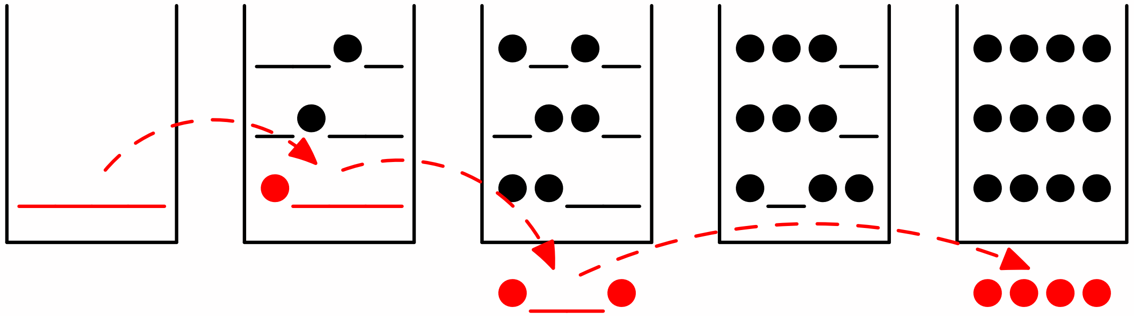
Search Aware Tuning for Machine Translation

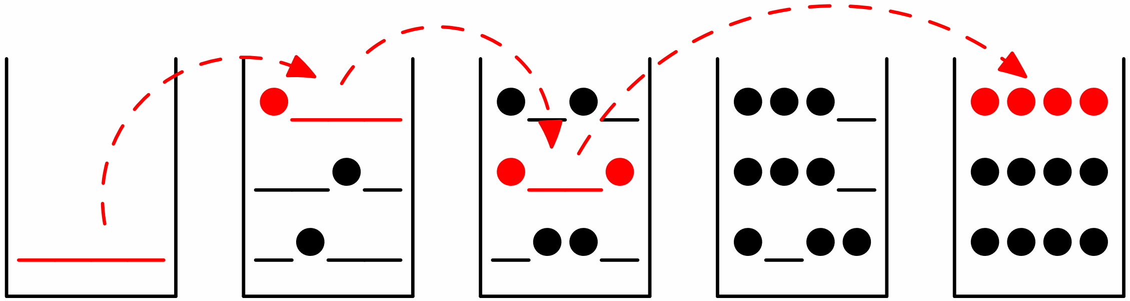


Lemao Liu   Liang Huang

City University of New York

EMNLP 2014. Presented by Taro Watanabe.

Search Aware Tuning for Machine Translation

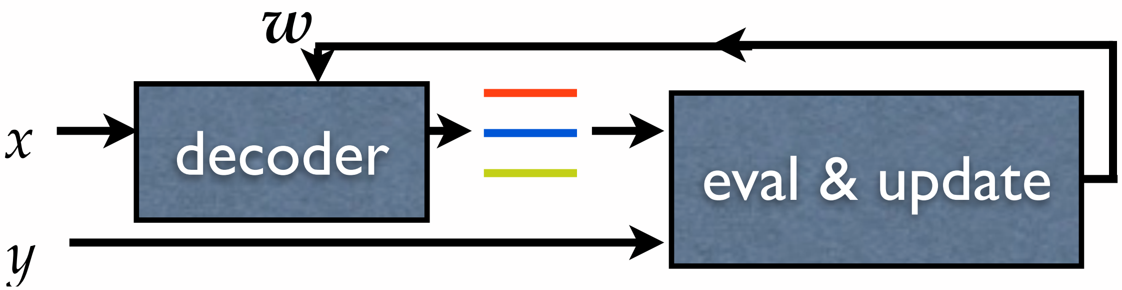


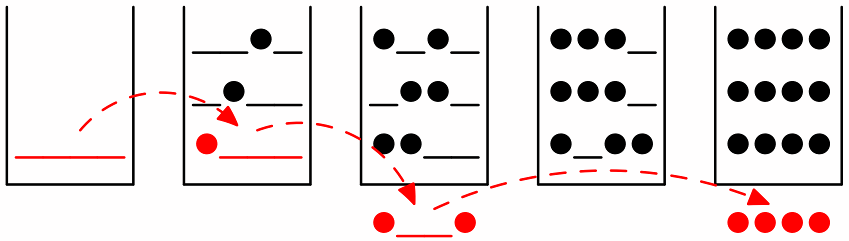
Lemao Liu Liang Huang

City University of New York

EMNLP 2014. Presented by Taro Watanabe.

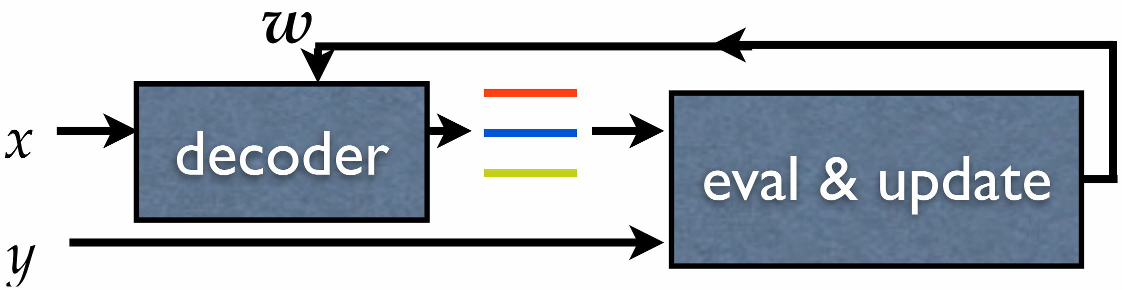
Parameter Tuning for MT





- most tuning methods view MT decoder as a black box
  - “search-agnostic” tuning (MERT, MIRA, PRO, ...)
- but actually search error is a main reason of bad quality
  - potentially good sub-translations pruned early in search
  - final *k*-best list also lacks diversity

Parameter Tuning for MT

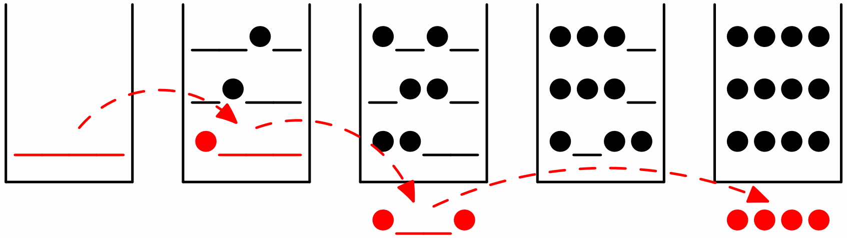

w

x

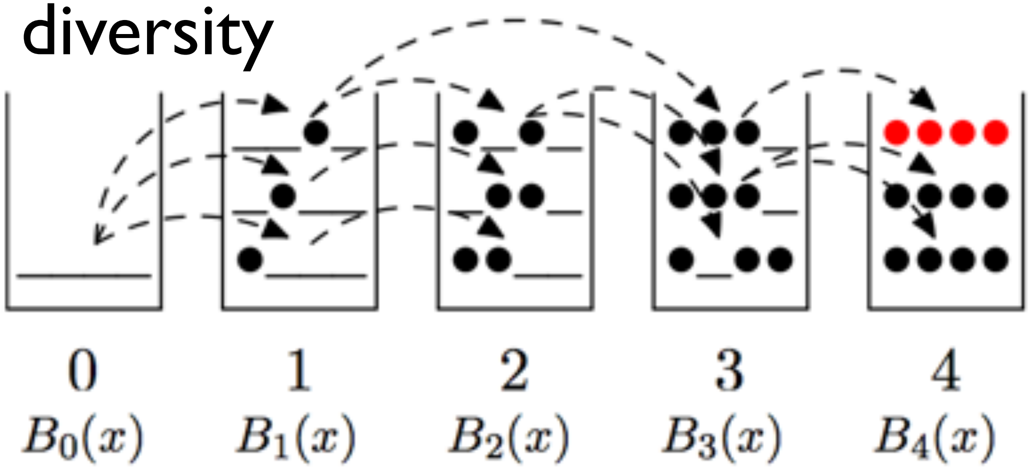
decoder

eval & update

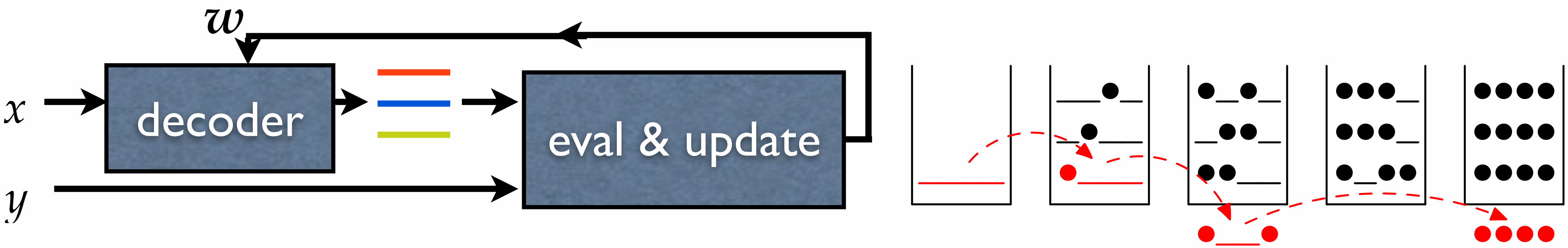
y



- most tuning methods view MT decoder as a black box
  - “search-agnostic” tuning (MERT, MIRA, PRO, ...)
- but actually search error is a main reason of bad quality
  - potentially good sub-translations pruned early in search
  - final *k*-best list also lacks diversity

