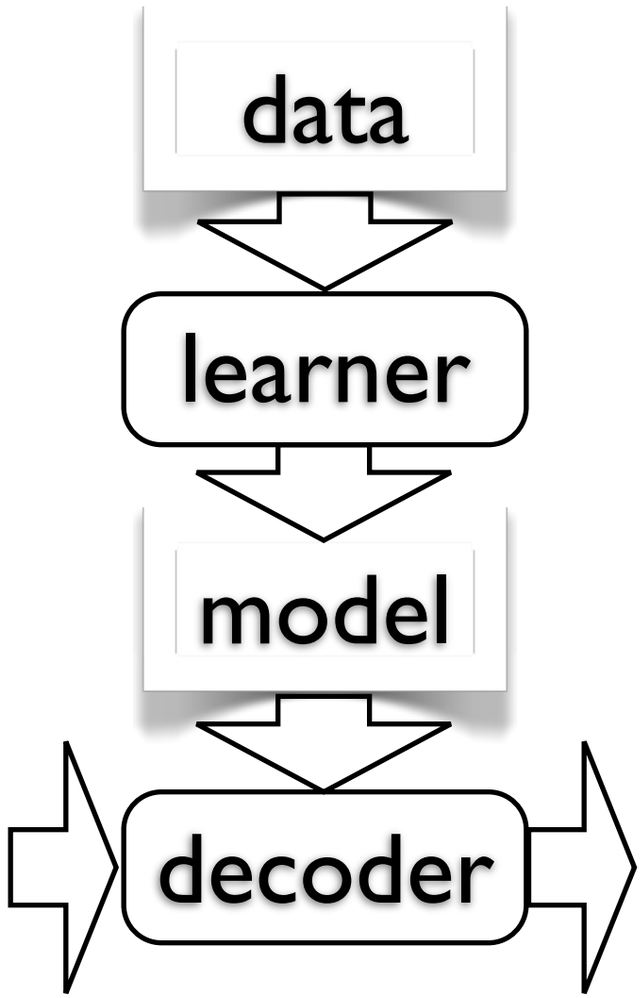
最適化問題としての機械翻訳

渡辺太郎

Machine Translation

黑山头口岸联检部门将原来要二至三天办完的出入境手续改为一天办完。



The United Inspection Department of Heishantou Port has shortened the procedures for leaving and entering the territory from originally 2 - 3 days to 1 day.

- A data-driven approach to MT
- We learn parameters from data assuming a “model”

Channel Model
+ noise

$$\begin{aligned}\hat{y} &= \operatorname*{argmax}_y Pr(y|x) \\ &= \operatorname*{argmax}_y \frac{Pr(x|y)Pr(y)}{Pr(x)} \\ &= \operatorname*{argmax}_y Pr(x|y)Pr(y)\end{aligned}$$

f = source

e = target

$$\hat{\mathbf{e}} = \operatorname*{argmax}_{\mathbf{e}} Pr(\mathbf{f}|\mathbf{e})Pr(\mathbf{e})$$

- Employed in: ASR, OCR, MT...

Research Questions

$$
\begin{aligned}
\hat{e} &= \arg\max_{e} \frac{\sum_{d} \exp\left(w^\top h(f, d, e)\right)}{\sum_{e', d'} \exp\left(w^\top h(f, d', e')\right)} \\
&\approx \arg\max w^\top h(f, d, e) \\
&\quad \langle e, d \rangle
\end{aligned}
$$

- How to model the process of translation?
- How to learn the parameters (given data + model)?
- How to decode (given model + parameters + input sentence)?

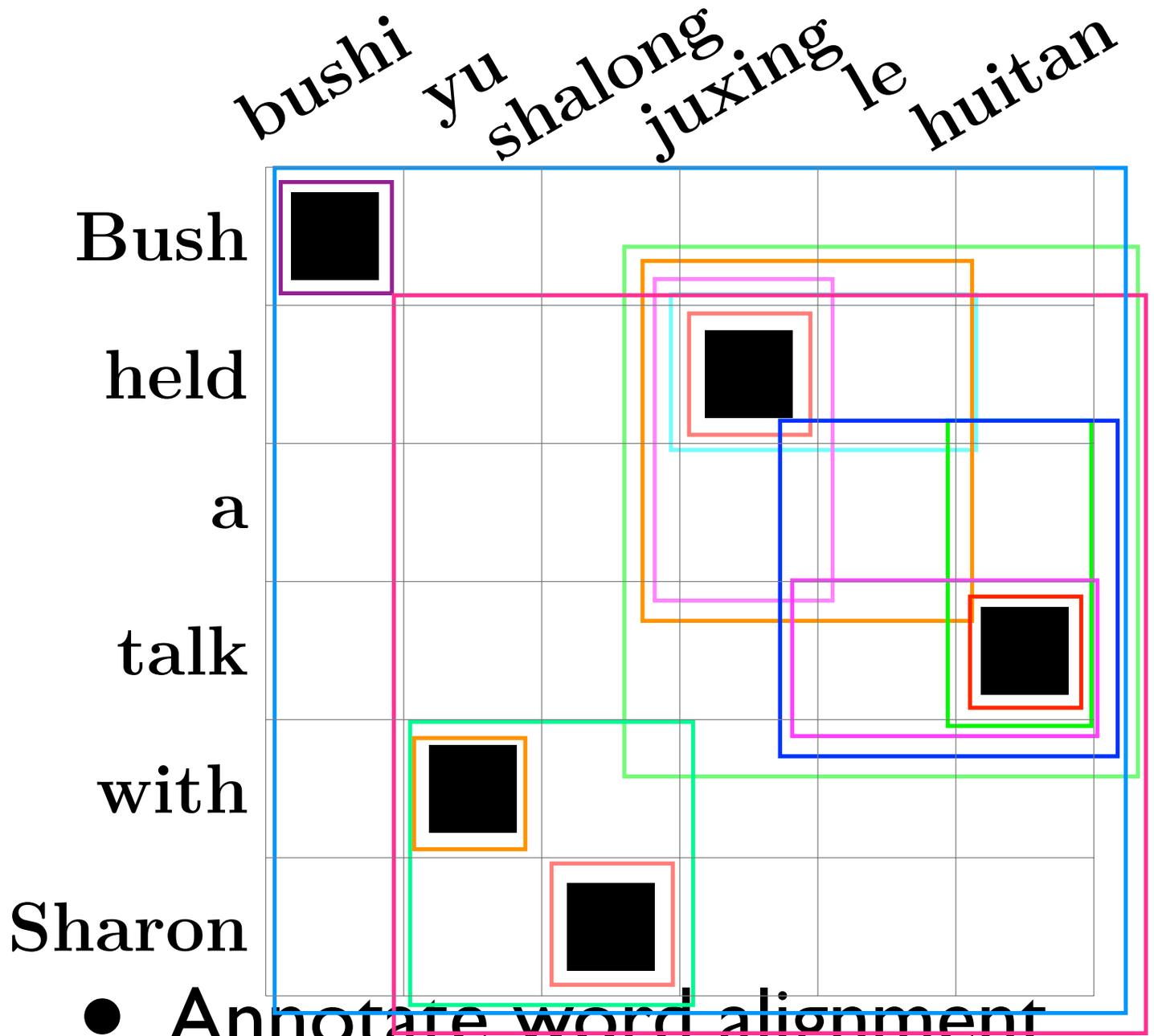
Bilingual Data

1.上海浦东开发与法制建设同步
2.新华社上海二月十日电（记者谢金虎、张持坚）
3.上海浦东近年来颁布实行了涉及经济、贸易、建设、规划、科技、文教等领域的七十一件法规性文件，确保了浦东开发的有序进行。
4.浦东开发开放是一项振兴上海，建设现代化经济、贸易、金融中心的跨世纪工程，因此大量出现的是以前不曾遇到过的新情况、新问题。
5.对此，浦东不是简单的采取“干一段时间，等积累了经验以后再制定法规条例”的做法，而是借鉴发达国家和深圳等特区的经验教训，聘请国内外有关专家学者，积极、及时地制定和推出法规性文件，使这些经济活动一出现就被纳入法制轨道。
6.去年初浦东新区诞生的中国第一家医疗机构药品采购服务中心，正因为一开始就比较规范，运转至今，成交药品一亿多元，没有发现一例回扣。

