

最適化

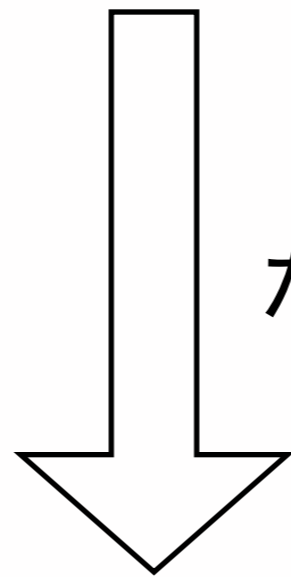
渡辺太郎

taro.watanabe at nict.go.jp



<https://sites.google.com/site/alaginmt2015/>

機械翻訳について勉強したい。



なるべく良い翻訳..

I want to study about machine translation.
I need to master machine translation.
machine translation want to study.

infobox)

infobox buddhist

道元は鎌倉時代初期のである。

道元（どうげん）は、鎌倉時代初期の禅僧。

曹洞禅の開祖

曹洞宗の開祖。

その生涯には kigen 名がある。

晩年に希玄という異称も用いた。

一般には宗と呼ばれることによって尊称は高僧がある。

同宗旨では高祖と尊称される。

死後にといった仏所伝灯国師、 joyo-daishi である。

諡は、仏性伝東国師、承陽大師_(僧)。

一般には道元禅師と呼ばれる。

一般には道元禅師と呼ばれる。

また～686の普及についての修行を tooth brushing、大峰、食事作法

cleaning として、日本にしている。

日本に歯磨き洗面、食事の際の作法や掃除の習慣を広めたといわれ

れる。

エラー

- 探索エラー: スコアの高い翻訳を出すのに失敗
- モデルエラー: スコアの高い翻訳が誤っている
- 学習データの問題: 小さい、異なる
- 手軽な対処: 最適化(チューニング)

k-best 翻訳

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

kbest翻訳

機械翻訳について勉強したい。

$$\log Pr(\phi|e) \log Pr(e) \log Pr(f, \alpha|\phi)$$

I want to study about machine translation

I need to master machine translation

machine translation want to study

I don't want to learn anything

-2	-3	-4	-9
-3	-4	-4	-11
-2	-5	-1	-8
-5	-2	-3	-10
0.5×-2	0.4×-3	0.2×-4	-3.0
0.5×-3	0.4×-4	0.2×-4	-3.9
0.5×-2	0.4×-5	0.2×-1	-3.2
0.5×-5	0.4×-2	0.2×-3	-3.9

重み付けにより並び替え

重み付け



$$\hat{e} = \arg \max_e Pr(\mathbf{f}, \alpha | \phi, \mathbf{e})^{0.2} Pr(\phi | \mathbf{e})^{0.5} Pr(\mathbf{e})^{0.4}$$

- より一般化:

$$\hat{e} = \arg \max_e \frac{\sum_d \exp(\mathbf{w}^\top \mathbf{h}(\mathbf{f}, d, \mathbf{e}))}{\sum_{e', d'} \exp(\mathbf{w}^\top \mathbf{h}(\mathbf{f}, d', e'))}$$

$$\approx \arg \max_{\langle \mathbf{e}, d \rangle} \mathbf{w}^\top \mathbf{h}(\mathbf{f}, d, \mathbf{e}) \quad (\text{Och and Ney, 2002})$$

最適化 = 最適な \mathbf{w} を識別学習



MTパイプライン

kyoto-train.{ja,en}

対訳データ

大量(低品質?)

翻訳モデル

言語モデル

デコーダ

重みパラメータ

(少量?)高品質

対訳データ

kyoto-dev.{ja,en}

一つの次元に着目

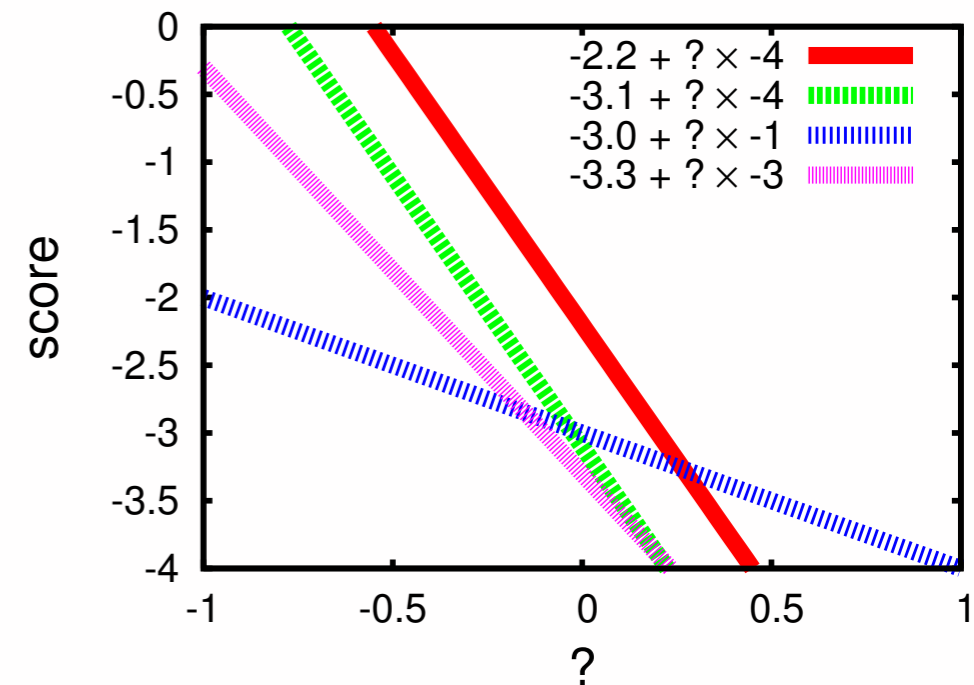


$$\hat{w} = \arg \min_w \ell_{\text{error}} \left(\left\{ \arg \max_e w^T \cdot h(f^{(s)}, e) \right\}_{s=1}^S, \left\{ e^{(s)} \right\}_{s=1}^S \right)$$

(Och, 2003)

0.5 × -2	0.4 × -3	? × -4
0.5 × -3	0.4 × -4	? × -4
0.5 × -2	0.4 × -5	? × -1
0.5 × -5	0.4 × -2	? × -3

