

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。

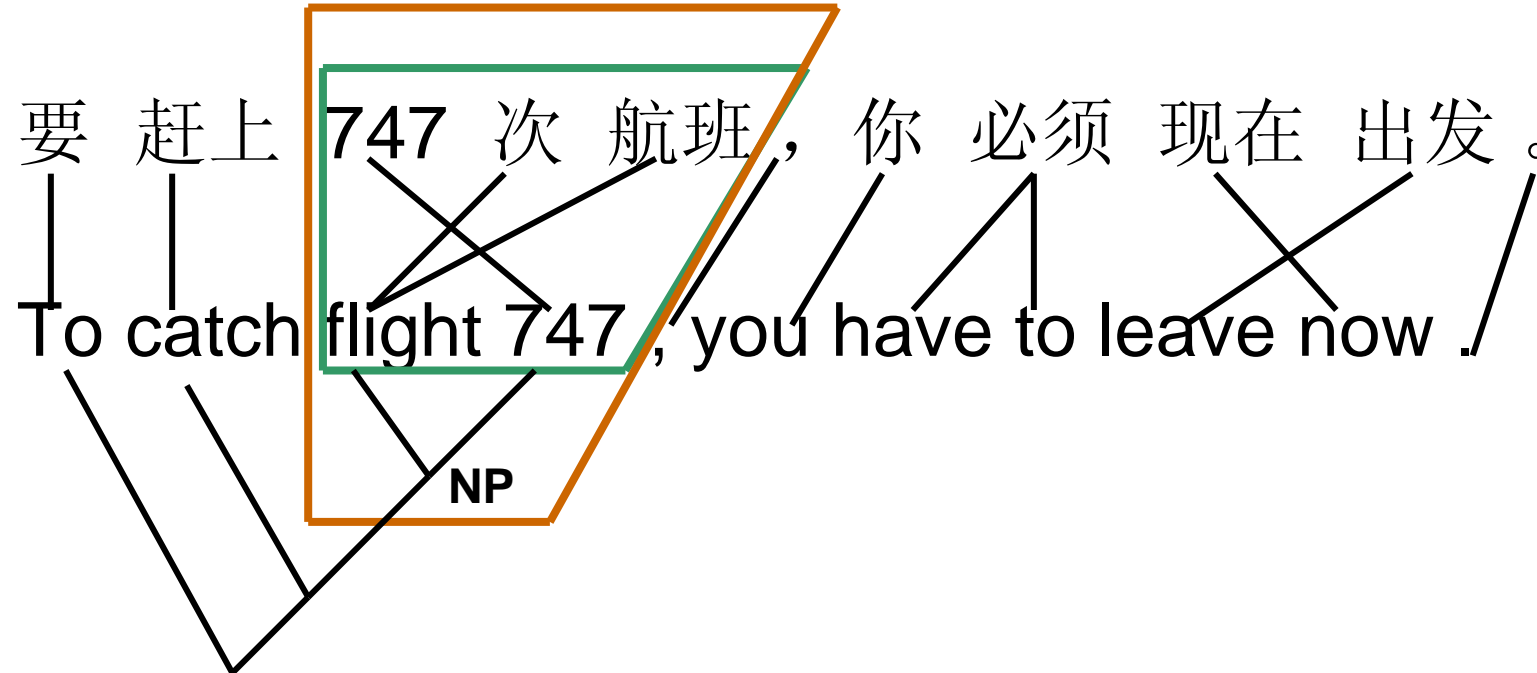
To catch flight 747 , you have to leave now !

A diagram illustrating word alignment between the Chinese sentence '要 赶上 747 次 航班， 你 必须 现在 出发。' and the English translation 'To catch flight 747 , you have to leave now !'. Vertical lines connect corresponding words: '要' to 'To', '赶上' to 'catch', '747' to '747', '次' to ',', '航班' to 'flight', '你' to 'you', '必须' to 'have to', '现在' to 'leave', and '出发' to 'now'. A red 'X' is drawn over the alignment between '747' and '747', and another red 'X' is drawn over the alignment between '现在' and 'now'. A green line connects '必须' to 'have to', and a yellow line connects '航班' to 'flight'.

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



Word alignment



- 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

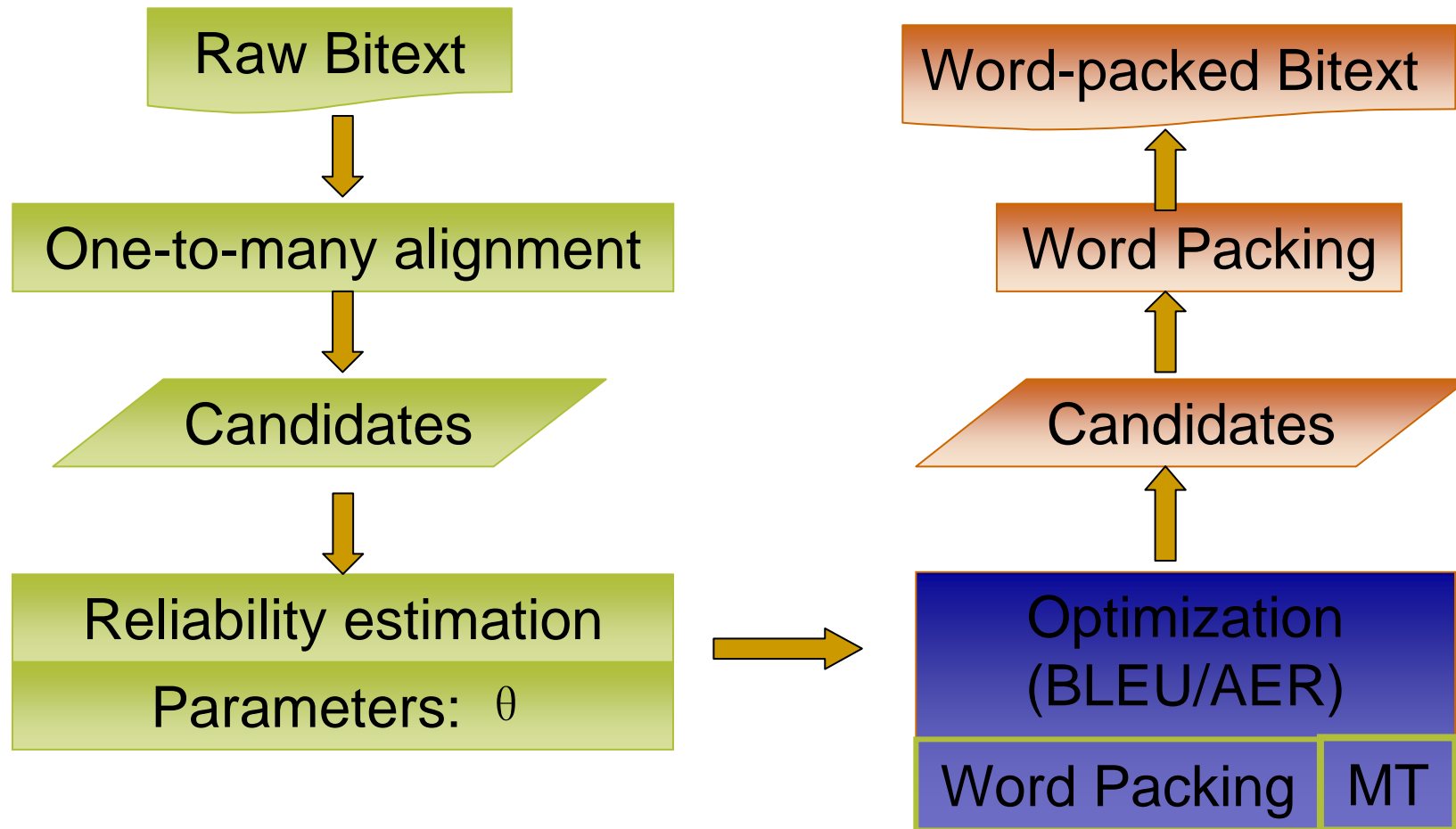
要赶上 747 次航班，你 必须 现在出发。

To catch flight 747, you have to leave now.

- Pack multiple consecutive words into one?

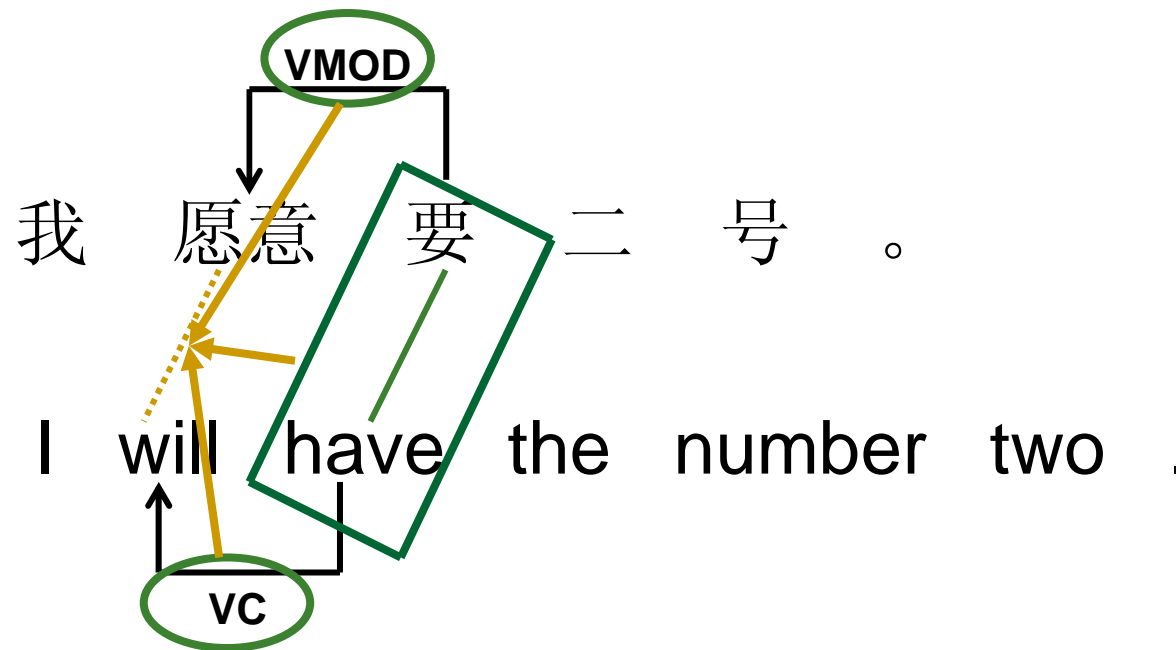


Word packing algorithm



Syntactic constraints (both sides)

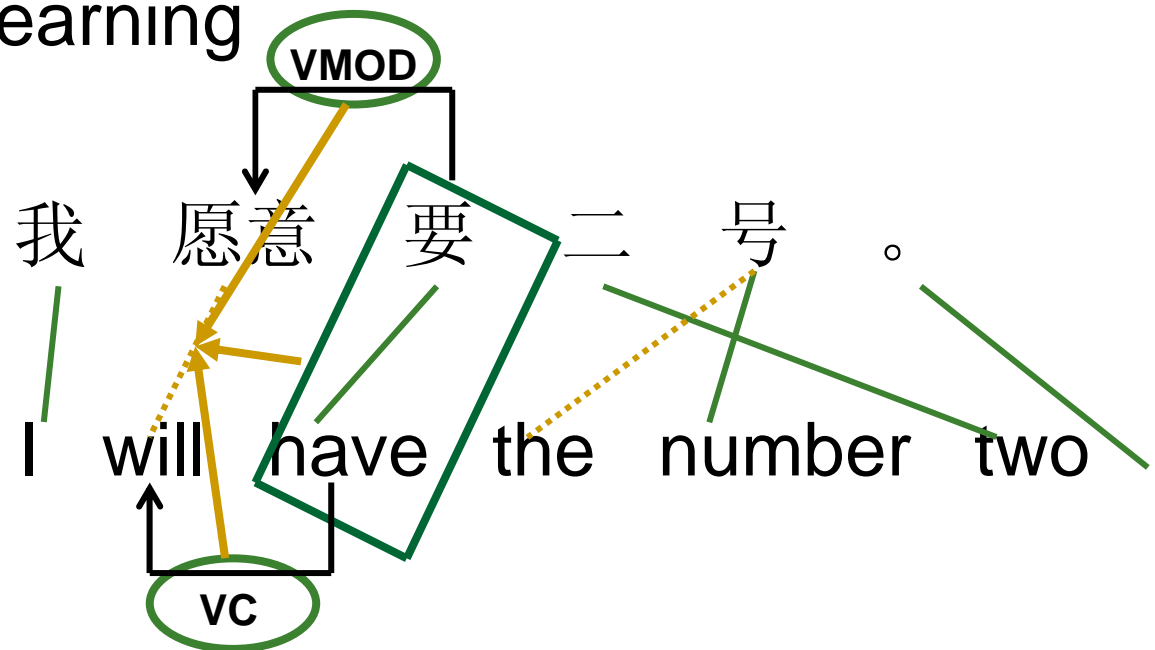
- Syntactic dependencies between words



Syntactic constraints (both sides)



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

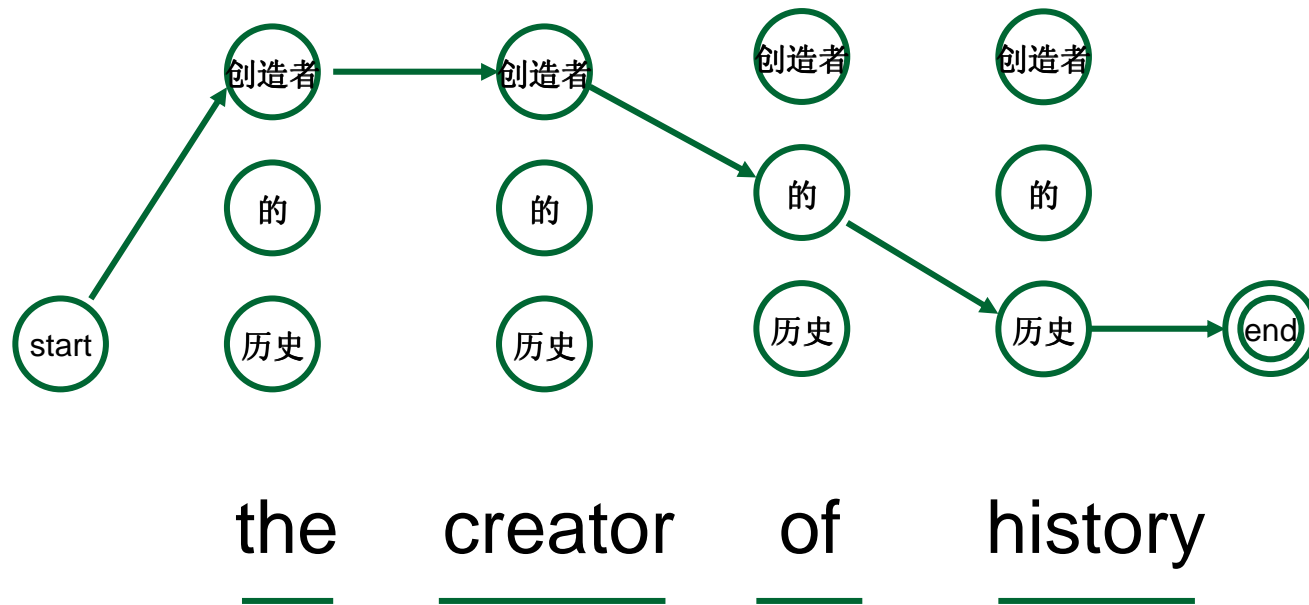


Syntactic constraints (target side)

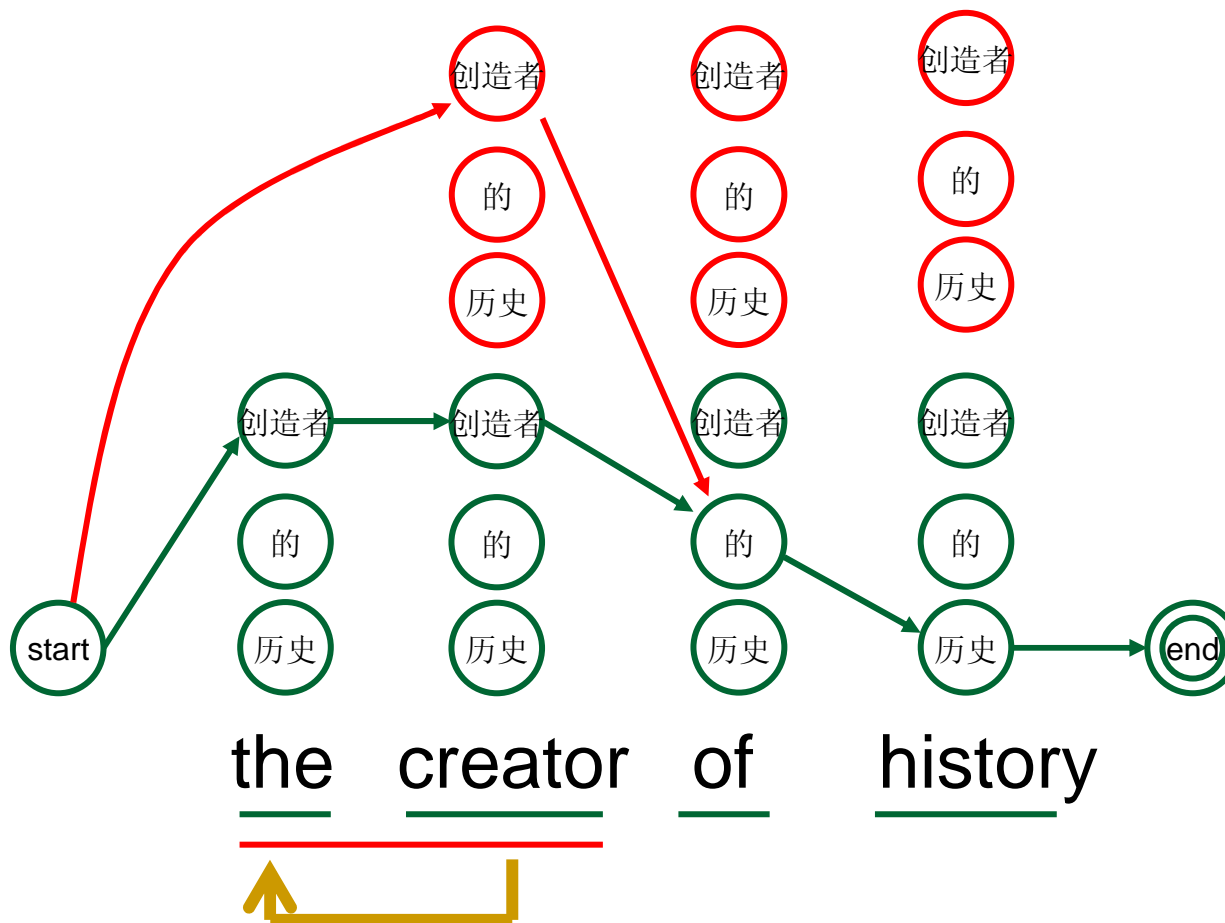


■ HMM alignment model

历史的创造者



历史的创造者



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)



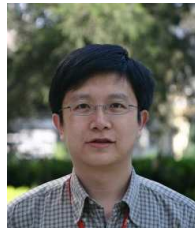
Conclusions



- 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...