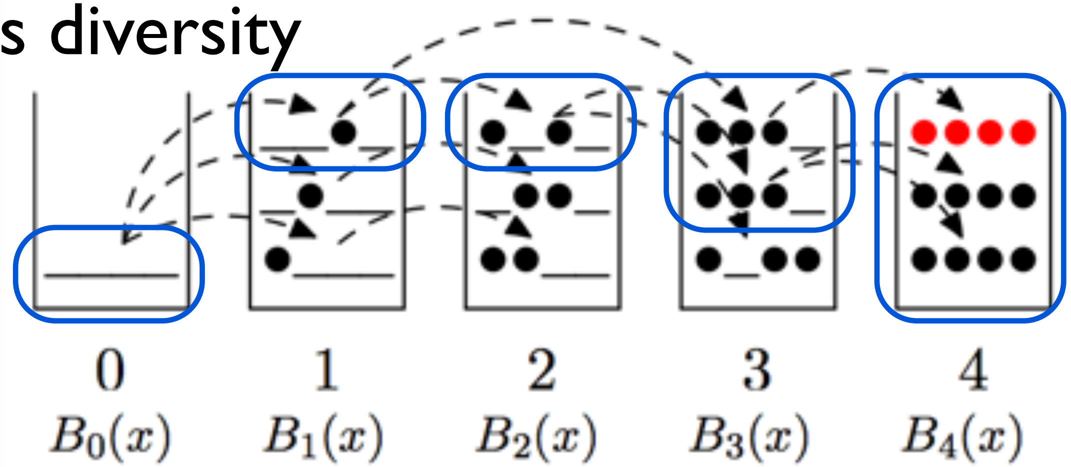


Parameter Tuning for MT

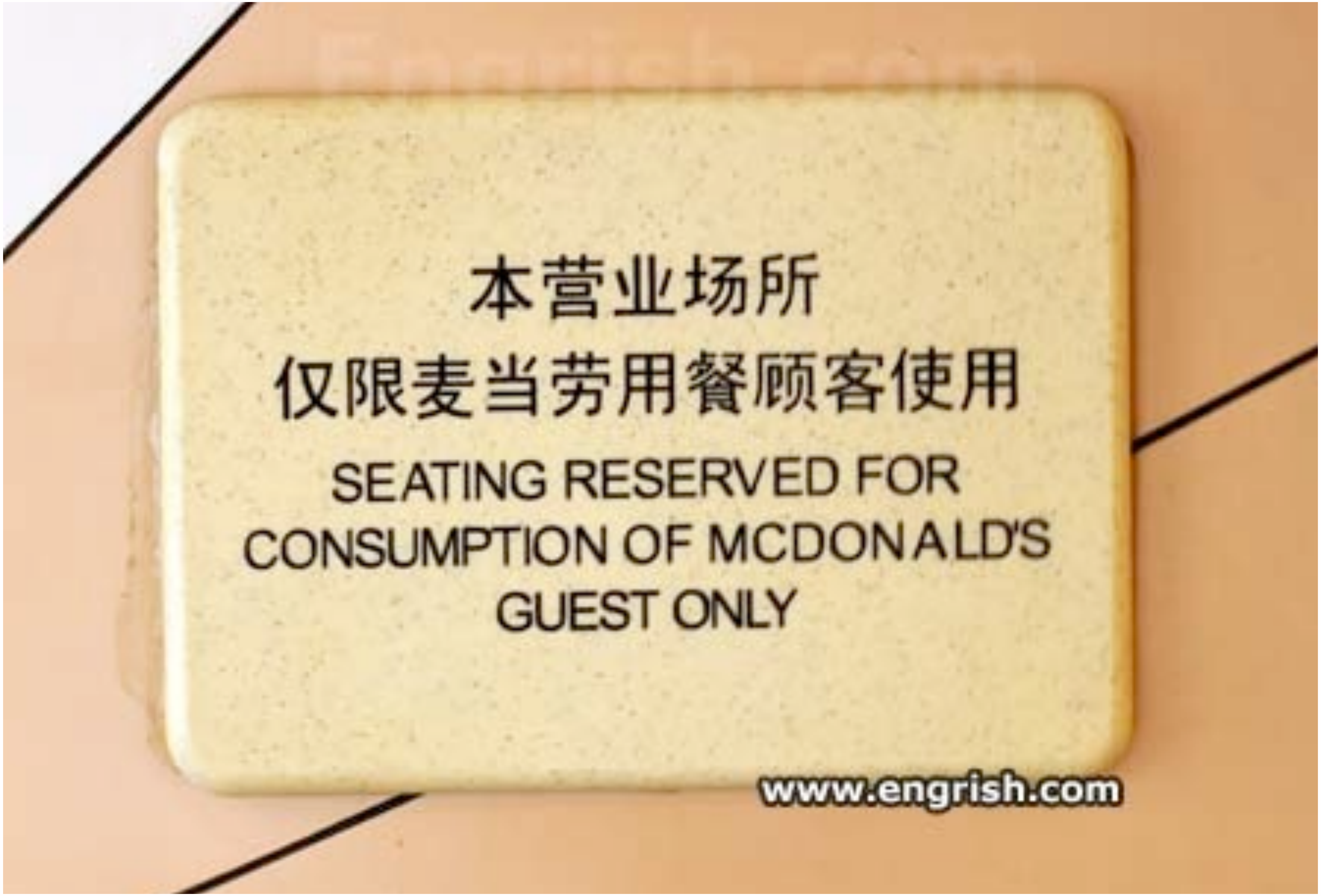


- most tuning methods view MT decoder as a black box
  - “search-agnostic” tuning (MERT, MIRA, PRO, ...)
- but actually search error is a main reason of bad quality
  - potentially good sub-translations pruned early in search
  - final *k*-best list also lacks diversity

cf.: Y-chromosome Adam
Mitochondria Eva



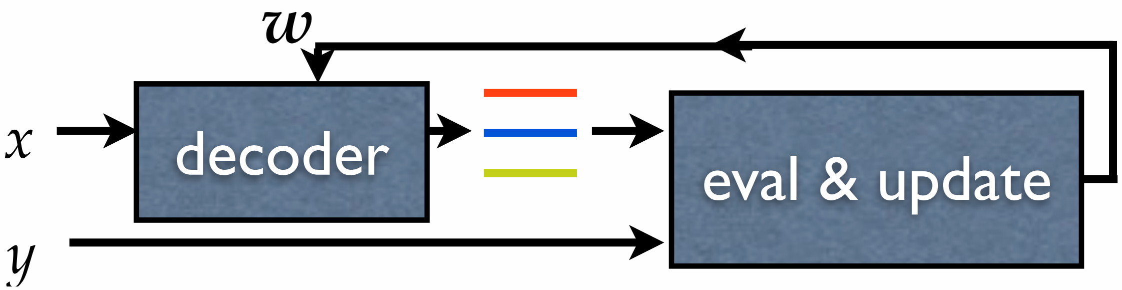
Search Error in MT

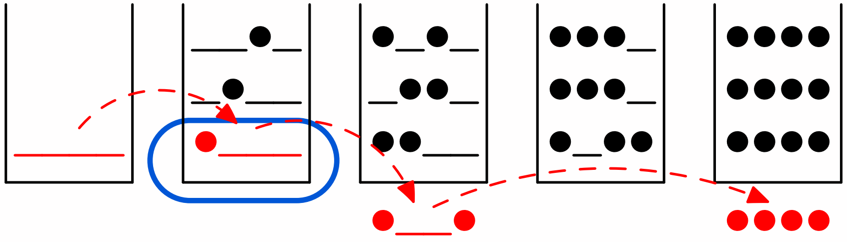


本营业场所
仅限麦当劳用餐顾客使用

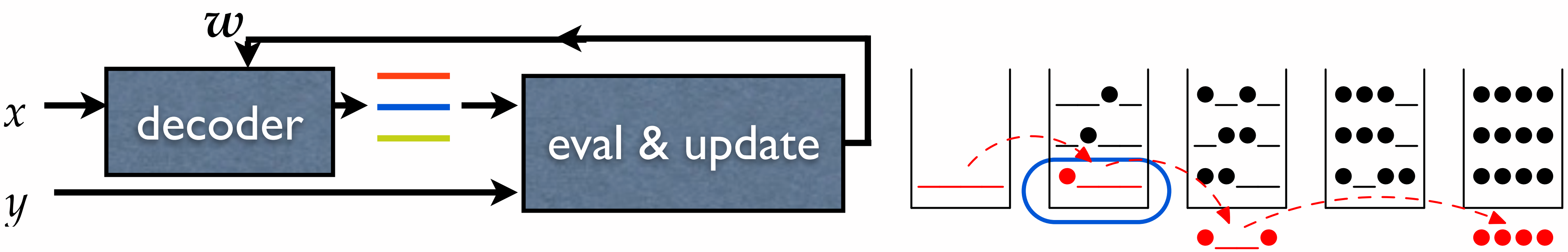
SEATING RESERVED FOR
CONSUMPTION OF MCDONALD'S
GUEST ONLY

Parameter Tuning for MT



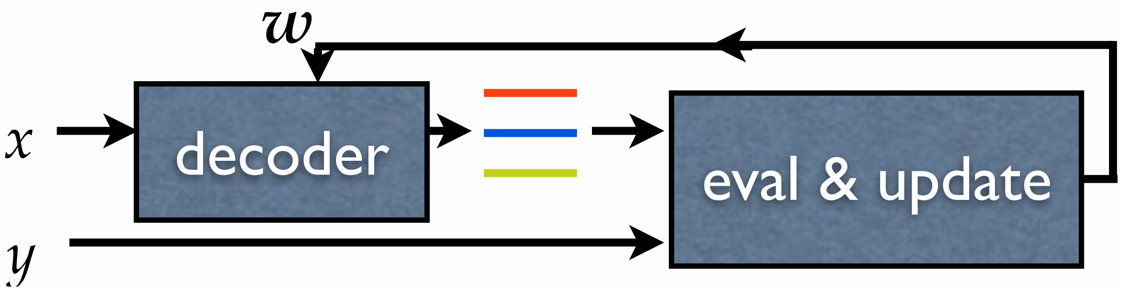


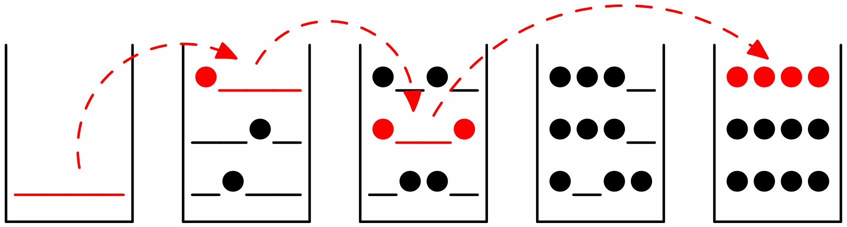
Parameter Tuning for MT



- most tuning methods view MT decoder as a black box
    - "search-agnostic" tuning (MERT, MIRA, PRO, ...)
- but actually search error is a main reason of bad quality
    - potentially good sub-translations pruned early in search

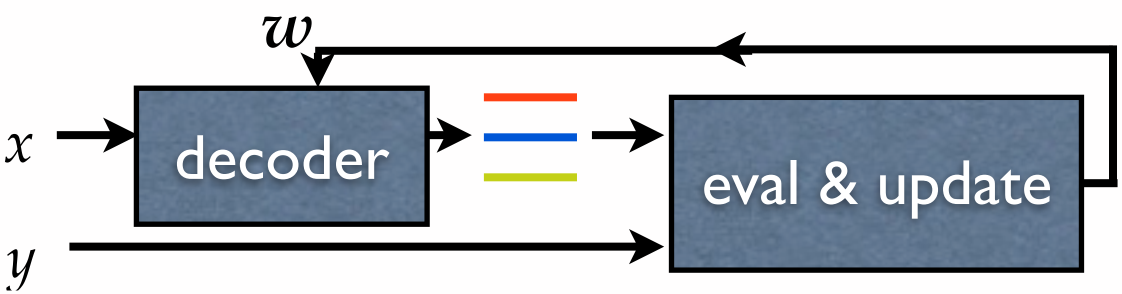
Parameter Tuning for MT

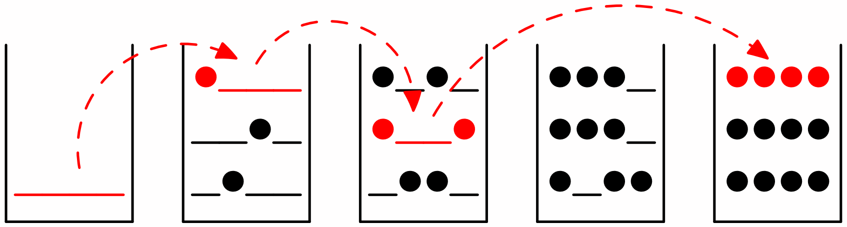




- most tuning methods view MT decoder as a black box
  - “search-agnostic” tuning (MERT, MIRA, PRO, ...)
- but actually search error is a main reason of bad quality
  - potentially good sub-translations pruned early in search

Parameter Tuning for MT



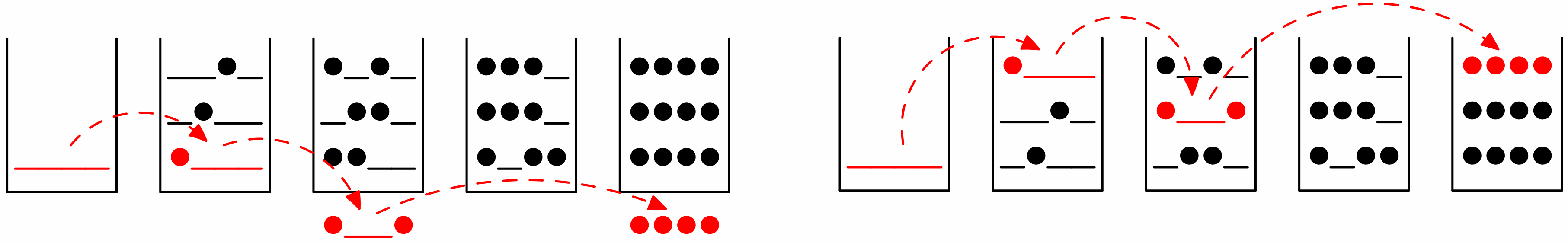


- most tuning methods view MT decoder as a black box
  - "search-agnostic" tuning (MERT, MIRA, PRO, ...)
- but actually search error is a main reason of bad quality
  - potentially good sub-translations pruned early in search
- Q: how to promote these promising sub-derivations?
- A: tune the ranking of non-final bins as well as final bin
  - "search-aware tuning" (SA-MERT, SA-MIRA, SA-PRO, ...)
  - Q: how to evaluate the "potential" of a sub-derivation?

Outline

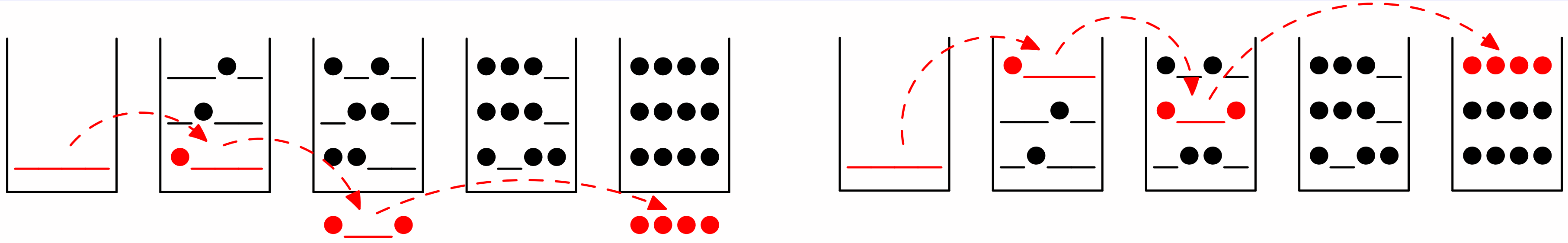
- Motivations
- Evaluating Partial Derivations
  - challenges
  - method 1: naive partial BLEU
  - method 2: novel potential BLEU
- Search-Aware MERT, MIRA, and PRO
- Experiments
  - consistent +1 BLEU improvement with dense features

Challenges in Partial Evaluation



- challenge 1: there is no "partial" references
- challenge 2: in phrase-based MT, partial translations in the same bin may cover different source words

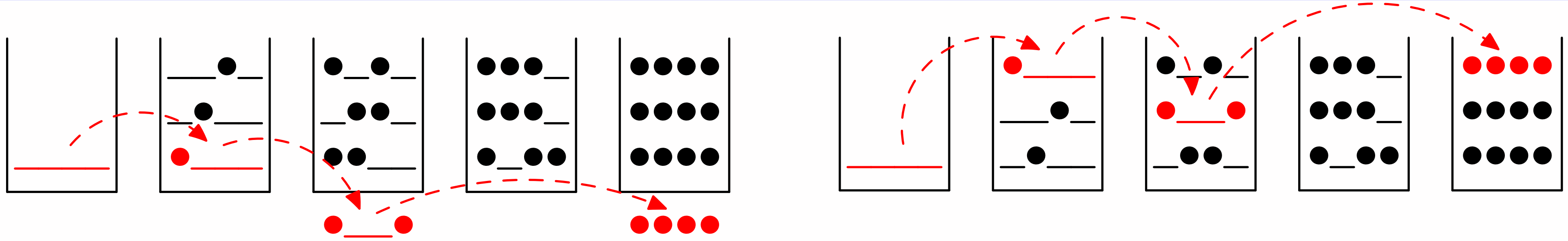
Challenges in Partial Evaluation



- challenge 1: there is no “partial” references
- challenge 2: in phrase-based MT, partial translations in the same bin may cover different source words

source: 我 从 上海 飞 到 北京

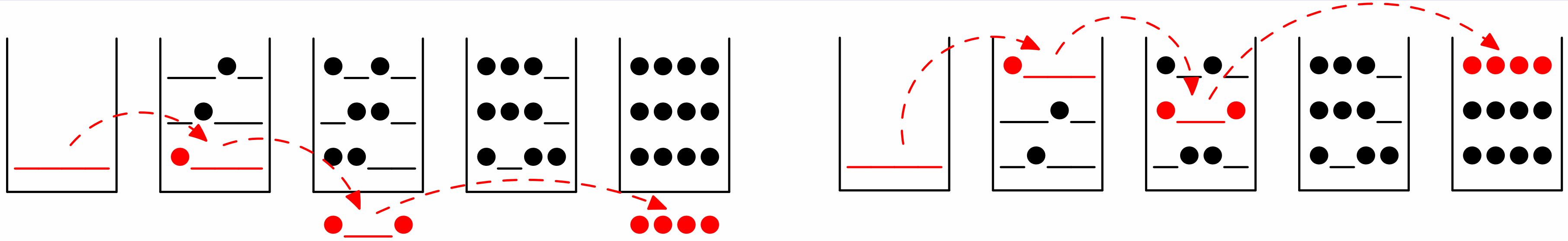
Challenges in Partial Evaluation



- challenge 1: there is no “partial” references
- challenge 2: in phrase-based MT, partial translations in the same bin may cover different source words

source:	我	从	上海	飞	到	北京
gloss:	I	from	Shanghai	fly	to	Beijing

Challenges in Partial Evaluation



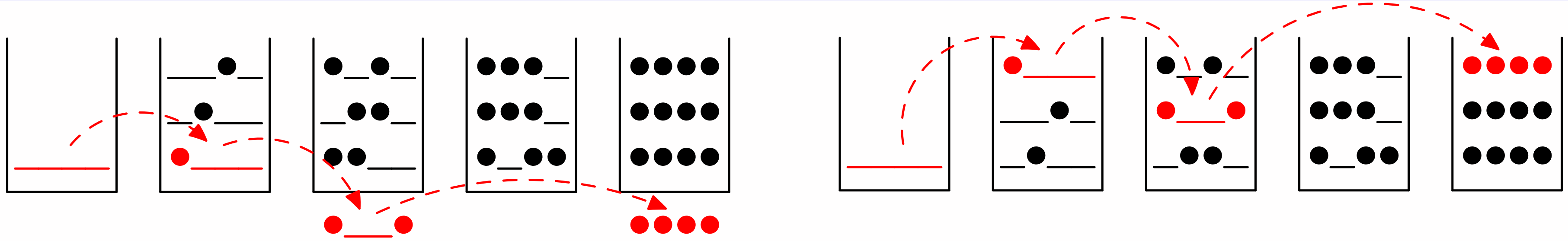
- challenge 1: there is no "partial" references
- challenge 2: in phrase-based MT, partial translations in the same bin may cover different source words

source: 我 从 上海 飞 到 北京

gloss: I from Shanghai fly to Beijing

reference: I flew from Shanghai to Beijing

Challenges in Partial Evaluation



- challenge 1: there is no “partial” references
- challenge 2: in phrase-based MT, partial translations in the same bin may cover different source words

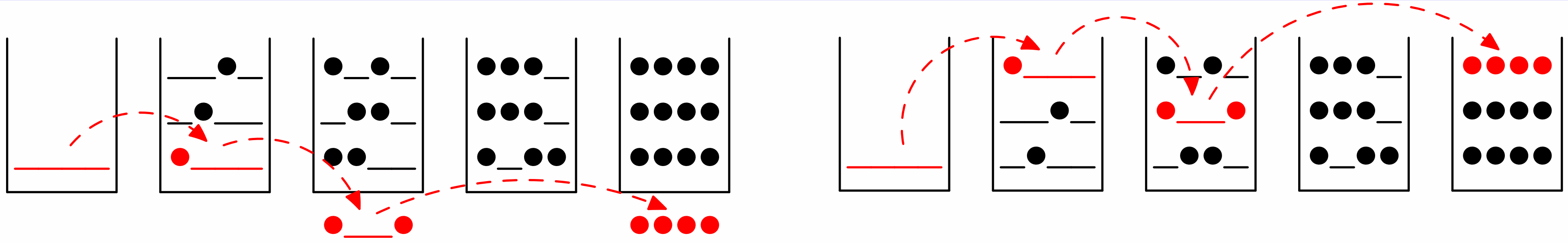
source: 我 从 上海 飞 到 北京

gloss: I from Shanghai fly to Beijing

reference: I flew from Shanghai to Beijing

partial 1: I from

Challenges in Partial Evaluation



- challenge 1: there is no "partial" references
- challenge 2: in phrase-based MT, partial translations in the same bin may cover different source words

source: 我 从 上海 飞 到 北京

gloss: I from Shanghai fly to Beijing

reference: I flew from Shanghai to Beijing

partial 1: I from
partial 2: I fly

Method 1: Naive Partial BLEU

- naive solution: just evaluate against the full reference
  - but using a prorated reference length
    - proportional to number of source words translated so far
  - inspired by oracle extraction (Li & Khudanpur 10; Chiang 12)
- problem: favoring those translating "easier" words first

source: 我 从 上海 飞 到 北京

gloss: I from Shanghai fly to Beijing

reference: I flew from Shanghai to Beijing

partial 1: I from unigram=2

partial 2: I fly unigram=1

Method 1: Naive Partial BLEU

- naive solution: just evaluate against the full reference
  - but using a prorated reference length
    - proportional to number of source words translated so far
  - inspired by oracle extraction (Li & Khudanpur 10; Chiang 12)
- problem: favoring those translating "easier" words first

source: 我 从 上海 飞 到 北京

gloss: I from Shanghai fly to Beijing

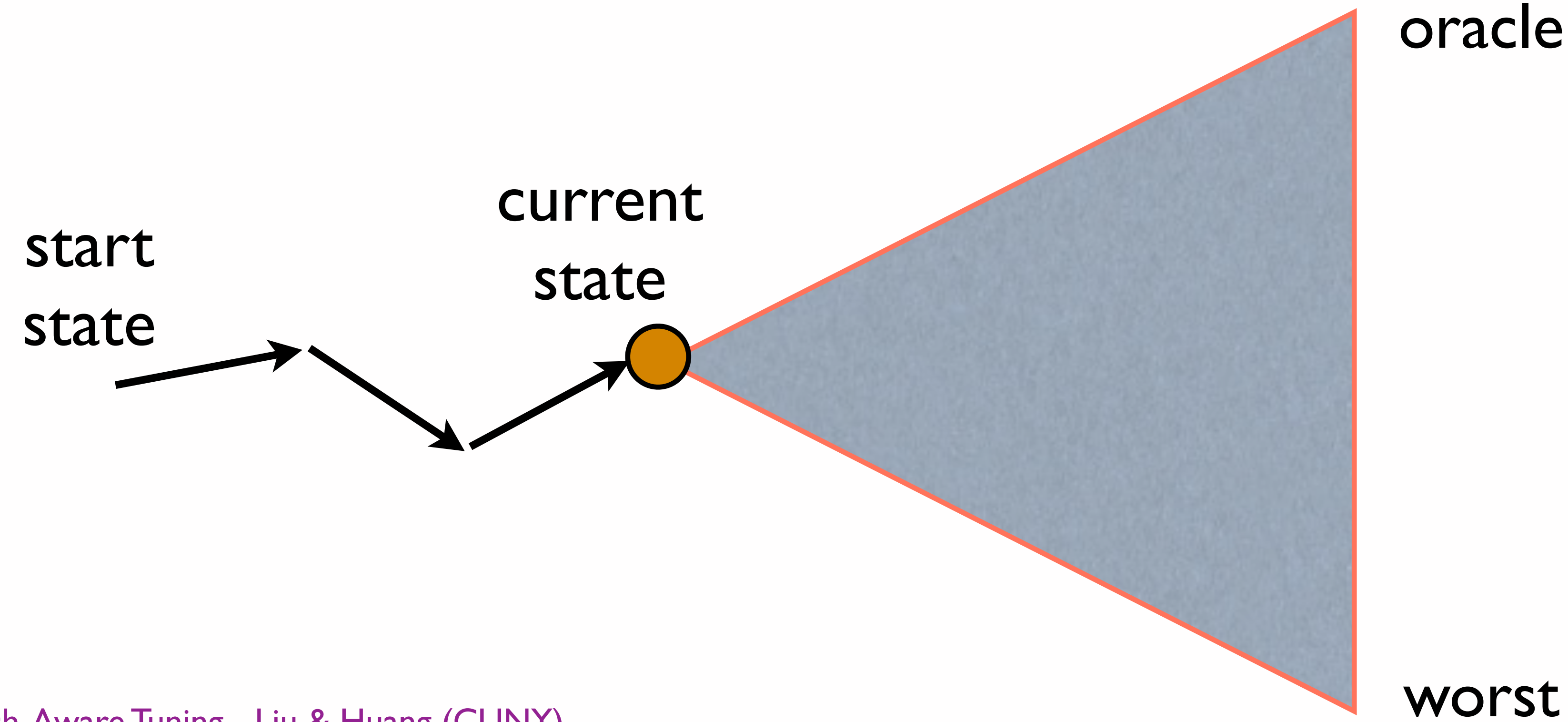
reference: I flew from Shanghai to Beijing

partial 1: I from unigram=2 ✔

partial 2: I fly unigram=1

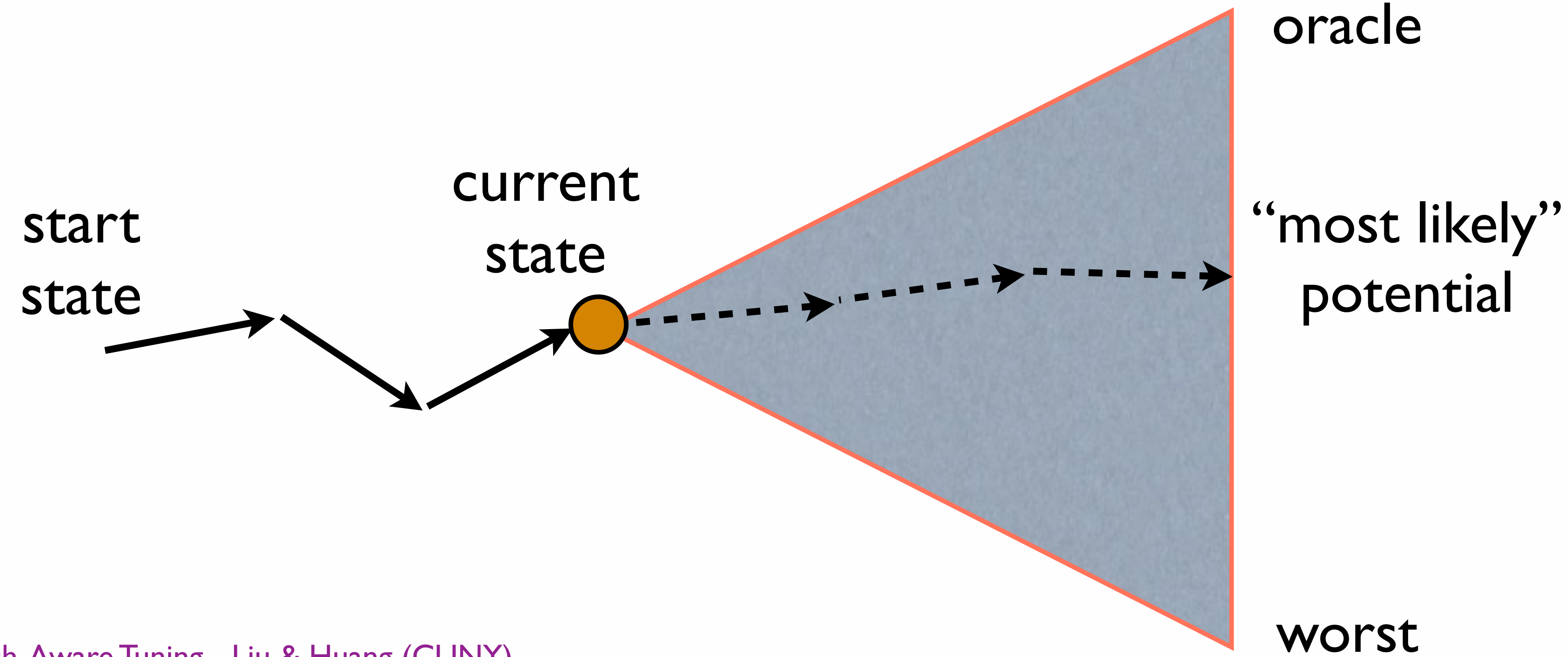
Evaluating the "Potential"

- better not evaluate partial translation as is, but its *potential*
- do we want the oracle (best) or average potential?
  - oracle is too hard to compute, and maybe not that useful
  - want the "most likely" potential given the current model



Evaluating the "Potential"

- better not evaluate partial translation as is, but its *potential*
- do we want the oracle (best) or average potential?
  - oracle is too hard to compute, and maybe not that useful
  - want the "most likely" potential given the current model



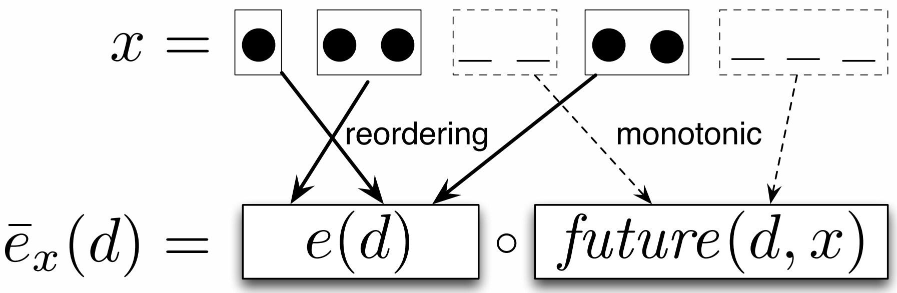
Method 2: Potential BLEU

- the “most likely potential” BLEU of a derivation
- extend partial derivation to cover uncovered words
    - using best monotonic translation for uncovered portions
    - inspired by “future cost” in phrase-based decoding
        - (inadmissible) A* heuristic computed by DP (Koehn, 2004)

source: 我 从 上海 飞 到 北京

gloss: I from Shanghai fly to Beijing

reference: I flew from Shanghai to Beijing



partial 1: I from

partial 2: I fly

Method 2: Potential BLEU

- the "most likely potential" BLEU of a derivation
- extend partial derivation to cover uncovered words
    - using best monotonic translation for uncovered portions
    - inspired by "future cost" in phrase-based decoding
        - (inadmissible) A* heuristic computed by DP (Koehn, 2004)

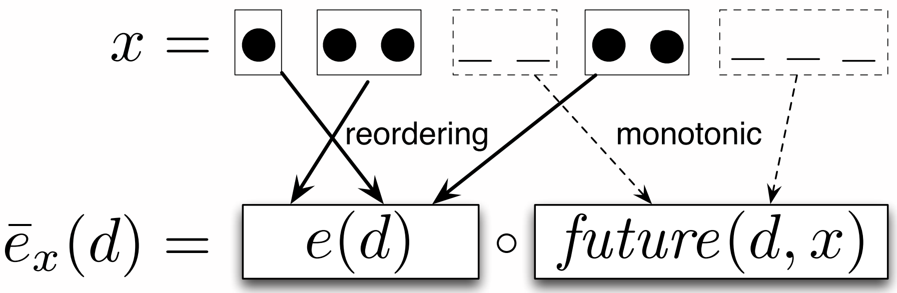
source: 我 从 上海 飞 到 北京

gloss: I from Shanghai fly to Beijing

reference: I flew from Shanghai to Beijing

partial 1: I from Shanghai fly to Beijing

partial 2: I fly



$$\bar{e}_x(d) = \boxed{e(d)} \circ \boxed{future(d, x)}$$

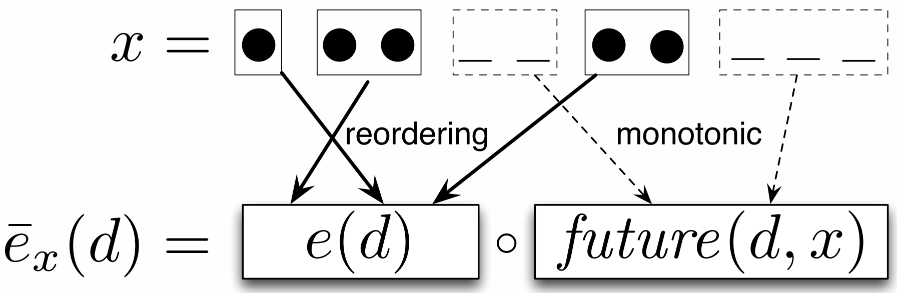
Method 2: Potential BLEU

- the “most likely potential” BLEU of a derivation
- extend partial derivation to cover uncovered words
  - using best monotonic translation for uncovered portions
  - inspired by “future cost” in phrase-based decoding
    - (inadmissible) A* heuristic computed by DP (Koehn, 2004)

source: 我 从 上海 飞 到 北京

gloss: I from Shanghai fly to Beijing

reference: I flew from Shanghai to Beijing



partial 1: I from Shanghai fly to Beijing

partial 2: I fly from Shanghai to Beijing

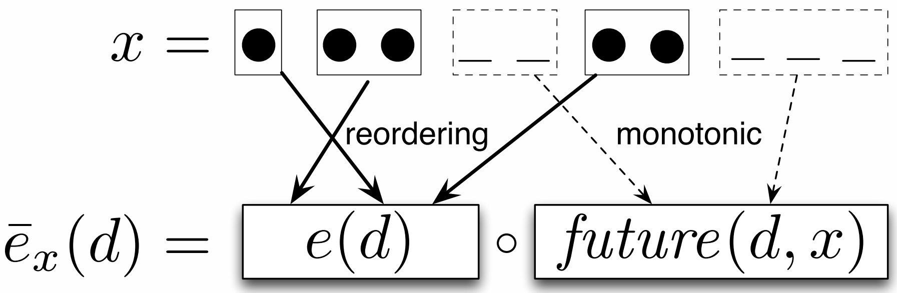
Method 2: Potential BLEU

- the "most likely potential" BLEU of a derivation
- extend partial derivation to cover uncovered words
  - using best monotonic translation for uncovered portions
  - inspired by "future cost" in phrase-based decoding
    - (inadmissible) A* heuristic computed by DP (Koehn, 2004)

source: 我 从 上海 飞 到 北京

gloss: I from Shanghai fly to Beijing

reference: I flew from Shanghai to Beijing



partial 1: I from Shanghai fly to Beijing

partial 2: I fly from Shanghai to Beijing

unigram=5, bi=2

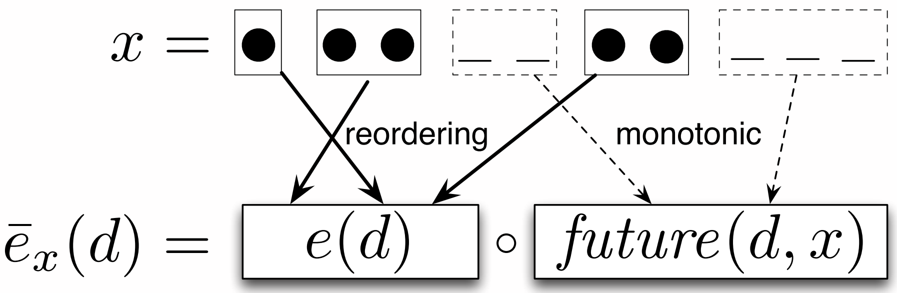
Method 2: Potential BLEU

- the “most likely potential” BLEU of a derivation
- extend partial derivation to cover uncovered words
  - using best monotonic translation for uncovered portions
  - inspired by “future cost” in phrase-based decoding
    - (inadmissible) A* heuristic computed by DP (Koehn, 2004)

source: 我 从 上海 飞 到 北京

gloss: I from Shanghai fly to Beijing

reference: I flew from Shanghai to Beijing



partial 1: I from \boxed{Shanghai fly to Beijing} unigram=5, bi=2

partial 2: I fly \boxed{from Shanghai to Beijing} unigram=5, bi=3, tri=2, 4gram=1

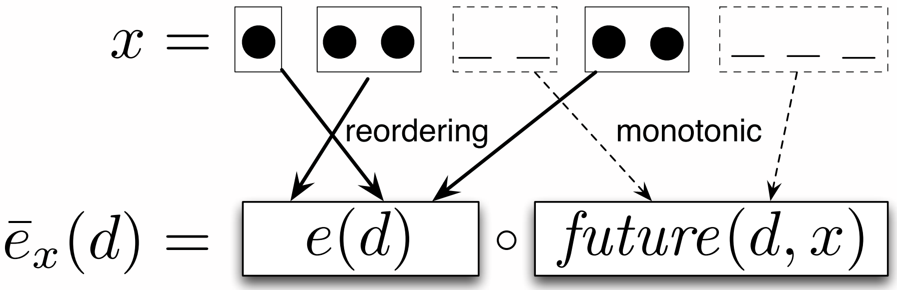
Method 2: Potential BLEU

- the "most likely potential" BLEU of a derivation
- extend partial derivation to cover uncovered words
  - using best monotonic translation for uncovered portions
  - inspired by "future cost" in phrase-based decoding
    - (inadmissible) A* heuristic computed by DP (Koehn, 2004)

source: 我 从 上海 飞 到 北京

gloss: I from Shanghai fly to Beijing

reference: I flew from Shanghai to Beijing



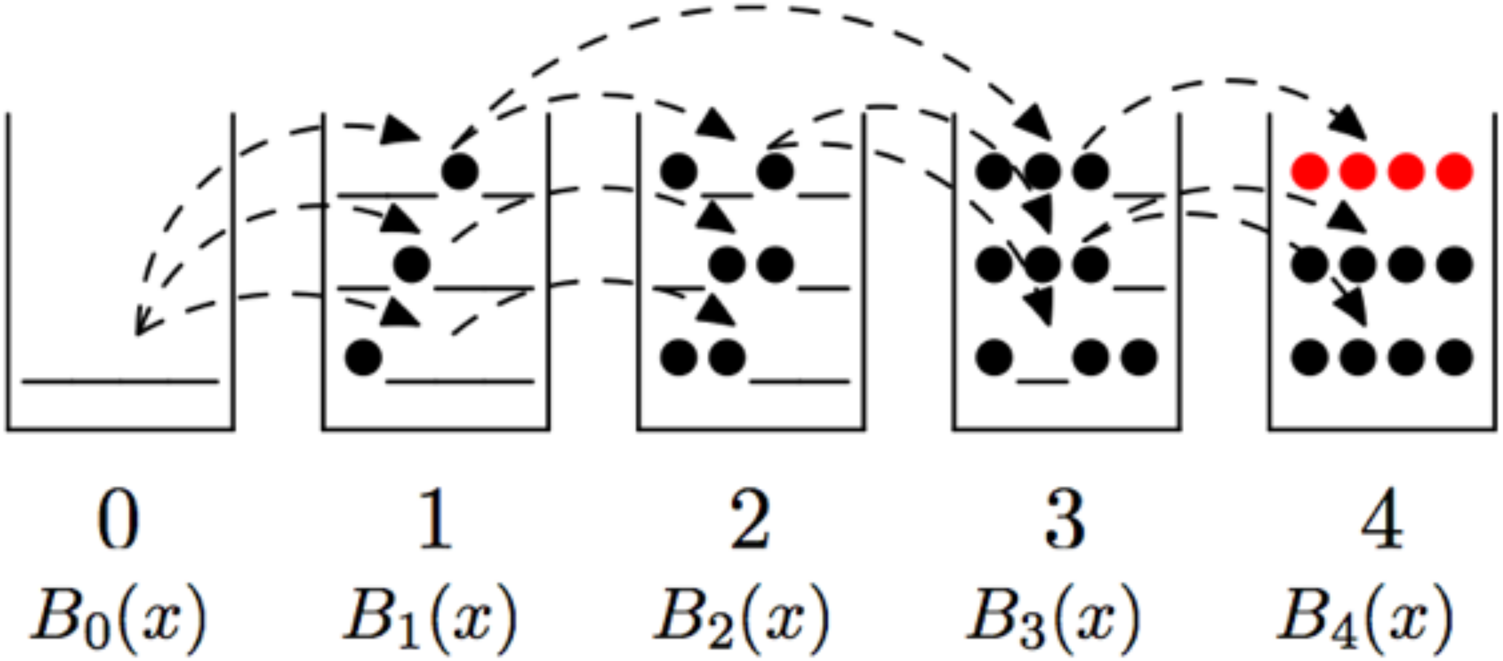
$x =$ ● ●● _ _ ●● _ _ _

$\bar{e}_x(d) = e(d) \circ future(d, x)$

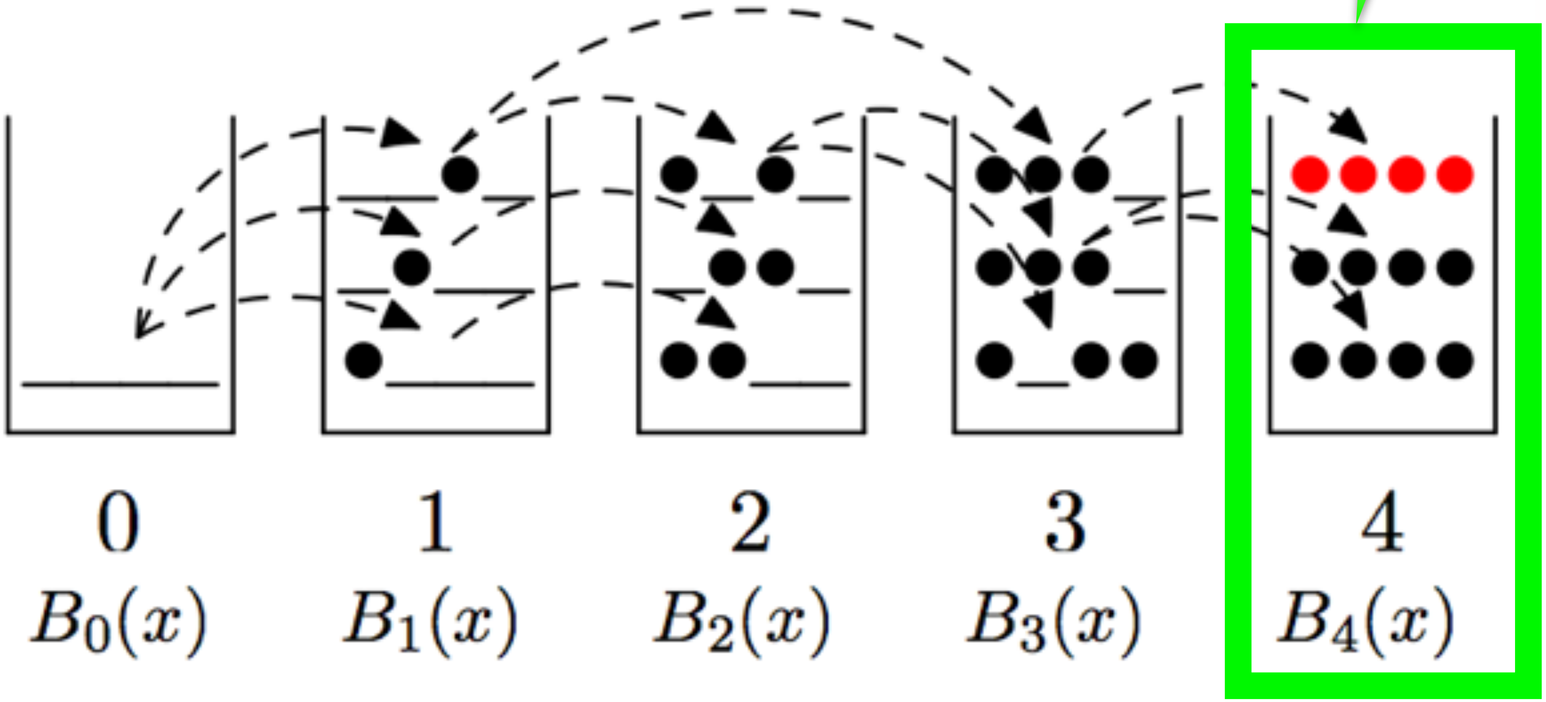
partial 1: I from | Shanghai fly to Beijing | unigram=5, bi=2

partial 2: I fly | from Shanghai to Beijing | unigram=5, bi=3, tri=2, 4gram=1 ✔

Towards Search-Aware Tuning

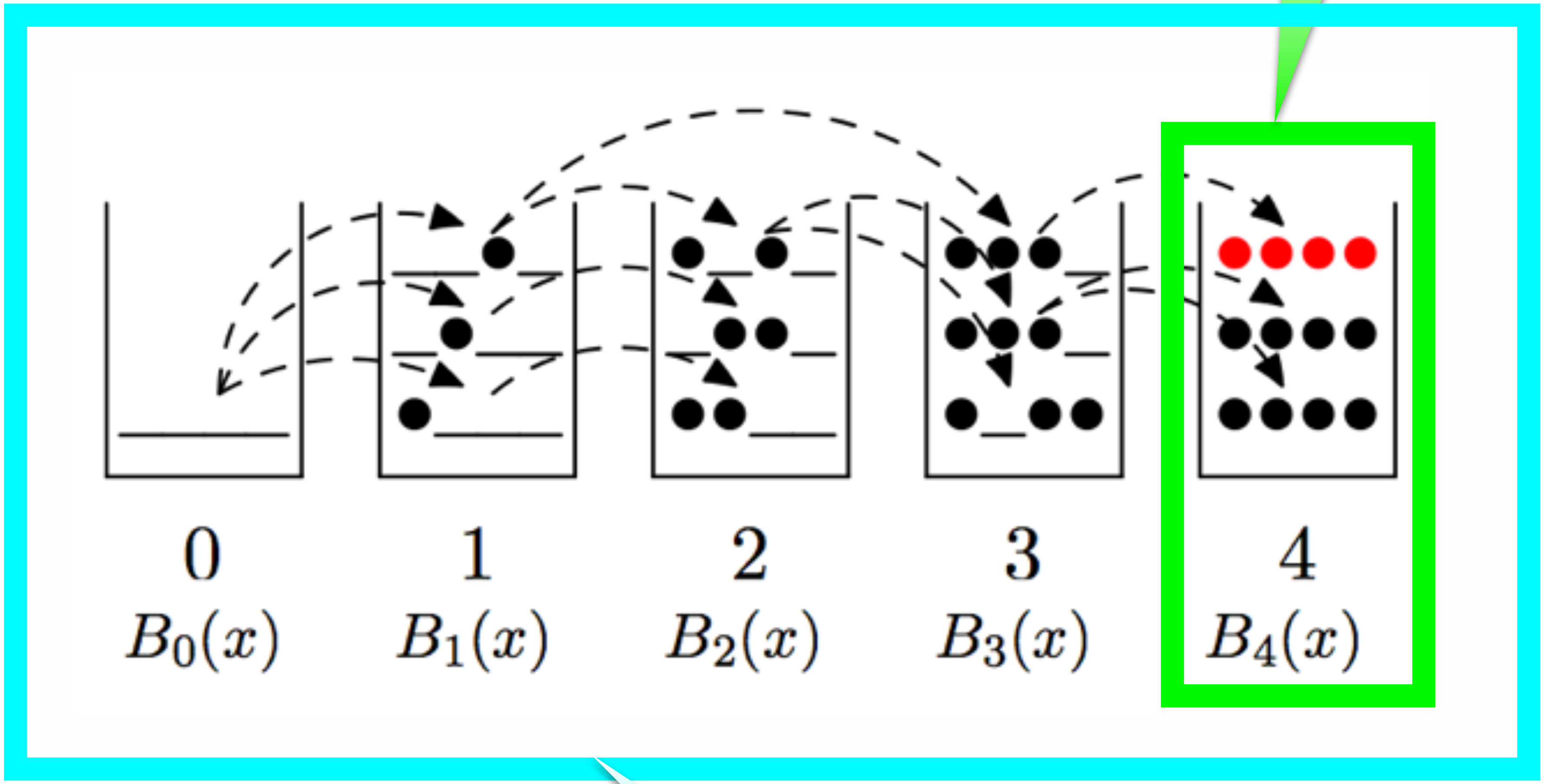


Towards Search-Aware Tuning



Traditional tuning
MERT/MIRA/PRO

Towards Search-Aware Tuning

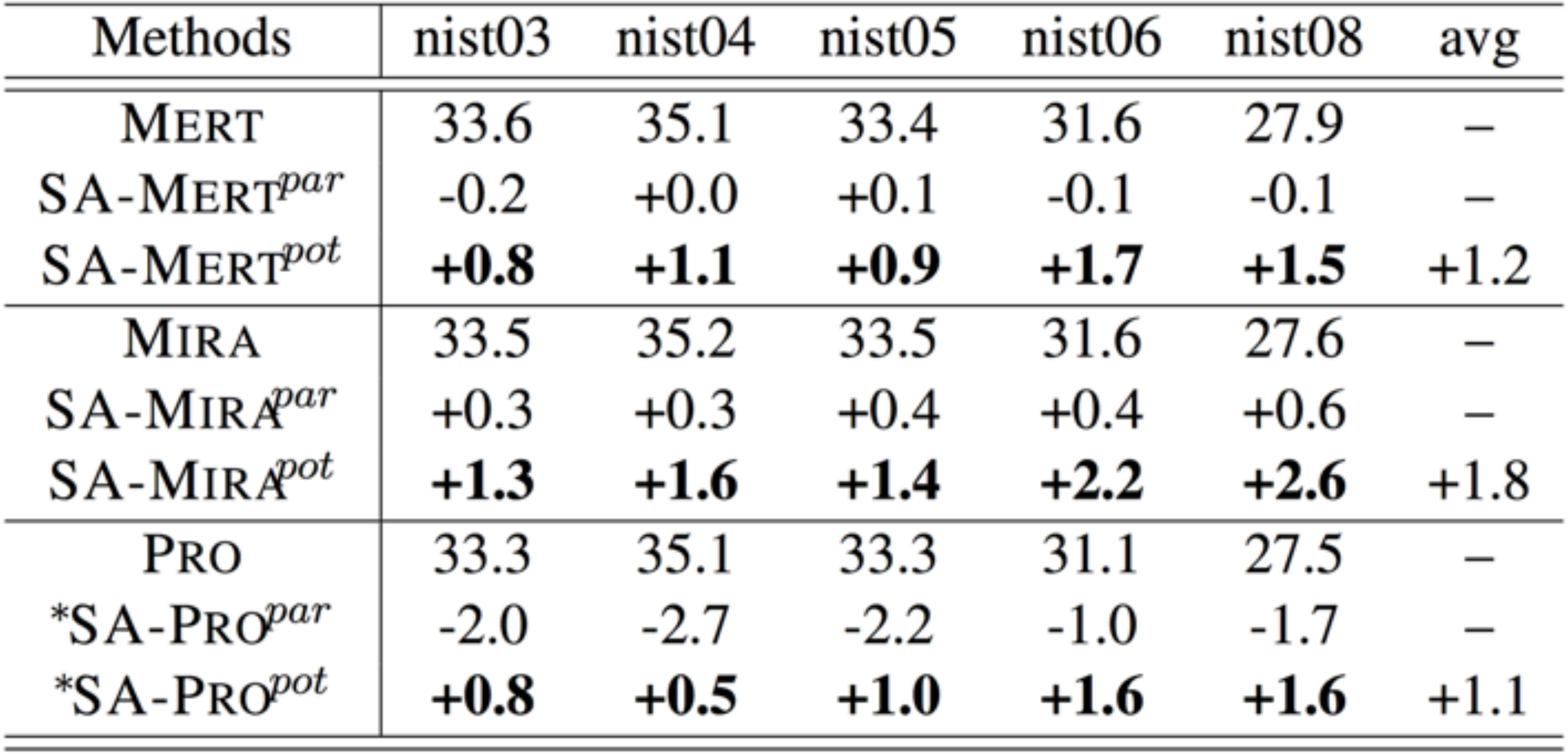


Traditional tuning
MERT/MIRA/PRO

Search-aware tuning

Experiments: Ch-to-En

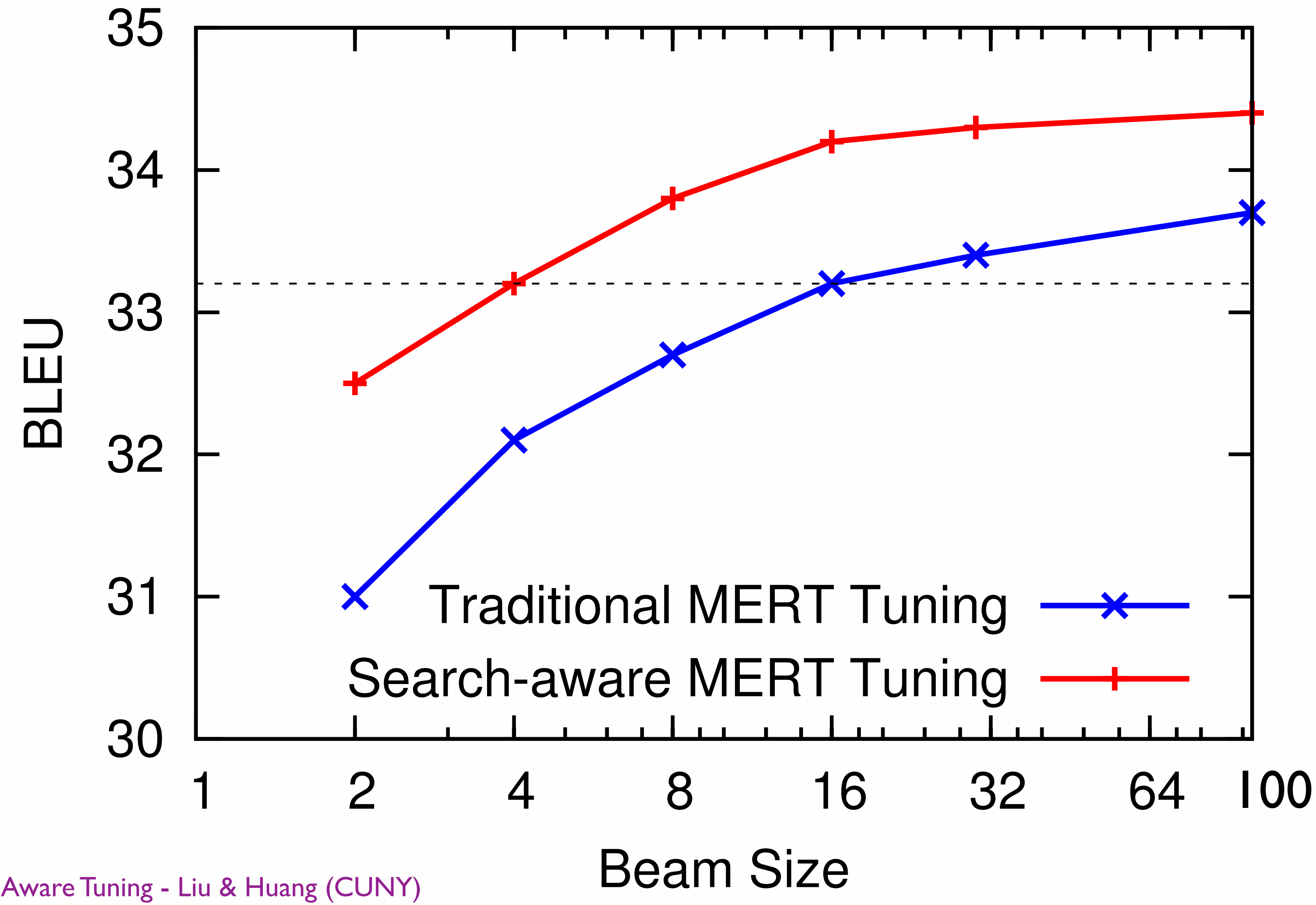
- on phrase-based decoder (Huang & Chiang 07; Yu et al 13)
  - partial BLEU not helpful, but potential BLEU very helpful
  - all experiments use only dense features



Methods	nist03	nist04	nist05	nist06	nist08	avg
MERT	33.6	35.1	33.4	31.6	27.9	–
SA-MERT^{par}	-0.2	+0.0	+0.1	-0.1	-0.1	–
SA-MERT^{pot}	**+0.8**	**+1.1**	**+0.9**	**+1.7**	**+1.5**	+1.2
MIRA	33.5	35.2	33.5	31.6	27.6	–
SA-MIRA^{par}	+0.3	+0.3	+0.4	+0.4	+0.6	–
SA-MIRA^{pot}	**+1.3**	**+1.6**	**+1.4**	**+2.2**	**+2.6**	+1.8
PRO	33.3	35.1	33.3	31.1	27.5	–
*SA-PRO^{par}	-2.0	-2.7	-2.2	-1.0	-1.7	–
*SA-PRO^{pot}	**+0.8**	**+0.5**	**+1.0**	**+1.6**	**+1.6**	+1.1