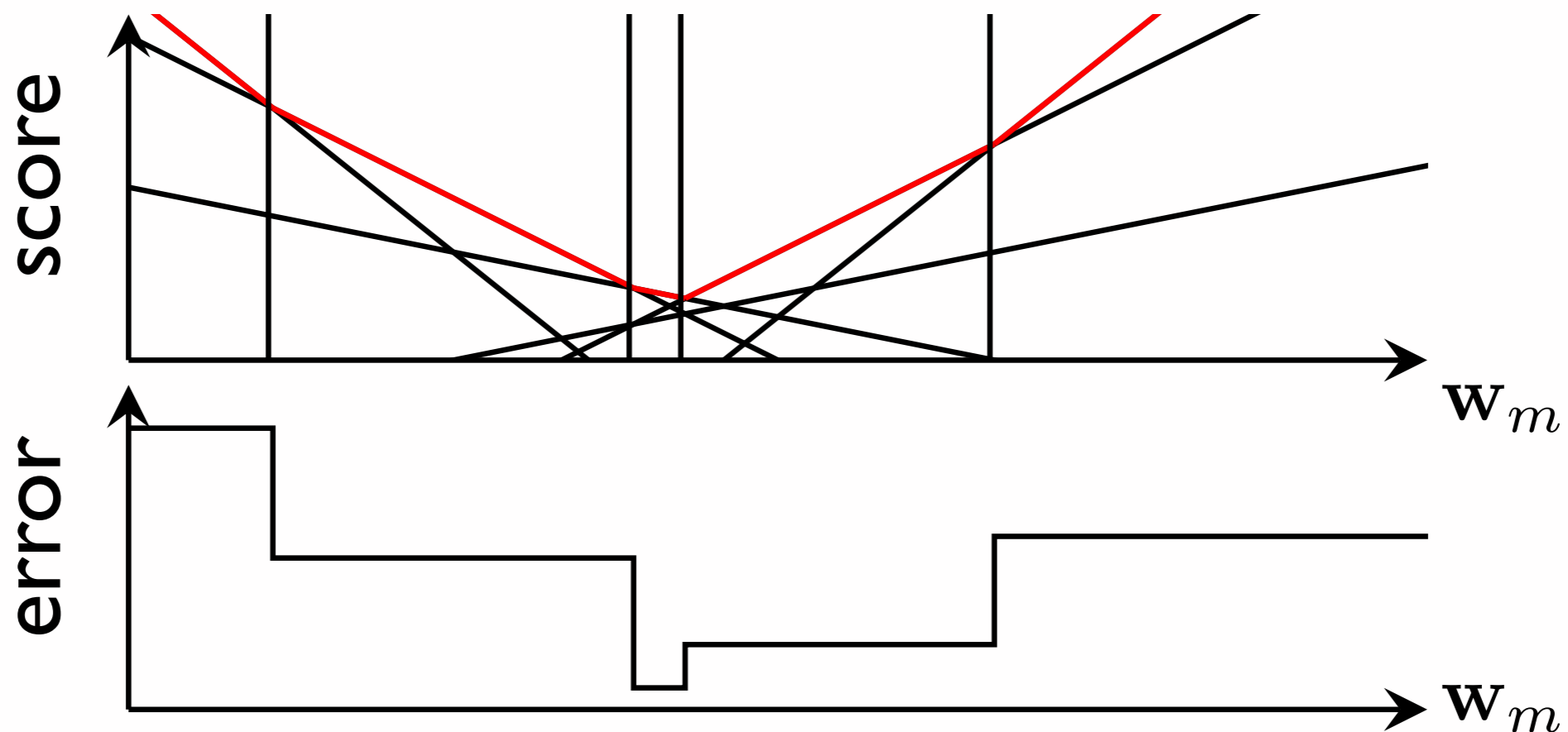
-2.2 + ? × -4
-3.1 + ? × -4
-3.0 + ? × -1
-3.3 + ? × -3



一つの次元に着目、線分探索

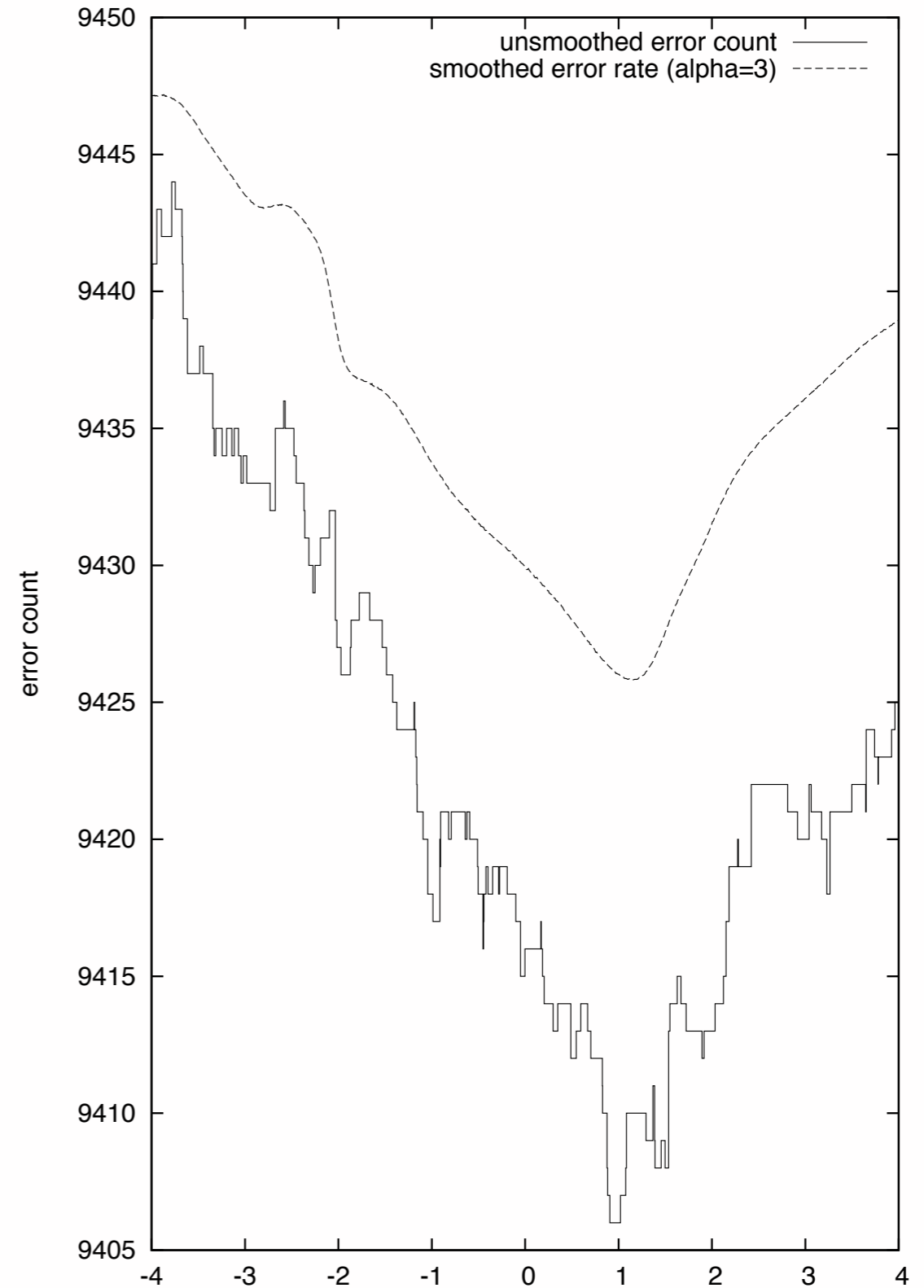
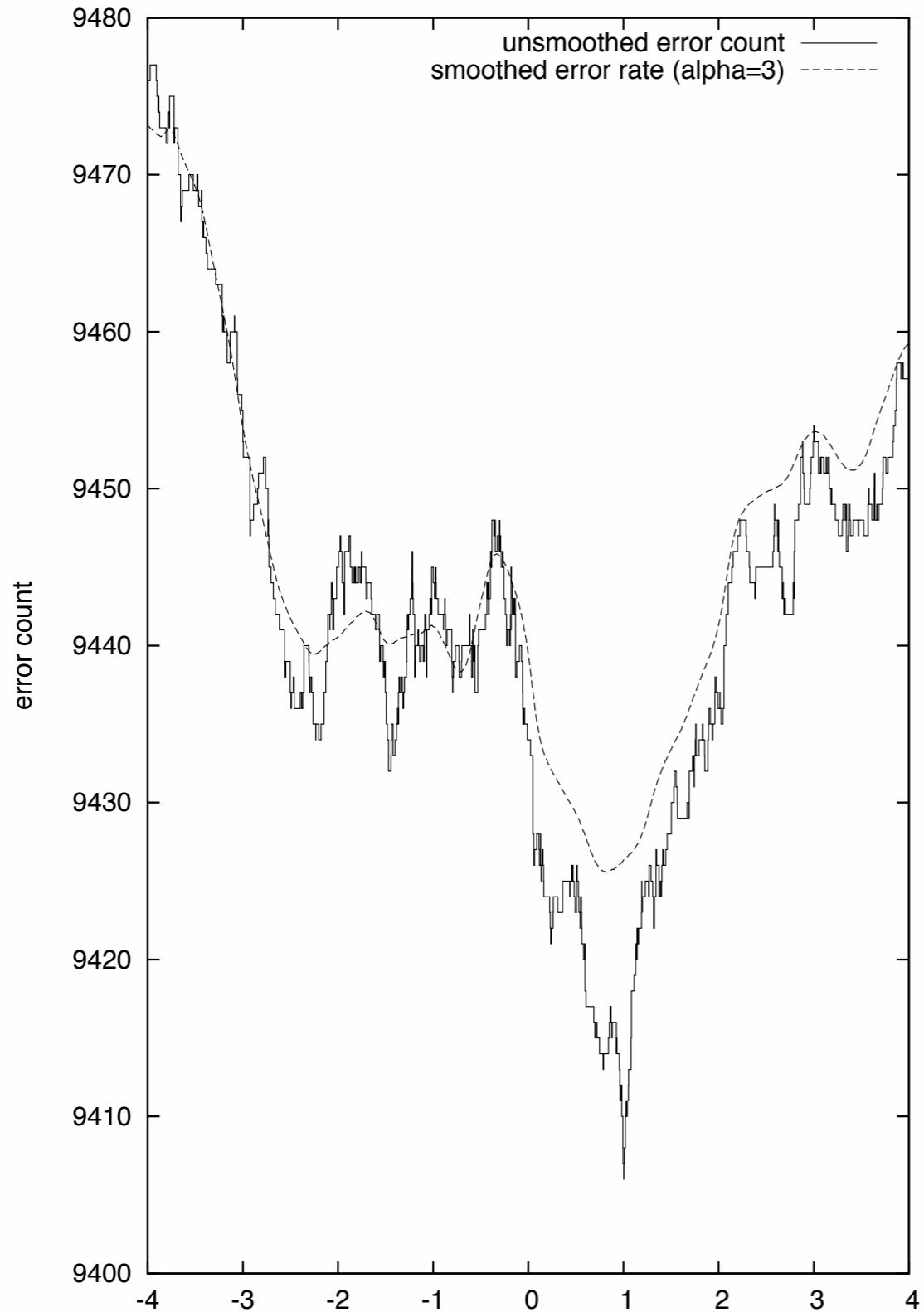
線分探索

