Introduction to

Machine Translation

Mu Li
Microsoft Research Asia

Introduction to Statistical Machine Translation

Mu Li
Microsoft Research Asia

Outline

- Machine translation overview
- Fundamental of SMT
- SMT Models
- SMT model training
- MT evaluation

An Example of Machine Translation

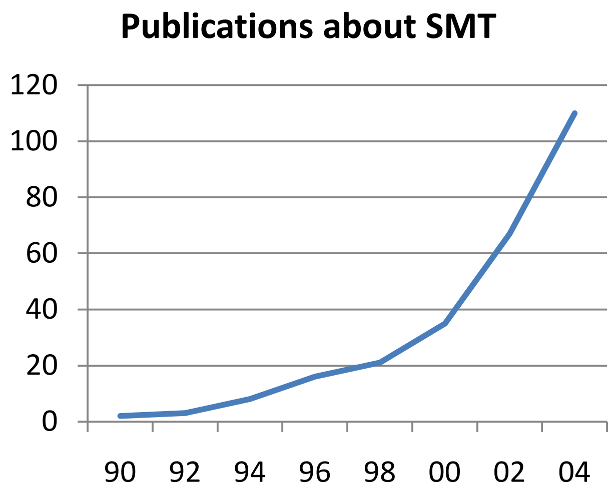
Definition of Machine Translation

- Machine Translation (MT)
  - Translate one language to another with computers
    - Mostly working at sentence level
  - Cheap way of access cross-lingual information
  - Classic problem of natural language processing research

- Application scenarios
  - Fully Automatic Machine Translation (全自动机器翻译)
  - Human Assisted Machine Translation (人助机译)
  - Computer Aided Translation (机助人译)

History of Machine Translation Research

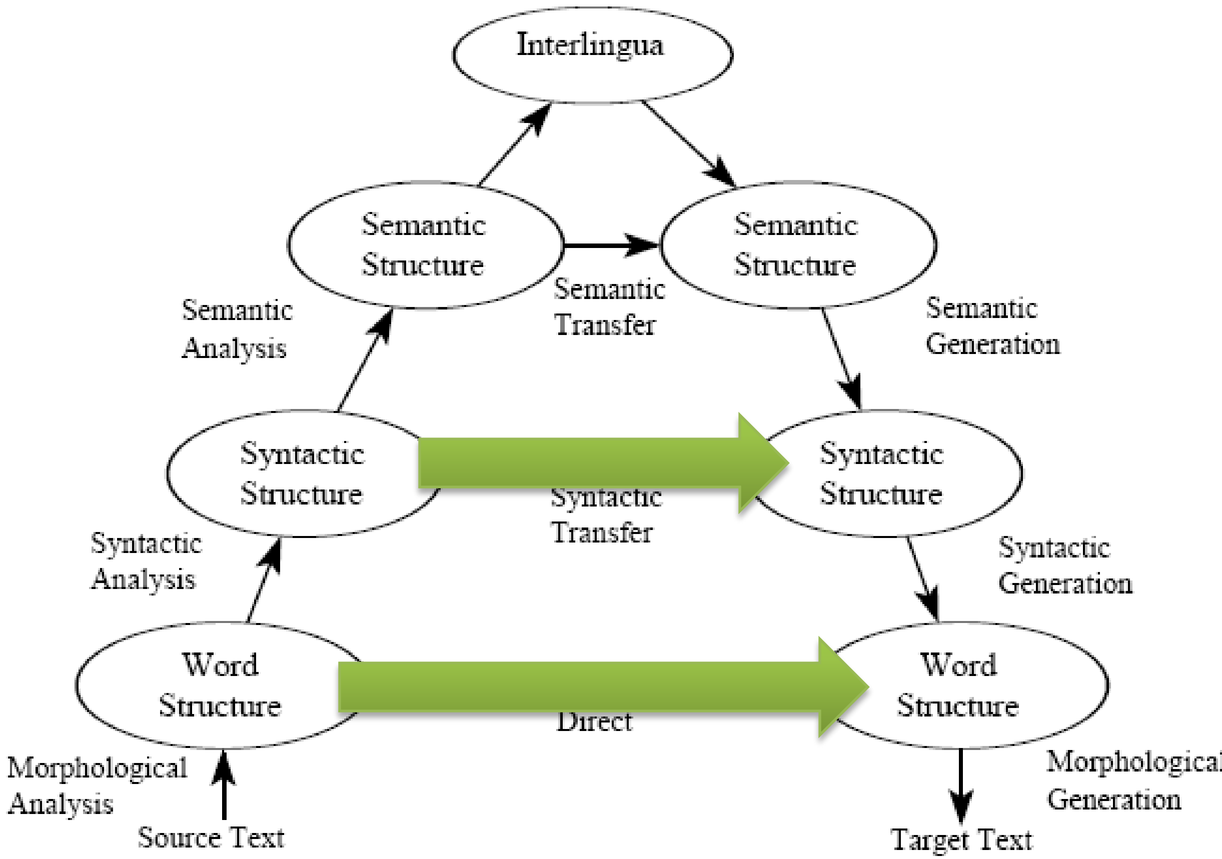
- 1946 –1954
  - The first MT system in Georgetown
  - Russian-English, 6 rules， 250 words， 50 sentences
- 1966
  - ALPAC Report
- 1970 – 1980
  - Fundamental research on natural language theory
    - Unification grammar, semantic network
- 1980 – 1990
  - Commercialized of rule-based MT systems
    - SYSTRAN
- 1988
  - Candie system @ IBM
- 1990 – 2000
  - Pioneer work on statistical machine translation
- 2000 – 2011
  - World-wide interest in statistical method in machine translation
  - Web service for machine translation using large scale data



Machine Translation Technologies

- RBMT (Rule-Based MT)
  - Word-by-word translation
  - Grammar-based direct transfer
  - Interlingua-based method
- EBMT (Example-Based MT)
  - Example as skeleton, top-down translation
- SMT (Statistical MT)
  - Assemble of translation units, bottom-up translation

Machine Translation Pyramid



Machine Translation Technologies

- RBMT (Rule-Based MT)
  - Word-by-word translation
  - Grammar-based direct transfer
  - Interlingua-based method
- EBMT (Example-Based MT)
  - Example as skeleton, top-down translation
- SMT (Statistical MT)
  - Assemble of translation units, bottom-up translation

RBMT: Source Language Analysis

她把一束花放在桌上。 ⇨ She put a bunch of flowers on the table.

⇩

Segmenter/POS tagger

⇩

她/r 把/p-q-v-n 一/m-d 束/q 花/n-v-a 放/v 在/p-d-v 桌/n 上/f-v 。/w

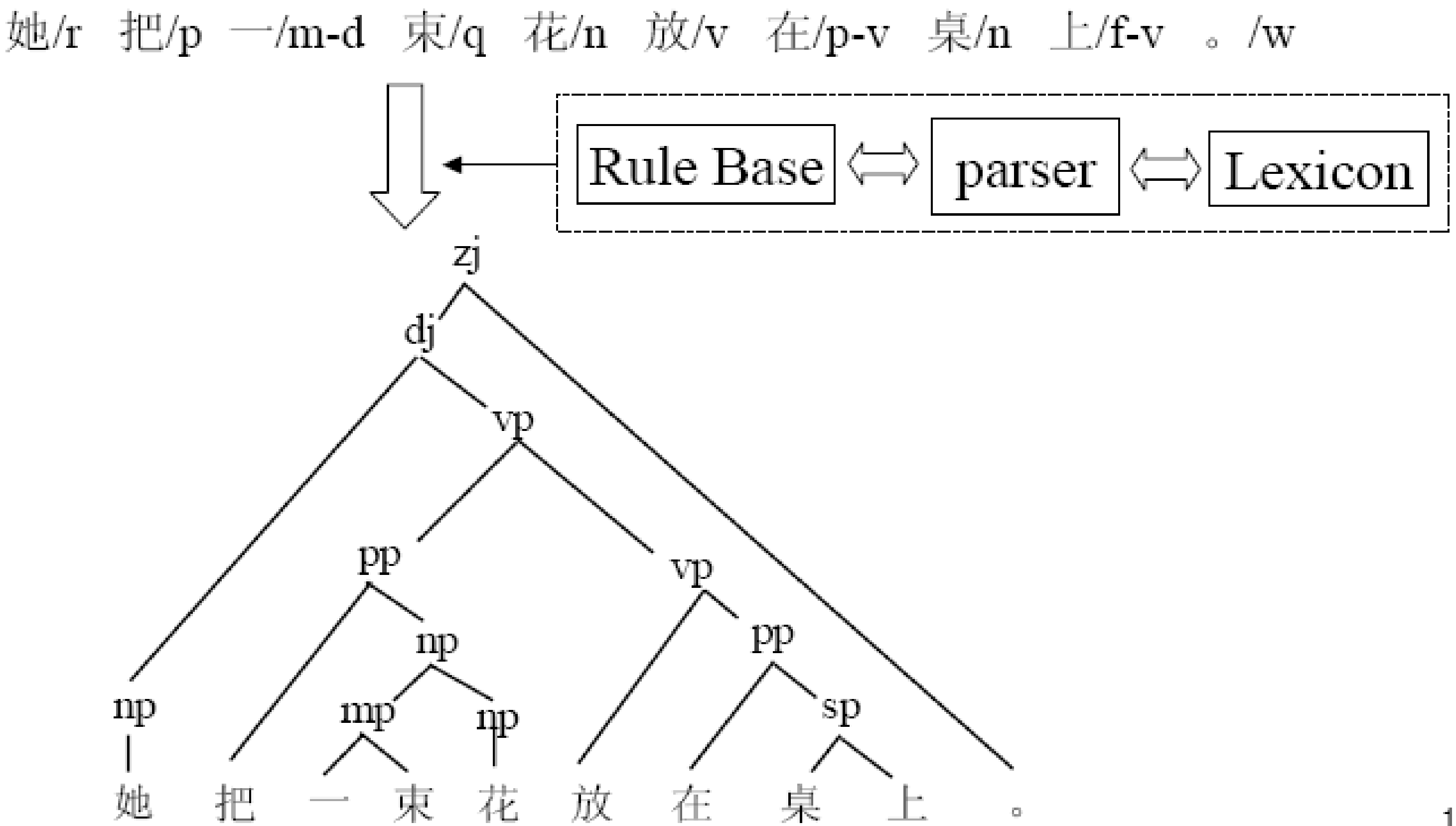
⇩

filter

⇩

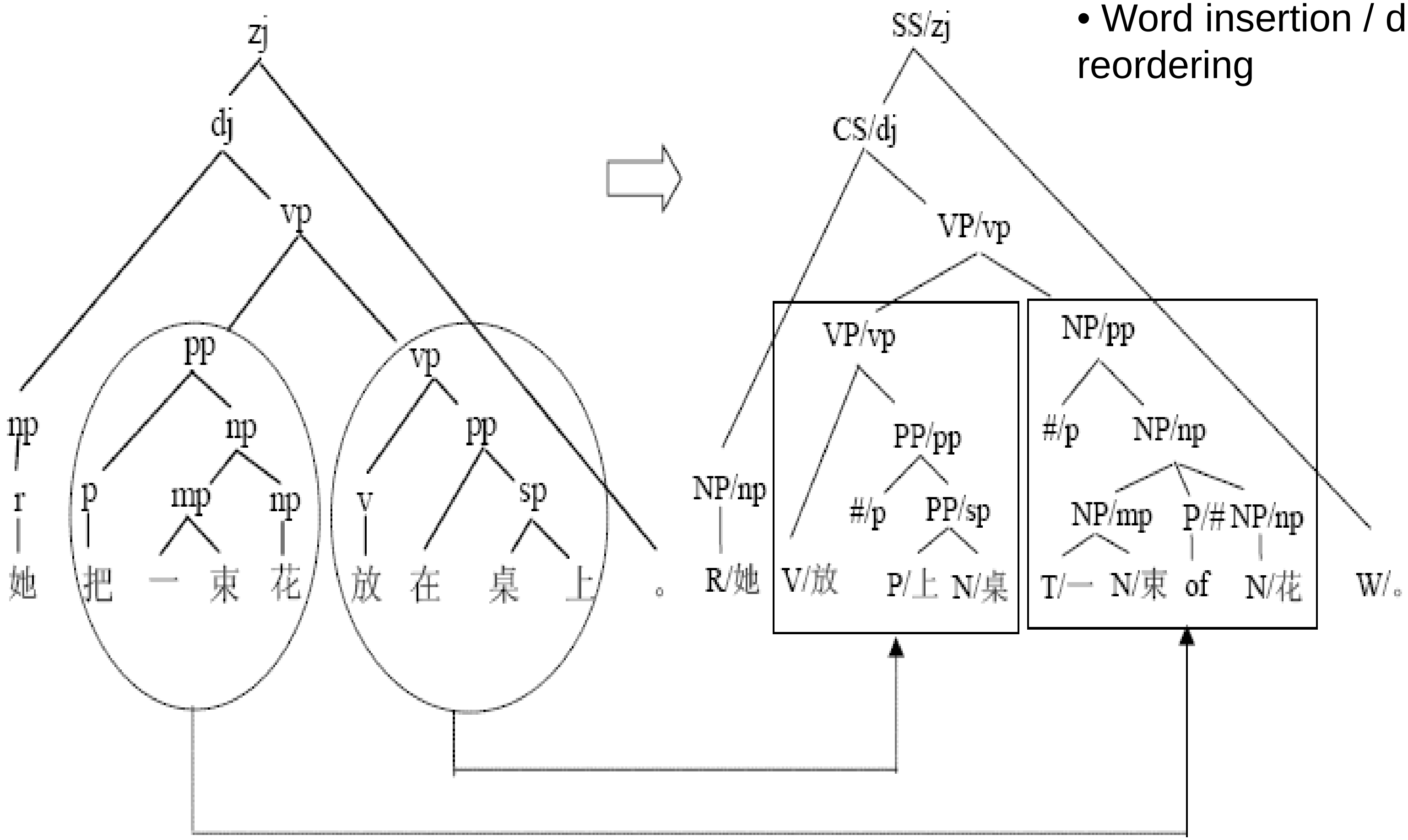
她/r 把/p 一/m-d 束/q 花/n 放/v 在/p-v 桌/n 上/f-v 。/w

RBMT: Source Language Parse Tree

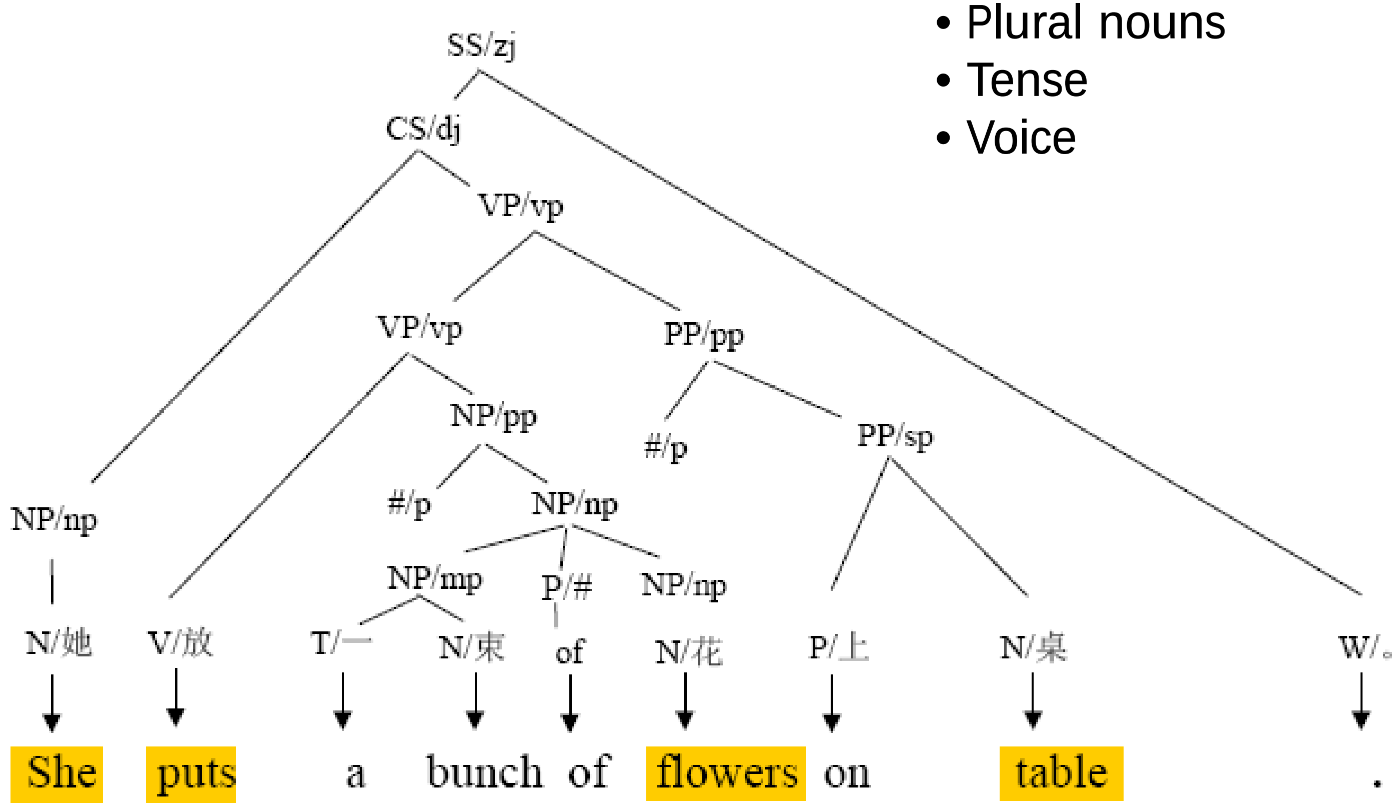


RBMT: Tree Transformation

- Phrase level
- Sentence level
- Word insertion / deletion / reordering



Target Language Generation



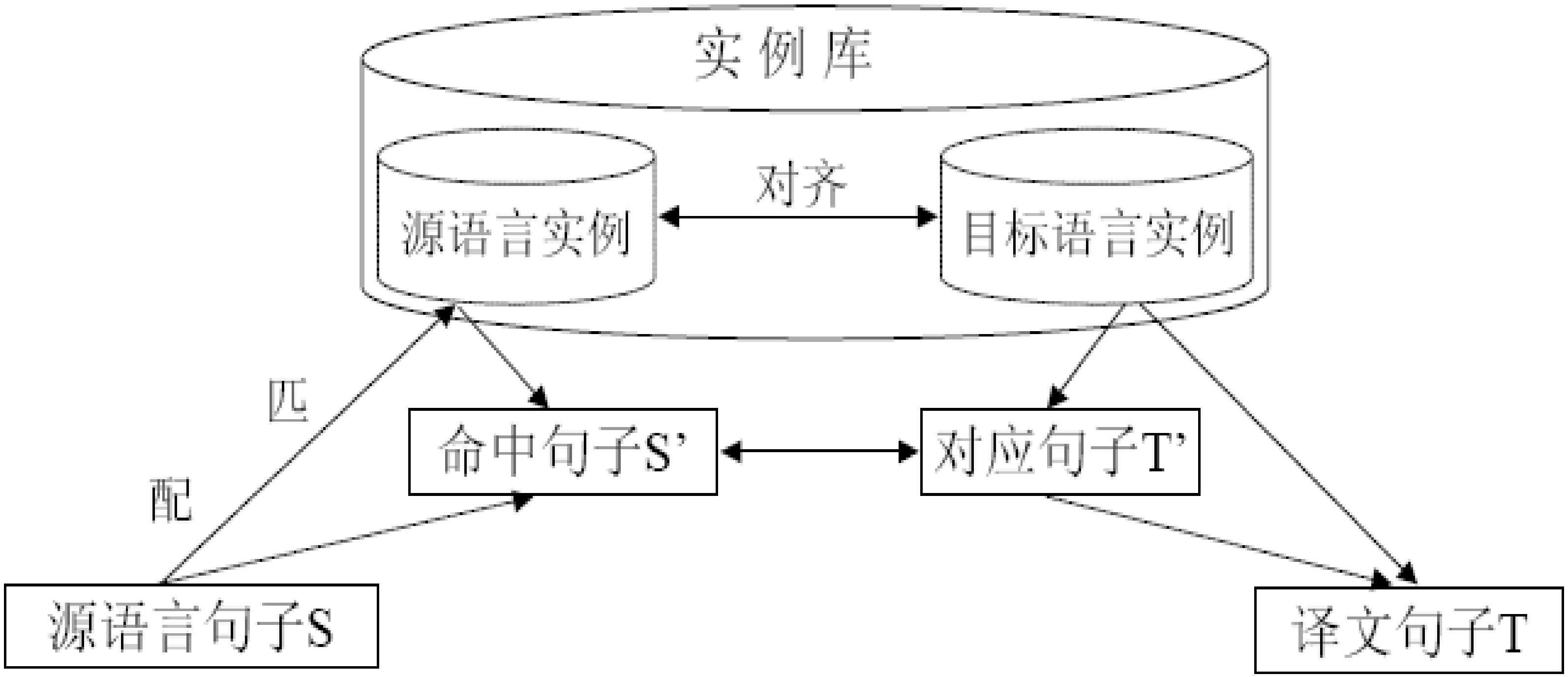
- Plural nouns
- Tense
- Voice

Comments on RBMT

- Pros: Easy to start
  - Intuitive
- Cons: Hard to improve (AI complete problem)
  - Require world knowledge
  - Knowledge-based maintenance

Example-based Machine Translation

- *Makoto Nagao*（*1984*）



Example-based Machine Translation

照猫画虎的机器翻译

英语实例	汉语实例
He eats vegetable	他吃蔬菜
Acid eats metal	酸腐蚀金属

输入： 输出：

I eat potatoes 我吃土豆

More about EBMT

- A data driven approach
- Learning from examples
  - Usually in a top-down manner
  - No principled and quantified method to choose examples
  - No principled way to find best translation

Statistical Machine Translation (SMT)

- MT as a statistical decision problem
  - Doing MT with statistical methods
    - Given a source sentence, use statistical data and metrics to decide what is the best translation
  - Learning from data
  - Quantized computation
  - Probabilistic predication
  - Herman Ney
    - SMT = linguistic modeling + statistical decision theory

SMT Chronicle

- 1988
  - IBM models
- 1999
  - JHU summer workshop
- 2002
  - Log-linear framework
  - SMT won in NIST evaluation
- 2003
  - Max BLEU training
- 2005
  - Power of web scale language model
- 2006
  - First large-scale success of syntax-based model in SMT

Why SMT?