1.The development of Shanghai's Pudong is in step with the establishment of its legal system
2.Xinhua News Agency, Shanghai, February 10, by wire (reporters Jinhu Xie and Chijian Zhang)
3.In recent years Shanghai's Pudong has promulgated and implemented 71 regulatory documents relating to areas such as economics, trade, construction, planning, science and technology, culture and education, etc., ensuring the orderly advancement of Pudong's development.
4.Pudong's development and opening up is a century-spanning undertaking for vigorously promoting Shanghai and constructing a modern economic, trade, and financial center. Because of this, new situations and new questions that have not been encountered before are emerging in great numbers.
5.In response to this, Pudong is not simply adopting an approach of "work for a short time and then draw up laws and regulations only after waiting until experience has been accumulated." Instead, Pudong is taking advantage of the lessons from experience of developed countries and special regions such as Shenzhen by hiring appropriate domestic and foreign specialists and scholars, by actively and promptly formulating and issuing regulatory documents, and by ensuring that these economic activities are incorporated into the sphere of influence of the legal system as soon as they appear.
6.Precisely because as soon as it opened it was relatively standardized, China's first drug purchase service center for medical treatment institutions, which came into being at the beginning of last year in the Pudong new region, in operating up to now, has concluded transactions for drugs of over 100 million yuan and hasn't had one case of kickback.

(part of LDC2007T02, English translation of Chinese treebank)

Alignment + Extraction



(Koehn et al., 2003)

- Annotate word alignment
- Exhaustively extract phrases

Features

$$\log p_\phi(\bar{f}|\bar{e}) = \log \frac{\text{count}(\bar{e}, \bar{f})}{\sum_{\bar{f}'} \text{count}(\bar{e}, \bar{f}')}$$

$$\log p_\phi(\bar{e}|\bar{f}) = \log \frac{\text{count}(\bar{e}, \bar{f})}{\sum_{\bar{e}'} \text{count}(\bar{e}', \bar{f})}$$

- Collect all the phrase pairs from the data, and ML estimates + other features

不 熟悉 ||| 'm not familiar ||| -1.4859937213 -7.2301988107 -0.3036824138 -3.0311892056

不 熟悉 ||| do n't know ||| -1.2064088591 -5.3571402084 -3.4402617349 -6.8870595804

不 熟悉 ||| i 'm not familiar ||| -2.522085653 -9.1804032749 -1.06784063 -3.0311892056

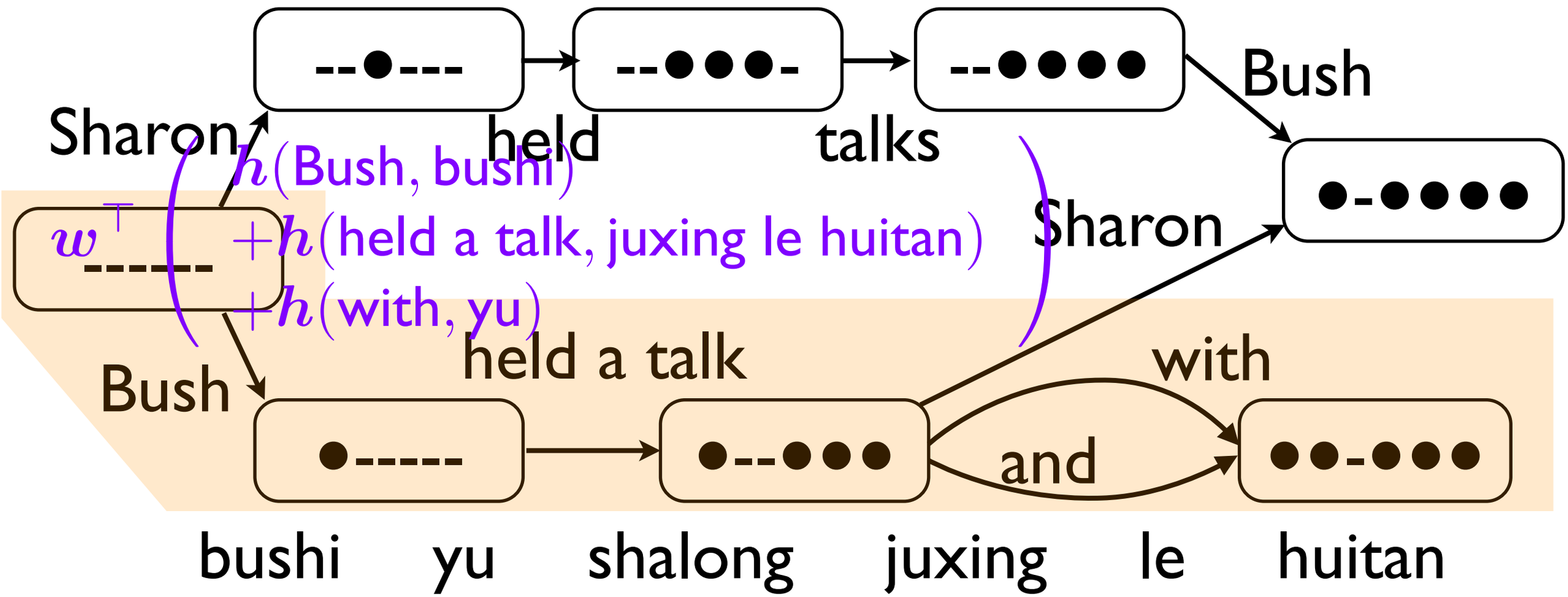
不 熟悉 ||| it will be great ||| -2.522085653 -20.871716142 0.0 -11.4593095552

不 熟悉 ||| not accustomed ||| -2.522085653 -5.5628513514 -0.6931471806 -2.2177906617

不 熟悉 ||| not accustomed to ||| -2.522085653 -8.5631752395 0.0 -2.2177906617

不 熟悉 ||| not familiar ||| -1.8754584881 -3.4150084505 -0.4212134651 -2.4210642434

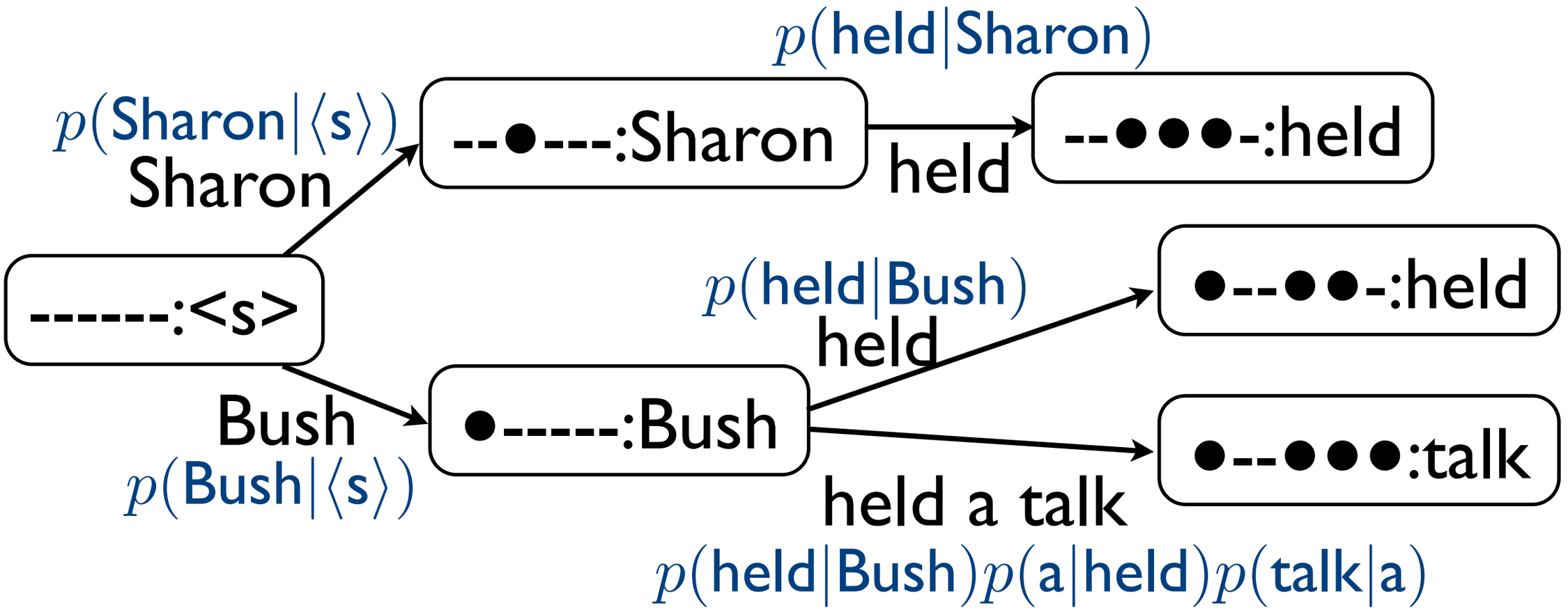
Decoding



(Koehn et al., 2003)