$$\hat{e} = \operatorname{argmax}_e \underbrace{w_m^\top \cdot h_m(f^{(s)}, d, e)}_{\text{傾き}} + \underbrace{w_{m_}^\top \cdot h_{m_}(f^{(s)}, d, e)}_{\text{切片}}$$

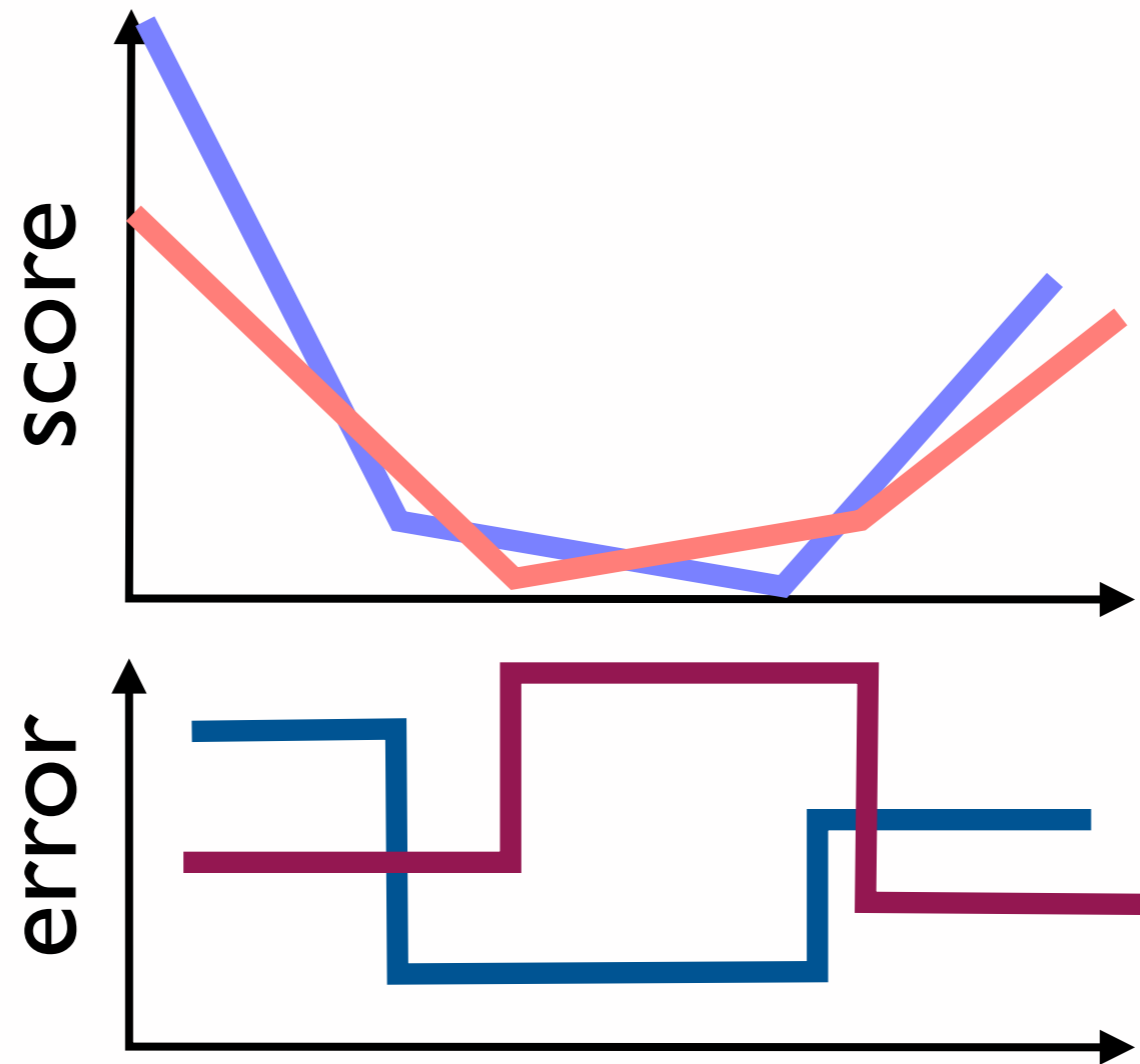


- 一つの次元を選択した場合、各候補を「線」としてみなせる
- 「線」の集合から、凸包(convex hull)を計算

エラー曲線



エラー空間への投射



- 複数の文の凸包を同一の軸へ写像
- 文書単位のエラー(i.e. BLEU)を計算可能

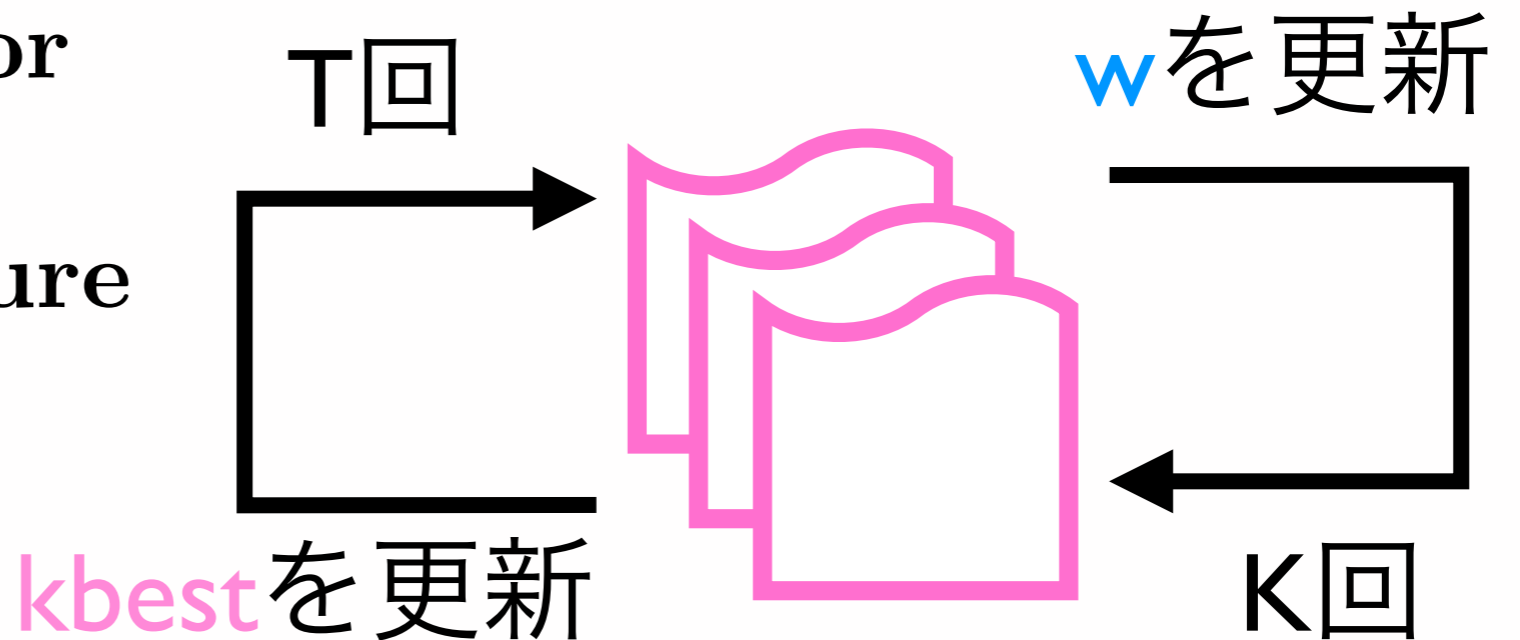
