE helps in MT systems

- Translation from original text



- Translation from DE-annotated and reordered text



- Reordering helps longer-distance distortion
- 的_{relc} captured the relative clause translation



MT experimental results

- BASELINE: standard Moses training
- WANG-NP: text reordered by NP rules in (Wang et al 07)
- DE-Annotated: with DE annotated and reordered text

	MT06(tune)	MT02	MT03	MT05
BASELINE	32.4	32.5	32.8	31.4
WANG-NP	32.8 (+0.4)	32.7 (+0.2)	33.2 (+0.4)	31.7 (+0.3)
DE-Annotated	33.4 (+1.0)	33.8 (+1.3)	33.6 (+0.8)	32.9 (+1.5)



Summary: disambiguating 的 (DE) helps MT

- Syntactic, semantic, lexical and discourse context are useful for disambiguating 的.
- Disambiguating “的” helps MT performance.
 - Put the word order closer to English
 - Mark the DEs based on the most likely translation



Conclusion

- Source-side linguistic processing helps Chinese-English MT
- Segmentation
 - Consistency
 - Word granularity best matched with English
- Chinese Grammatical Relations
 - Helps reordering models decide word orders
- 的 (DE) disambiguation
 - Understanding the most common Chinese word helps MT
 - Might be worth investigating other Chinese specific function words