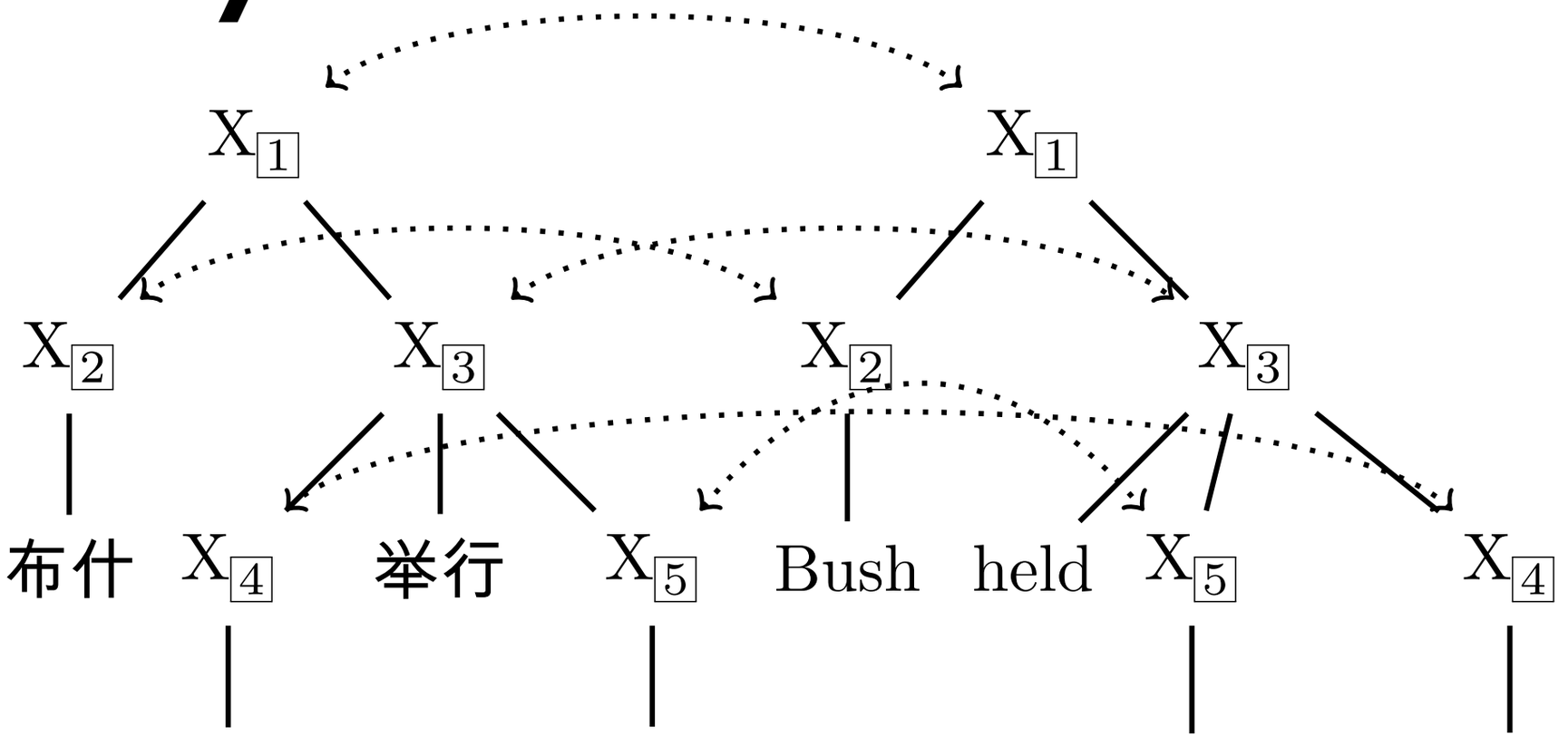
- Node: bit-vector representing covered source words
- Edge: phrasal translations, strictly left-to-right + score
- DP-based Search space: $O(2^n)$, Time: $O(2^n n^2)$

Non-local features



- Features that requires scoring out of phrases: bigram language model
- We usually employ 5-gram LM (4th order Markov)
- Space: $O(2^n V^{m-1})$, Time: $O(2^n V^{m-1} n^2)$ for m-gram LM

synchronous-CFG



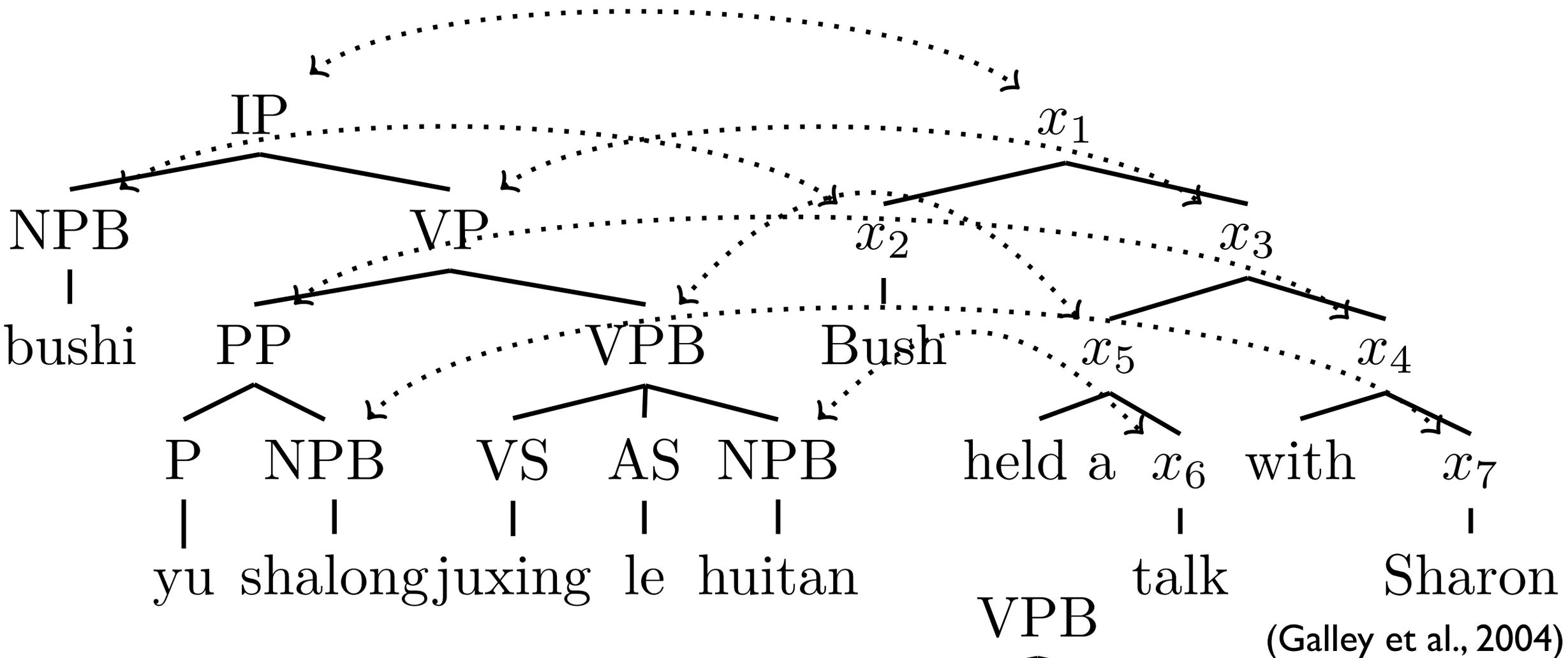
(Chiang, 2007)

$$\mathrm{X} \rightarrow \langle \mathrm{X}_{\boxed{1}} \text{ 举行 } \mathrm{X}_{\boxed{2}}, \text{hold } \mathrm{X}_{\boxed{2}} \ \mathrm{X}_{\boxed{1}} \rangle$$

$$\mathrm{X} \rightarrow \langle \text{与 沙龙}, \text{with Sharon} \rangle$$

- paired-CFG: two RHS, but single LHS
- Reordering represented by swapped non-terminals

synchronous-TSG



(Galley et al., 2004)

- Bilingual version of Tree Substitution Grammar

$$_{11} \rightarrow \text{held a } x_1$$

Evaluation: ngram precision

Well , I 'd like to stay five nights beginning October twenty-fifth to thirty .

- I 'd like to stay there for five nights , from October twenty fifth to the thirtieth .
- I want to stay for five nights , from October twenty fifth to the thirtieth .
- I 'd like to stay for five nights , from October twenty fifth to the thirtieth .
- I would like to reserve a room for five nights , from October twenty fifth to the thirtieth .

Evaluation: ngram precision

Well , I 'd like stay five nights beginning October twenty-fifth to thirty .

$$p_1 = \frac{11}{15}$$

- I 'd like to stay there for five nights , from October twenty fifth to the thirtieth .
- I want to stay for five nights , from October twenty fifth to the thirtieth .
- I 'd like to stay for five nights , from October twenty fifth to the thirtieth .
- I would like to reserve a room for five nights , from October twenty fifth to the thirtieth .

Evaluation: ngram precision

Well , I 'd like to stay five nights beginning October twenty-fifth to thirty .

$$p_1 = \frac{11}{15} \qquad p_2 = \frac{5}{14}$$

- I 'd like to stay there for five nights , from October twenty fifth to the thirtieth .
- I want to stay for five nights , from October twenty fifth to the thirtieth .
- I 'd like to stay for five nights , from October twenty fifth to the thirtieth .
- I would like to reserve a room for five nights , from October twenty fifth to the thirtieth .

Evaluation: ngram precision

Well , I 'd like to stay five nights beginning October twenty-fifth to thirty .

$$p_1 = \frac{11}{15} \qquad p_2 = \frac{5}{14} \qquad p_3 = \frac{3}{13}$$

- I 'd like to stay there for five nights , from October twenty fifth to the thirtieth .
- I want to stay for five nights , from October twenty fifth to the thirtieth .
- I 'd like to stay for five nights , from October twenty fifth to the thirtieth .
- I would like to reserve a room for five nights , from October twenty fifth to the thirtieth .

Evaluation: ngram precision

Well , I 'd like to stay five nights beginning October twenty-fifth to thirty .

$$p_1 = \frac{11}{15} \qquad p_2 = \frac{5}{14} \qquad p_3 = \frac{3}{13} \qquad p_4 = \frac{2}{12}$$

- I 'd like to stay there for five nights , from October twenty fifth to the thirtieth .
- I want to stay for five nights , from October twenty fifth to the thirtieth .
- I 'd like to stay for five nights , from October twenty fifth to the thirtieth .
- I would like to reserve a room for five nights , from October twenty fifth to the thirtieth .

Evaluation: BLEU

$$\exp\left(\sum_{n=1}^{4}\frac{1}{4}\log p_n\right)\cdot\min\left\{\exp\left(1-\frac{r}{c}\right),1\right\}$$

- (Uniformly) weighted combination of precision (Papineni et al., 2002)
- brevity penalty: penalize too short sentences
  - r = reference length, c = candidate length
  - If we have multiple "r", choose the closest-shortest reference to "c"
- Both factors are computed over the whole document

Why BLEU?

- Used as a standard metric for more than 10 years: Progress of MT is due by BLEU!
  - Good indication of you "progress" and probably not for comparing systems with different philosophy
- However, non-linear decomposition into sentences: corpus-wise metric, thus, harder to optimize
- BP-problem(Chiang et al., 2009): You can generate spuriously long translations together with a highly confident short translations

A Bad Example

"we come from the land of the ice and snow"
"from the midnight sun where the hot springs flow"

system 1

xxx xxx xxx xxx land xxx xxx ice xxx snow
xxx xxx midnight xxx xxx xxx hot xxx flow

system 2

x come x x land x x ice x snow x x x x x
from xxx sun xxx

- Both shared the same # of words, and the same # of matches

MT Research @ NICT

- Better model, search, and optimization
  - Sophisticated models using "syntax" (or, better "parsing" with complex models)
    - Unsupervised learning with structured hidden variables (via Bayesian models)
    - Better optimization (via discriminative training)
- Today's focus: Optimization

Nightmares of Optimization

Optimization

- How to learn w? $\langle \hat{e}, \hat{d} \rangle = \underset{\langle e,d \rangle}{\arg\max}\, \boldsymbol{w}^\top \boldsymbol{h}(\boldsymbol{f}, \boldsymbol{d}, \boldsymbol{e})$
- Assume a loss function ℓ and minimize the "risk" over bilingual data, F and E (usually small... WHY?)
- Since true distribution is unknown, minimize a regularized empirical risk
- NOTE: $\ell \neq$ error (i.e. 1 - BLEU)

$$\begin{aligned} \hat{\boldsymbol{w}} &= \underset{\boldsymbol{w} \in \mathcal{W}}{\arg\min}\, \mathbb{E}_{Pr(F,E)}[\ell(F, E; \boldsymbol{w})] \\ &= \underset{\boldsymbol{w} \in \mathcal{W}}{\arg\min}\, \ell(F, E; \boldsymbol{w}) + \lambda \Omega(\boldsymbol{w}) \end{aligned}$$

Nightmares

$$\hat{\boldsymbol{w}} = \arg\min_{\boldsymbol{w} \in \mathcal{W}} \ell(F, E; \boldsymbol{w}) + \lambda \Omega(\boldsymbol{w})$$

- Problem 1: Many alternative translations (e) possible with many alternative derivations (d)
  - We cannot enumerate all possible hidden variables
- Problem 2: ℓ and/or error are corpus-wise, not sentence-wise (i.e. BLEU)
  - We cannot assume a convex function, or gradients are available
- Problem 3: Enormous search space and high risk of search errors

Learning Strategies

(k-best) Batch Learning

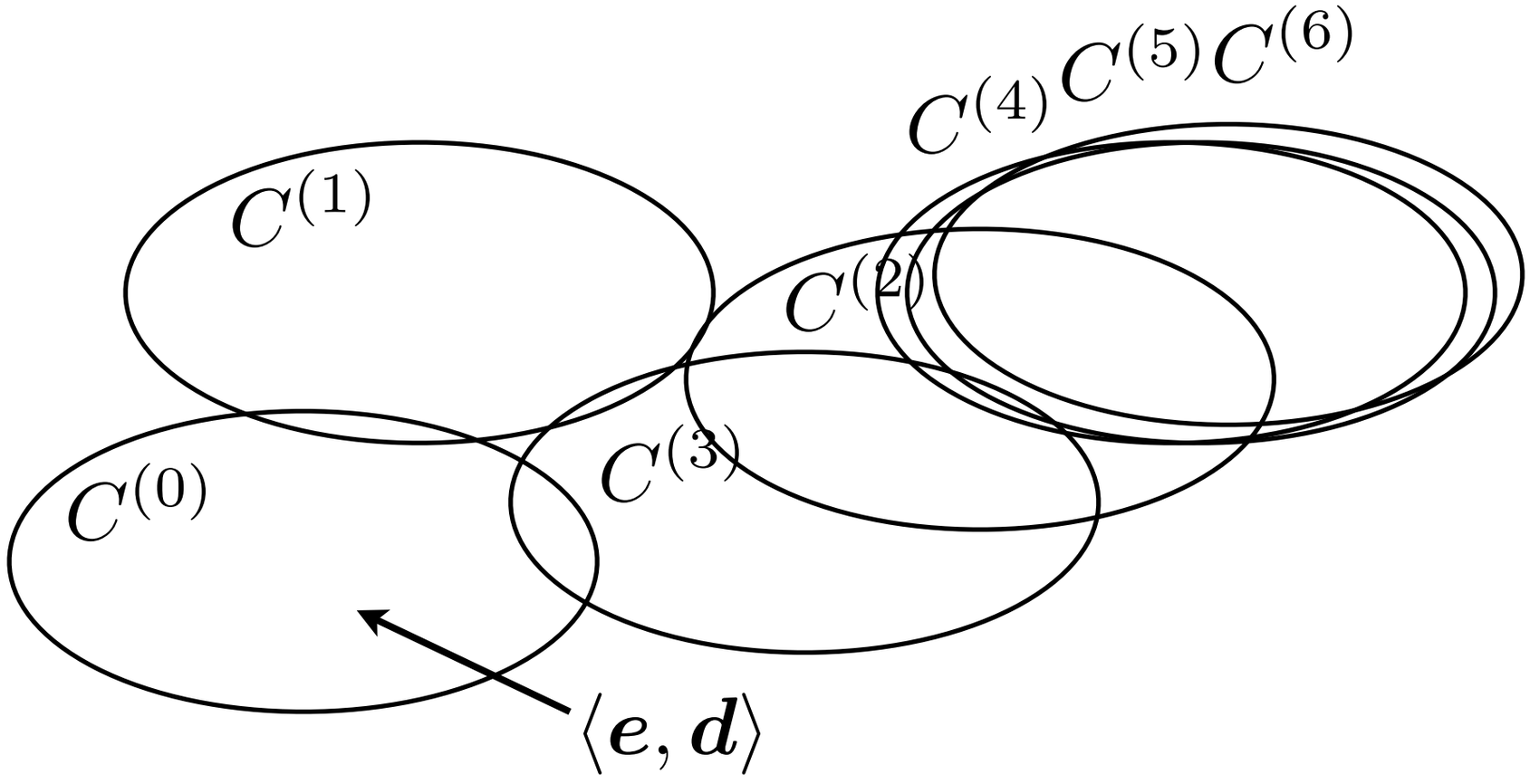
1: **procedure** BATCHLEARN($\langle F, E \rangle = \left\{ \langle \boldsymbol{f}^{(i)}, \boldsymbol{e}^{(i)} \rangle \right\}_{i=1}^{N}$)
2: $\quad \boldsymbol{w}^{(0)} \leftarrow \emptyset$
3: $\quad C = \left\{ \boldsymbol{c}^{(i)} \equiv \emptyset \right\}_{i=1}^{N}$ ▷ k-best list
4: $\quad$ **for** $t \in \{1 \dots T\}$ **do**
5: $\quad\quad$ **for** $i \in \{1 \dots N\}$ **do**
6: $\quad\quad\quad kbest^{(i)} \leftarrow \mathrm{GEN}(\boldsymbol{f}^{(i)}, \boldsymbol{w}^{(t-1)})$ ▷ decode by $\boldsymbol{w}^{(t-1)}$
7: $\quad\quad\quad \boldsymbol{c}^{(i)} \leftarrow \boldsymbol{c}^{(i)} \cup kbest^{(i)}$ ▷ merge k-best list
8: $\quad\quad$ **end for**
9: $\quad\quad \boldsymbol{w}^{(t)} \leftarrow \arg\min_{\boldsymbol{w} \in \mathcal{W}} \ell(F, E, C; \boldsymbol{w}) + \lambda \Omega(\boldsymbol{w})$ ▷ optimize
10: $\quad$ **end for**
11: $\quad$ **return** $\boldsymbol{w}^{(T)}$
12: **end procedure**

- Approximate the document-wise candidate space by k-best merging using decoding step + optimization step (Och and Ney, 2002)

k-best

this is the kyoto kanko hotel , front desk . ||| -48.68790464
this is the kyoto kanko hotel , front desk . ||| -48.85902546
this is the kyoto kanko hotel , front desk . ||| -49.90369084
this is the kyoto kanko hotel , front desk . ||| -50.07481166
this is the kyoto kanko hotel , front desk . ||| -50.32856858
kyoto kanko hotel , front desk . ||| -51.13501382
kyoto kanko hotel , front desk . ||| -51.30613464
this is the kyoto kanko hotel , front desk . ||| -51.54435478
kyoto kanko hotel , front desk . ||| -52.35080002
kyoto kanko hotel , front desk . ||| -52.52192084
kyoto kanko hotel , front desk . ||| -52.71186262
kyoto kanko hotel , front desk . ||| -52.77567776
kyoto kanko hotel , front desk . ||| -52.88298344
hello , this is the kyoto kanko hotel . front desk . may i help you as soon as possible . ||| -53.77178844
hello , this is the kyoto kanko hotel , front desk . may i help you as soon as possible . ||| -53.90754257
hello , this is the kyoto kanko hotel . front desk . may i help you as soon as possible . ||| -53.92571267
kyoto kanko hotel , front desk . ||| -53.92764882
kyoto kanko hotel , front desk . ||| -53.99146396
hello , this is the kyoto kanko hotel , front desk . may i help you as soon as possible . ||| -54.06146681
kyoto kanko hotel , front desk . ||| -54.09876964
kyoto kanko hotel , front desk . ||| -54.35252656
hello , this is the kyoto kanko hotel . front desk . may i help you as soon as possible . ||| -54.98757464

Batch Candidate Space



- Perform optimization over the merged k-bests
- Eventually converge, but merely an approximation of all the translation candidates

Objectives

- Minimum Error Rate (MERT) (Och, 2003)
- Log linear (softmax) (Och and Ney, 2002; Blunsom et al., 2008)
- Pair-wise Rank Optimization (PRO) (Hopkins and May, 2011)
- Expected BLEU (xBLEU) (Pauls et al., 2009; Rosti et al., 2010; Rosti et al., 2011)

MERT

$$\ell_{\text{error}}(F, E, C; \boldsymbol{w}) = \text{error}\left(E, \left\{ \underset{\langle \boldsymbol{e}, \boldsymbol{d} \rangle \in \boldsymbol{c}^{(i)}}{\arg\max}\, \boldsymbol{w}^\top \boldsymbol{h}(\boldsymbol{f}^{(i)}, \boldsymbol{d}, \boldsymbol{e}) \right\}_{i=1}^{N}\right)$$

- Due to non-linear error function, non-convex and impossible to compute derivatives
- Gradient-free optimization: Powell's method or Downhill-simplex method
- An efficient line search by exploiting piece-wise linear function of k-best list (Och, 2002)

softmax

$$\ell_{\mathrm{softmax}}(F, E, C; \boldsymbol{w}) = \prod_{i=1}^{N} \frac{\sum_{\langle \boldsymbol{e}^*, \boldsymbol{d}^* \rangle \in \boldsymbol{o}^{(i)}} \exp(\boldsymbol{w}^\top \boldsymbol{h}(\boldsymbol{f}^{(i)}, \boldsymbol{d}^*, \boldsymbol{e}^*))}{\sum_{\langle \boldsymbol{e}, \boldsymbol{d} \rangle \in \boldsymbol{c}^{(i)}} \exp(\boldsymbol{w}^\top \boldsymbol{h}(\boldsymbol{f}^{(i)}, \boldsymbol{d}, \boldsymbol{e}))}$$

- A set of oracle candidates (o) are softly separated (Blunsom et al., 2008), thus not strictly-convex
- Due to non-linear error function, oracles are computed from k-bests by hill-climbing (Watanabe, 2012)

PRO

$$\ell_{\text{hinge}}(F, E, C; \boldsymbol{w}) = \sum_{i=1}^{N} \sum_{\langle \boldsymbol{e}^*, \boldsymbol{d}^* \rangle \in \boldsymbol{c}^{(i)}} \sum_{\substack{\langle \boldsymbol{e}, \boldsymbol{d} \rangle \in \boldsymbol{c}^{(i)}, \\ \text{error}(\boldsymbol{e}^*, \boldsymbol{e}) > 0}}$$

$$\max \left\{ 0, 1 - \boldsymbol{w}^\top \left(\boldsymbol{h}(\boldsymbol{f}^{(i)}, \boldsymbol{e}^*, \boldsymbol{d}^*) - \boldsymbol{h}(\boldsymbol{f}^{(i)}, \boldsymbol{e}, \boldsymbol{d}) \right) \right\}$$

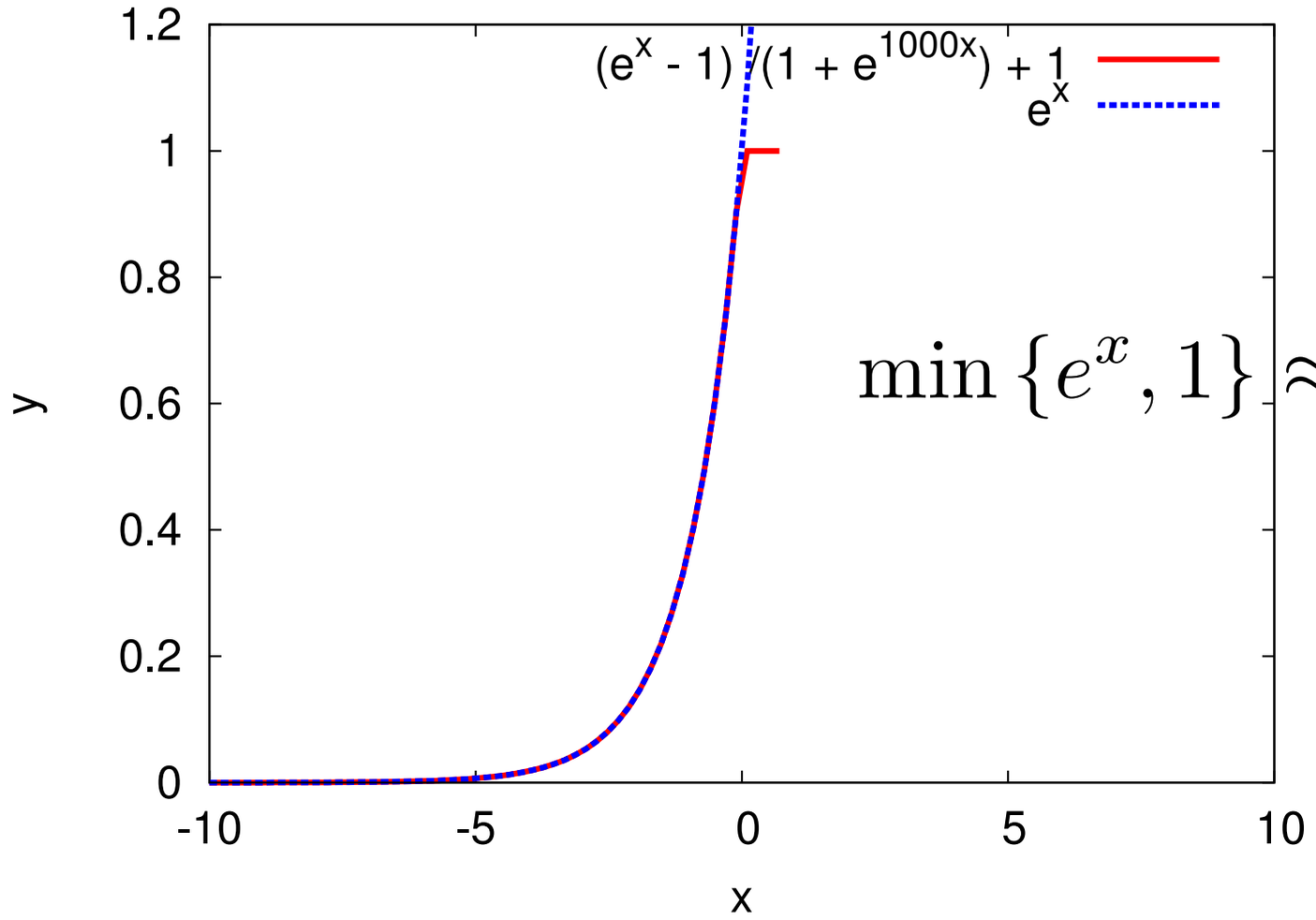
- We suffer hinge-loss through exhaustive pair-wise comparisons
- Sample small # of pairs (Hopkins and May, 2011) or sample from a set of good translations and a set of bad translations (Watanabe, 2012; Semianer et al., 2012)

xBLEU

$$\prod_{n=1}^{4} \left(\frac{\min\left\{ \sum_s \sum_i \sum_{g_n \in \boldsymbol{e}_s^i} \mathbb{E}_{\gamma,\boldsymbol{w}}[c(g_n)], c^*(g_n) \right\}}{\sum_s \sum_i \sum_{g_n \in \boldsymbol{e}_s^i} \mathbb{E}_{\gamma,\boldsymbol{w}}[c(g_n)]} \right)^{\frac{1}{4}} \times \min\left\{ \exp\left(1 - \frac{\sum_s r_s}{\sum_s \sum_i \sum_{g_1 \in \boldsymbol{e}_s^i} \mathbb{E}_{\gamma,\boldsymbol{w}}[c(g_1)]} \right), 1 \right\}$$

- Minimize BLEU risk
  - by the expected ngram count (Pauls et al., 2009; Rosti et al., 2010; Rosti et al., 2011)
  - not by the expected sentence BLEU (Li and Eisner, 2009)

BP?



$$\min\{e^x, 1\} \approx \frac{e^x - 1}{1 + e^{1000x}} + 1$$

- They tried many alternatives by matlab (Rosti et al., 2010; Rosti et al., 2011)
- Ignore BP(Tromble et al., 2008)
- Ignore min(Pauls et al., 2009)

Online Learning

(k-best) Online Learning

procedure ONLINELEARN($\langle F, E \rangle = \left\{ \langle \boldsymbol{f}^{(i)}, \boldsymbol{e}^{(i)} \rangle \right\}_{i=1}^{N}$)
  $\boldsymbol{w}^{(0)} \leftarrow \emptyset$
  $j \leftarrow 1$
  for $t \in \{1 \dots T\}$ **do**
    Choose $B_t = \{\boldsymbol{b}_1^{(t)}, \cdots \boldsymbol{b}_M^{(t)}\}$ ▷ randomly choose M batch
    for $\boldsymbol{b} \in B_t$ **do** ▷ $\boldsymbol{b} = \{\dots, \langle \boldsymbol{f}, \boldsymbol{e} \rangle, \dots\}$
      $\boldsymbol{c} \leftarrow \text{GEN}(\boldsymbol{b}, \boldsymbol{w}^{(j-1)})$ ▷ decode using $\boldsymbol{w}^{(j-1)}$
      $\boldsymbol{w}^{(j)} \leftarrow \arg\min_{\boldsymbol{w} \in \mathcal{W}} \ell(\boldsymbol{b}, \boldsymbol{c}; \boldsymbol{w}) + \lambda \Omega(\boldsymbol{w})$ ▷ optimize
      $j \leftarrow j + 1$
    end for
  end for
  return $\boldsymbol{w}^{(T \cdot M)}$
end procedure

- Introduce online approximation for objectives by optimizating over a "batch" with decoding step + optimization step

Optimize Updates

- Perceptron for MT (Liang et al., 2006)
- MIRA (or PA-I) for MT (Watanabe et al., 2007; Chiang et al., 2008)
- SGD + projection for MT(Watanabe, 2012; Semianer et al., 2012)

MIRA

$$\arg\min_{\boldsymbol{w}} \frac{\lambda}{2}\|\boldsymbol{w} - \boldsymbol{w}^{(j-1)}\|_2^2 + \sum_{(\boldsymbol{f}^{(i)},\boldsymbol{e}^{(i)})\in\boldsymbol{b}} \sum_{\substack{\boldsymbol{e}^* \in \boldsymbol{o}^{(i)}, \\ \boldsymbol{e}' \in \boldsymbol{c}(i) \setminus \boldsymbol{o}^{(i)}}} \xi_{\boldsymbol{f}^{(i)},\boldsymbol{e}^*,\boldsymbol{e}'}$$

$$\begin{aligned}
\boldsymbol{w}^\top \Delta\boldsymbol{h}(\boldsymbol{f}^{(i)},\boldsymbol{e}^*,\boldsymbol{e}') &\geq \Delta\,\mathrm{error}(\boldsymbol{e}^{(i)},\boldsymbol{e}^*,\boldsymbol{e}') - \xi_{\boldsymbol{f}^{(i)},\boldsymbol{e}^*,\boldsymbol{e}'} \\
\Delta\boldsymbol{h}(\boldsymbol{f}^{(i)},\boldsymbol{e}^*,\boldsymbol{e}') &= \boldsymbol{h}(\boldsymbol{f}^{(i)},\boldsymbol{e}^*) - \boldsymbol{h}(\boldsymbol{f}^{(i)},\boldsymbol{e}') \\
\Delta\,\mathrm{error}(\boldsymbol{e}^{(i)},\boldsymbol{e}^*,\boldsymbol{e}') &= \mathrm{error}(\boldsymbol{e}^{(i)},\boldsymbol{e}') - \mathrm{error}(\boldsymbol{e}^{(i)},\boldsymbol{e}^*)
\end{aligned}$$

- Large margin principle: a good translation is separated from a bad translation by a margin
- Minimize the progress by regularizer (Crammer et al., 2006)
- Easily overfit: stop iterations, averaging etc. (Watanabe et al., 2007; Chiang et al., 2008)

SGD + Projection

$$\arg\min_{\boldsymbol{w}} \frac{\lambda}{2}\|\boldsymbol{w}\|_2^2 + \ell(\boldsymbol{w}; \boldsymbol{b})$$

$$\begin{aligned}
\boldsymbol{w}^{(j-\frac{1}{2})} &\leftarrow (1-\lambda\eta_j)\boldsymbol{w}^{(j-1)} + \eta_j \Delta\ell(\boldsymbol{w}; \boldsymbol{b}) \\
\boldsymbol{w}^{(j)} &\leftarrow \min\left\{1, \frac{1/\sqrt{\lambda}}{\|\boldsymbol{w}^{(j-\frac{1}{2})}\|_2}\right\}\boldsymbol{w}^{(j-\frac{1}{2})}
\end{aligned}$$

- Solve a "batch local" objective in each update
- Set learning rate (η) + update by a sub-gradient (L_2 and $\Delta\ell$) + projection into a L_2-ball (Shalev-Shwartz et al., 2007)

Intricacy of BLEU

- Optimization for a sentence-wise BLEU ≠optimal for a document-wise BLEU
- BLEU on a larger batch: better document-wise BLEU estimates
- However, requiring more iterations
- Previous work: Pseudo-document, Decayed BLEU (Watanabe et al., 2007, Chiang et al., 2008)

Optimized Update

$$\begin{aligned}
\boldsymbol{w}^{(j-\frac{3}{4})} &\leftarrow (1-\lambda\eta_j)\boldsymbol{w}^{(j-1)} \\
\boldsymbol{w}^{(j-\frac{1}{2})} &\leftarrow \arg\min_{\boldsymbol{w}} \frac{1}{2}\|\boldsymbol{w}-\boldsymbol{w}^{(j-\frac{3}{4})}\|_2^2 + \eta_j \Delta\ell(\boldsymbol{w};\boldsymbol{b})
\end{aligned}$$

- 2-step update: suffer sub-gradient from L_2 + solve a QP (Watanabe, 2012)
- Similar to MIRA: global L_2 + directly use the learning rate as a hyperparameter

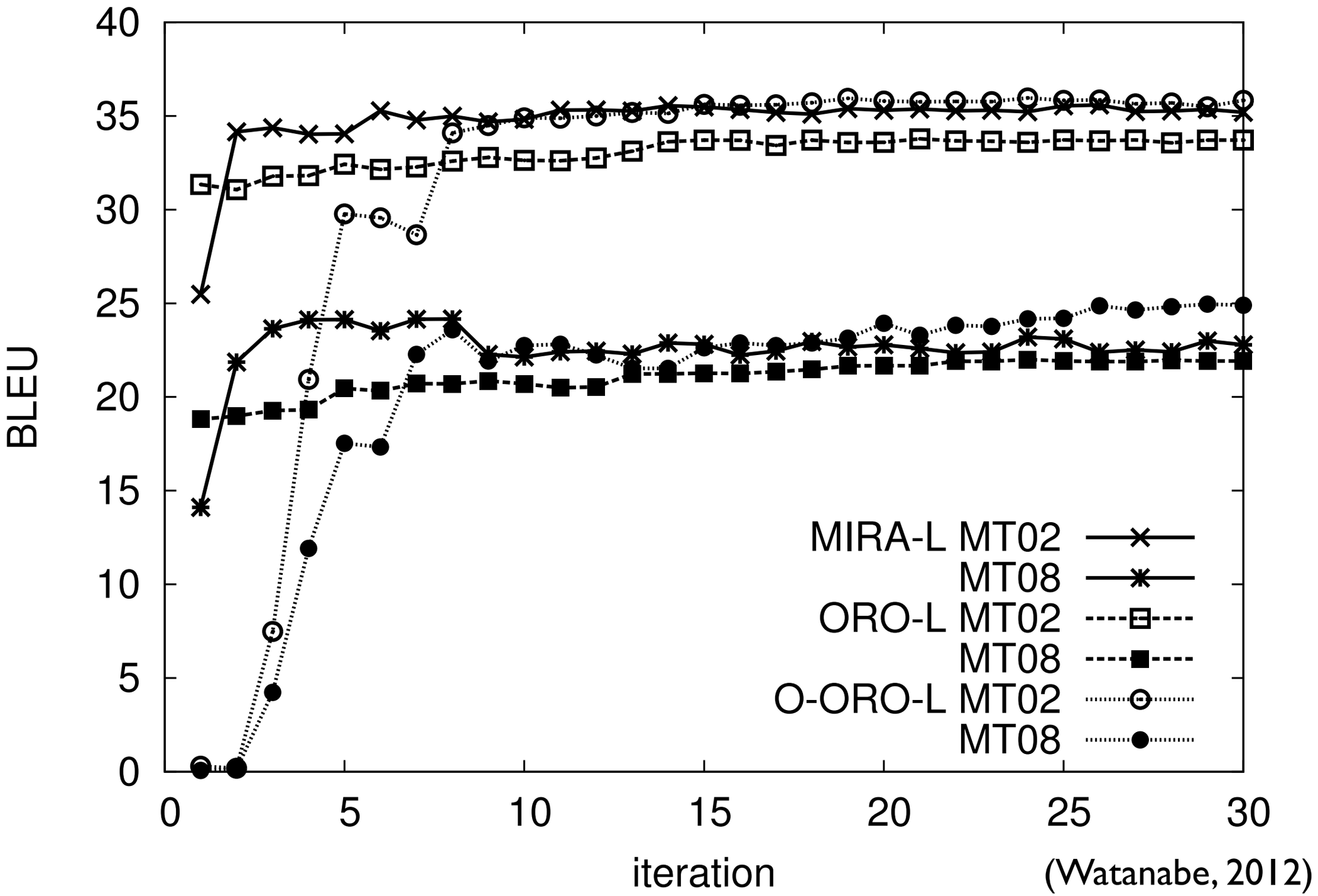
Experiments

	MT06	MT08
MERT	31.45†	24.13†
PRO	31.76†	24.43†
MIRA-L	31.42†	24.15†
ORO-L_{hinge}	29.76	21.96
O-ORO-L_{hinge}	**32.06**	**24.95**
ORO-$L_{softmax}$	30.77	23.07
O-ORO-$L_{softmax}$	31.16†	23.20

(Watanabe, 2012)

- NIST Chinese-to-English translation task
- Tune on MT02, development testing on MT06, testing on MT08

Learning Curves



(Watanabe, 2012)

Experiments: Batch Size

	MT08			
batch size	1	4	8	16
MIRA-L	23.46	23.97†	**24.58**	24.15†
ORO-L_{hinge}	23.63	23.12	22.07	21.96
O-ORO-L_{hinge}	23.72	24.02†	24.28†	**24.95**
ORO-$L_{softmax}$	19.27	23.59	23.50	23.07
O-ORO-$L_{softmax}$	23.62	23.31	23.03	23.20

(Watanabe, 2012)

Conclusion

Summary

- Some concepts from MT
- Optimization for MT
  - Intricacy of an evaluation metric
  - k-best batch approximation for candidate space
  - Online approximation of objectives

Outlook

- Feature selection (Simianer et al., 2012)
- Larger data for optimization (Xiao et al., 2011)
- Bayesian models
  - Phrasal model (DeNero et al., 2008; Neubig et al., 2011)
  - Syntactic models (Blunsom et al., 2009; Cohn and Blunsom, 2009; Levenberg et al., 2012)
- Deep learning (Le et al., 2012)

References

Blunsom, P., Cohn, T., Dyer, C., & Osborne, M. (2009, August). A gibbs sampler for phrasal synchronous grammar induction. In *Proc. of acl/ijcnlp 2009* (pp. 782--790). Suntec, Singapore: Association for Computational Linguistics. Retrieved from http://www.aclweb.org/anthology/P/P09/P09-1088

Blunsom, P., Cohn, T., & Osborne, M. (2008, June). A discriminative latent variable model for statistical machine translation. In *Proc. of acl-08: Hlt* (pp. 200--208). Columbus, Ohio. Retrieved from http://www.aclweb.org/anthology/P/P08/P08-1024

Chiang, D. (2007, June). Hierarchical phrase-based translation. *Computational Linguistics*, *33*(2), 201--228. Retrieved from http://dx.doi.org/10.1162/coli.2007.33.2.201 doi: 10.1162/coli.2007.33.2.201

Chiang, D., DeNeefe, S., Chan, Y. S., & Ng, H. T. (2008, October). Decomposability of translation metrics for improved evaluation and efficient algorithms. In *Proc. of emnlp 2008* (pp. 610--619). Honolulu, Hawaii. Retrieved from http://www.aclweb.org/anthology/D08-1064

Chiang, D., Marton, Y., & Resnik, P. (2008, October). Online large-margin training of syntactic and structural translation features. In *Proc. of emnlp 2008* (pp. 224--233). Honolulu, Hawaii. Retrieved from http://www.aclweb.org/anthology/D08-1024

Cohn, T., & Blunsom, P. (2009, August). A Bayesian model of syntax-directed tree to string grammar induction. In *Proc. of emnlp 2009* (pp. 352--361). Singapore. Retrieved from http://www.aclweb.org/anthology/D/D09/D09-1037

Crammer, K., Dekel, O., Keshet, J., Shalev-Shwartz, S., & Singer, Y. (2006, March). Online passive-aggressive algorithms. *Journal of Machine Learning Research*, *7*, 551--585.

DeNero, J., Bouchard-Côté, A., & Klein, D. (2008, October). Sampling alignment structure under a Bayesian translation model. In *Proc. of emnlp 2008* (pp. 314--323). Honolulu, Hawaii: Association for Computational Linguistics. Retrieved from http://www.aclweb.org/anthology/D08-1033

Galley, M., Hopkins, M., Knight, K., & Marcu, D. (2004, May 2 - May 7). What's in a translation rule? In D. M. Susan Dumais & S. Roukos (Eds.), *Proc. of hlt-naacl 2004* (pp. 273--280). Boston,

Massachusetts, USA: Association for Computational Linguistics.
Hopkins, M., & May, J. (2011, July). Tuning as ranking. In *Proc. of emnlp 2011* (pp. 1352--1362). Edinburgh, Scotland, UK.: Association for Computational Linguistics. Retrieved from http://www.aclweb.org/anthology/D11-1125
Koehn, P., Och, F. J., & Marcu, D. (2003, May-June). Statistical pharse-based translation. In *Proc. of HLT-NAACL 2003* (pp. 48--54). Edmonton.
Le, H.-S., Allauzen, A., & Yvon, F. (2012, June). Continuous space translation models with neural networks. In *Proceedings of the 2012 conference of the north american chapter of the association for computational linguistics: Human language technologies* (pp. 39--48). Montréal, Canada: Association for Computational Linguistics. Retrieved from http://www.aclweb.org/anthology/N12-1005
Levenberg, A., Dyer, C., & Blunsom, P. (2012, July). A bayesian model for learning scfgs with discontiguous rules. In *Proceedings of the 2012 joint conference on empirical methods in natural language processing and computational natural language learning* (pp. 223--232). Jeju Island, Korea: Association for Computational Linguistics. Retrieved from http://www.aclweb.org/anthology/D12-1021
Li, Z., & Eisner, J. (2009, August). First- and second-order expectation semirings with applications to minimum-risk training on translation forests. In *Proc. of emnlp 2009* (pp. 40--51). Singapore. Retrieved from http://www.aclweb.org/anthology/D/D09/D09-1005
Neubig, G., Watanabe, T., Sumita, E., Mori, S., & Kawahara, T. (2011, June). An unsupervised model for joint phrase alignment and extraction. In *Proc. of acl-hlt 2011* (pp. 632--641). Portland, Oregon, USA: Association for Computational Linguistics. Retrieved from http://www.aclweb.org/anthology/P11-1064
Och, F. J. (2003, July). Minimum error rate training in statistical machine translation. In *Proceedings of the 41st annual meeting of the association for computational linguistics* (pp. 160--167). Sapporo, Japan: Association for Computational Linguistics. Retrieved from http://www.aclweb.org/anthology/P03-1021 doi: 10.3115/1075096.1075117
Och, F. J., & Ney, H. (2002, July). Discriminative training and maximum entropy models for statistical machine translation. In *Proc. of acl 2002* (pp. 295--302). Philadelphia, Pennsylvania, USA. Retrieved from http://www.aclweb.org/anthology/P02-1038 doi: 10.3115/1073083.1073133

Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2002, July). Bleu: a method for automatic evaluation of machine translation. In *Proc. of acl 2002* (pp. 311--318). Philadelphia, Pennsylvania, USA: Association for Computational Linguistics. Retrieved from http://www.aclweb.org/anthology/P02-1040 doi: 10.3115/1073083.1073135

Pauls, A., Denero, J., & Klein, D. (2009, August). Consensus training for consensus decoding in machine translation. In *Proc. of emnlp 2009* (pp. 1418--1427). Singapore. Retrieved from http://www.aclweb.org/anthology/D/D09/D09-1147

Rosti, A.-V., Zhang, B., Matsoukas, S., & Schwartz, R. (2010, July). Bbn system description for wmt10 system combination task. In *Proceedings of the joint fifth workshop on statistical machine translation and metricsmatr* (pp. 321--326). Uppsala, Sweden: Association for Computational Linguistics. Retrieved from http://www.aclweb.org/anthology/W10-1748

Rosti, A.-V., Zhang, B., Matsoukas, S., & Schwartz, R. (2011, July). Expected bleu training for graphs: Bbn system description for wmt11 system combination task. In *Proceedings of the sixth workshop on statistical machine translation* (pp. 159--165). Edinburgh, Scotland: Association for Computational Linguistics. Retrieved from http://www.aclweb.org/anthology/W11-2119

Shalev-Shwartz, S., Singer, Y., & Srebro, N. (2007). Pegasos: Primal estimated sub-gradient solver for svm. In *Proc. of icml '07* (pp. 807--814). Corvalis, Oregon. Retrieved from http://doi.acm.org/10.1145/1273496.1273598 doi: http://doi.acm.org/10.1145/1273496.1273598

Simianer, P., Riezler, S., & Dyer, C. (2012, July). Joint feature selection in distributed stochastic learning for large-scale discriminative training in smt. In *Proceedings of the 50th annual meeting of the association for computational linguistics (volume 1: Long papers)* (pp. 11--21). Jeju Island, Korea: Association for Computational Linguistics. Retrieved from http://www.aclweb.org/anthology/P12-1002

Tromble, R., Kumar, S., Och, F., & Macherey, W. (2008, October). Lattice Minimum Bayes-Risk decoding for statistical machine translation. In *Proc. of emnlp 2008* (pp. 620--629). Honolulu, Hawaii. Retrieved from http://www.aclweb.org/anthology/D08-1065

Watanabe, T. (2012, June). Optimized online rank learning for machine translation. In *Proceedings of the 2012 conference of the north american chapter of the association for computational linguistics: Human language technologies* (pp. 253--262). Montréal, Canada: Association for Computational Linguistics.

Retrieved from http://www.aclweb.org/anthology/N12-1026

Watanabe, T., Suzuki, J., Tsukada, H., & Isozaki, H. (2007, June). Online Large-Margin Training for Statistical Machine Translation. In *Proc. of emnlp-conll 2007* (pp. 764--773). Prague, Czech Republic.

Xiao, X., Liu, Y., Liu, Q., & Lin, S. (2011, July). Fast generation of translation forest for large-scale smt discriminative training. In *Proceedings of the 2011 conference on empirical methods in natural language processing* (pp. 880--888). Edinburgh, Scotland, UK.: Association for Computational Linguistics. Retrieved from http://www.aclweb.org/anthology/D11-1081