- 知识获取瓶颈
  - 基于规则的翻译面临知识获取的困难。统计机器翻译从双语对照文本中自动学习翻译知识
  - 虽然在建立统计机器翻译模型时要花费很大的人力，但是在开拓一个新语言对的时候，代价相对基于规则的方法要小很多。
- 知识表达的颗粒度
  - 由于统计机器翻译是数据驱动，可获细小颗粒度的知识并且可以获得上下文有关的约束，因此译文质量要好于粗颗粒度的基于规则的方法。
- 系统的可维护性和可扩展性
  - 规则系统利用专家手工知识比较困难。而统计方法利用数据驱动易于维护和扩展。
- 但是，如果双语的数据少，比如对某些语言对来说，双语的数据很难获得，则统计翻译方法会变得无效。那时，基于规则的方法要好很多。

名人名言

- A word is a world.
  - **Douglas Lenat, founder of CYC project**

- It must be recognized that the notion “probability of a sentence” is an entirely useless one, under any known interpretation of this term.
  - **Noam Chomsky, 1969**

- Whenever I fire a linguist our system performance improves.
  - **Frederick Jelinek, 1988**

Translation as Decoding

人们会自然地认为翻译的问题实际上可以看作是一个密码破译的问题。当我看到一本用俄语写的书，我认为它实际上是英语写的，只不过是用一些奇快的符号编码。我只要想想如何破解即可。

Warren Weaver, 1947

Fundamental Problems of SMT

- Modeling
  - What translation to find
- Searching / decoding
  - How to find translation
- Training / learning
  - Model parameter estimation

An Old Story – Source-Channel Modeling

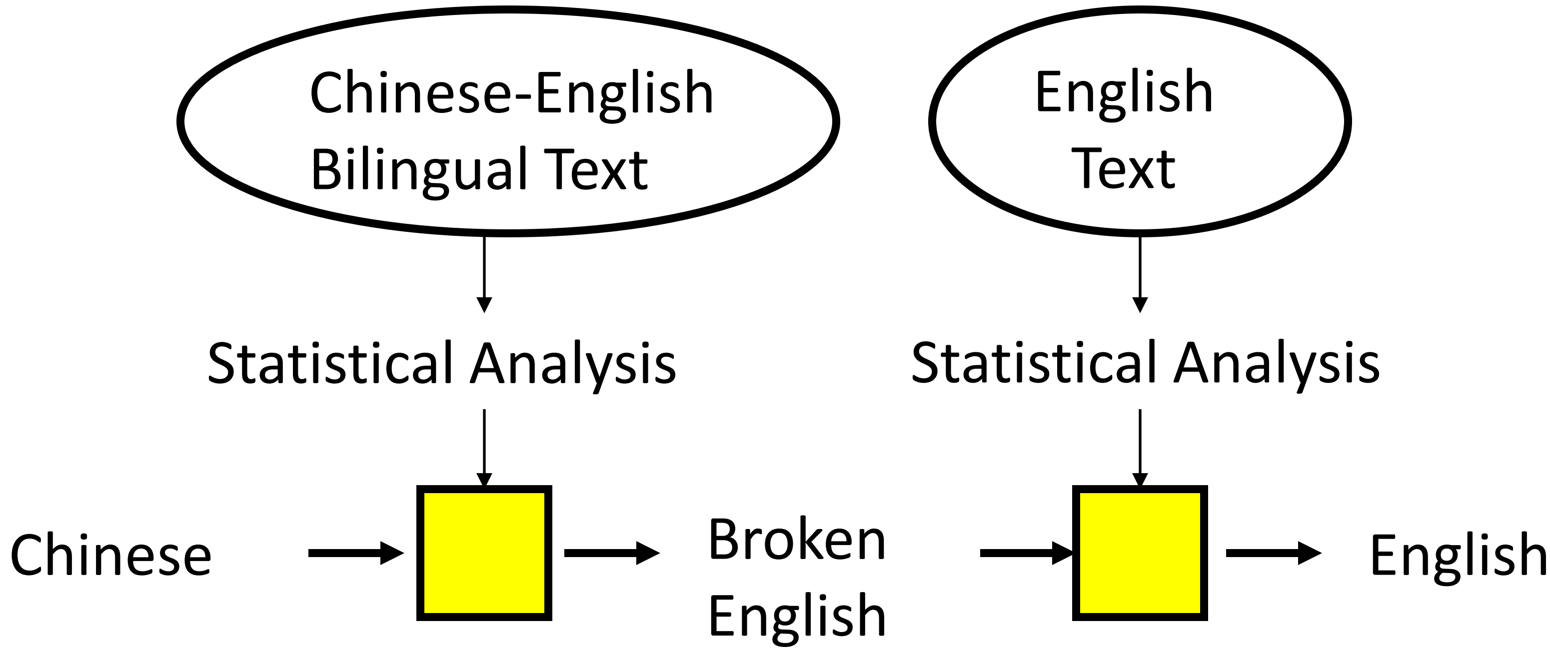
$$p(e|f) \sim P(\cdot|\cdot)$$

$$e^* = \mathrm{argmax}_e \; P(e|f)$$

$$P(e|f) = \frac{P(e)P(f|e)}{P(f)}$$

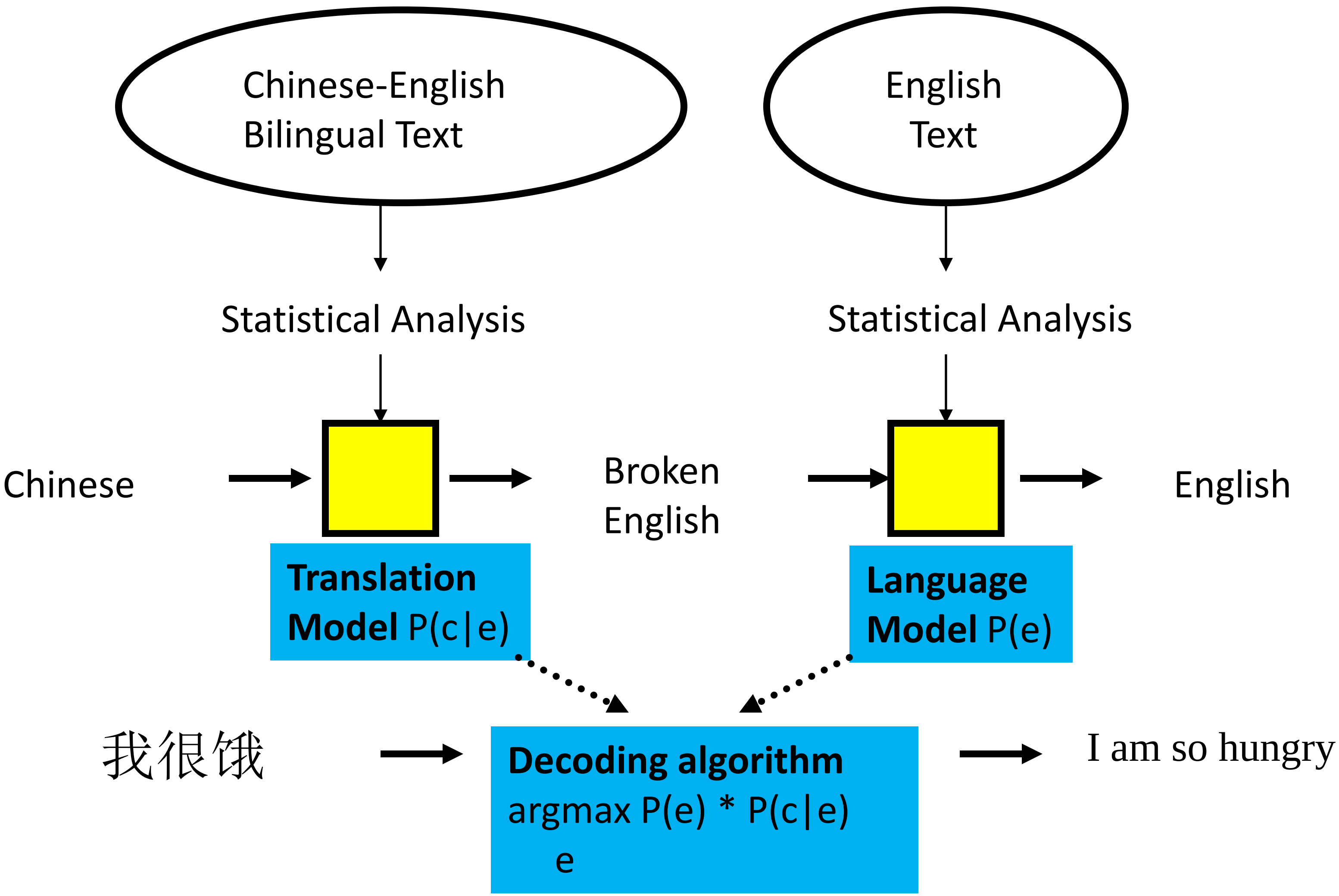
$$e^* = \mathrm{argmax}_e \; \underbrace{P(e)}_{\substack{\text{Language Model} \\ \text{(Source Model)}}} \; \underbrace{P(f|e)}_{\substack{\text{Translation Model} \\ \text{(Channel Model)}}}$$

Source-Channel SMT System



我很饿 → What hunger have I,
Hungry I am so,
I am so hungry,
Have I that hunger … → I am so hungry

Source-Channel SMT System



More on Modeling

- N-gram language model
  - $\boldsymbol{e} = e_1^m = e_1 \dots e_m$

$$P(\boldsymbol{e}) = P(e_1)P(e_2|e_1)\dots P(e_{n-1}|e_1 \dots e_{n-2}) \prod_{i=n}^{m} P(e_i|e_{i-n+1} \dots e_{i-1})$$

- Translation model
  - $P(\boldsymbol{f}|\boldsymbol{e}) = \prod P(f_j|e_{a_j})$

More on Language Model

Site	BLEU
Google	0.3531
ISI	0.3073

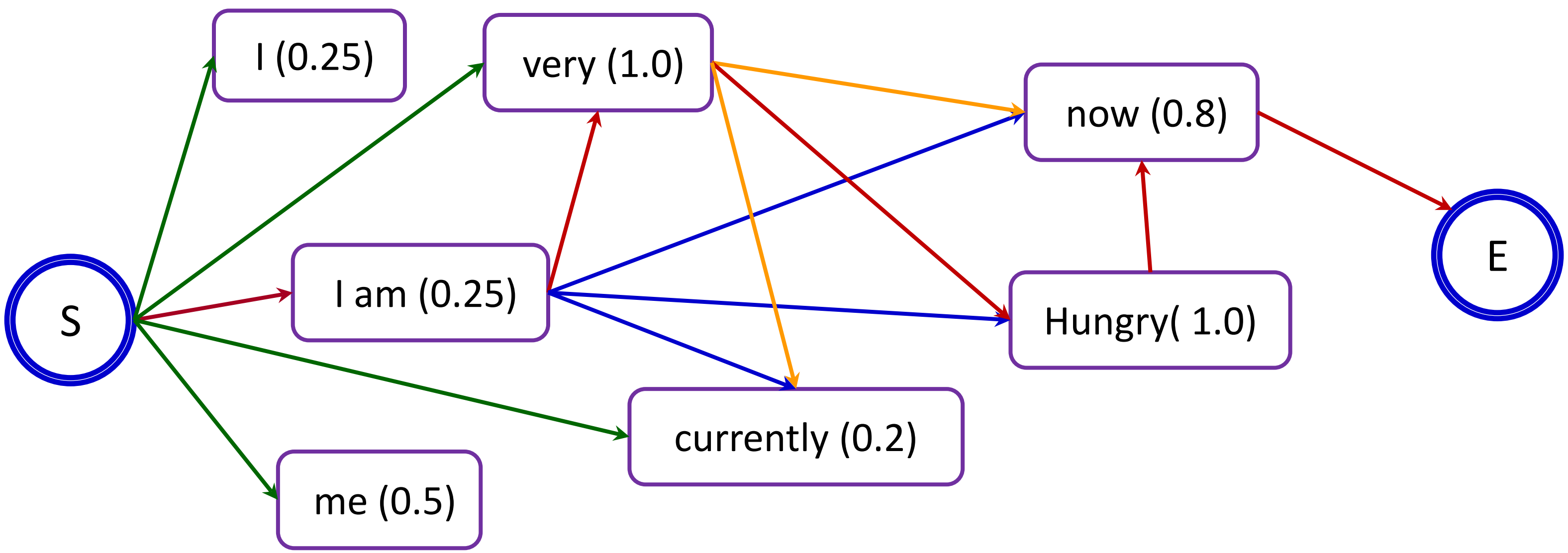
NIST 2005 Chinese-English MT Evaluation results

ngram	BLEU
2	31.9
3	36.7
4	38.4
5	38.8
6	38.8

# words	BLEU
30 M	35.58
60 M	36.51
120 M	37.43
250 M	38.80
400 M	39.39

Search in Word Graph / Lattice

Input: 我 现在 很 饿

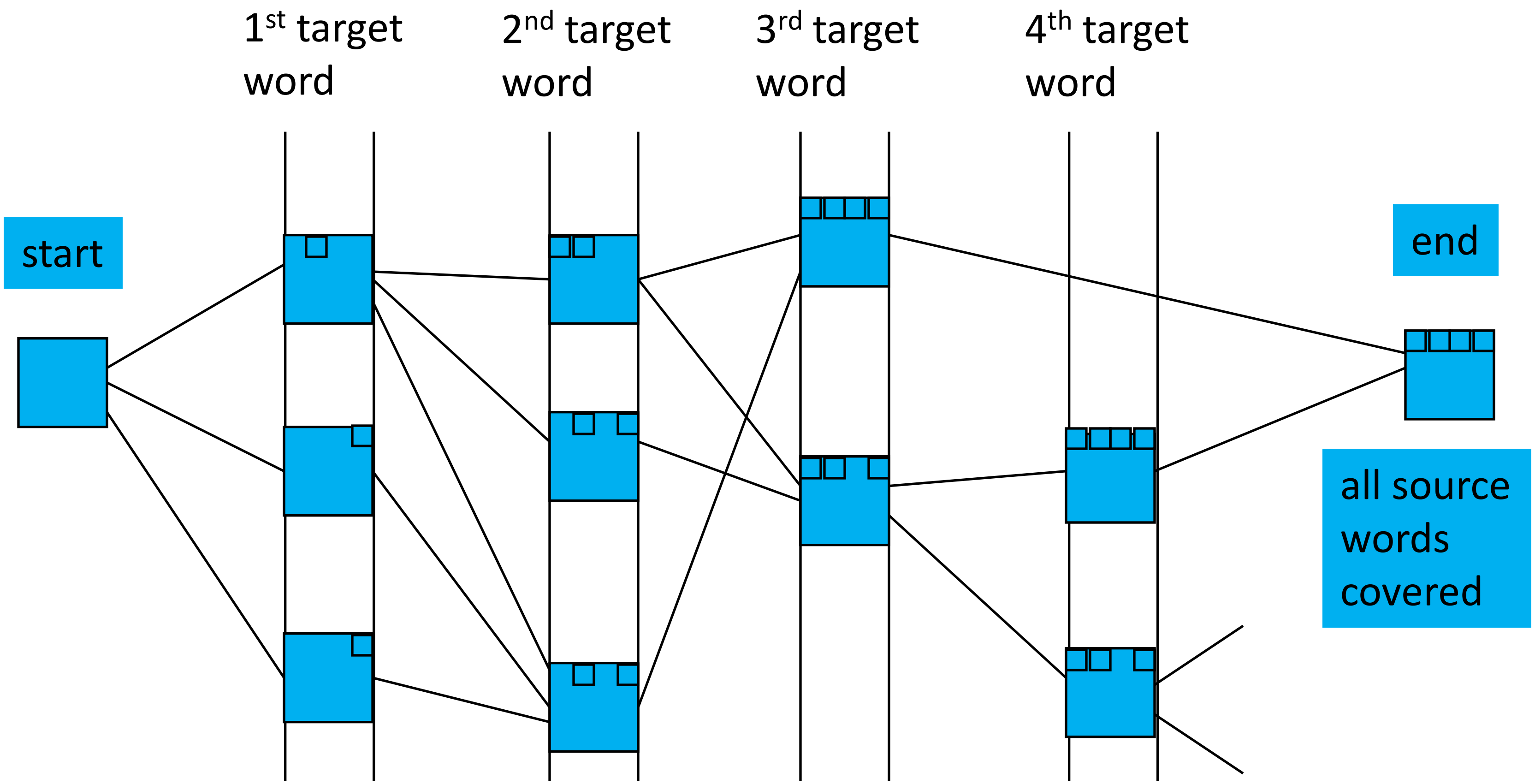


- Translation for each word as nodes
- A link exists between every two nodes not from the same source word
- Find best path from start to end node
- Cost of path determined by translation and language models

Decoding for Classic SMT Models

- Of all conceivable English word strings, find the one maximizing $P(e)P(e|f)$
- Decoding is an NP-complete challenge (Knight, 1999)
  - n! permutations for an English sentence with n words
  - Each potential English output is called a *hypothesis*.
- Several speed-up strategies are available
  - Dynamic programming
  - Histogram pruning
    - Limited number of candidates in each bucket/stack
  - Threshold pruning
    - Abandon candidates with low score

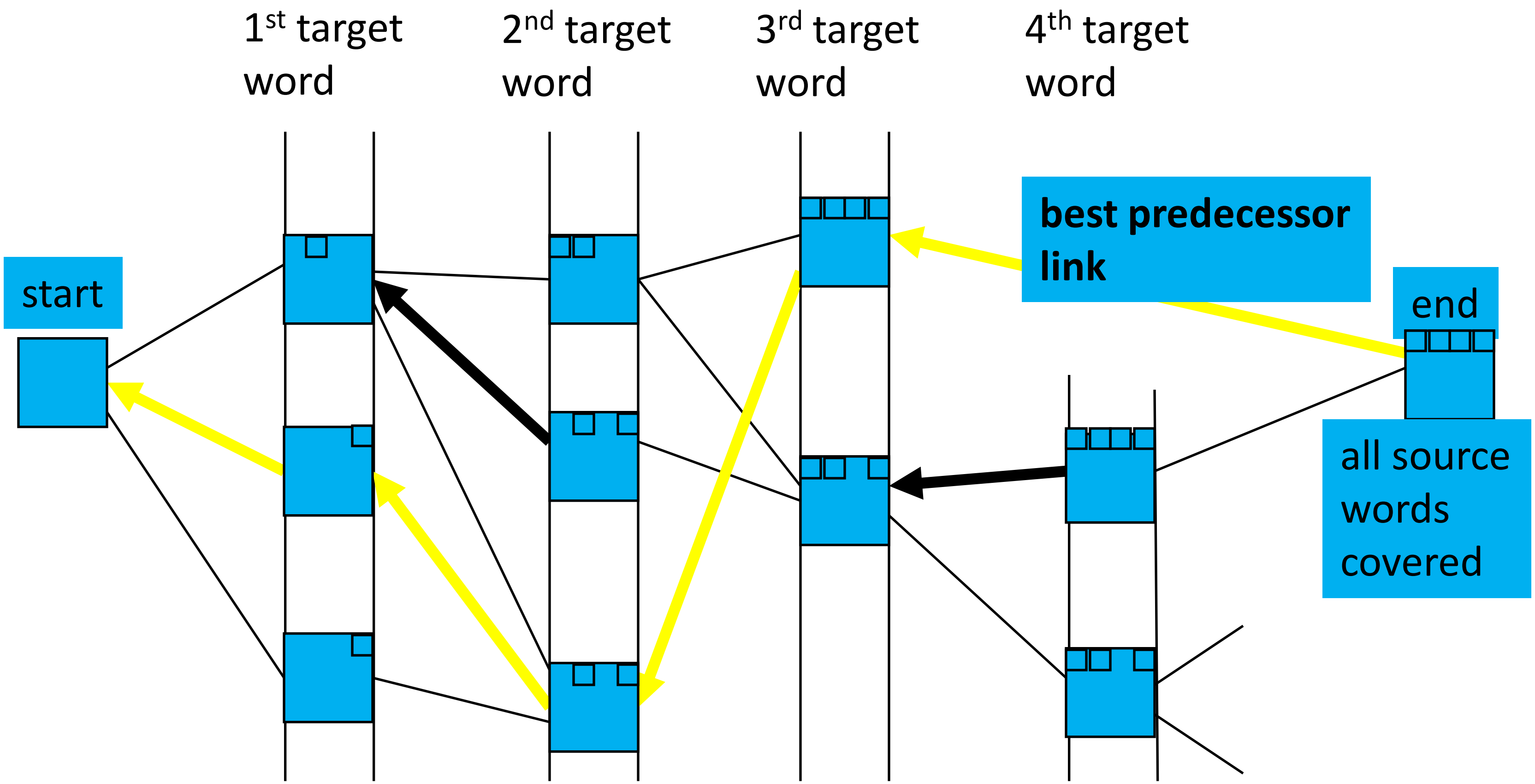
Dynamic Programming Beam Search



Each partial translation hypothesis contains:
- Last English word chosen + source words covered by it
- Next-to-last English word chosen
- Entire coverage vector (so far) of source sentence
- Language model and translation model scores (so far)

[Jelinek, 1969;
Brown et al, 1996 US Patent;
(Och, Ueffing, and Ney, 2001]

Dynamic Programming Beam Search



Each partial translation hypothesis contains:
- Last English word chosen + source words covered by it
- Next-to-last English word chosen
- Entire coverage vector (so far) of source sentence
- Language model and translation model scores (so far)

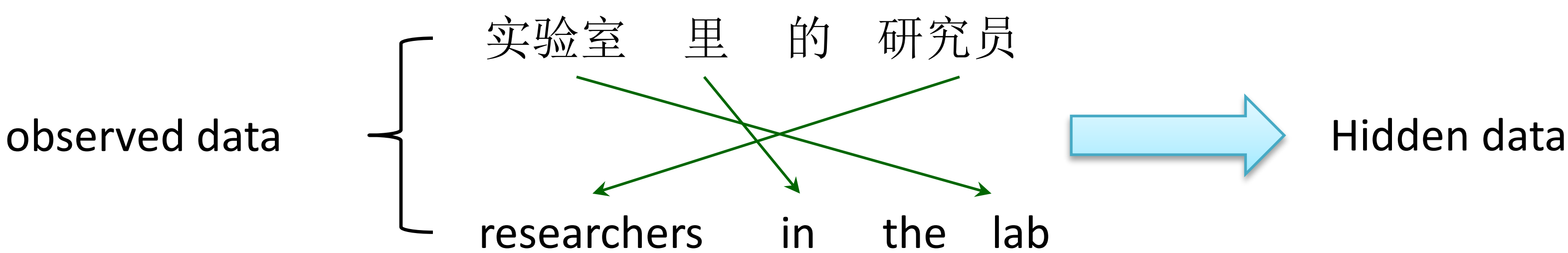
[Jelinek, 1969;
Brown et al, 1996 US Patent;
(Och, Ueffing, and Ney, 2001]

Translation Models and Word Alignment

- IBM Models
  - Model 1 ~ 5 dealing with estimating $P(\boldsymbol{f}|\boldsymbol{e})$
- HMM model
  - Improvement over IBM Model 2