MERT

$$\hat{w} = \arg \min_w l_{\text{error}} \left(\left\{ \arg \max_e w^\top \cdot h(f^{(s)}, e) \right\}_{s=1}^S, \left\{ e^{(s)} \right\}_{s=1}^S \right)$$

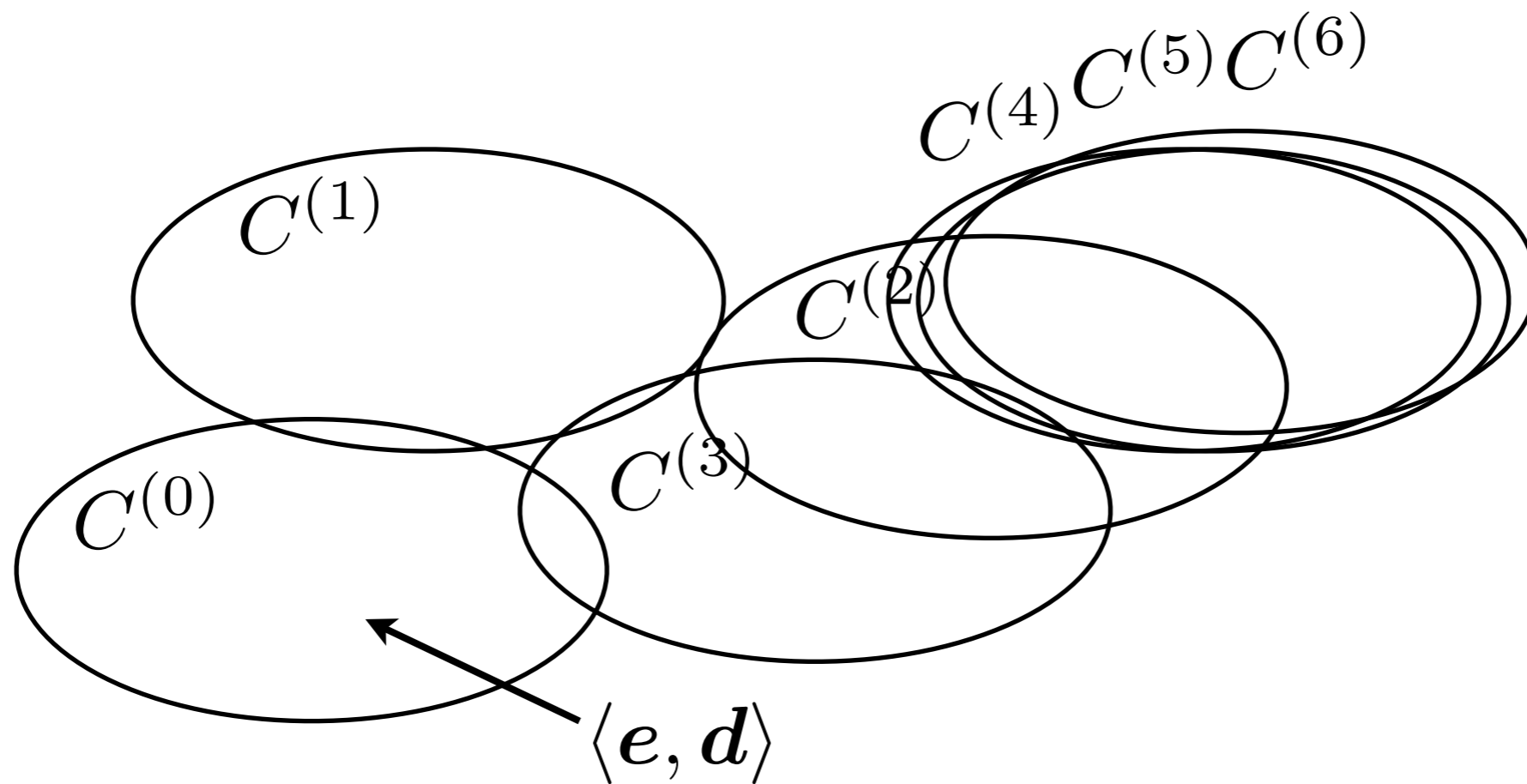
- MERT (Minimum Error Rate Training) (Och, 2003)
- $l_{\text{error}}(\cdot)$ に対して、様々なエラー関数を使用可能 (例えば、1 - BLEU)
- w を更新するたびに、argmax を計算: kbest 近似

おおきなループ

```
1: procedure MERT( $\left\{ (f^{(s)}, e^{(s)}) \right\}_{s=1}^S$ )
2:   for  $n = 1 \dots T$  do ▷ 繰り返し  $T$ 
3:      $w$  でデコード、 $kbest$  リストを生成
4:      $kbest$  リストを結合
5:     for  $k = 1 \dots K$  do ▷ 繰り返し  $K$ 
6:       for  $m = 1 \dots M$  do
7:         次元  $m$  で線分探索、 $w$  を更新
8:       end for
9:     end for
10:  end for
11: end procedure
```

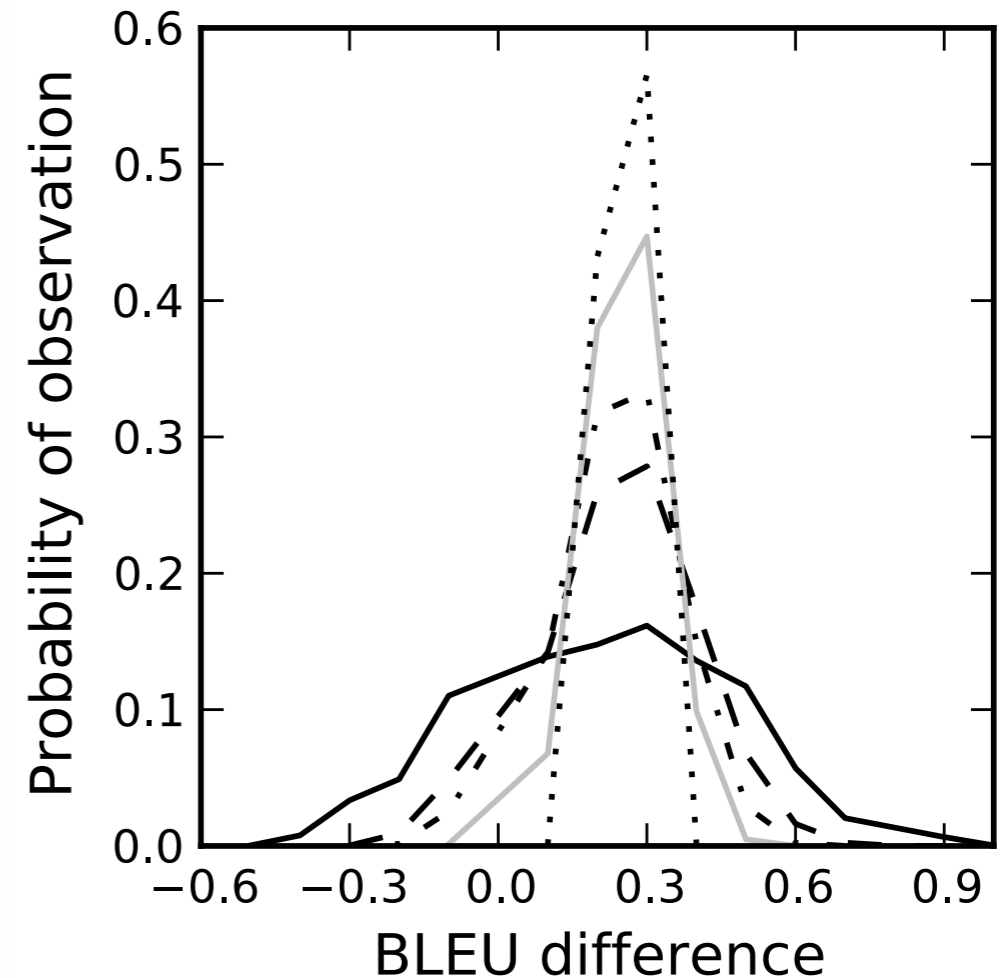
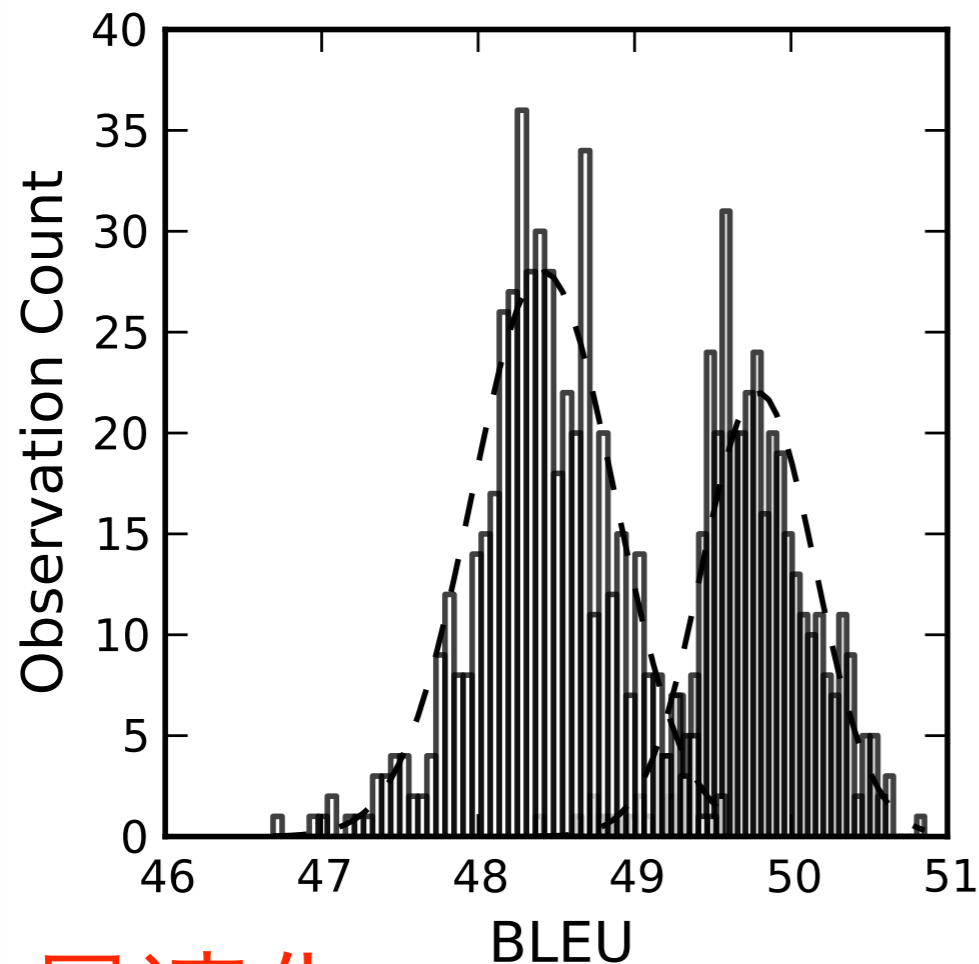
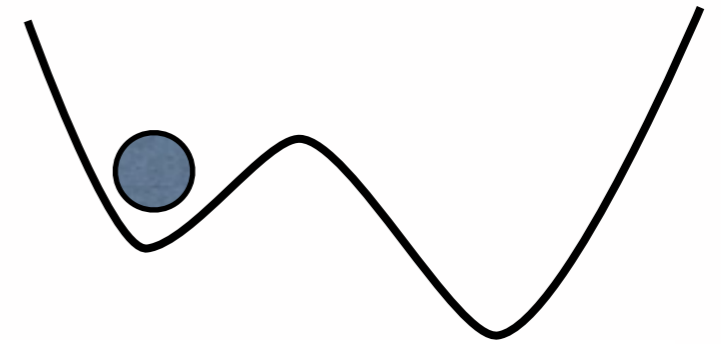


kbest近似最適化



- 各繰り返しでk-bestを結合しつつ学習
- 収束はするけど... 局所解

局所解



(Clark et al., 2011)

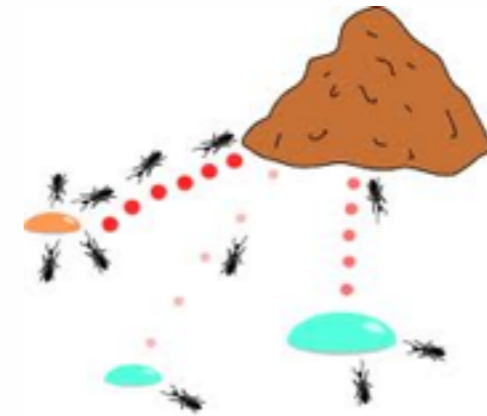
最適化

- ~~MERT~~ は、パラメータの初期値に大きく依存

最適化

- ランダムな初期値から複数 ~~MERT~~

最適化



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

- ある損失関数 ℓ を仮定し、対訳データ (F, E) に対するリスクを最小化□□
- 真の分布は未知なため、正則化 (Ω) された経験リスクを最小化

kbest近似最適化

```
1: procedure LEARN( $\langle F, E \rangle$ )
2:    $C = \{\}$ 
3:   for  $t = 1 \dots T$  do
4:      $w$ でデコード、kbestリストを生成
5:     kbestリストを $C$ へ結合
6:      $w^{(t+1)} = \arg \min_w \ell(F, E, C; w) + \lambda \Omega(w)$ 
7:   end for
8: end procedure
```

- ℓ_{error} (MERT) 以外の ℓ ? (できれば、凸関数)
- Ω は... L_1 あるいは L_2 正則化

確率モデル

e_1 I want to study about machine translation

e_2 I need to master machine translation

e_3 machine translation want to study

e_4 I don't want to learn anything

$$\begin{aligned} &Pr(e_1 \text{ or } e_2 | e_1 \text{ or } e_2 \text{ or } e_3 \text{ or } e_4) \\ &> Pr(e_3 \text{ or } e_4 | e_1 \text{ or } e_2 \text{ or } e_3 \text{ or } e_4) \end{aligned}$$

- kbestのうち、良さそうな翻訳(オラクル翻訳)の確率が高くなるように最適化(Och and Ney, 2002)

確率モデル

$$\hat{\boldsymbol{w}} = \arg \min_{\boldsymbol{w}} \ell_{\text{softmax}}(F, E, C) + \lambda \Omega(\boldsymbol{w})$$

$$\ell_{\text{softmax}}(F, E, C) = - \sum_{s=1}^S \frac{\sum_{\boldsymbol{e}^* \in \boldsymbol{o}^{(s)}} \boldsymbol{w}^\top \boldsymbol{h}(f^{(s)}, \boldsymbol{e}^*)}{\sum_{\boldsymbol{e}' \in \boldsymbol{c}^{(s)}} \boldsymbol{w}^\top \boldsymbol{h}(f^{(s)}, \boldsymbol{e}')}$$

(Och and Ney, 2002; Blunsom et al., 2008)

- オラクル翻訳($\boldsymbol{o}^{(s)}$)は山登り法などで決定
- CGやLBFGS、SGDで最適化

PRO

e_1 I want to study about machine translation

e_2 I need to master machine translation

e_3 machine translation want to study

e_4 I don't want to learn anything

$$\text{error}(e_1) < \text{error}(e_3) \iff w^\top h(f, e_1) > w^\top h(f, e_3)$$

- Pairwise Rank Optimization(PRO) (Hopkins and May, 2011)
- 各翻訳候補の全順序関係を見るのではなく、ペア単位で比較

PRO

$$\hat{\boldsymbol{w}} = \arg \min_{\boldsymbol{w}} \ell_{\text{hinge}}(F, E, C) + \lambda \Omega(\boldsymbol{w})$$

$$\ell_{\text{hinge}}(F, E, C) =$$

$$\sum_{s=1}^S \sum_{\boldsymbol{e} \in \mathbf{c}^{(s)}} \sum_{\substack{\boldsymbol{e}' \in \mathbf{c}^{(s)} \\ \text{error}(\boldsymbol{e}) < \text{error}(\boldsymbol{e}')}} \max \left\{ 0, 1 - \left(\boldsymbol{w}^{\top} \boldsymbol{h}(\boldsymbol{f}^{(s)}, \boldsymbol{e}) - \boldsymbol{w}^{\top} \boldsymbol{h}(\boldsymbol{f}^{(s)}, \boldsymbol{e}') \right) \right\}$$

- 網羅的にペア単位に比較(あるいはサンプリング)、誤ったランクに対するペナルティ
- 二値分類問題へと帰着: 一般的な分類器

期待値

e_1 I want to study about machine translation

e_2 I need to master machine translation

e_3 machine translation want to study

e_4 I don't want to learn anything

$$\mathbb{E} \begin{bmatrix} \text{error}(e_1) \text{ or} \\ \text{error}(e_2) \text{ or} \\ \text{error}(e_3) \text{ or} \\ \text{error}(e_4) \end{bmatrix}$$

- 翻訳エラーの期待値を最小化

期待値

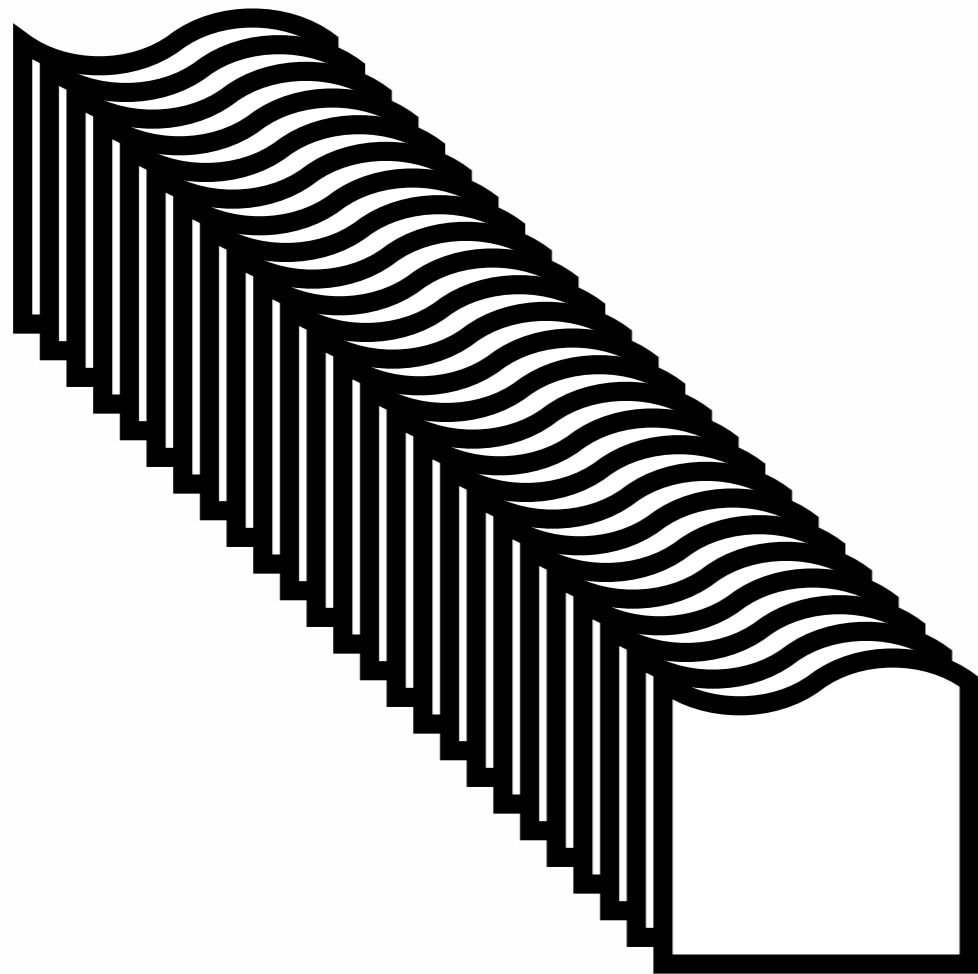
$$\ell_{\text{expectation}}(F, E, C) = \sum_{s=1}^S \sum_{e \in c^{(s)}} \text{error}(e) P_{\lambda, \mathbf{w}}(e | \mathbf{f}^{(s)})$$

$$P_{\lambda, \mathbf{w}}(e | \mathbf{f}^{(s)}) = \frac{\lambda \mathbf{w}^\top \mathbf{h}(\mathbf{f}^{(s)}, e)}{\sum_{e' \in c^{(s)}} \lambda \mathbf{w}^\top \mathbf{h}(\mathbf{f}^{(s)}, e')}$$

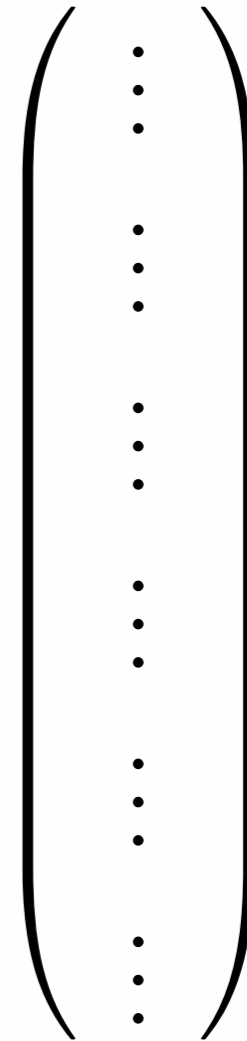
(Smith and Eisner, 2006)

- 文単位の翻訳エラーの期待値: テイラー展開 (Tromble et al., 2008) や ngram の期待値 (Pauls et al., 2009; Rosti et al., 2010; Rosti et al., 2011) により、文単位の尺度を近似
- CG や LBFGS、SGD で最適化

大規模化

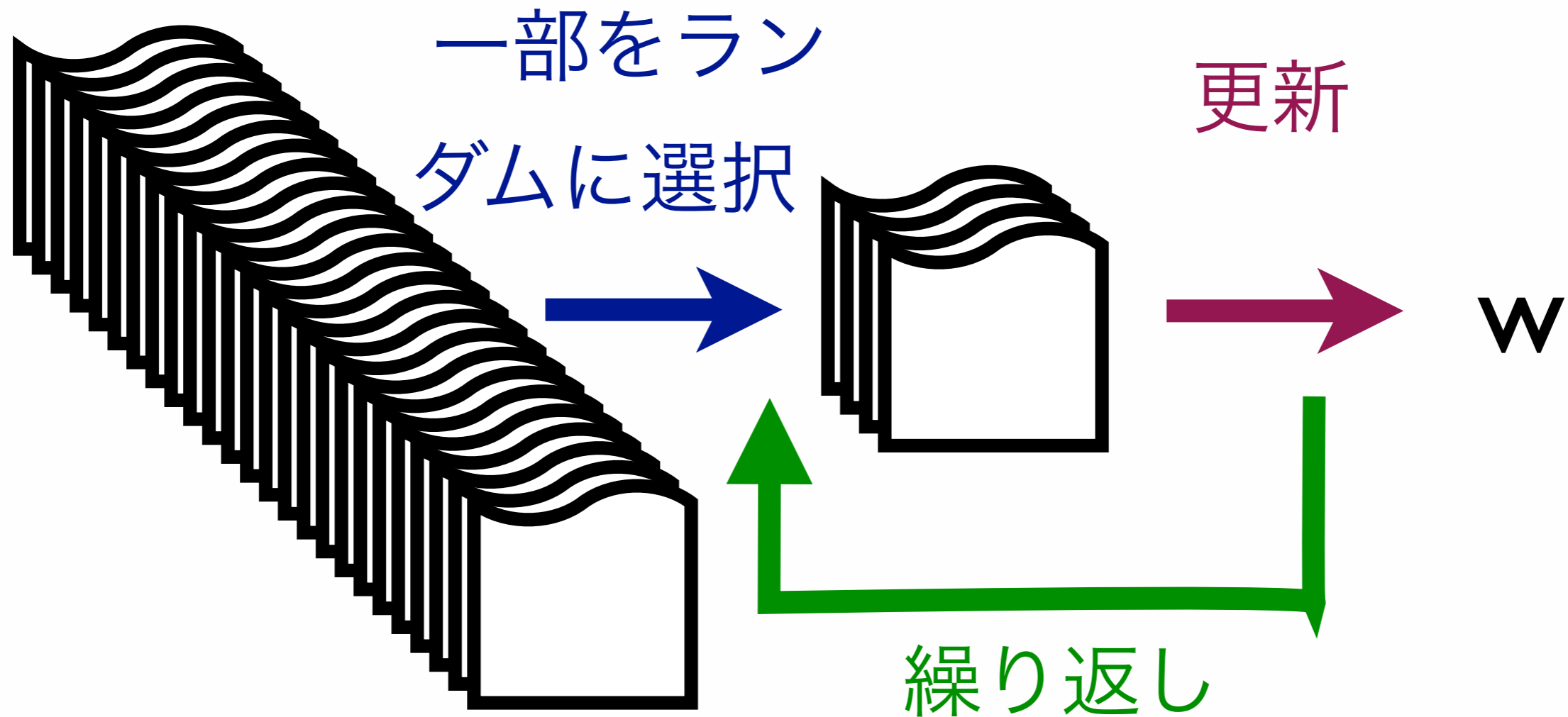


$$h(f, d, e) =$$



- データが多い、高次元な素性ベクトル
- 効率のよい学習手法?

どうせ近似するなら...



- 学習データを近似
- 非常に簡単なアルゴリズムで実現

オンライン学習

```
1: procedure ONLINELEARN( $\langle F, E \rangle = \left\{ \langle f^{(i)}, e^{(i)} \rangle \right\}_{i=1}^N$ )
2:    $w^{(1)} \leftarrow \emptyset$ 
3:   for  $t \in \{1, \dots, T\}$  do
4:      $\langle \tilde{F}^{(t)}, \tilde{E}^{(t)} \rangle \subseteq \langle F, E \rangle$ 
5:      $\tilde{C}^{(t)} \leftarrow \text{GEN}(\tilde{F}^{(t)}, w^{(t)})$ 
6:      $w^{(t+1)} \leftarrow \arg \min_{w \in \mathcal{W}} \ell(\tilde{F}^{(t)}, \tilde{E}^{(t)}, \tilde{C}^{(t)}; w) + \lambda \Omega(w)$ 
7:   end for
8:   return  $w^{(T+1)}$ 
9: end procedure
```

ランダムにサンプル
リング: mini batch

デコード

更新

- 自然言語処理ではよく使われる: 形態素解析、構文解析 etc.
- 当然、機械翻訳にも使えます (Watanabe et al., 2007)

問題: BLEU



e_1 I want to study about machine translation

e_2 I need to master machine translation

e_3 machine translation want to study

e_4 I don't want to learn anything

- k-bestのランキングに基づく最適化
- コーパス単位のBLEU \neq 文単位のBLEUの平均
- 文単位では、正しく学習するのは困難

BLEUの近似

$$\text{GEN}(f^{(s)}, w)$$

$$e^{(1)}, \dots, \begin{pmatrix} c_1^{(s)} \\ \vdots \\ c_i^{(s)} \\ \vdots \\ c_K^{(s)} \end{pmatrix}, \dots, e^{(S)}$$

- 今までの各文に対するBLEUの統計量を保存(I-bestあるいはoracle) (Watanabe et al., 2007)

減衰によるBLEUの近似

$$\mathbf{b} \leftarrow 0.9 \times (\mathbf{b} + \mathbf{c}(e))$$

$$l \leftarrow 0.9 \times (l + |\mathbf{f}|)$$

$$B(e) = (l + |\mathbf{f}|) \times \text{Bleu}(\mathbf{b} + \mathbf{c}(e))$$

$$\hat{e}^{(s)} = \underset{e}{\operatorname{argmax}} -B(e) + \mathbf{w}^\top \mathbf{h}(\mathbf{f}^{(s)}, e)$$

$$\dot{e}^{(s)} = \underset{e}{\operatorname{argmax}} +B(e) + \mathbf{w}^\top \mathbf{h}(\mathbf{f}^{(s)}, e)$$

- sentence-BLEUに対して、今までのBLEUの履歴 ($\times 0.9$)を加える (Chiang et al., 2008)

パーセプトロン

```
1: procedure PERCEPTRON( $\langle F, E \rangle = \left\{ \langle f^{(i)}, e^{(i)} \rangle \right\}_{i=1}^N$ )
2:    $w^{(1)} \leftarrow \mathbf{0}$ 
3:   for  $t \in \{1 \dots T\}$  do
4:      $\langle f, e \rangle \sim \langle F, E \rangle$ 
5:      $\tilde{c} \leftarrow \text{GEN}(f, w^{(t)})$ 
6:      $\langle \hat{e}, \hat{d} \rangle \in \tilde{c}$ 
7:      $\langle e^*, d^* \rangle \in \tilde{o} \subseteq \tilde{c}$ 
8:     if  $\langle e^*, d^* \rangle \neq \langle \hat{e}, \hat{d} \rangle$  then
9:        $w^{(t+1)} \leftarrow w^{(t)} + h(f, e^*, d^*) - h(f, \hat{e}, \hat{d})$ 
10:    end if
11:  end for
12:  return  $w^{(T+1)}$ 
13: end procedure
```

一文だけサンプル

I-best

オラクル翻訳

間違ったらペナルティ

最後のパラメータ or 平均値

パーセプトロン

e_3 machine translation want to study

e_2 I need to master machine translation

e_1 I want to study about machine translation

e_4 I don't want to learn anything

$$\arg \min_w \max \{0, 0 - (w^\top h(f, e_1) - w^\top h(f, e_3))\}$$

- e_1 のスコアが、 e_3 のスコアよりも、大きくなるように、 w を更新

hinge損失

e_3 machine translation want to study

e_2 I need to master machine translation

e_1 I want to study about machine translation

e_4 I don't want to learn anything

$$\arg \min_w \max \{0, 1 - (w^\top h(f, e_1) - w^\top h(f, e_3))\}$$

- e_1 のスコアが、 e_3 のスコアよりも、**1以上**大きくなるように、 w を更新

MIRA

$$\arg \min_w \lambda \frac{1}{2} \|w - w^{(t)}\|^2 + \Delta \text{error}(e_1, e_3) - (w^\top h(f, e_1) - w^\top h(f, e_3))$$

(Crammer et al., 2006)

- 以前のパラメータ $w^{(t)}$ との差分を小さくしつつ、 e_1 のスコアが、 e_3 のスコアよりも、**翻訳のエラー以上**大きくなるように、 w を更新

MIRA

$$\mathbf{w}^{(t+1)} \leftarrow \mathbf{w}^{(t)} + \alpha^{(t)} \Delta \mathbf{h}(\mathbf{f}, \mathbf{e}_1, \mathbf{e}_3)$$

$$\alpha^{(t)} = \min \left\{ \frac{1}{\bar{\lambda}}, \frac{\Delta \text{error}(\mathbf{e}_1, \mathbf{e}_3) - \mathbf{w}^{(t)\top} \Delta \mathbf{h}(\mathbf{f}, \mathbf{e}_1, \mathbf{e}_3)}{\|\Delta \mathbf{h}(\mathbf{f}, \mathbf{e}_1, \mathbf{e}_3)\|^2} \right\}$$

$$\Delta \mathbf{h}(\mathbf{f}, \mathbf{e}_1, \mathbf{e}_3) = \mathbf{h}(\mathbf{f}, \mathbf{e}_1) - \mathbf{h}(\mathbf{f}, \mathbf{e}_3)$$

- 非常に簡単な更新: $\alpha^{(t)}$ により更新の量を調整
- $\alpha^{(t)}=1$: パーセプトロン

他にも、CWやAROWなど様々な更新式

まとめ

- kbestを結合しながらバッチ学習
 - MERT: 線分探索による最適化
- 他にも: 確率モデル、PRO、期待エラー
 - MERTと比較して、大量の素性を最適化可能
- オンライン学習: 大規模データ、高次元素性

参考文献

- Phil Blunsom, Trevor Cohn, and Miles Osborne. 2008. A discriminative latent variable model for statistical machine translation. In *Proc. of ACL-08: HLT*, pages 200-208, Columbus, Ohio, June.
- David Chiang, Yuval Marton, and Philip Resnik. 2008. Online large-margin training of syntactic and structural translation features. In *Proc. of EMNLP 2008*, pages 224-233, Honolulu Hawaii, October.
- David Chiang, Kevin Knight, and Wei Wang. 2009. 11,001 new features for statistical machine translation. In *Proc. of NAACL-HLT 2009*, pages 218-226, Boulder, Colorado, June
- Jonathan H. Clark, Chris Dyer, Alon Lavie, and Noah A Smith. 2011. Better hypothesis testing for statistical machine translation: Controlling for optimizer instability. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 176-181, Portland, Oregon, USA, June.
- Koby Crammer, Ofer Dekel, Joseph Keshet, