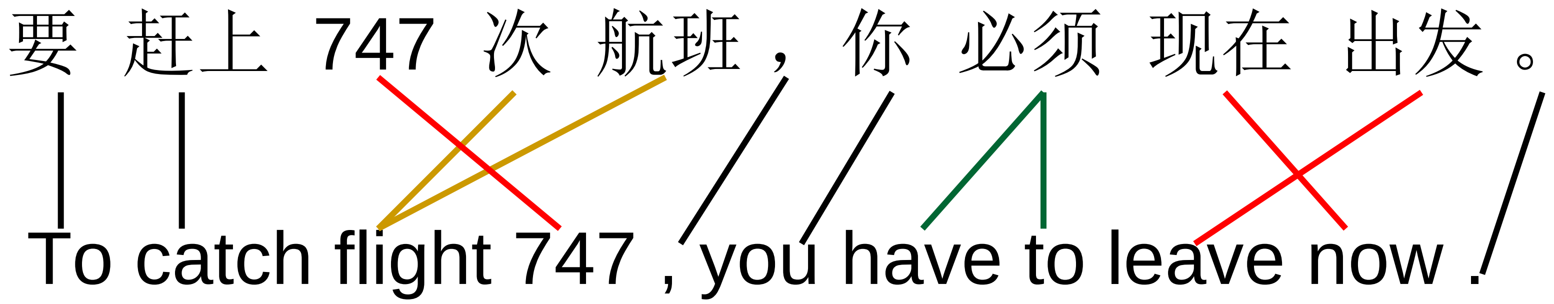
Constrained Word Alignment Models for Statistical Machine Translation

Yanjun Ma
Centre for Next Generation Localization and School of Computing
Dublin City University
Supervisor: Prof. Andy Way

Word alignment

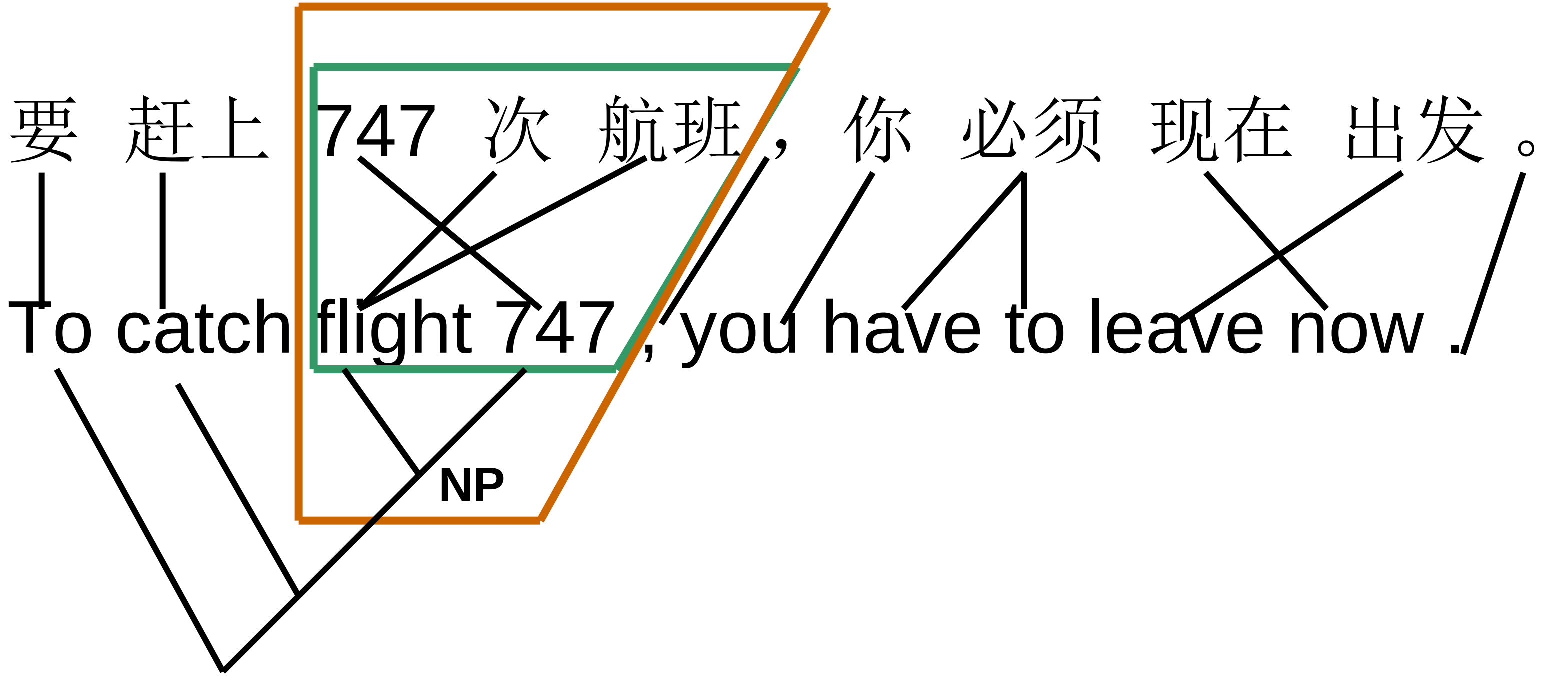


要 赶上 747 次 航班 ， 你 必须 现在 出发 。

To catch flight 747 , you have to leave now .

- The translation between two languages are secretly encoded in word alignment
- A fundamental component underpinning the success of Statistical Machine Translation

Word alignment


要 赶上 747 次 航班 ， 你 必须 现在 出发 。
To catch flight 747 , you have to leave now .
NP

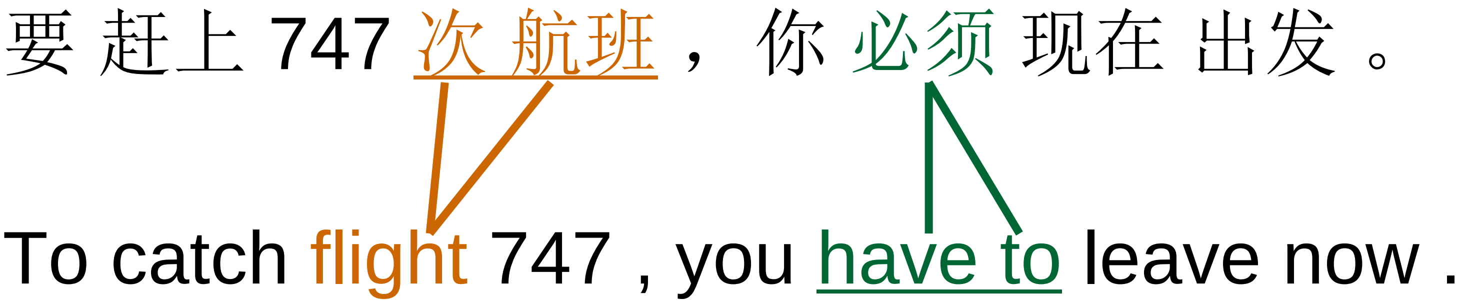
- Quality is key
- Alignment is complex process, linguistically motivated, fine-grained constraints can improve the quality.

Constrained alignment models

- Lexical constraints: bootstrapping word alignment via word packing (ACL 07; ACM TALIP 09; EACL09)
- Syntactic constraints: discriminative word alignment with syntactic dependencies (ACL08-SSST-2; EAMT09)
- Syntactic constraints: syntactically constrained HMM word-to-phrase alignment (forthcoming)

Lexical constraints

- One-to-many correspondences

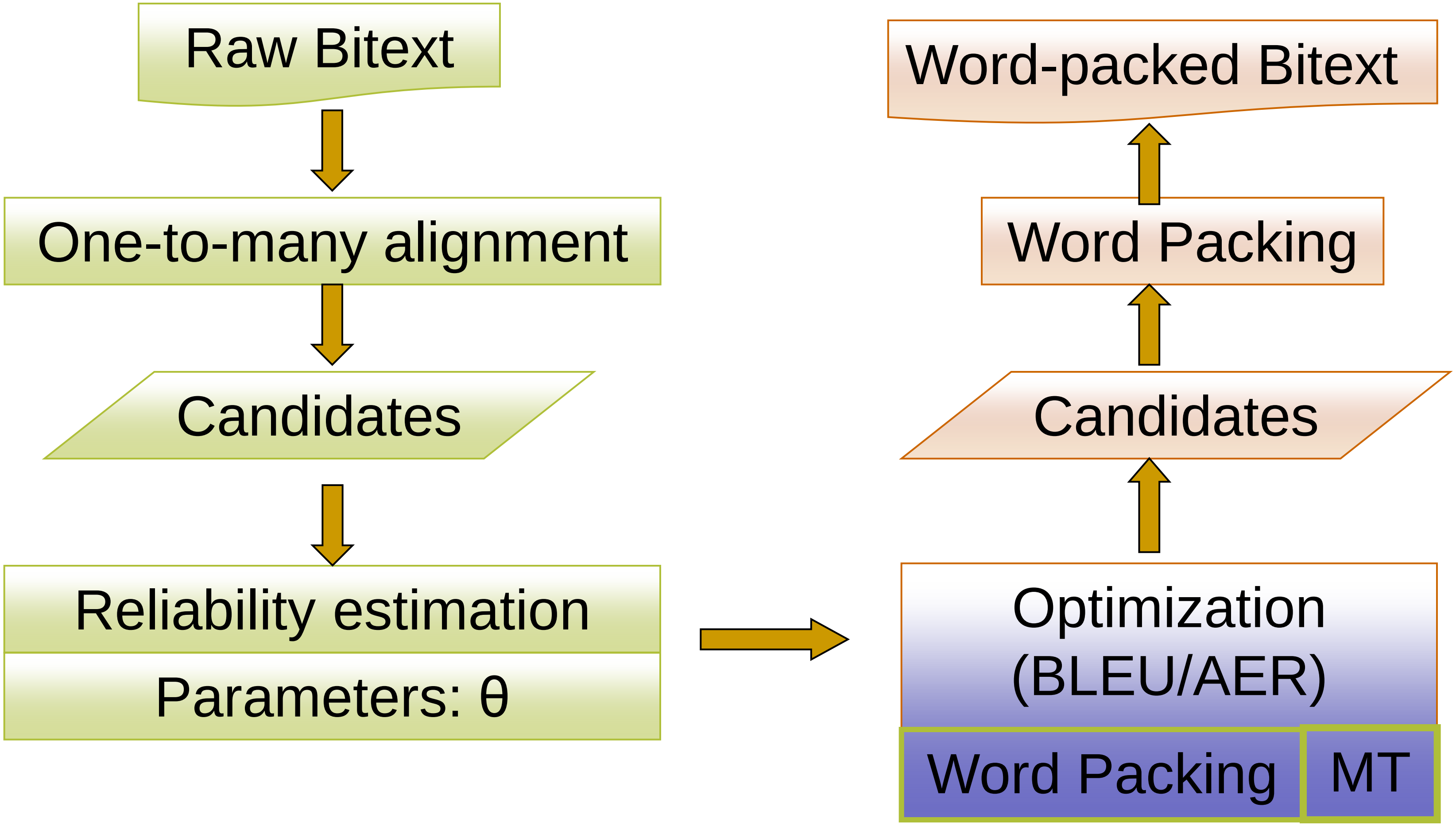


要 赶上 747 次 航班 ， 你 必须 现在 出发 。

To catch flight 747 , you have to leave now .

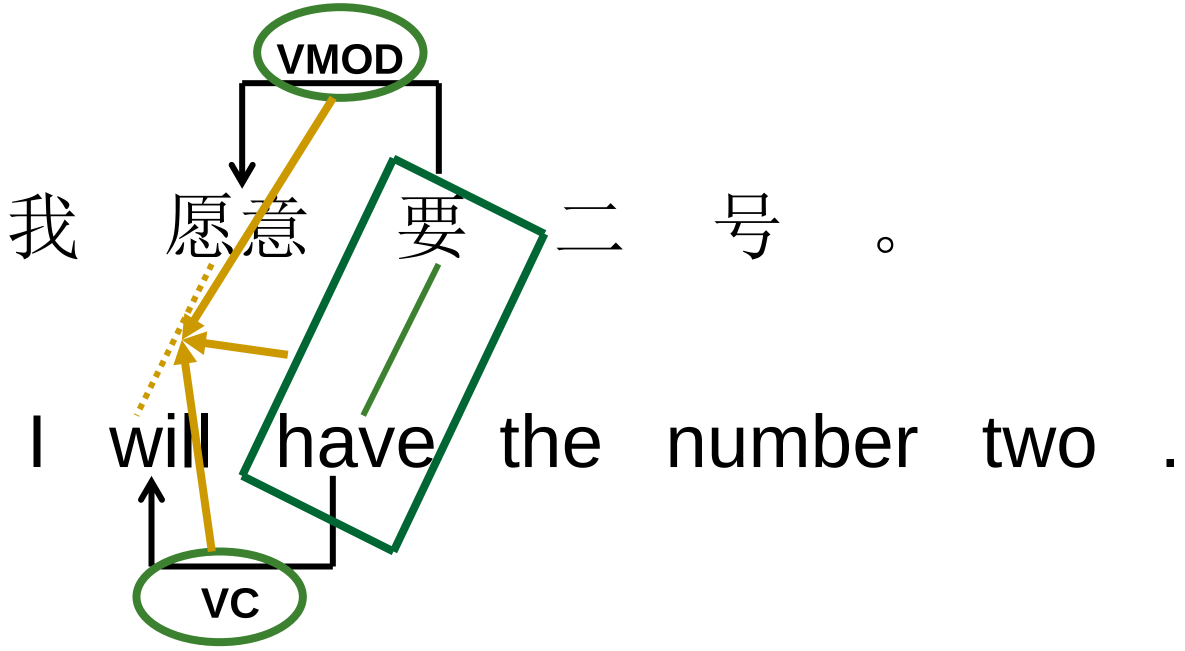
- Pack multiple consecutive words into one?

Word packing algorithm



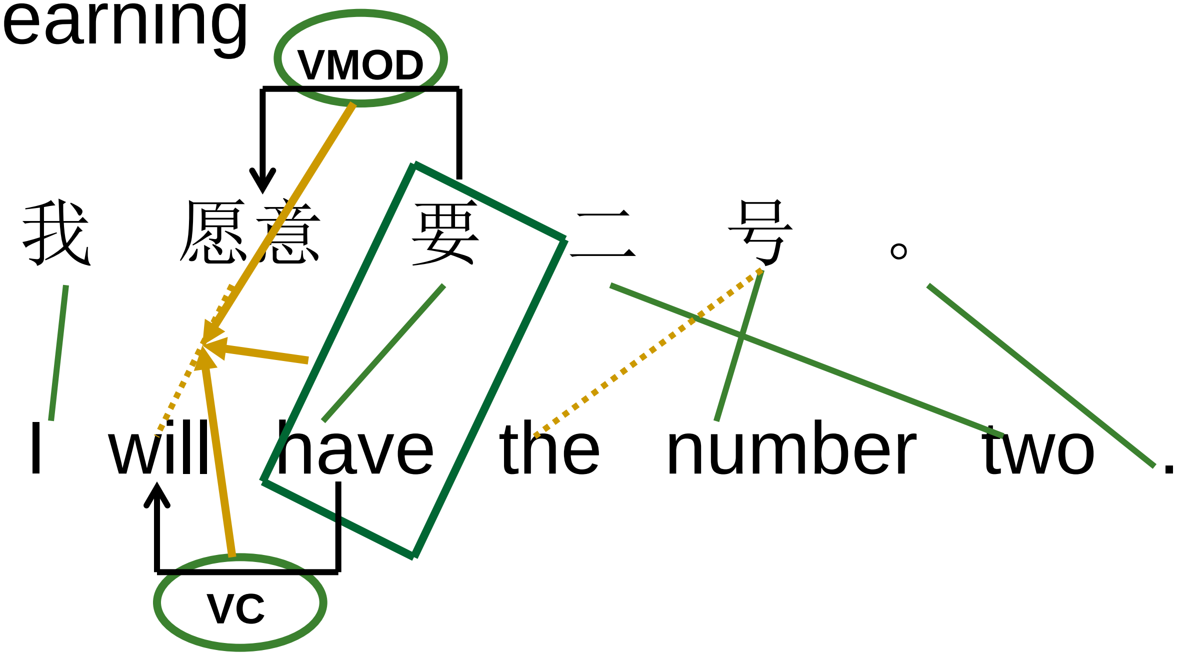
Syntactic constraints (both sides)

- Syntactic dependencies between words



Syntactic constraints (both sides)

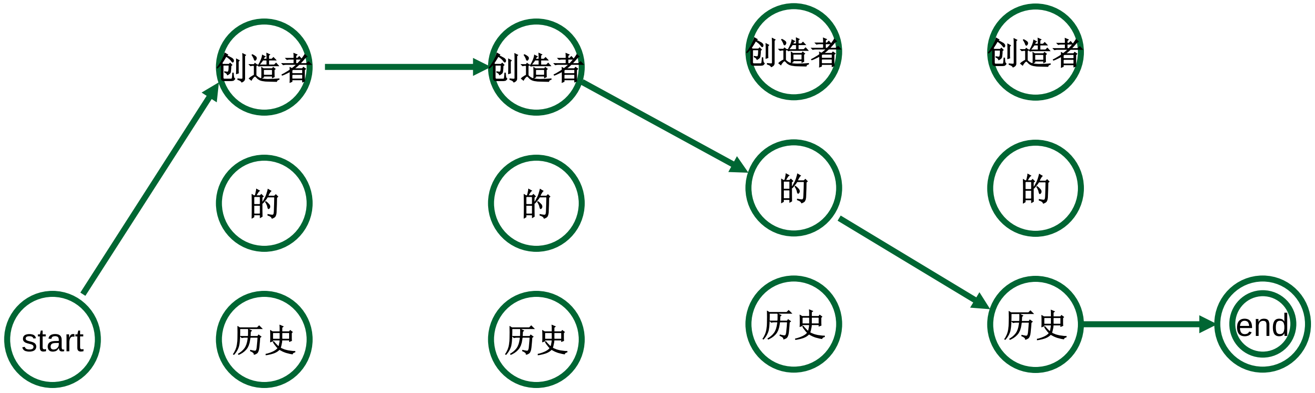
- A two-phase framework
  - Anchor word alignment
  - Non-anchor word alignment: discriminative learning



Syntactic constraints (target side)

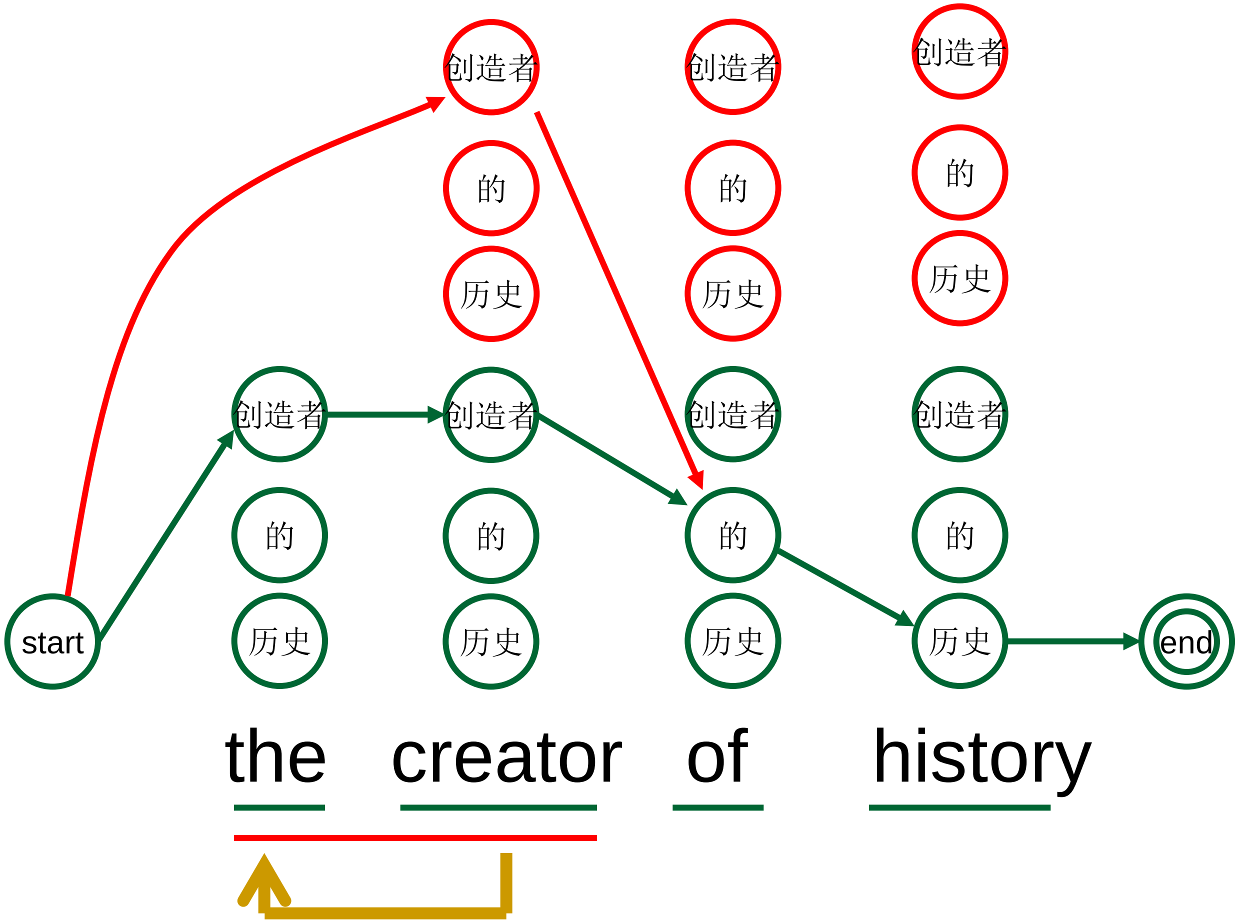
- HMM alignment model

历史　的　创造者



the　creator　of　history

历史 的 创造者



the creator of history

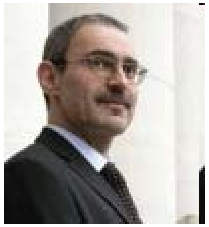
How these models work

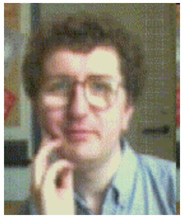
- Automatic evaluation, e.g. BLEU
- Lexical constraints (word packing): 5.44% relative improvement over IBM model 4
- Syntactic constraints for discriminative model: 5.41% relative improvement over IBM model 4
- Syntactic constraints for generative model: consistent gains over baseline HMM word-to-phrase alignment model (2.82% relative improvement over IBM model 4)

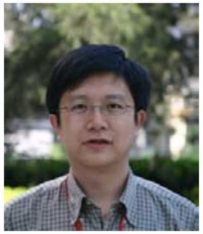
Conclusions

- Adding linguistically motivated, fine-grained constraints can boost the performance of alignment models
- However, for long sentences and/or radically different language pairs, the quality of word alignment is still far from satisfactory

Thank you!







Among others…

Results I

- Automatic evaluation (IWSLT 2007)

System	BLEU
Baseline (IBM Model 4)	33.85
Word Packing step 1	35.02
Word Packing step 2	35.69

- Translation example

在 巴黎 出 了 交通 事故 。

Gloss:	*in Paris happen traffic accident*
Reference:	I was involved in a traffic accident in Paris .
Baseline:	In Paris out a traffic accident .
Word Packing:	In Paris there is a traffic accident .

Results II

- Automatic evaluation (IWSLT 2007)

System	BLEU
Baseline (IBM Model 4)	33.85
Syntactic	35.67

- Translation example

	您 是 在 这儿 用餐 还是 带走 ？
Gloss:	*you are here eat or take-out*
Reference:	Is that for here or take-out ?
Baseline:	Are you here meal or take-out ?
Syntactic:	Are you eat here or take it out ?

Results III

- Automatic evaluation (NIST2006)

System	Small Data Set	Large Data Set
Baseline (HMM word-to-phrase)	14.18	26.09
+Syntactic constraints	14.64	26.24
IBM Model 4	14.58	25.52

- Translation example

南非 太空 观光 客 结束 太空 之 旅 返 抵 地球

Gloss: *south africa space tourist end space tour* back to earth

Reference: The South African space tourist back to earth after his space travel

Baseline: The South African space space trip to the visitors’ end backed earth

+Syntactic: The South African space tourism’ end space trip back to benefit the earth