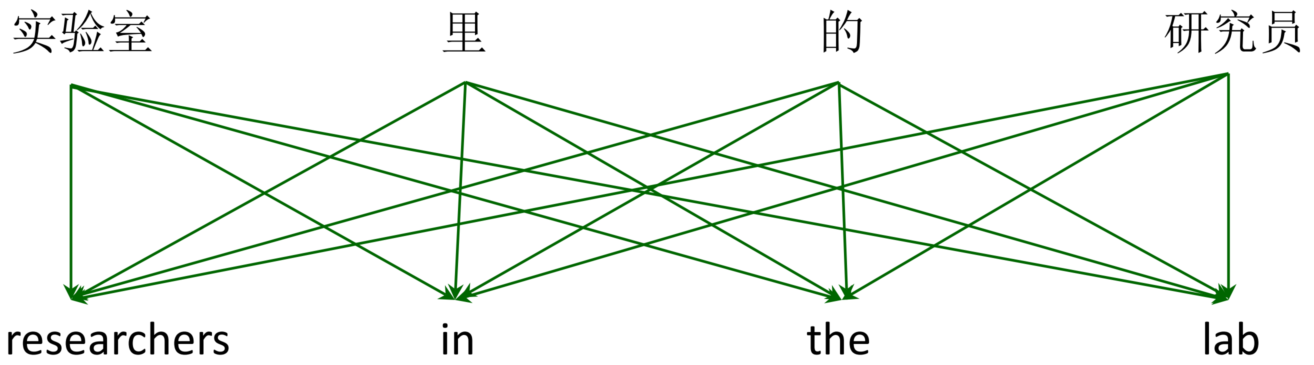
EM Training for Word Alignment

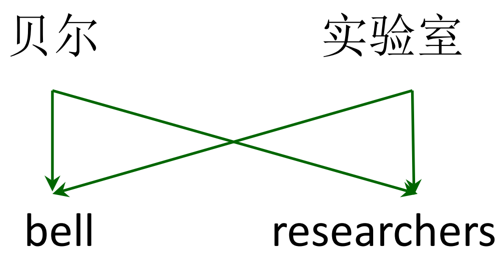
- EM in a couple of slides
  - Expectation Maximization, one optimization method
  - Unsupervised method working on incomplete data
  - Interactively optimize the objective function



EM Training for Word Alignment

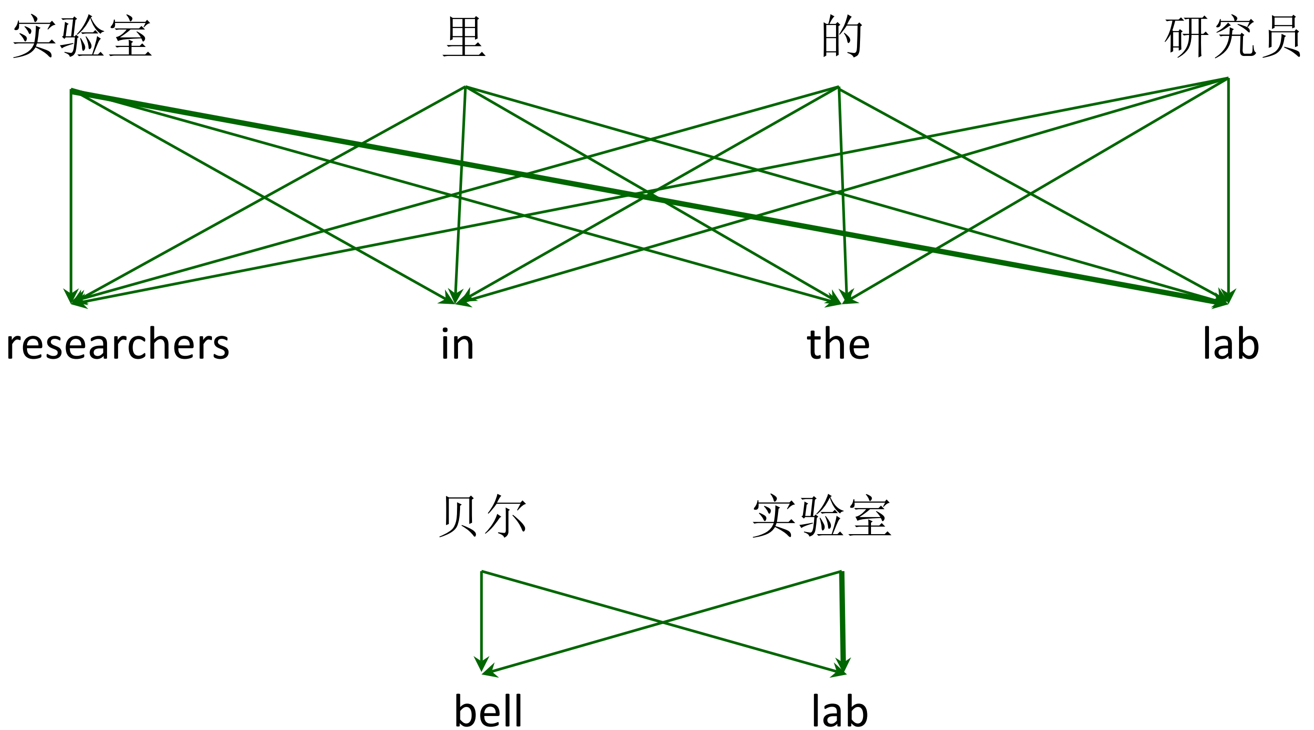
- Initial step: all alignment links are equally likely





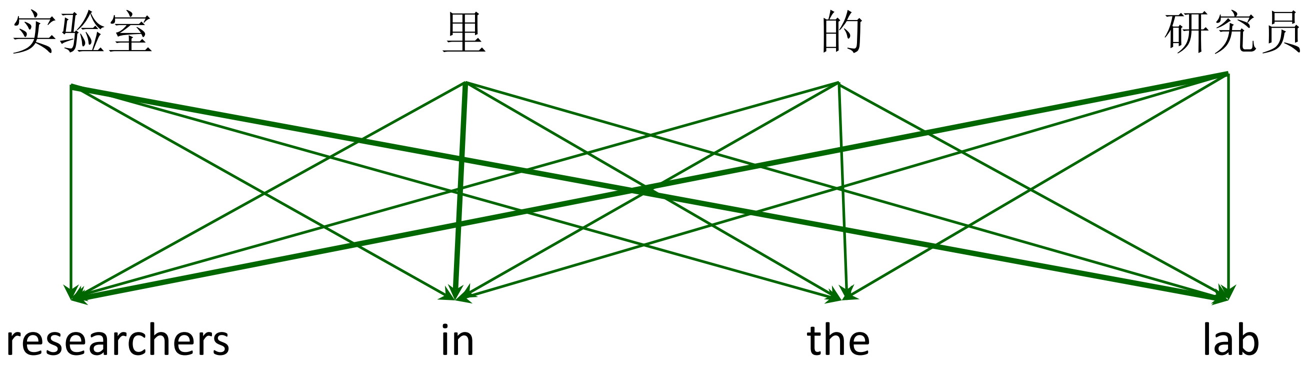
EM Training for Word Alignment

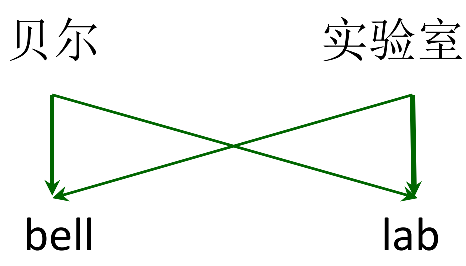
- After one iteration, link between 实验室 and lab becomes stronger



EM Training for Word Alignment

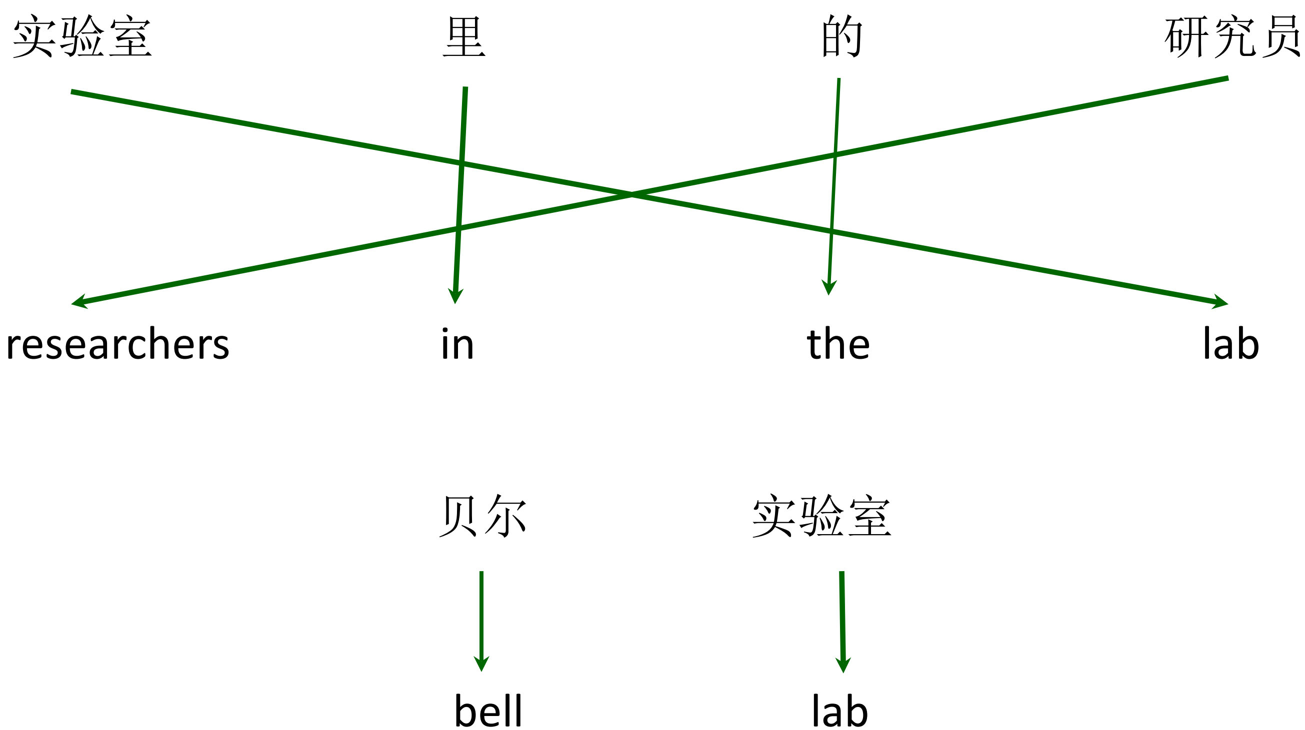
- After two iterations, mores links between words become stronger





EM Training for Word Alignment

- Finally the algorithm will converge with some hidden structure



Notations

$$\boldsymbol{e} = e_1^l = e_1 \dots e_l \qquad \boldsymbol{f} = f_1^m = f_1 \dots f_m$$

$$\boldsymbol{a} = a_1^m = a_1 \dots a_m \qquad (0 \leq a_i \leq l)$$

$$a_j = i \Rightarrow (f_j, e_i)$$

$$P(\boldsymbol{f}|\boldsymbol{e}) = \sum_{\boldsymbol{a}} P(\boldsymbol{f}, \boldsymbol{a}|\boldsymbol{e})$$

$$P(\boldsymbol{f}, \boldsymbol{a}|\boldsymbol{e}) = P(m|\boldsymbol{e}) \prod_{j=1}^{m} P(a_j|a_1^{j-1}, f_1^{j-1}, m, \boldsymbol{e}) P(f_j|a_1^j, f_1^{j-1}, m, \boldsymbol{e})$$

IBM Model 1

$$P(\boldsymbol{f}, \boldsymbol{a}|\boldsymbol{e}) = \underset{\substack{\downarrow \\ \epsilon}}{P(m|\boldsymbol{e})} \prod_{j=1}^{m} \underset{\substack{\downarrow \\ \frac{1}{l+1}}}{P(a_j|a_1^{j-1}, f_1^{j-1}, m, \boldsymbol{e})} \underset{\substack{\downarrow \\ t(f_j|e_{a_j})}}{P(f_j|a_1^{j}, f_1^{j-1}, m, \boldsymbol{e})}$$

$$P(\boldsymbol{f}, \boldsymbol{a}|\boldsymbol{e}) = \frac{\epsilon}{(l+1)^m} \prod_{j=1}^{m} t(f_j|e_{a_j})$$

$$P(\boldsymbol{f}|\boldsymbol{e}) = \frac{\epsilon}{(l+1)^m} \sum_{a_1=0}^{l} \cdots \sum_{a_m=0}^{l} \prod_{j=1}^{m} t(f_j|e_{a_j})$$

Goal: maximize $P(\boldsymbol{f}|\boldsymbol{e})$ subject to $\sum_f t(f|e) = 1$ for all $\boldsymbol{e}$

IBM Model 1

$$h(t,\lambda) = \frac{\epsilon}{(l+1)^m} \sum_{a_1=0}^{l} \cdots \sum_{a_m=0}^{l} \prod_{j=1}^{m} t(f_j|e_{a_j}) - \sum_{e} \lambda_e \left(\sum_{f} t(f|e) - 1\right)$$

$$\frac{\partial h}{\partial t(f|e)} = \frac{\epsilon}{(l+1)^m} \sum_{a_1=0}^{l} \cdots \sum_{a_m=0}^{l} \sum_{j=1}^{m} \delta(f, f_j)\delta(e, e_{a_j}) t(f|e)^{-1} \prod_{k=1}^{m} t(f_k|e_{a_k}) - \lambda_e$$

$$t(f|e) = \lambda_e^{-1} \frac{\epsilon}{(l+1)^m} \sum_{a_1=0}^{l} \cdots \sum_{a_m=0}^{l} \sum_{j=1}^{m} \delta(f, f_j)\delta(e, e_{a_j}) \prod_{k=1}^{m} t(f_k|e_{a_k})$$

$$t(f|e) = \lambda_e^{-1} \sum_{\boldsymbol{a}} P(\boldsymbol{f}, \boldsymbol{a}|\boldsymbol{e}) \boxed{\sum_{j=1}^{m} \delta(f, f_j)\delta(e, e_{a_j})}$$

number of times e connects to f in $\boldsymbol{a}$

IBM Model 1

$$c(f|e;\boldsymbol{f},\boldsymbol{e}) = \sum_{\boldsymbol{a}} P(\boldsymbol{a}|\boldsymbol{e},\boldsymbol{f}) \sum_{j=1}^{m} \delta(f,f_j)\delta(e,e_{a_j})$$

$$\sum_{a_1=0}^{l} \cdots \sum_{a_m=0}^{l} \prod_{j=1}^{m} t(f_j|e_{a_j}) = \prod_{j=1}^{m} \sum_{i=0}^{l} t(f_j|e_i)$$

$$P(\boldsymbol{f}|\boldsymbol{e}) = \frac{\epsilon}{(l+1)^m} \prod_{j=1}^{m} \sum_{i=0}^{l} t(f_j|e_i)$$

$$c(f|e;\boldsymbol{f},\boldsymbol{e}) = \frac{t(f|e)}{t(f|e_0)+\cdots+t(f|e_l)} \boxed{\sum_{j=1}^{m} \delta(f,f_j)} \boxed{\sum_{i=0}^{l} \delta(e,e_i)}$$

IBM Model 2

$$P(\boldsymbol{f}, \boldsymbol{a}|\boldsymbol{e}) = P(m|\boldsymbol{e}) \prod_{j=1}^{m} P(a_j|a_1^{j-1}, f_1^{j-1}, m, \boldsymbol{e}) P(f_j|a_1^{j}, f_1^{j-1}, m, \boldsymbol{e})$$

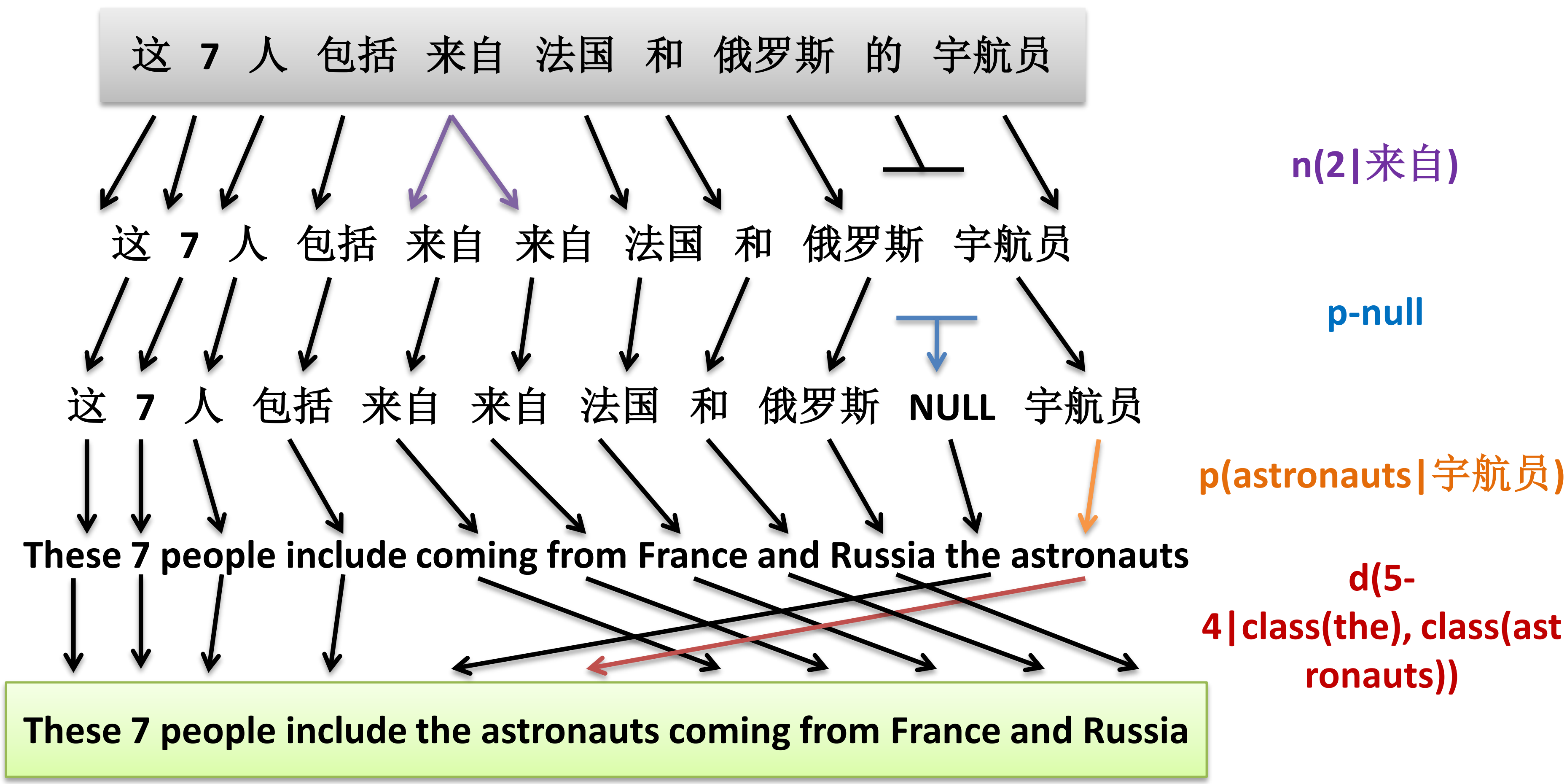
ϵ $\quad a(i|j, m, l) \quad t(f_j|e_{a_j})$

Goal: maximize $P(\boldsymbol{f}|\boldsymbol{e})$ subject to $\sum_{i=0}^{l} a(i|j,m,l) = 1$ for each (j, m, l)

IBM Model 3, 4 and 5

- Model 3
  - Adds fertility model
- Model 4
  - Dealing with distortion (relative reordering model)
- Model 5
  - Removing deficiency

IBM Model 4



这 7 人 包括 来自 法国 和 俄罗斯 的 宇航员
这 7 人 包括 来自 来自 法国 和 俄罗斯 宇航员
这 7 人 包括 来自 来自 法国 和 俄罗斯 NULL 宇航员
These 7 people include coming from France and Russia the astronauts
These 7 people include the astronauts coming from France and Russia
n(2|来自)
p-null
p(astronauts|宇航员)
d(5-4|class(the), class(astronauts))

HMM

$$P(\boldsymbol{f}, \boldsymbol{a}|\boldsymbol{e}) = P(m|\boldsymbol{e}) \prod_{j=1}^{m} P(a_j|a_1^{j-1}, f_1^{j-1}, m, \boldsymbol{e}) P(f_j|a_1^{j}, f_1^{j-1}, m, \boldsymbol{e})$$

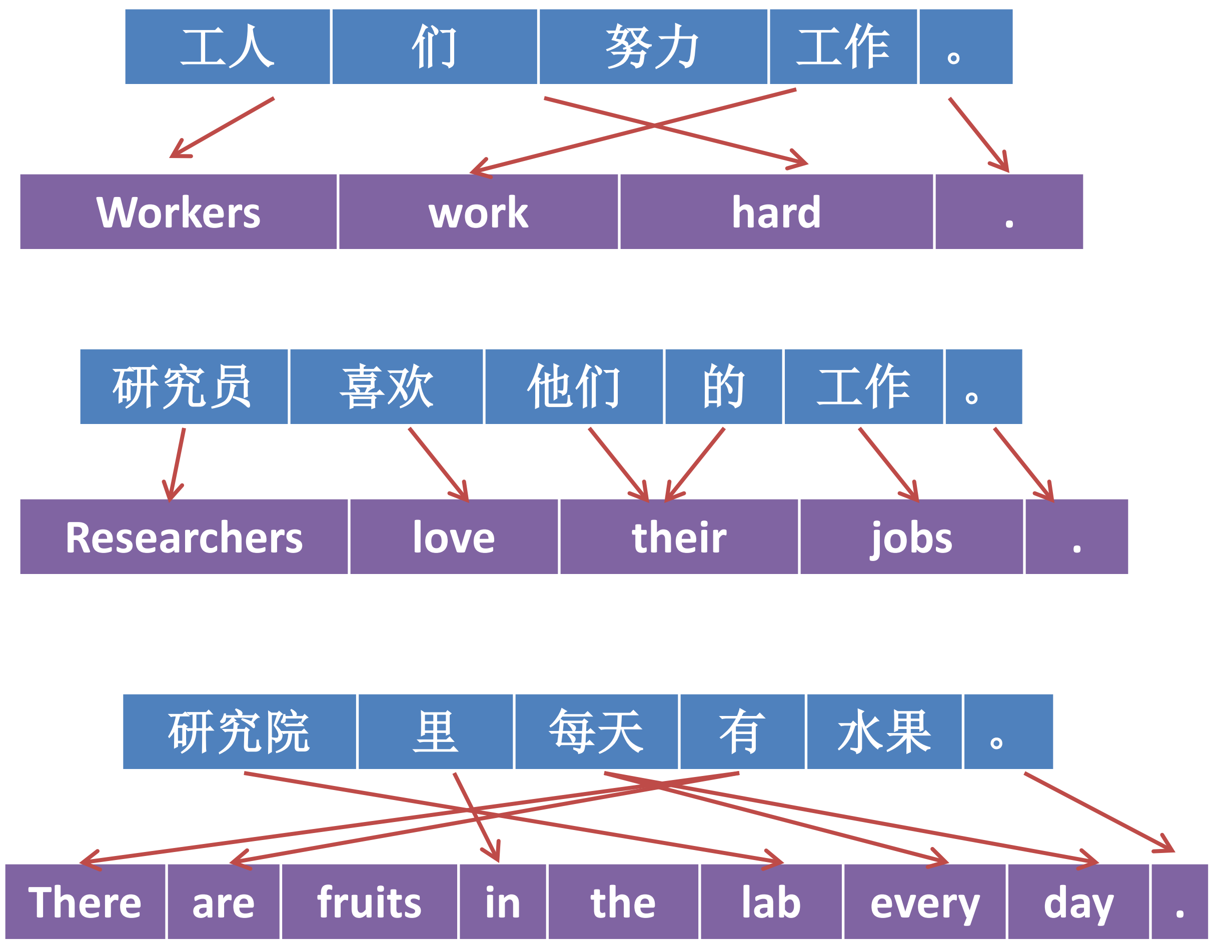
$$\epsilon \qquad\qquad P(a_j|a_{j-1}, l) \qquad\qquad t(f_j|e_{a_j})$$

$$\frac{s(a_i - a_{i-1})}{\sum_{j=1}^{l} s(l - a_{i-1})}$$

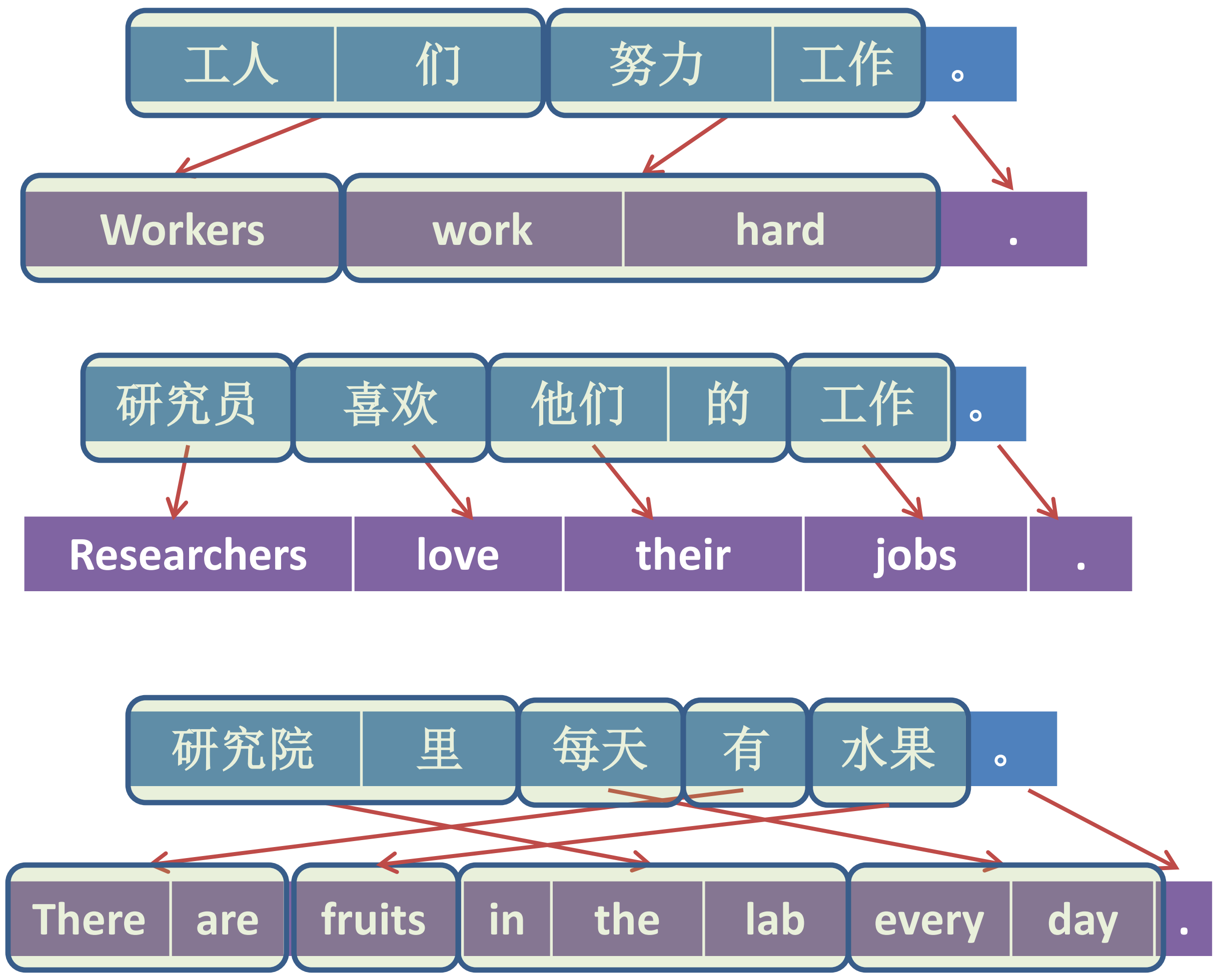
From Word To Phrase

- Words are not always natural units for translation
  - with all due respect ==> 恕我直言
  - you are welcome ==> 不必客气
- Language model is powerful
  - But not as powerful as imagined
- Solution
  - Remember more beyond word translations

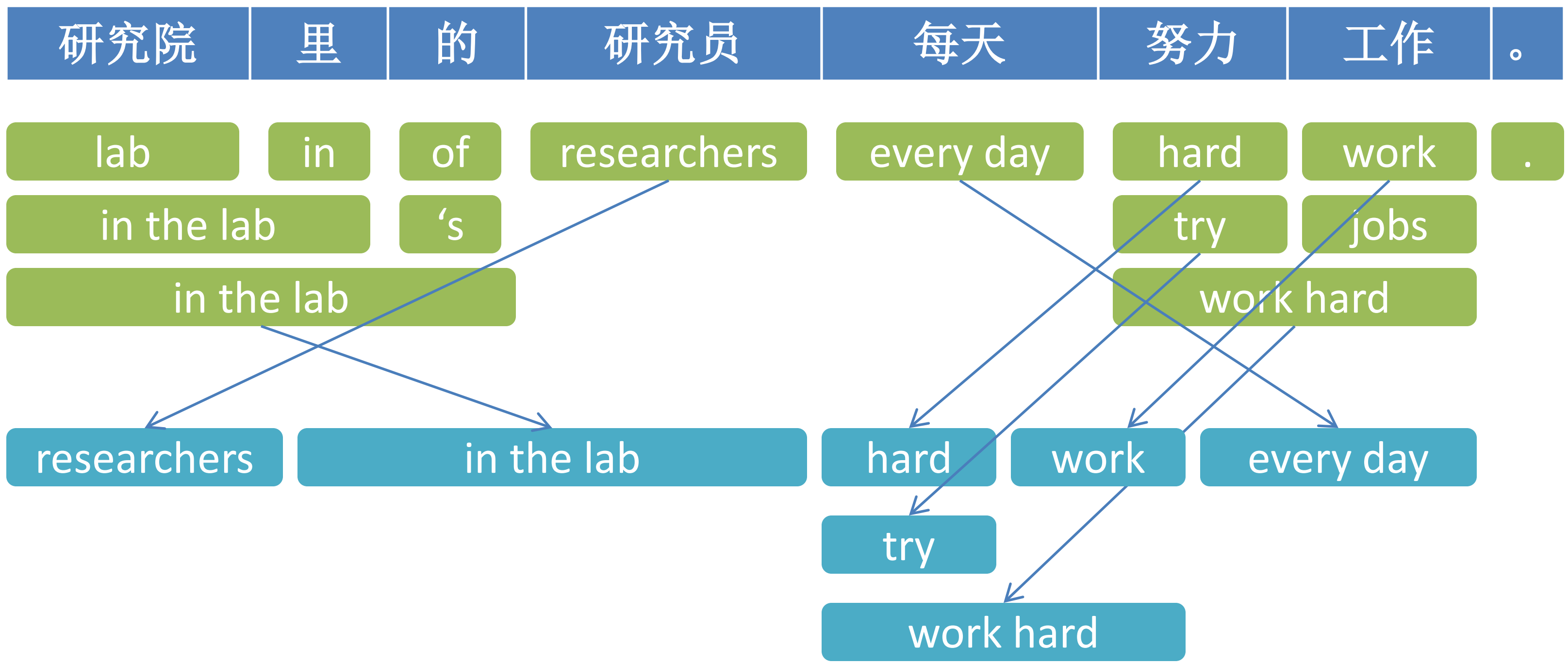
Word Alignment



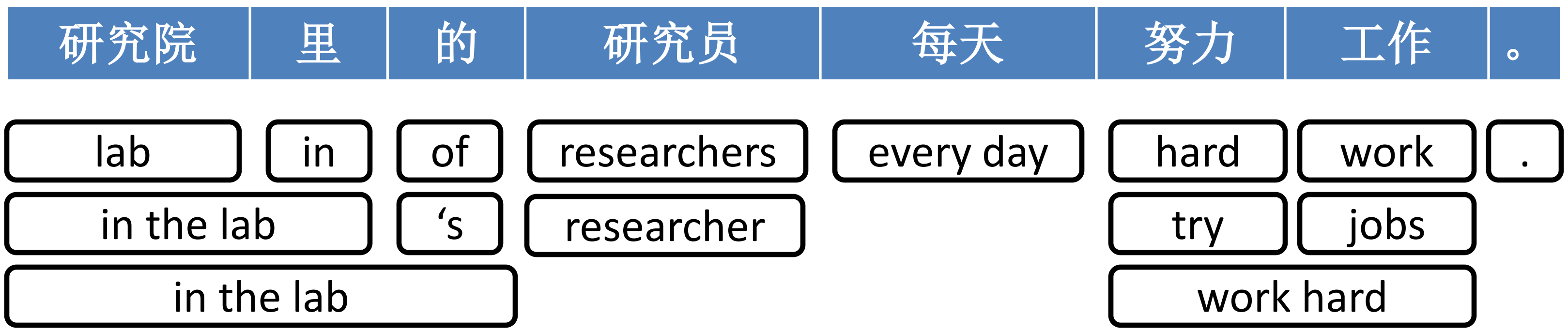
Phrase Extraction



Decoding for Phrase Graph



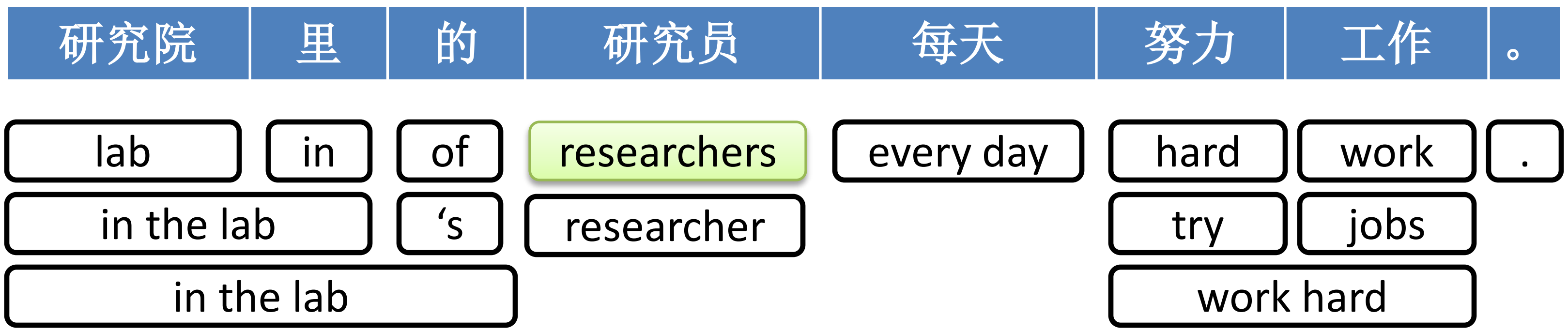
Decoding for Phrase Graph

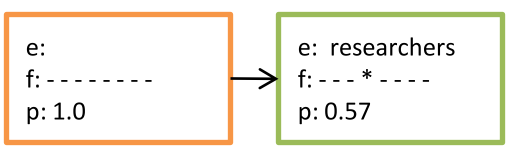


e:
f: - - - - - - - -
p: 1.0

- Start with empty hypothesis

Decoding for Phrase Graph

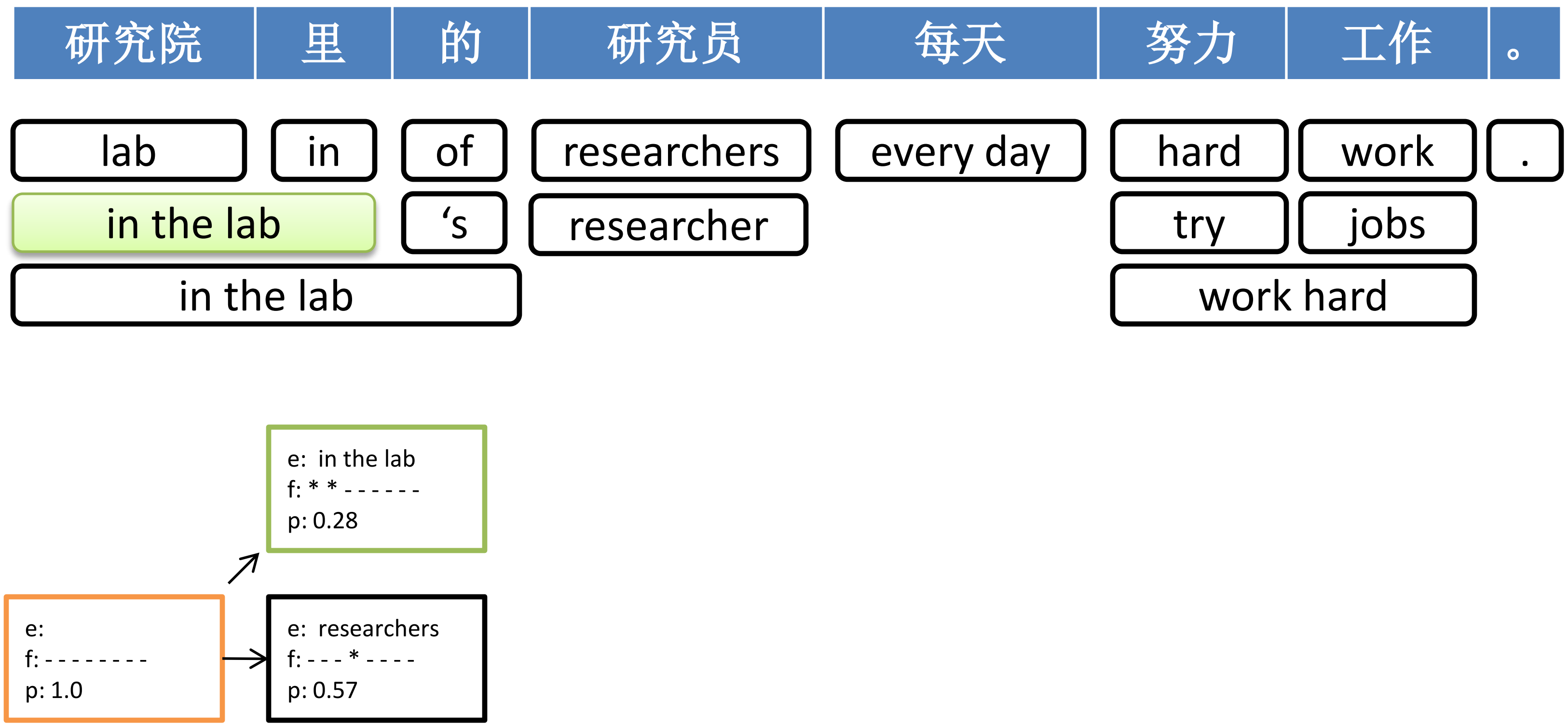




- Pick translation option
- Create hypothesis
  - ✓ add target phrase researchers
  - ✓ cover the forth foreign word
  - ✓ assign probability 0.57

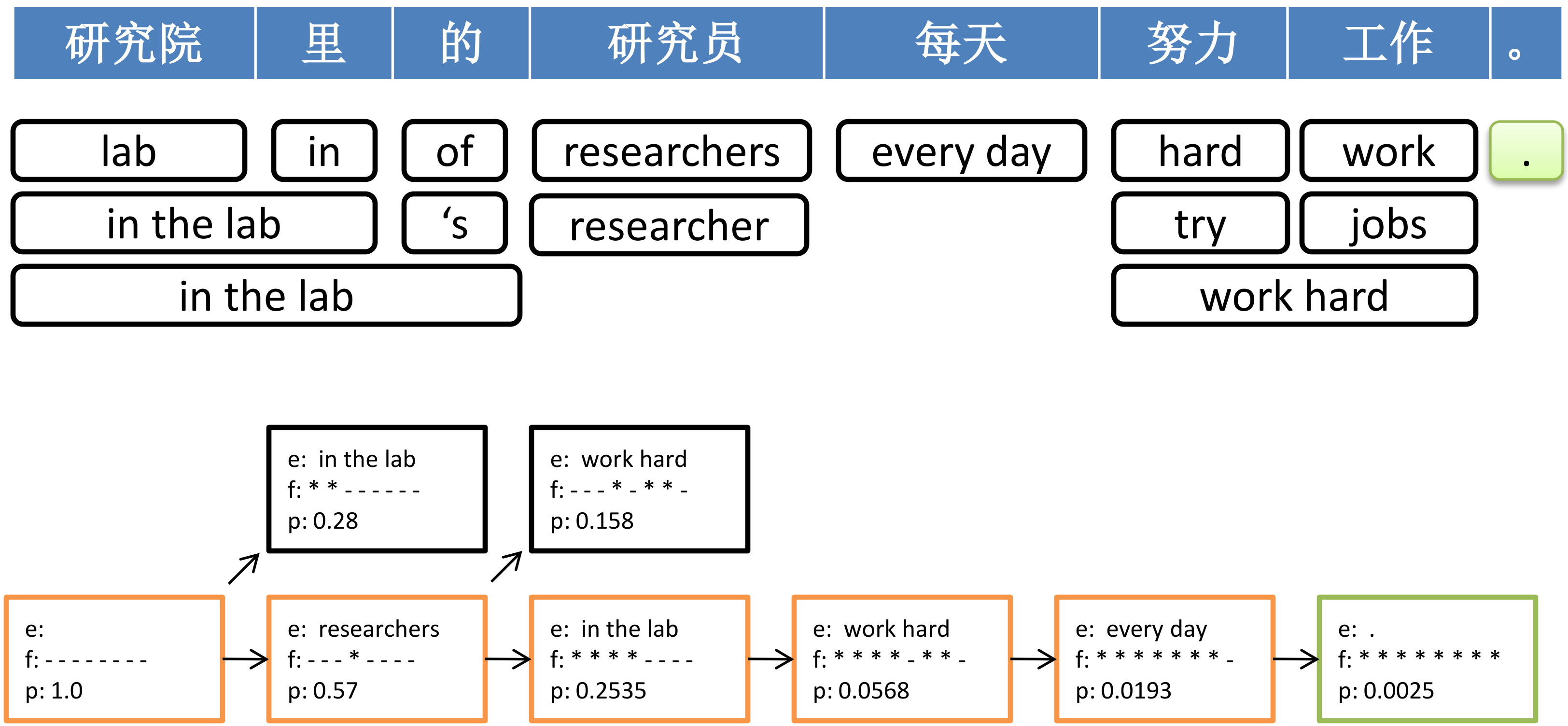
Adapted from Philip Koehn's tutorial on SMT

Decoding for Phrase Graph



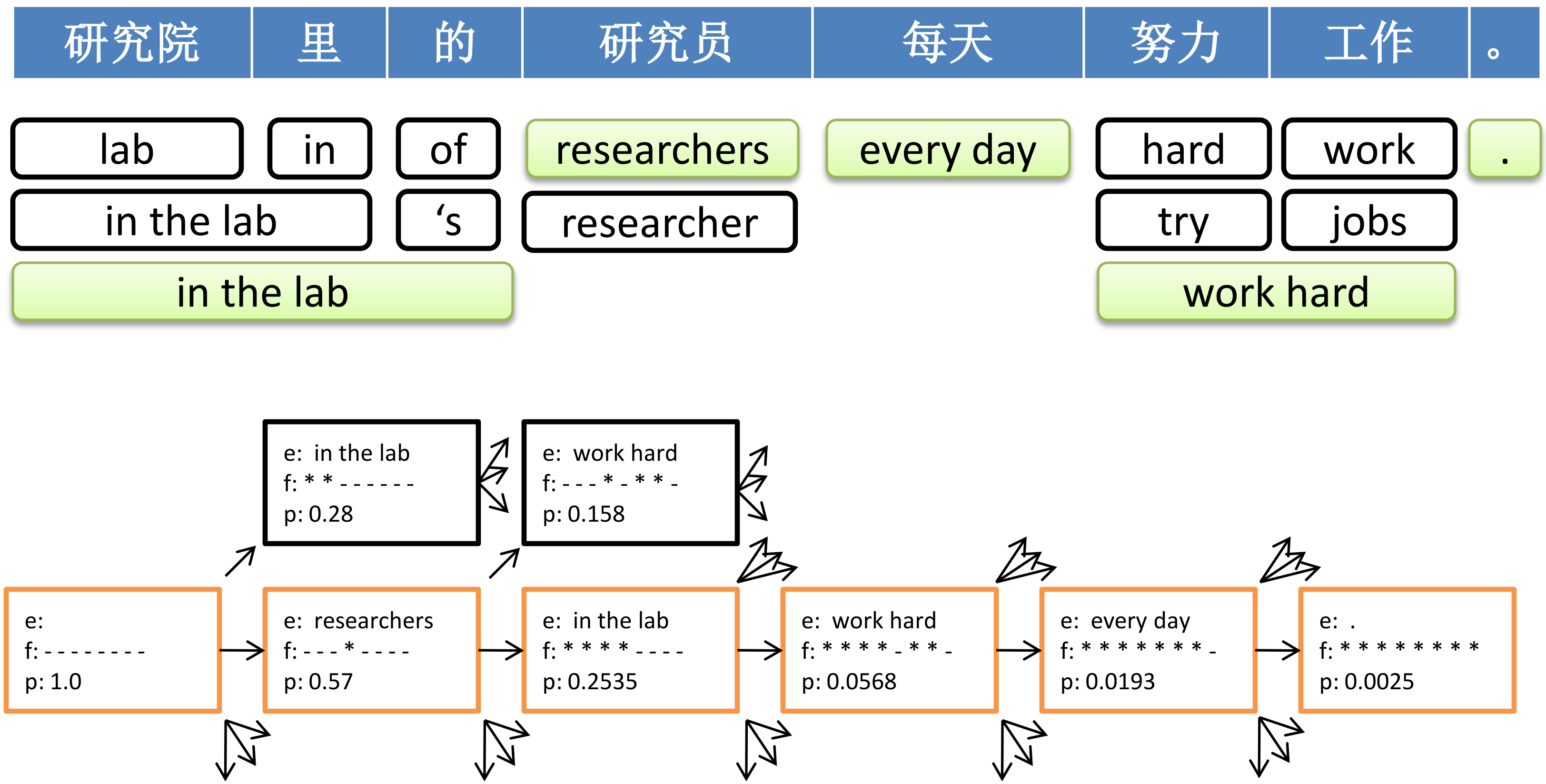
- Add another hypothesis

Decoding for Phrase Graph



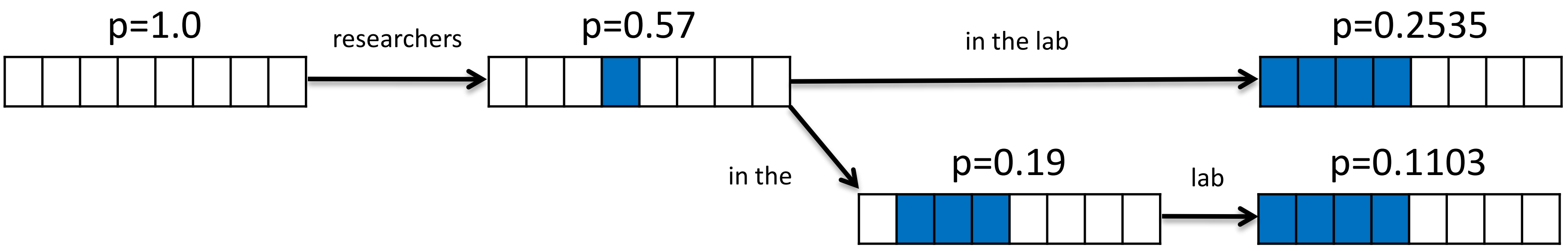
- Further hypothesis expansion

Decoding for Phrase Graph



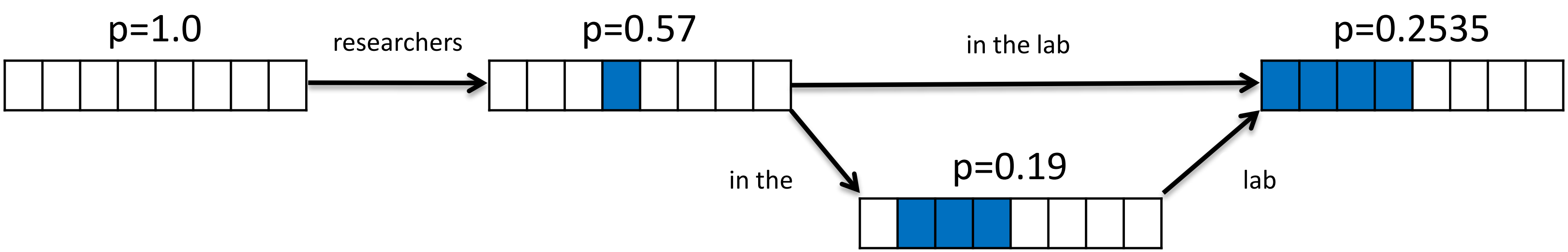
- … until all foreign words covered
  - ✓ find best hypothesis that covers all foreign words
  - ✓ backtrack to read off the translation

Hypothesis Recombination



- Different paths to the ***same*** partial hypothesis

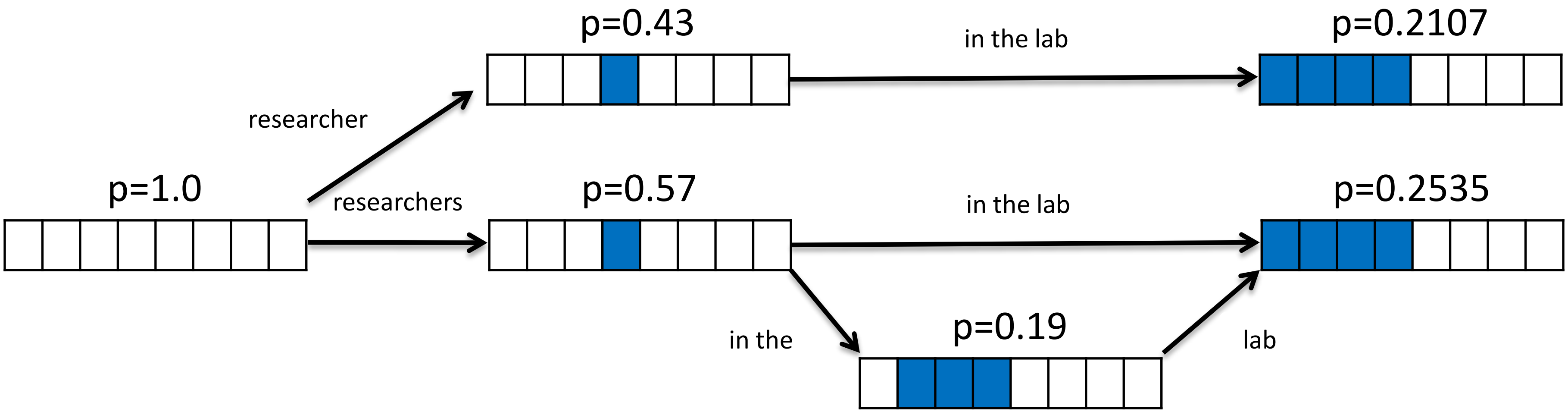
Hypothesis Recombination



- Different paths to the ***same*** partial hypothesis

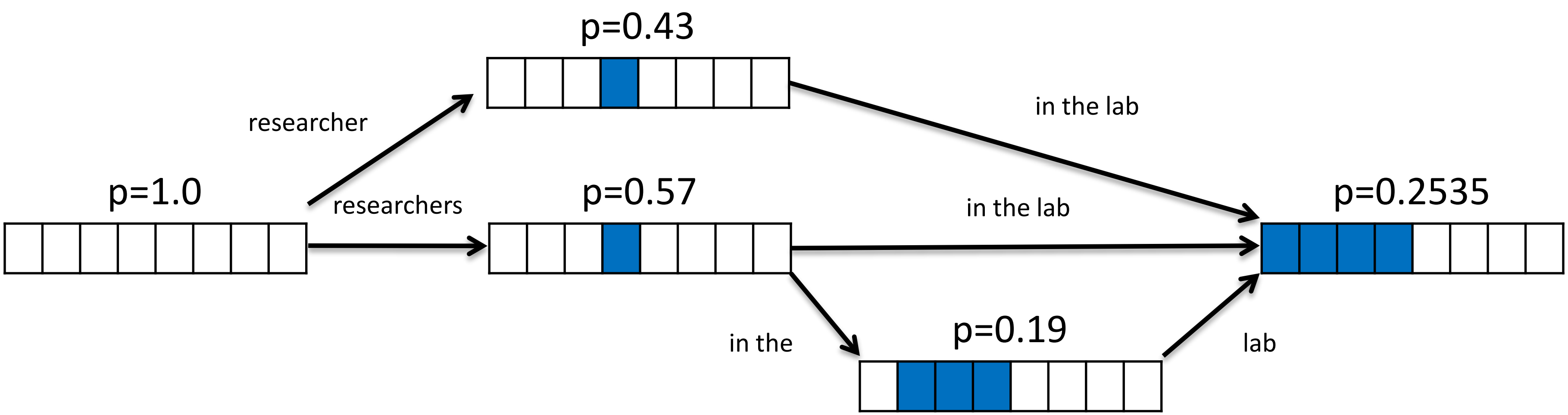
- Hypothesis recombination ***combines*** the path
  - ✓ ***drop weaker*** path
  - ✓ keep pointer from weaker (for lattice generation)

Hypothesis Recombination



- Recombined hypotheses do ***not*** have to match completely
- No matter what is added, weaker path can be dropped, if:
  - ✓ ***last (n-1) target words*** match (for language model computation)
  - ✓ ***foreign word coverage vectors*** match (effects future path)

Hypothesis Recombination



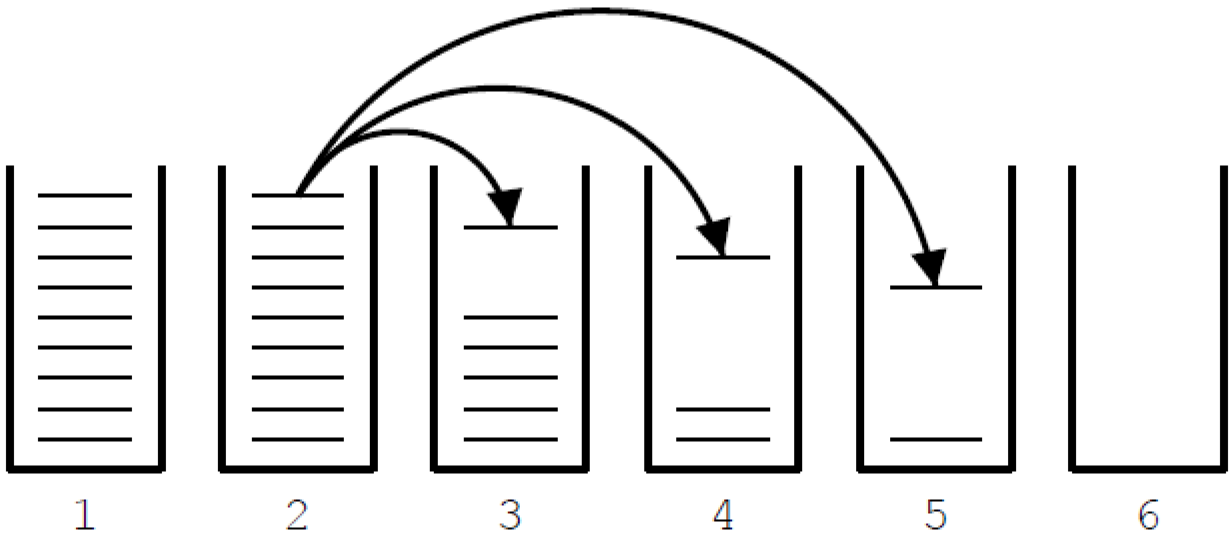
- Recombined hypotheses do ***not*** have to match completely

- No matter what is added, weaker path can be dropped, if:
  - ✓ ***last (n-1) target words*** match (for language model computation)
  - ✓ ***foreign word coverage vectors*** match (effects future path)

Pruning

- Heuristically discard weak hypothesis early
- Organize hypothesis in stacks (buckets), e.g. by
  - ***same*** foreign words covered
  - ***same number*** of source words covered
  - ***same number*** of target words covered
- Compare hypotheses in stacks, discard bad ones
  - ***histogram pruning***: keep top n hypotheses in each stack
  - ***threshold pruning***: keep hypotheses that are at most α times the cost of the best hypothesis in stack

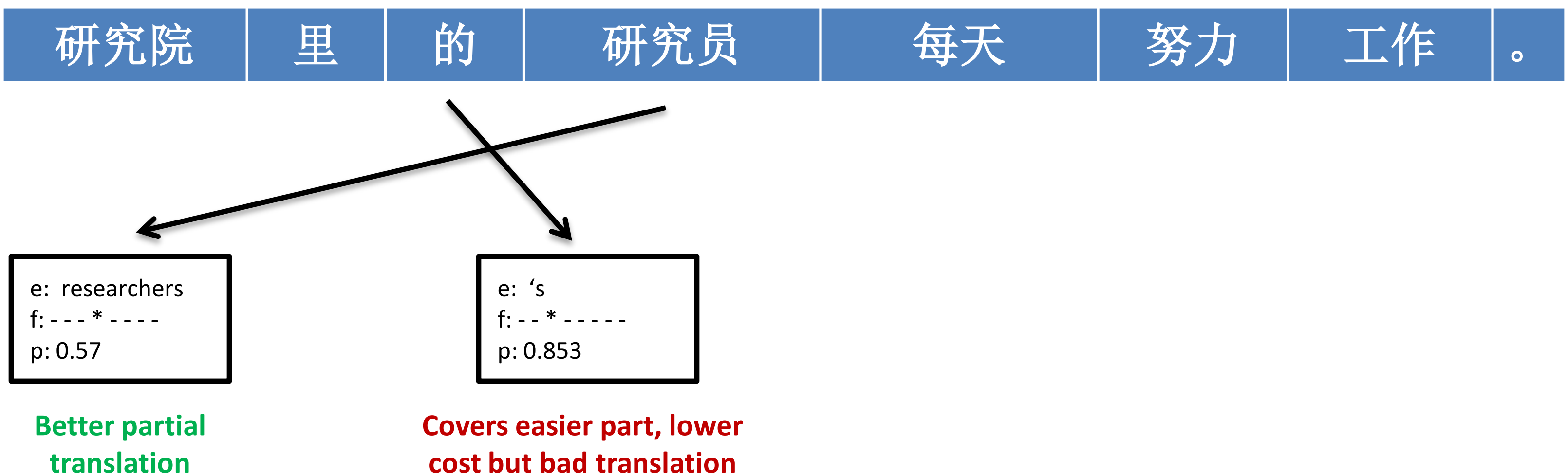
Hypothesis Stack



- Organization of hypothesis into stacks
  - in this example, based on ***number of foreign words*** translated
  - during translation, all hypotheses from one stack are expanded
  - expanded hypotheses are placed into stacks

Comparing Hypothesis

- Comparing hypotheses with same number of foreign words covered



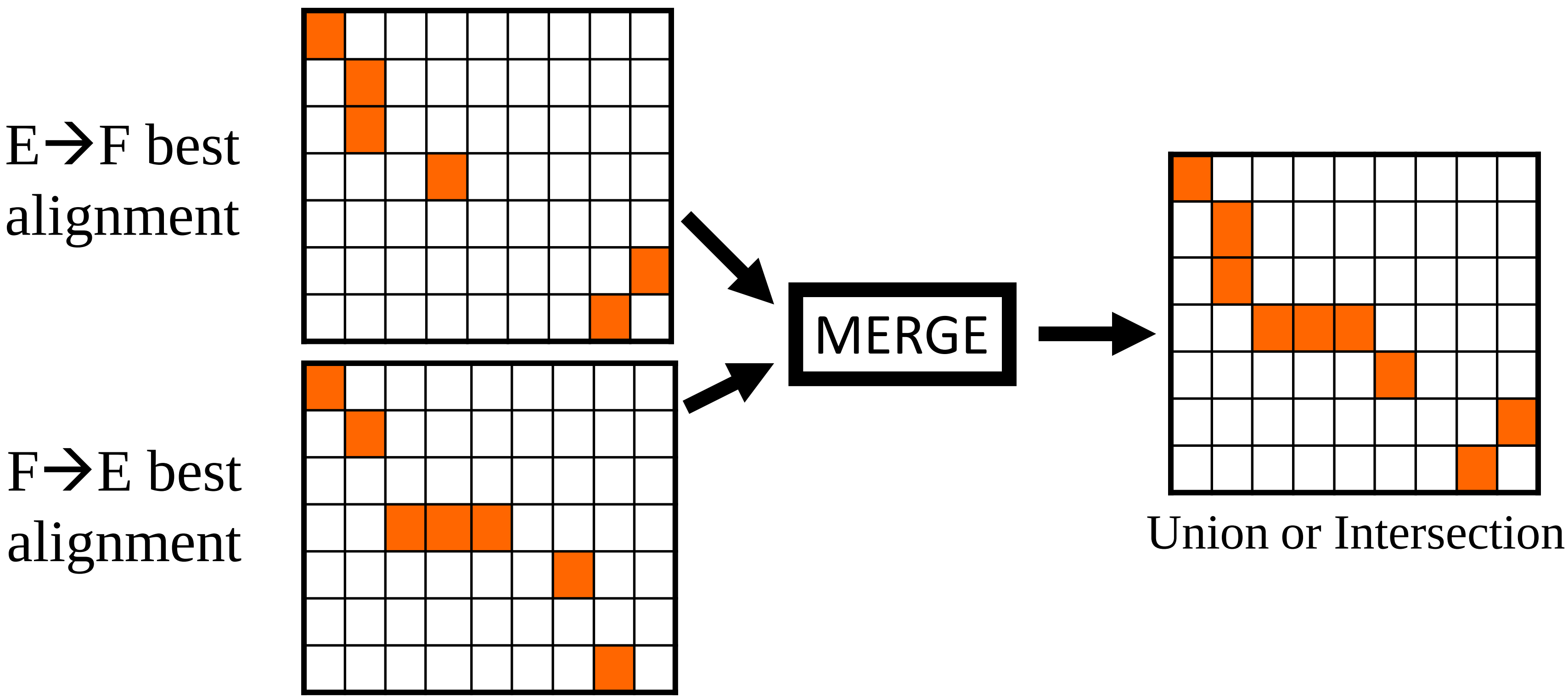
- Hypothesis that covers easy part of sentence is preferred
  - **Solution**: to consider ***future cost*** of uncovered parts

Advantages of Phrase-Based SMT

- Still simple enough
  - but much better than word-based models
- n-to-n mappings can handle non-compositional phrases
  - with all due respect ==> 恕我直言
  - as far as I know ==> 据我所知
- Local context is very useful for disambiguating
  - interest rate ==> 利率
  - interest in ==> … 方面的兴趣
- The more data, the longer the learned phrases
  - even whole sentences

IBM Models are 1-to-Many

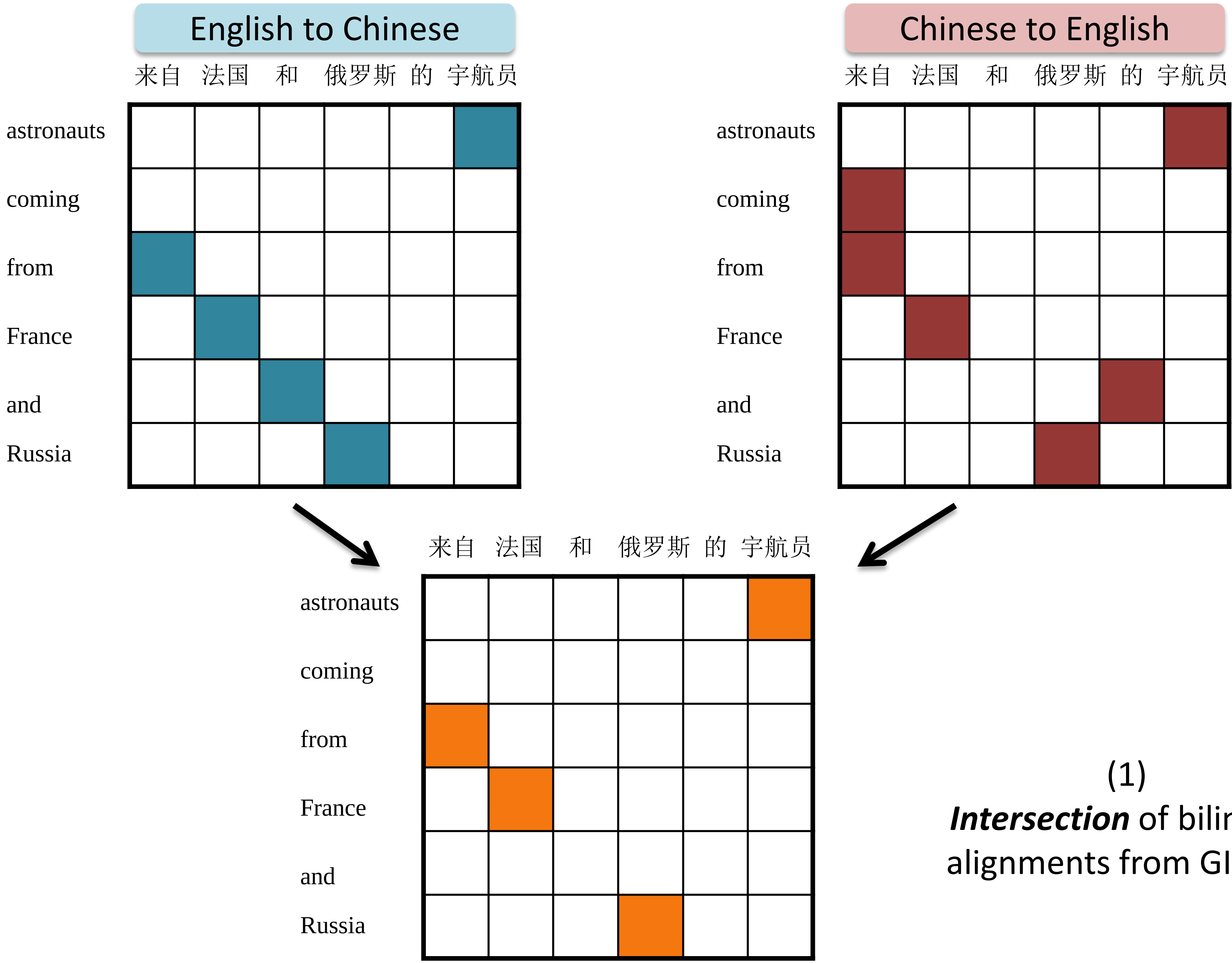
- Run IBM-style aligner both directions, then merge:



Merge Heuristics

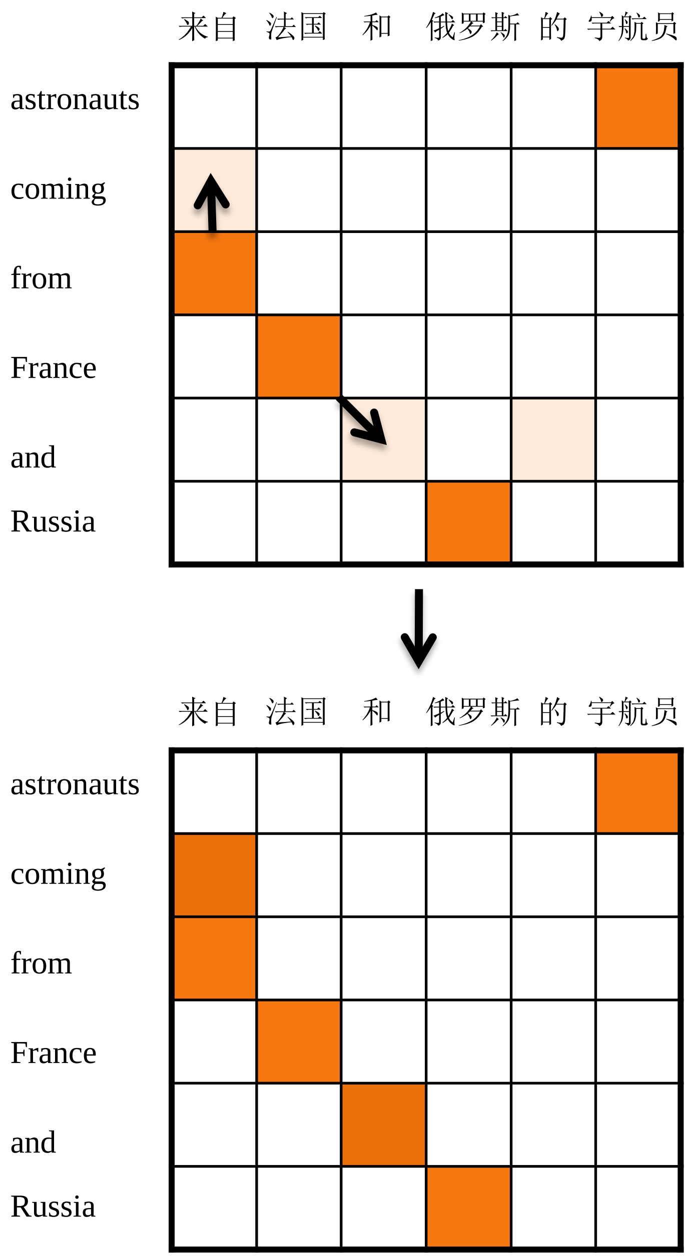
- GDF (intersection-Grow-Diag-Finalization)
  - Better precision
- Union-Reduce
  - Better recall
- Works better than using EM at phrase level

Symmetrizing Word Alignments



(1)
Intersection of bilingual alignments from GIZA++

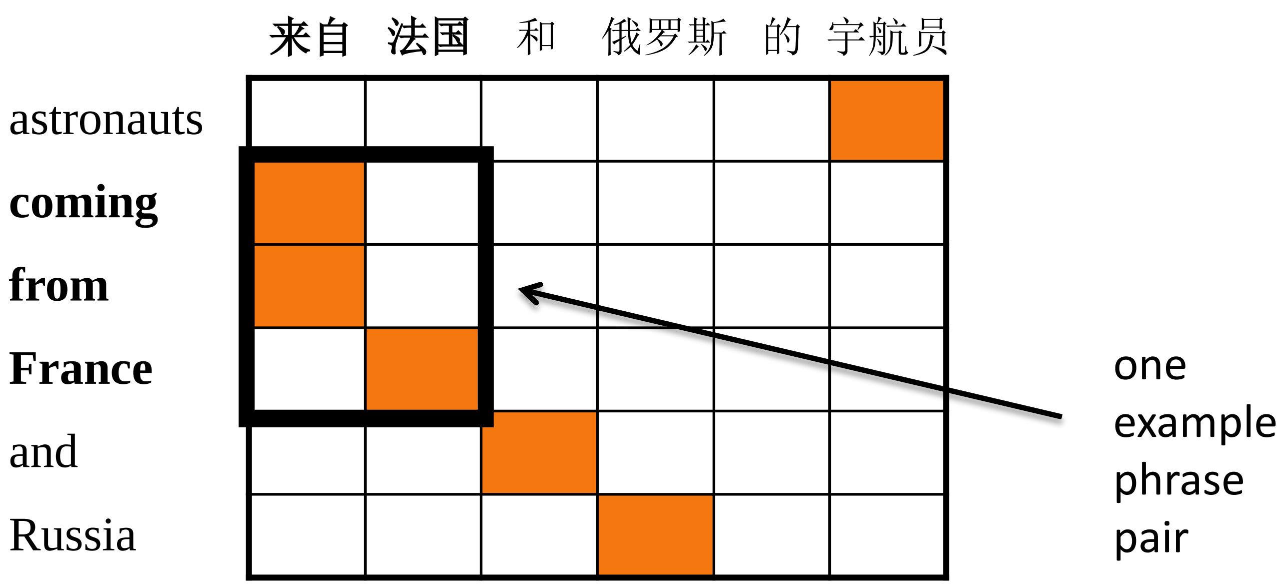
Symmetrizing Word Alignments



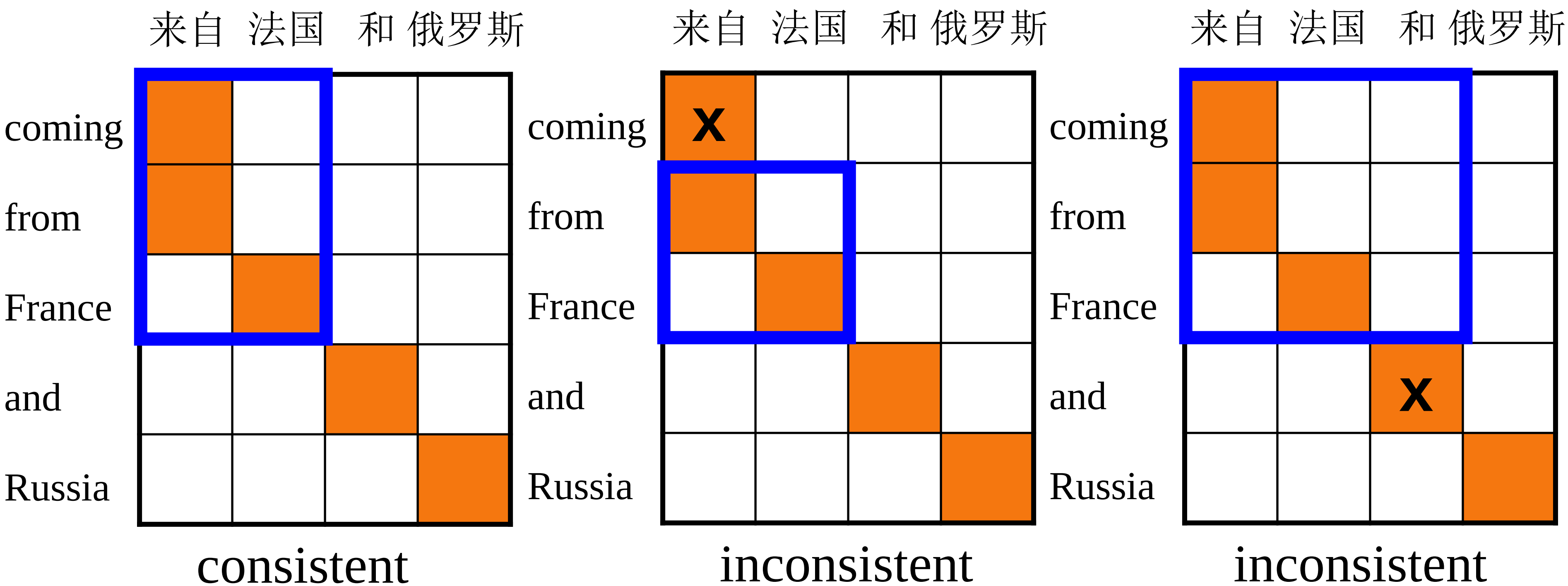
(2)
Grow additional alignment points

How to Learn the Phrase Translation Table?

- Collect all phrase pairs *that are consistent with the word alignment*

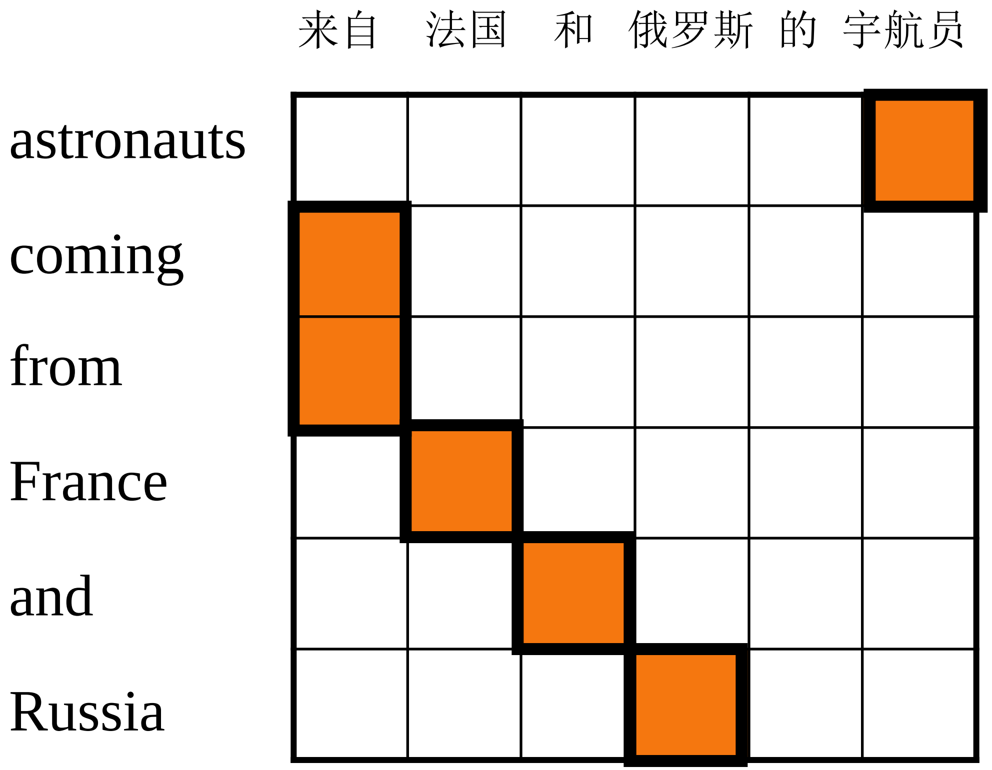


Consistent with Word Alignment



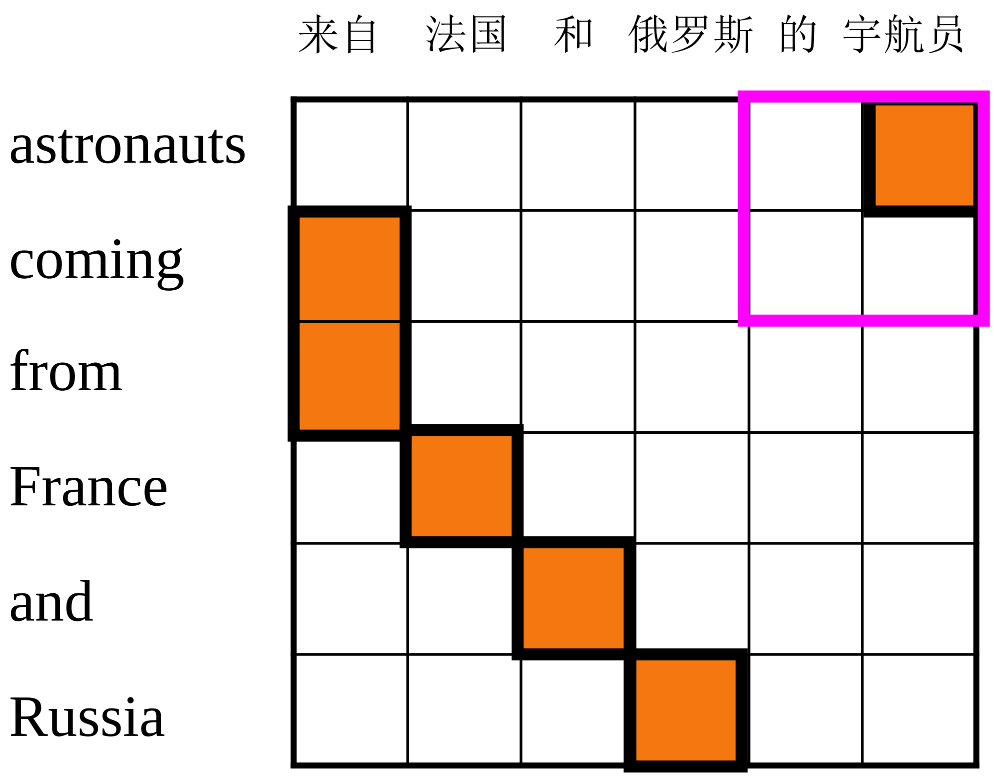
Phrase alignment must contain all alignment points for all the words in both phrases!

Word Alignment Induced Phrases



(宇航员, astronauts) (来自, coming from) (法国, France) (和, and) (俄罗斯, Russia)

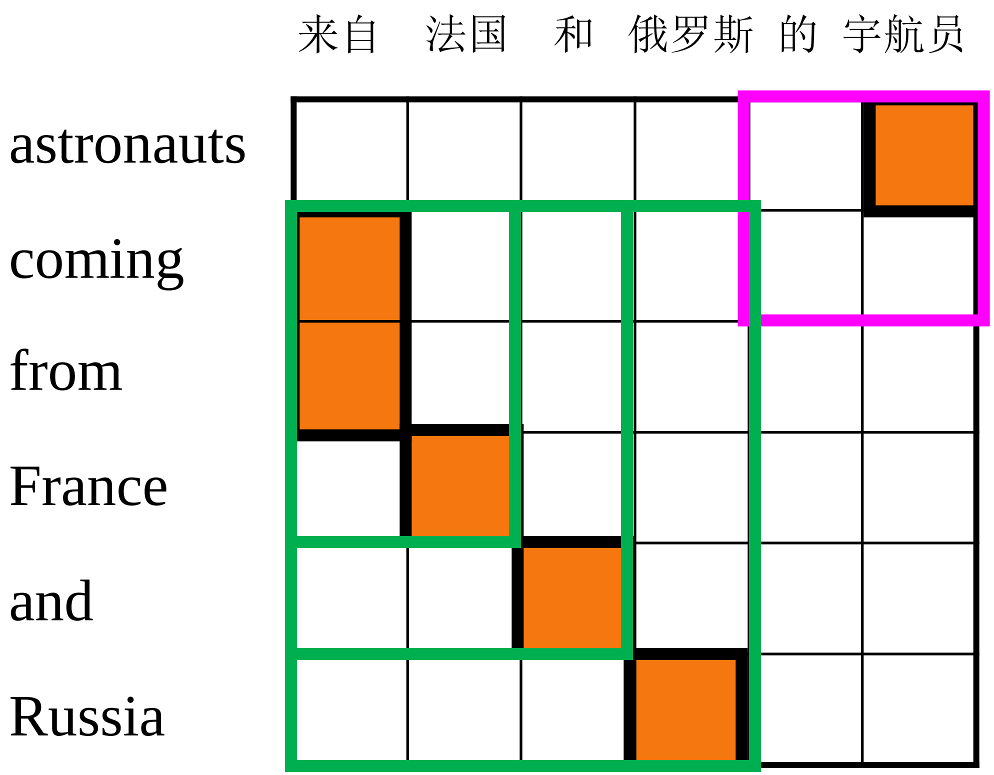
Word Alignment Induced Phrases



(宇航员, astronauts) (来自, coming from) (法国, France) (和, and) (俄罗斯, Russia)

(的 宇航员, astronauts) ...

Word Alignment Induced Phrases



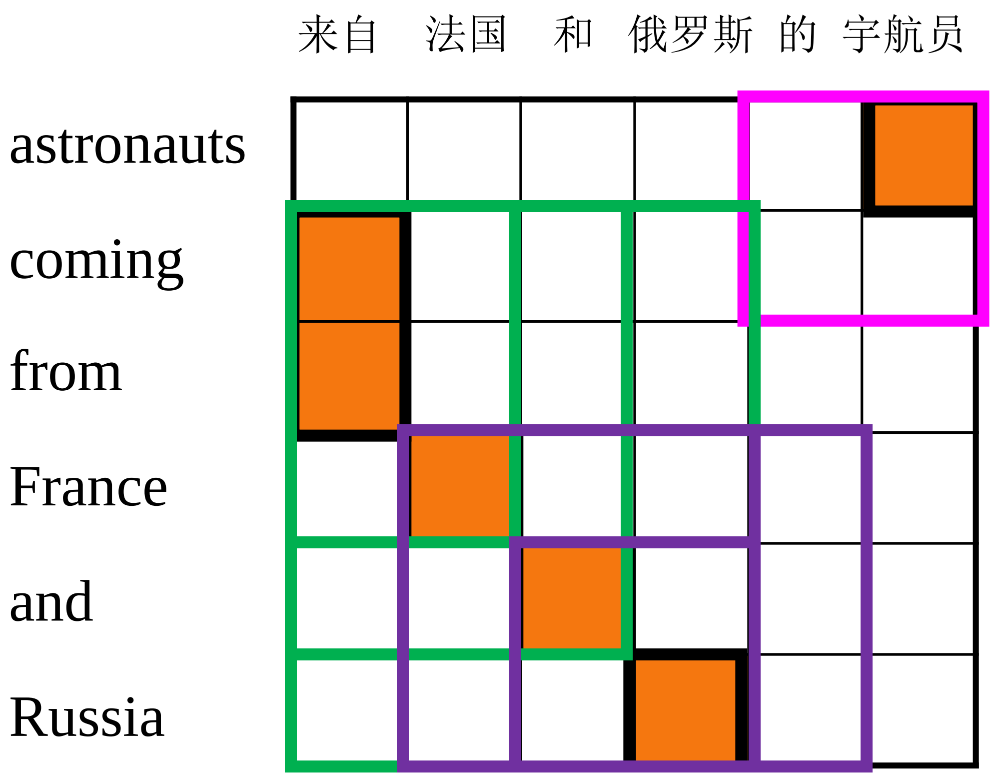
(宇航员, astronauts) (来自, coming from) (法国, France) (和, and) (俄罗斯, Russia)

(的 宇航员, astronauts) …

(来自 法国, coming from France) (来自 法国 和, coming from France and)

(来自 法国 和 俄罗斯, coming from France and Russia) …

Word Alignment Induced Phrases



(宇航员, astronauts) (来自, coming from) (法国, France) (和, and) (俄罗斯, Russia)

(的 宇航员, astronauts) …

(来自 法国, coming from France) (来自 法国 和, coming from France and)

(来自 法国 和 俄罗斯, coming from France and Russia) …

(和 俄罗斯, and Russia) (法国 和 俄罗斯, France and Russia)

(法国 和 俄罗斯 的, France and Russia) …

Phrase Pair Probabilities

- A certain phrase pair $(f_1 f_2 f_3, e_1 e_2 e_3)$ may appear many times across the bilingual corpus.
  - And we hope so
- Then there is a vast list of phrase pairs and their frequencies – how to assign probabilities?

Phrase Pair Probabilities

- Basic idea:
  - Relative frequency:
    - $P(f_1 f_2 f_3, e_1 e_2 e_3) = \frac{\#(f_1 f_2 f_3, e_1 e_2 e_3)}{\#(e_1 e_2 e_3)}$
- Some important refinements:
  - Smooth using word probs $P(f|e)$ for individual words connected in the word alignment
    - Some low count phrase pairs now have high probability, others have low probability
  - Discount for ambiguity
    - If phrase $(e_1 e_2 e_3)$ can map to 5 different French phrases, due to the ambiguity of unaligned words, each pair gets a 1/5 count
  - Count **BAD** events too
    - If phrase $(e_1 e_2 e_3)$ doesn't map onto *any* contiguous French phrase, increment event $\#(\mathbf{BAD}, e_1 e_2 e_3)$

Log-linear Model for SMT

- Mis-use of translation probability
  - $e^* = \text{argmax}_e \quad P(e) \cdot P(f|e)$
  - $P(f|e) = P_{\text{TM}}(e|f) \cdot P_{\text{LM}}(e)$
- Model scaling in source-channel model
  - $P(f|e) = P_{\text{LM}}(e)^{\lambda_1} \cdot P_{\text{TM}}(f|e)^{\lambda_2}$
  - $\log P(f|e) = \lambda_1 \log P_{\text{LM}}(e) + \lambda_2 \log P_{\text{TM}}(f|e)$
- Generalized log-linear model?
  - $\log P(f|e) = \lambda_1 \log P_{\text{LM}}(e) + \lambda_2 \log P_{\text{TM}}(f|e) + \lambda_3 \log P_{\text{TM}}(e|f)$

Log-linear Model for SMT

- Maximum entropy model for SMT

$$P(e|f) = \frac{1}{Z}\exp\left(\sum_i \lambda_i h_i(f,e)\right)$$

$$e^* = \mathrm{argmax}_e P(e|f) = \mathrm{argmax}_e \sum_i \lambda_i h_i(f,e)$$

- Features
  - Log language model probability
  - Log translation probability (both directions)
  - Number of phrases
  - Number of target words
  - Distortion cost

Model Training

- GIS algorithms used to learn feature weights
- Problems:
  - Correct translation
    - Oracle translation approximation
  - Normalization factor computation
    - N-best approximation for solution space
  - N-best translations sensitive to model parameter
    - Iterative decoding

Minimum Error Rate Training

- Optimization criteria
  - Data likelihood
    - $$\boldsymbol{\lambda}^* = \operatorname*{argmax}_{\boldsymbol{\lambda}} \left\{ \sum_{s=1}^{S} \log p_{\boldsymbol{\lambda}}(e_s|f_s) \right\}$$
  - Error count
    - $$\boldsymbol{\lambda}^* = \operatorname*{argmin}_{\boldsymbol{\lambda}} \left\{ \sum_{s=1}^{S} E(r_s, e^*(f_s, \boldsymbol{\lambda})) \right\}$$
- Smoothed error count
  - $$\boldsymbol{\lambda}^* = \operatorname*{argmin}_{\boldsymbol{\lambda}} \left\{ \sum_{s,k} E(e_{s,k}) \frac{p(e_{s,k}|f)^{\alpha}}{\sum_k p(e_{s,k}|f)^{\alpha}} \right\}$$

Unsmoothed vs. Smoothed Error Count

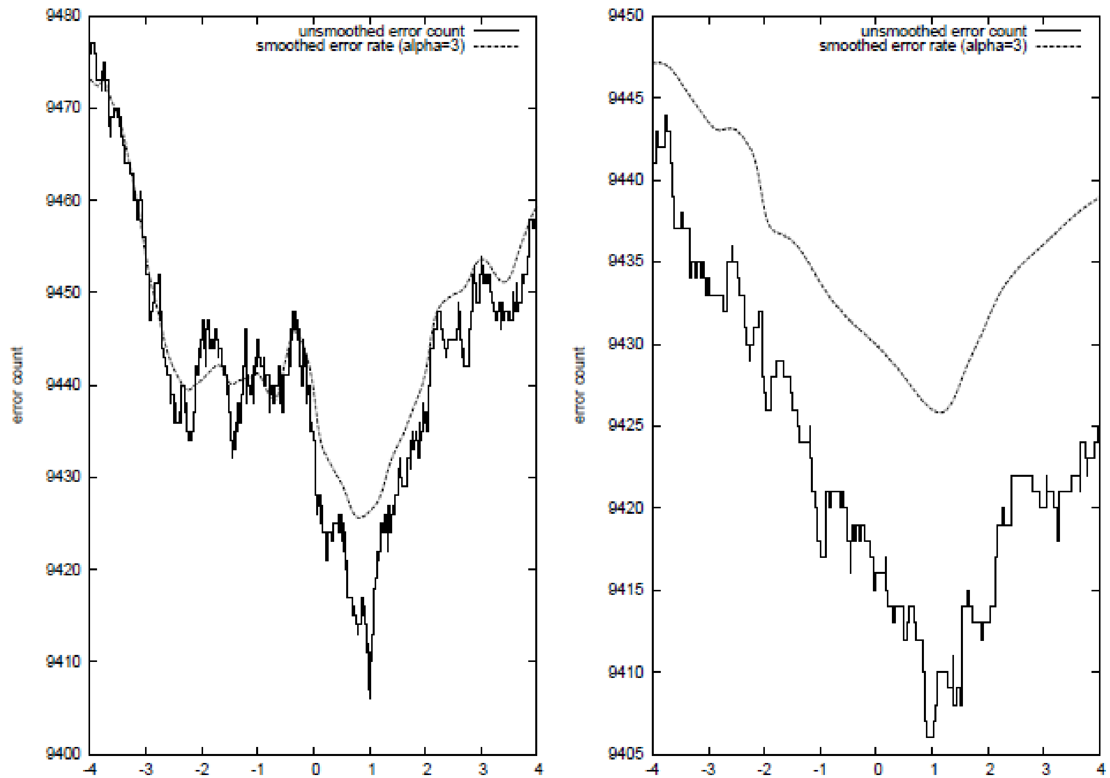
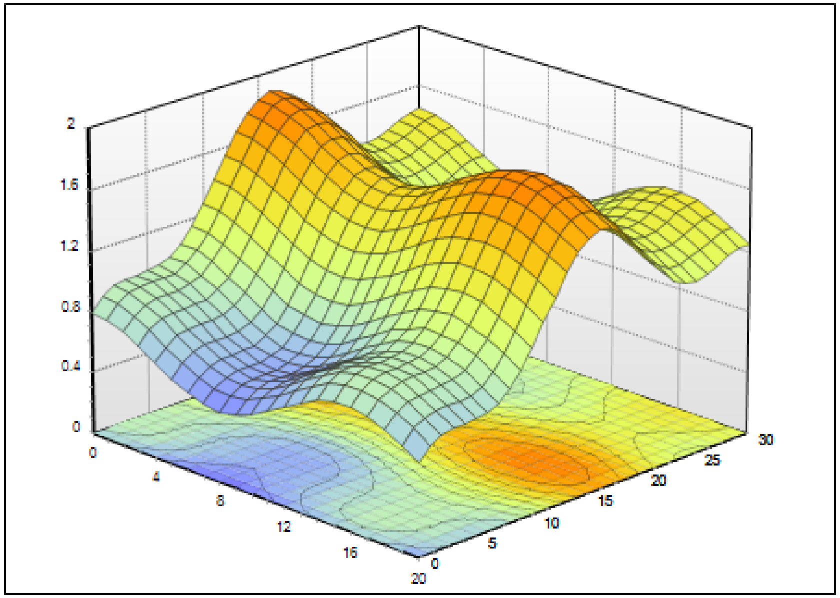


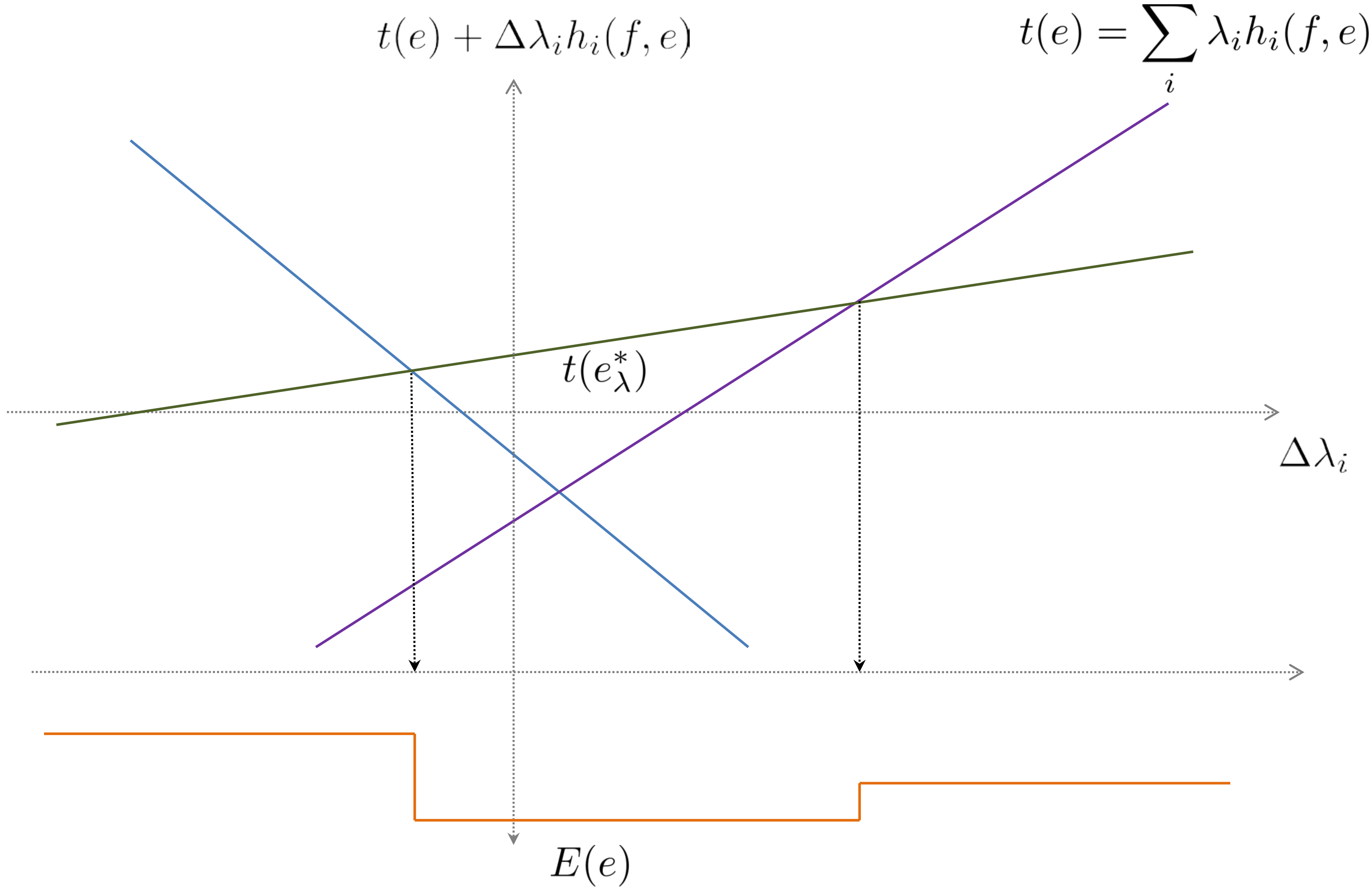
Figure 1: Shape of error count and smoothed error count for two different model parameters. These curves have been computed on the development corpus (see Section 7, Table 1) using $1,600$ alternatives per source sentence. The smoothed error count has been computed with a smoothing parameter $\alpha = 3$.

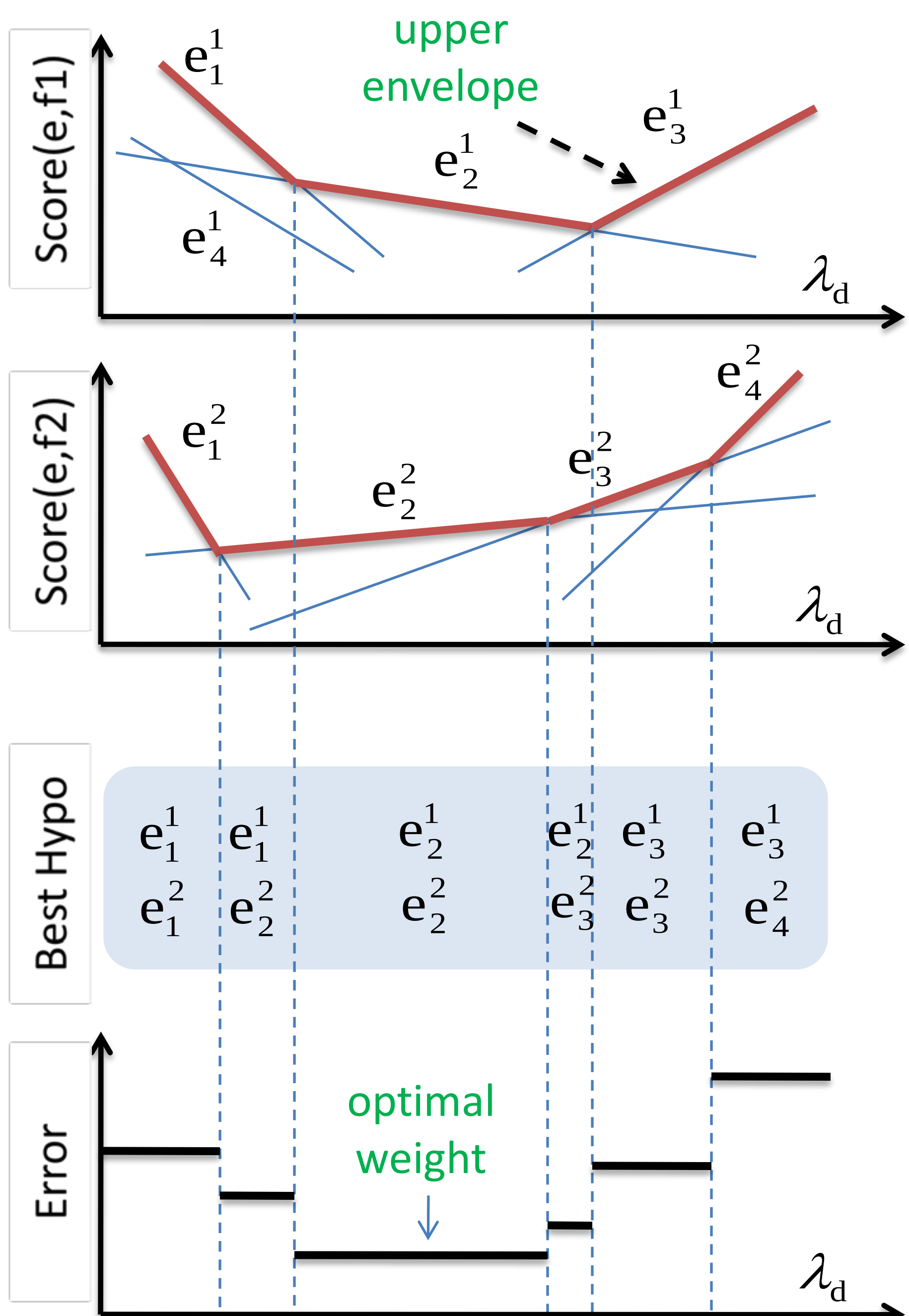
Grid-base Line Search

- Issues
  - Local maximum
  - Efficiency



Error Surface



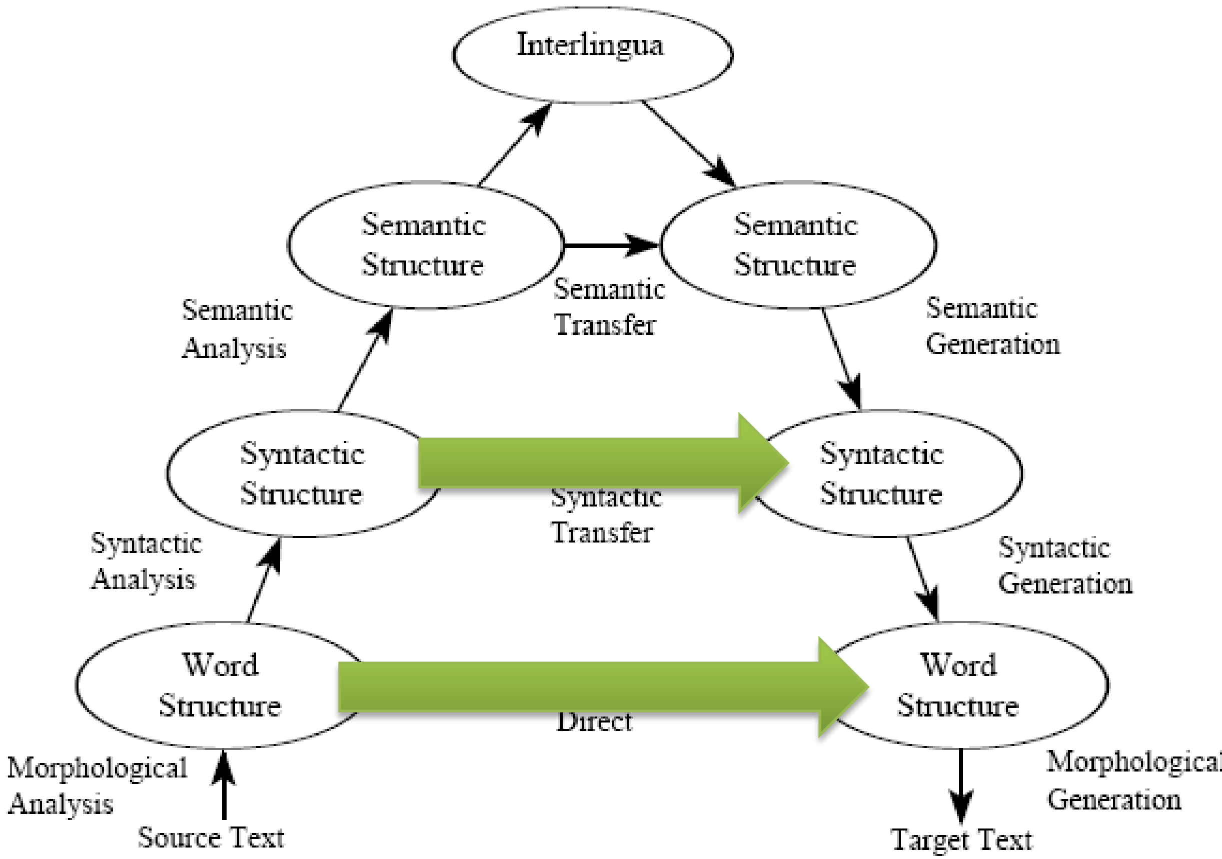


upper envelope
Score(e,f1)
e_1^1
e_2^1
e_3^1
e_4^1
λ_d
Score(e,f2)
e_1^2
e_2^2
e_3^2
e_4^2
λ_d
Best Hypo
e_1^1 e_1^2
e_1^1 e_2^2
e_2^1 e_2^2
e_2^1 e_3^2
e_3^1 e_3^2
e_3^1 e_4^2
Error
optimal weight
λ_d

Comments on MERT

- Random start to work around local maxima
- Effective when
  - Solution space is limited
  - Feature space is small (< 20)
- Still need iterative decoding

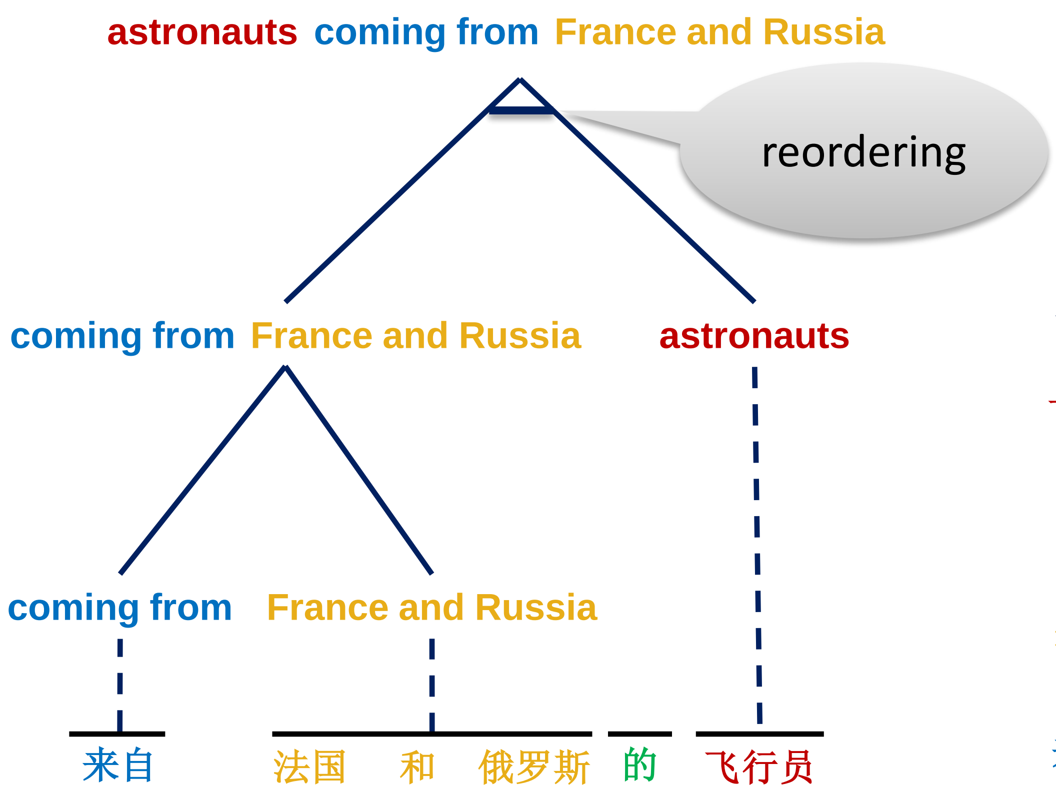
Machine Translation Pyramid



Syntax-based SMT Models

- Formal synchronous syntax
  - BTG (Bracketing Transduction Grammar)
    - $X \rightarrow [XX] \quad X \rightarrow < XX > \quad X \rightarrow \alpha, \gamma$
  - Hiero rules
    - $X \rightarrow < \gamma, \alpha, \sim >$
    - X → 与 X1 有 X2, have X2 with X1
- Linguistic synchronous syntax
  - $\mathrm{NP} \rightarrow \mathrm{VP}_1$ (NP (NNS (*fei-xing-yuan*))), astaurants VP_1

BTG decoding



$X_1\ X_2$ ||| $X_2\ X_1$

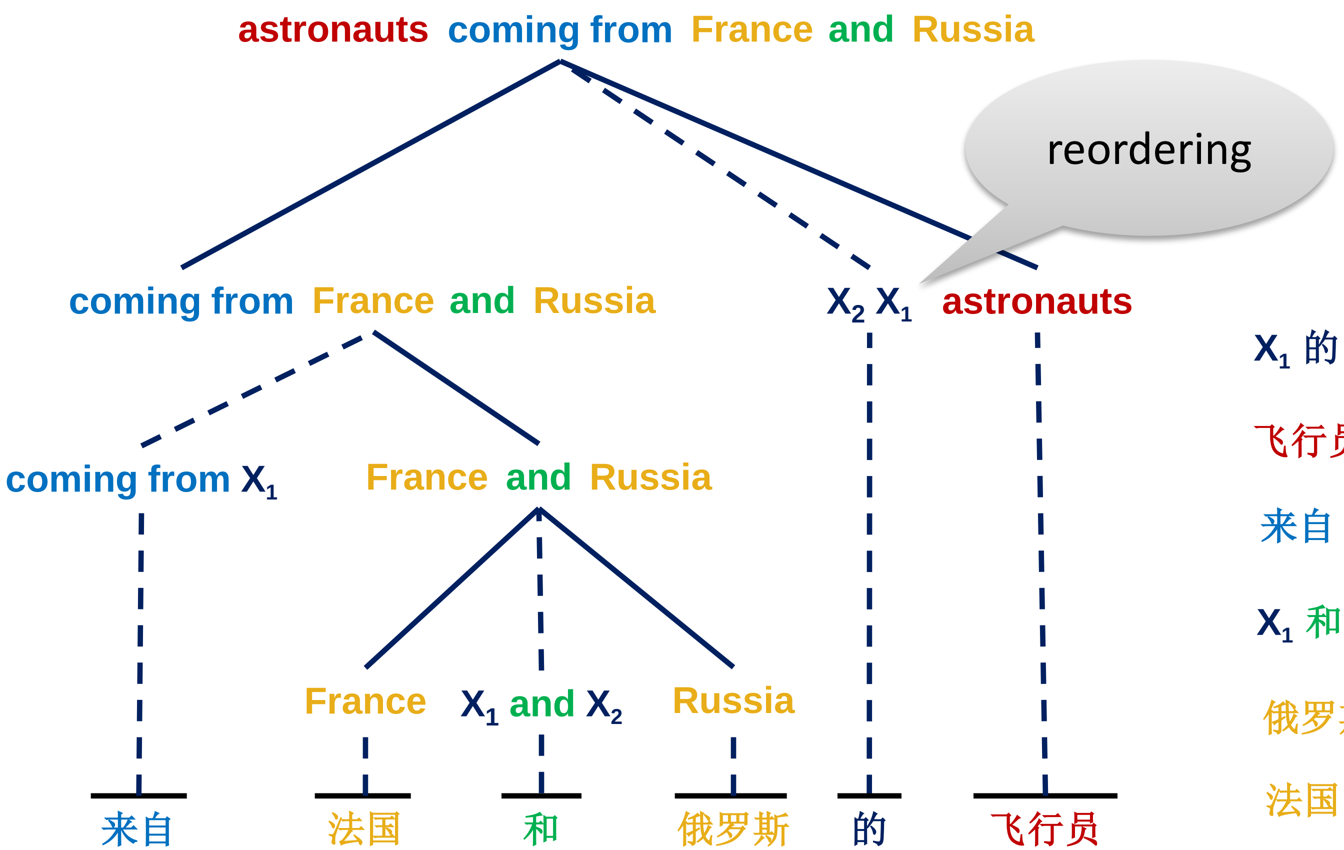
飞行员 ||| astronauts

$X_1\ X_2$ ||| $X_1\ X_2$

法国 和 俄罗斯 ||| France and Russia

来自 ||| coming from

Hiero decoding



X_1 的 X_2 ||| X_2 X_1

飞行员 ||| astronauts

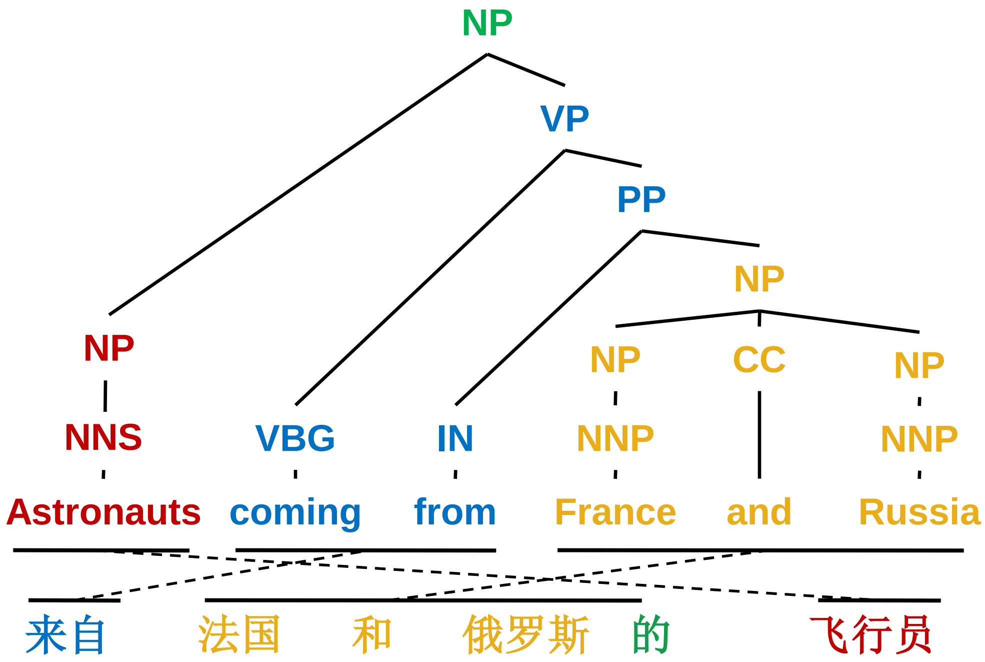
来自 X_1 ||| coming from X_1

X_1 和 X_2 ||| X_1 and X_2

俄罗斯 ||| Russia

法国 ||| France

Syntax-based Model



SMT Bet

- String models vs. syntax-based models

	Arabic-English	Chinese-English
Google	42.81	33.16
ISI	39.08	33.93

NIST 2006 MT Evaluation results

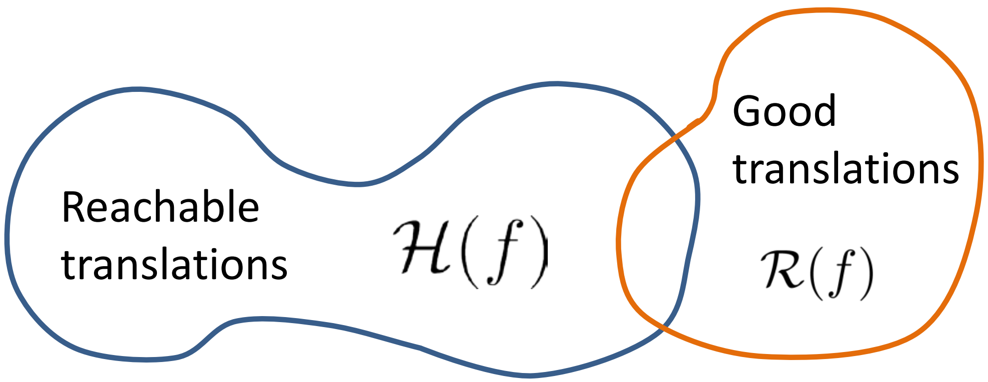
Consensus-based SMT Decoding

- Minimum Bayes Risk (MBR) decoding
  - Consensus decoding using one system
- System combination
  - Consensus decoding using multiple systems

MAP Decoding

- Maximum A Posteriori (MAP) decision rule

$$e^* = \mathrm{argmax}_{e \in \mathcal{H}(f)} P(e|f)$$



Minimum Bayes-Risk Decoding

Finding consensus within one system

- Risk function
  - $R(e) = \sum_{r \in \mathcal{R}(f)} L(e,r) P(r|f) \qquad \mathcal{R}(f) = (r_1, r_2, \ldots, r_n)$
- MBR decision Rule
  - Ideal MBR

$$e^* = \mathrm{argmin}_e R(e) = \mathrm{argmin}_{e \in \mathcal{H}(f)} \sum_{r \in \mathcal{R}(f)} L(e,r) P(r|f)$$

  - MBR in practice

$$e^* = \mathrm{argmin}_e R(e) = \mathrm{argmin}_{e \in \mathcal{H}(f)} \sum_{e' \in \mathcal{H}(f)} L(e,e') P(e'|f)$$

Minimum Bayes-Risk Decoding

Finding consensus within one system

- Loss vs. Gain
  - $L(e, e') = C(f) - G(e, e')$
  - $e^* = \mathrm{argmax}_{e \in \mathcal{H}(f)} \sum_{e' \in \mathcal{H}(f)} G(e, e') P(e'|f)$
- Consensus measure
  - Unigram overlapping
  - N-gram overlapping
  - Structure overlapping

Literature of MBR Decoding

- Speech recognition
  - Bickel and Doksum, 1977
- Statistical machine translation
  - Kumar and Byrne, 2004
  - Tromble et al., 2008
  - DeNero et al., 2009
  - Li et al., 2009

System Combination

Finding consensus between systems

- Combining outputs from multiple machine translation systems
  - Rosti et al., NAACL 2007
  - Sentence-level, phrase-level, word-level

- Relation between word-level combination consensus decoding
  - $c(e|f) = \sum_{w_i \in e = w_1, \ldots, w_n} c(w_i) + \mu N_{null}(e) \quad c(w_i) = \sum_{j}^{N_s} \lambda_j c(w_i, j)$

- More work
  - Confusion network decoding
    - Bangalore et al., 2001
    - Matusov et al., 2006
    - Sim et al., 2007
  - Better word alignment
    - Rosti et al., 2008
    - Xiaodong He et al., 2008
    - Li et al., 2009

Combination Approach

- Word-level system combination

  - N-best translation generation
  - Skeleton translation selection
  - Confusion network construction
  - Confusion network decoding

Combination Method

- Word-level system combination

  - N-best translation generation
  - Skeleton translation selection
  - Confusion network construction
  - Confusion network decoding

Combination Method

- Word-level system combination

➢ N-best translation generation

➢ Skeleton translation selection

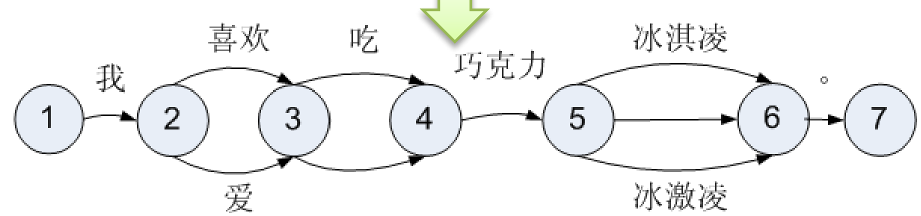
➢ Confusion network construction

➢ Confusion network decoding

I like eating chocolate icecream.

我 喜欢 巧克力 冰激凌 。

我	喜欢	~	巧克力	冰激凌	。
我	喜欢	吃	巧克力	~	。
我	爱	吃	巧克力	冰淇凌	。
我	爱	~	巧克力	冰激凌	。



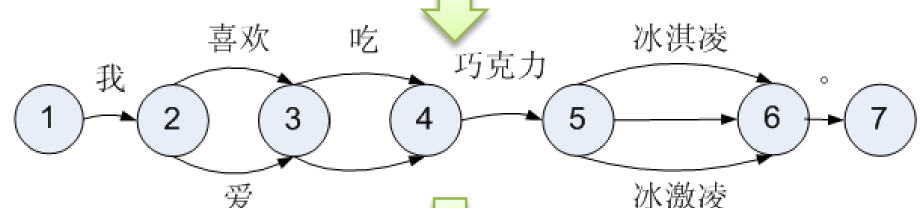
Combination Method

- Word-level system combination

I like eating chocolate icecream.

我 喜欢 巧克力 冰激凌 。

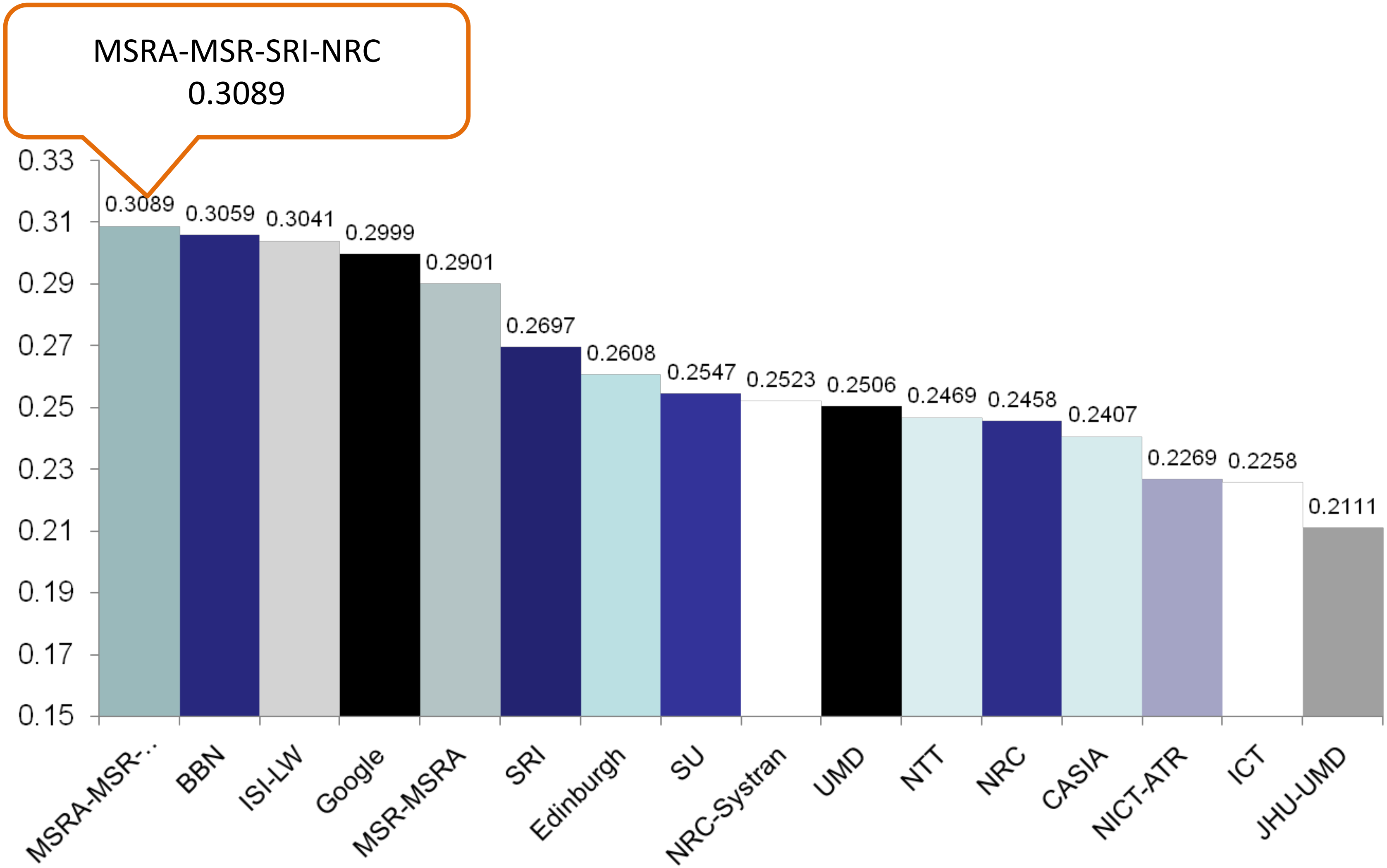
我	喜欢	~	巧克力	冰激凌	。
我	喜欢	吃	巧克力	~	。
我	爱	吃	巧克力	冰淇淩	。
我	爱	~	巧克力	冰激凌	。



我 爱 吃 巧克力 冰激凌 。

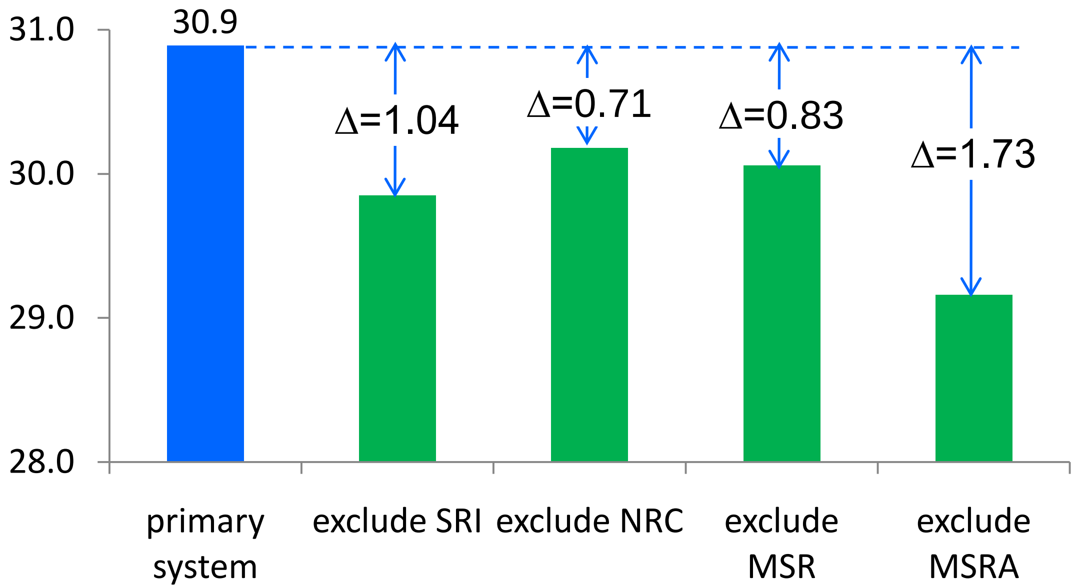
> N-best translation generation
> Skeleton translation selection
> Confusion network construction
> Confusion network decoding

Chinese-English Results



Impact of Systems of Each Site

- Each site provided individual systems for combination
  - MSR: 3 systems, MSRA: 3 systems, NRC:1 system, SRI: 1 system
- The impact is measured by the BLEU score loss due to excluding system(s) of that site



MSR-MSRA-NRC-SRI joint entry

Cited from X He, et al, NIST2008 SMT workshop

Machine Translation Evaluation

- Human evaluation
- 信达雅
    - Adequacy
        - I cannot agree you more → 我不能同意你更多
    - Fluency
        - How old are you → 怎么老是你
  - High cost
- Automatic evaluation
  - Convenient
  - May not be consistent with human preference

BLEU Evaluation Metric

(Papineni et al, ACL-2002)

Reference (human) translation:

The U.S. island of Guam is maintaining a high state of alert after the Guam airport and its offices both received an e-mail from someone calling himself the Saudi Arabian Osama bin Laden and threatening a biological/chemical attack against public places such as the airport .

Machine translation:

The American [?] international airport and its the office all receives one calls self the sand Arab rich business [?] and so on electronic mail , which sends out ; The threat will be able after public place and so on the airport to start the biochemistry attack , [?] highly alerts after the maintenance.

- N-gram precision (score is between 0 & 1)
  - What percentage of machine n-grams can be found in the reference translation?
    - An n-gram is an sequence of n words
  - Not allowed to use same portion of reference translation twice (can't cheat by typing out "the the the the the")

- Brevity penalty
  - Can't just type out single word "the" (precision 1.0!)

*** Amazingly hard to "game" the system (i.e., find a way to change machine output so that BLEU goes up, but quality doesn't)

slide from Kevin Knight's tutorial

BLEU Metric

$$BLEU = BP \bullet \exp\left(\sum_{1}^{N} w_n \log p_n\right)$$

$$BP = \begin{cases} 1 & if\ c > r \\ e^{(1-r/c)} & if\ c \le r \end{cases}$$

$$\log BLEU = \min(1 - \frac{r}{c}, 0) + \sum_{1}^{N} w_n \log p_n$$

$$N = 4, w_n = 1/N$$

BLEU: An Example

- **Candidate 1: the book is on the desk**

- **Reference 1: there is a book on the desk**
- **Reference 2: the book is on the table**

unigram:	**bigram:**	**trigram:**
$Count_{clip}(the) = 2$	$Count_{clip}(the, book) = 1$	$Count_{clip}(the, book, is) = 1$
$Count_{clip}(book) = 1$	$Count_{clip}(book, is) = 1$	$Count_{clip}(book, is, on) = 1$
$Count_{clip}(is) = 1$	$Count_{clip}(is, on) = 1$	$Count_{clip}(is, on, the) = 1$
$Count_{clip}(on) = 1$	$Count_{clip}(on, the) = 1$	$Count_{clip}(on, the, desk) = 1$
$Count_{clip}(desk) = 1$	$Count_{clip}(the, desk) = 1$	
$\sum_{unigram \in C} Count(unigram) = 6$	$\sum_{bigram \in C} Count(bigram) = 5$	$\sum_{trigram \in C} Count(trigram) = 4$
$p_1 = 1$	$p_2 = 1$	$p_3 = 1$

$$\left.\begin{array}{l} c = 6 \\ r = 6 \end{array}\right\} = e^{1-\frac{r}{c}} = e^0 = 1 = BP$$

$$BLEU = BP \bullet \exp\left(\sum_{n=1}^{N} w_n \log p_n\right)$$

$$= \exp\left[\frac{1}{3}(\log 1 + \log 1 + \log 1)\right] = \sqrt[3]{1 \cdot 1 \cdot 1} = 1$$

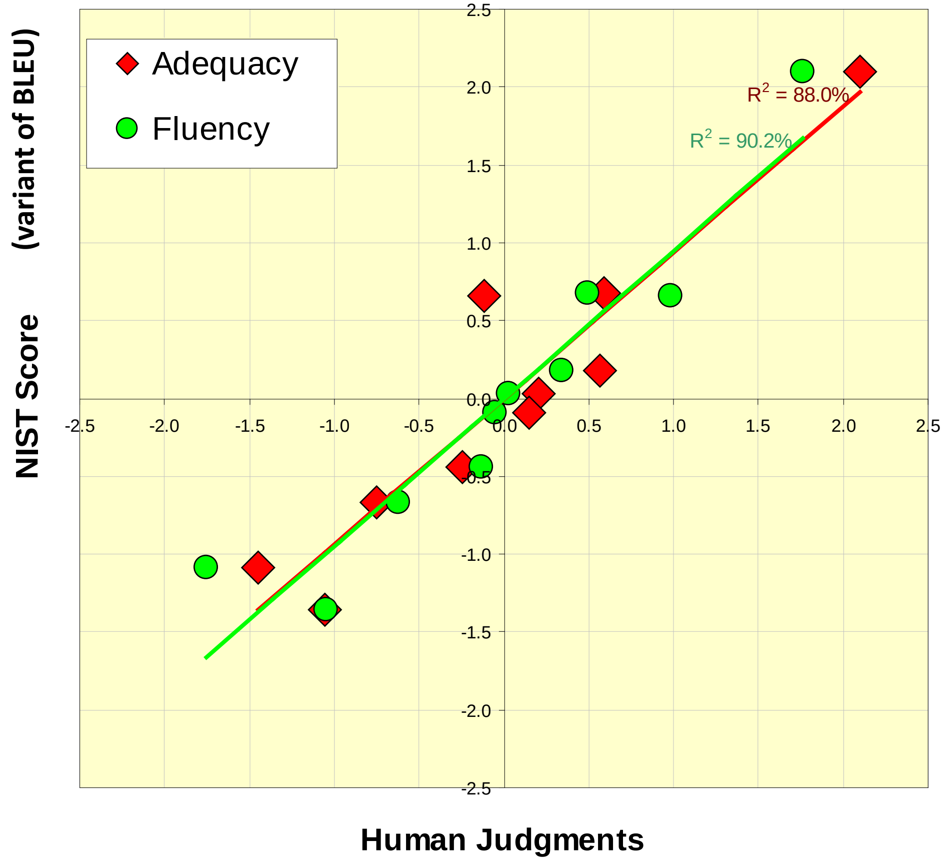
BLEU

I do not speak French . reference

I not speak French . output

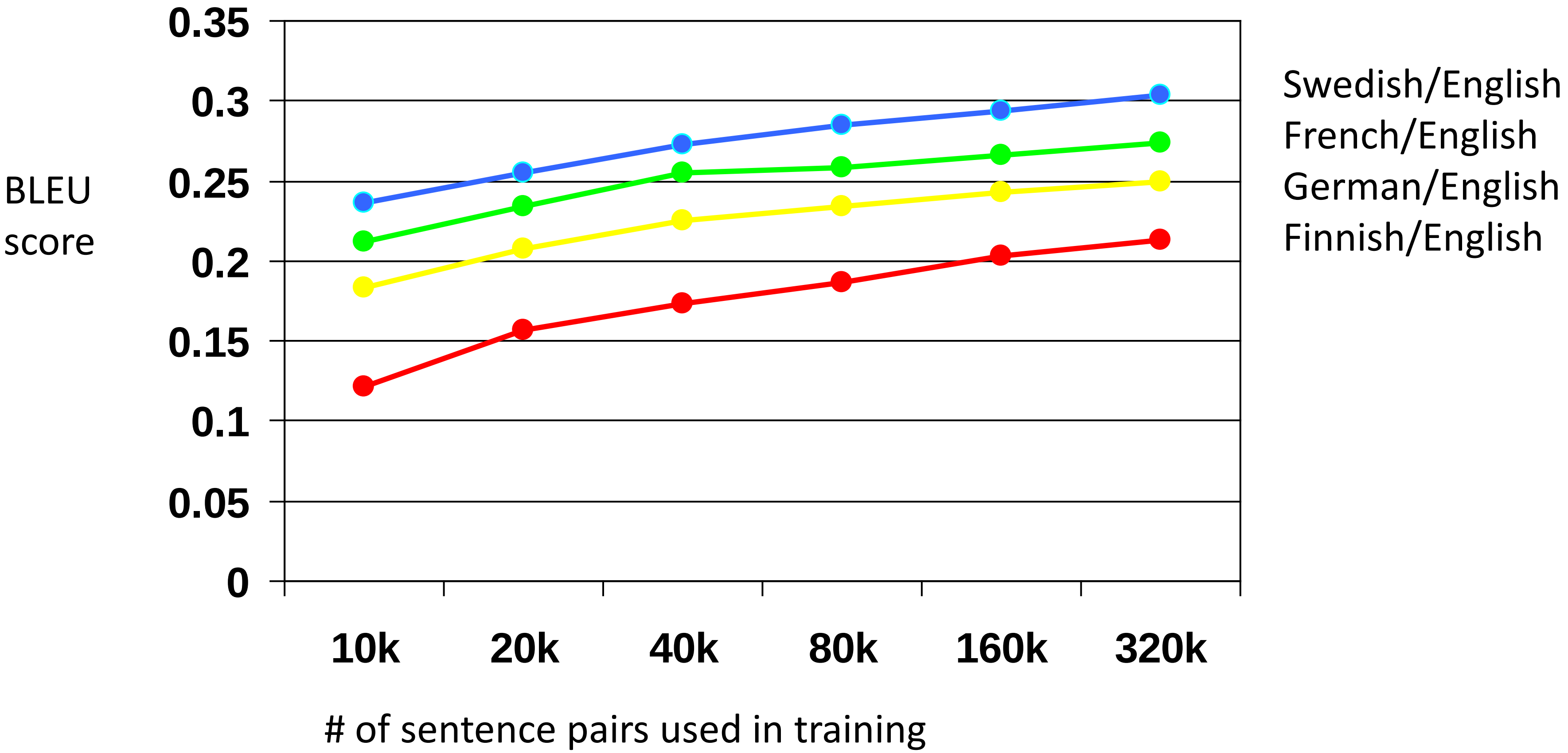
$$\Big(\underbrace{\frac{5}{5}}_{\text{1-grams}} \times \underbrace{\frac{3}{4}}_{\text{2-grams}} \times \underbrace{\frac{2}{3}}_{\text{3-grams}} \times \underbrace{\frac{1}{4}}_{\text{4-grams}}\Big)^{\frac{1}{4}} \times \underbrace{e^{1-5/6}}_{\text{brevity penalty}}$$

BLEU Tends to Predict Human Judgments



slide from G. Doddington (NIST)

Observing Learning Curves using Bleu



Experiments by Philipp Koehn

More Comments on BLEU

- Cannot be used to evaluate human translators
- Not work well for heterogeneous systems
  - E.g. a statistical and a rule-based system
- Test set should be large enough
- Marginal improvements may not be meaningful

Other Metrics

- Metrics based on lexical similarity
  - (most of the metrics!)
- Edit Distance
  - WER, PER, TER
- Precision
  - BLEU, NIST, WNM
- Recall
  - ROUGE, CDER
- Precision/Recall
  - GTM, METEOR, BLANC, SIA

Recommend Readings

1. A Statistical MT Tutorial Workbook. Kevin Knight. 1999.
   Very good introduction to word-based statistical machine translation.
   Written in an informal, understandable, tutorial oriented style.
2. The Mathematics of Statistical Machine Translation:
   Parameter Estimation. P. F. Brown, S. A. Della Pietra,
   V. J. Della Pietra and R.L. Mercer. 1993.
3. Phrase based statistical MT:
   Statistical Phrase-Based Translation.
   Philipp Koehn, Franz Jasof Ock and Daniel Marcu. 2003.
4. Discriminative Training and Maximum Entropy Models for Statistical Machine Translation.
   Och and Ney. 2002.
5. BLEU: A Method for Automatic Evaluation of Machine Translation.
   Papineni, Roukos, Ward and Zhu. 2001.