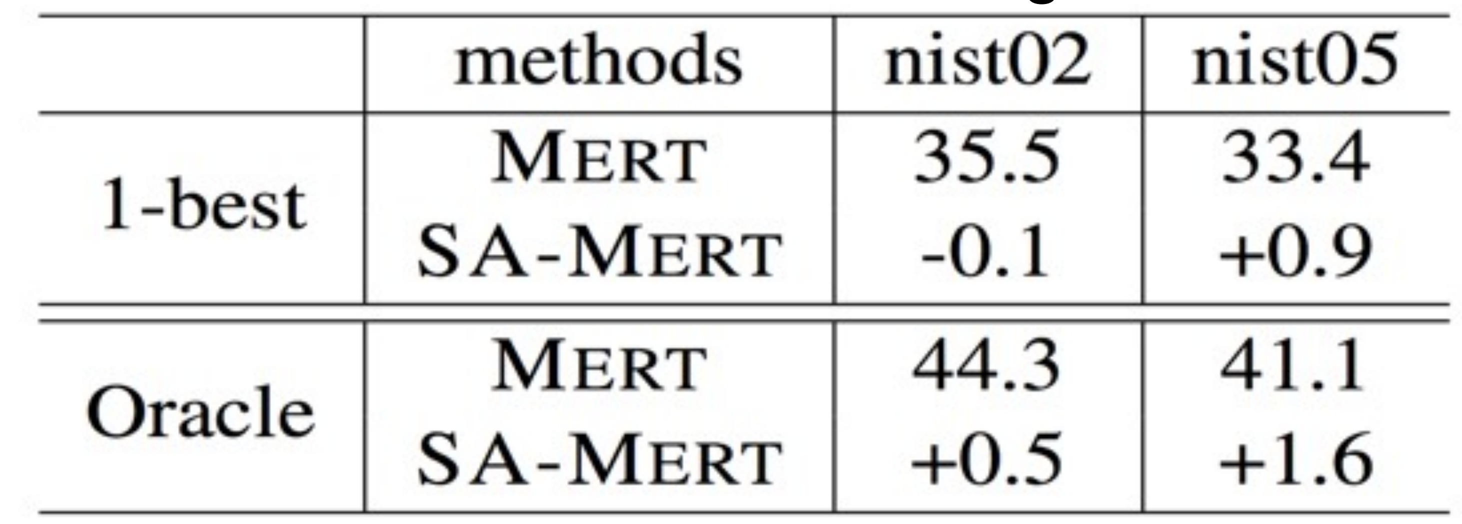
Beam Size

- helps more in smaller beam sizes

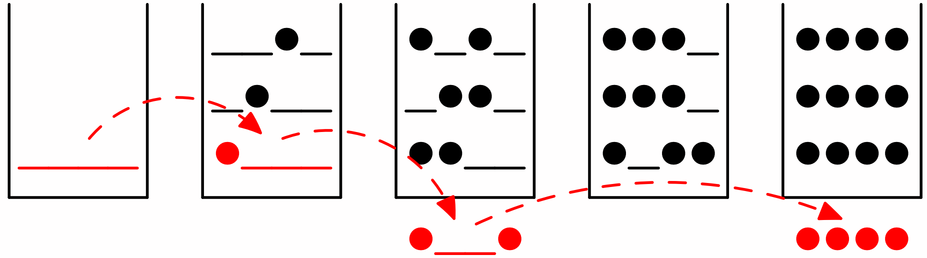


Oracle Improvement

- search-aware tuning improves *k*-best oracle in final bin
  - quality of *k*-best list improves more than 1-best
  - more improvement on test than tuning

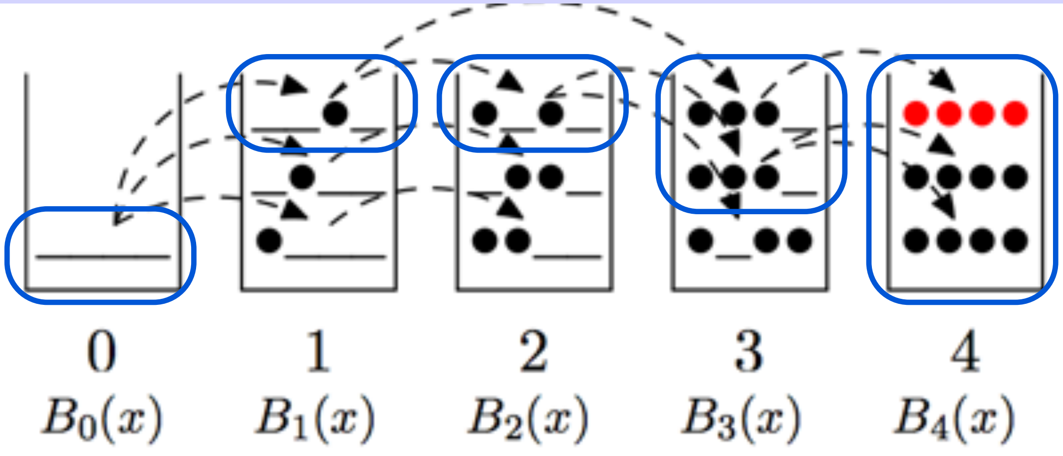


		tuning	test
	methods	nist02	nist05
1-best	MERT	35.5	33.4
	SA-MERT	-0.1	+0.9
Oracle	MERT	44.3	41.1
	SA-MERT	+0.5	+1.6



More Diversity in the Final Bin

cf.: Y-chromosome Adam
Mitochondria Eva



- search-aware tuning does promote diversity
  - even though we do *not* include diversity in the objectives
  - adapt n-gram diversity metric (Gimpel et al 2013) with modifications

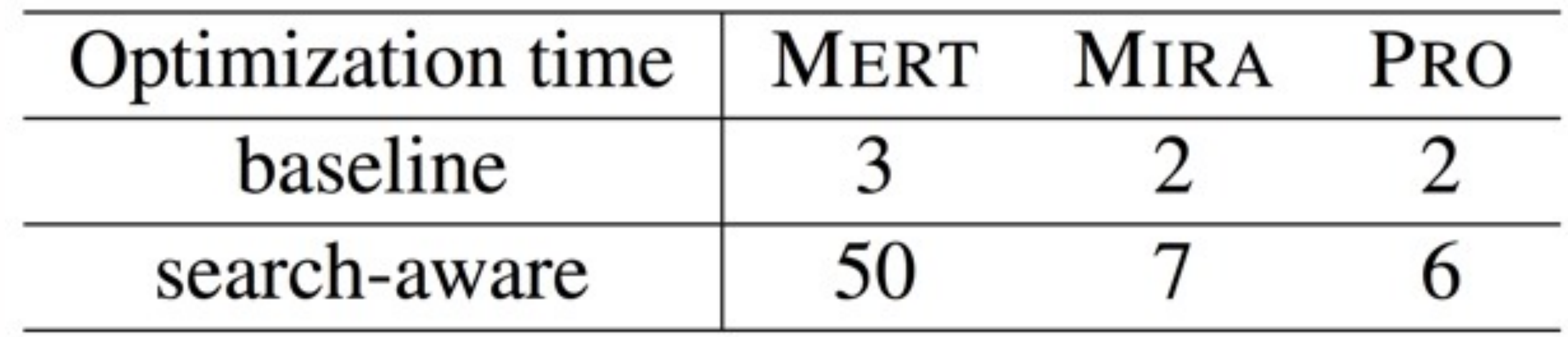
$$d(y, y') = -\sum_{i=1}^{|y|-q} \sum_{j=1}^{|y'|-q} [\![y_{i:i+q} = y'_{j:j+q}]\!]$$

$$d'(y, y') = 1 - \frac{2 \times d(y, y')}{d(y, y) + d(y', y')}$$

Diversity	nist02	nist05
MERT	0.216	0.204
SA-MERT	0.227	0.213

Drawback: Slow Optimization

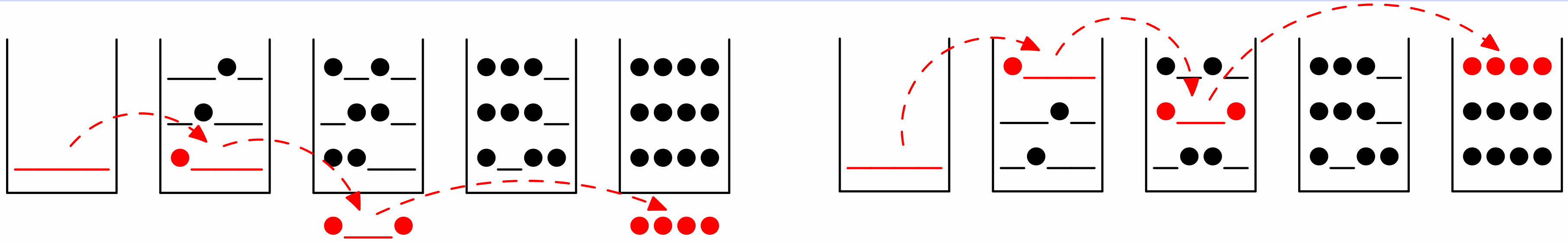
- search-aware tuning does slow down optimization
- but decoding is the bottle-neck in tuning
  - though parallelizable
- overall slowdown is not significant for MIRA/PRO



Optimization time	MERT	MIRA	PRO
baseline	3	2	2
search-aware	50	7	6

decoding time: 20 min. on single CPU

Conclusions



- search error is a major reason for bad translation
  - search-agnostic tuning does not address this problem
- our search-aware tuning promotes promising translations
- potential BLEU is a good evaluator for sub-translations
  - also works for TER and other metrics
- very simple framework; applies to MERT/MIRA/PRO...
  - first consistent ~1 BLEU point improvement with dense features
  - only drawback: slower optimization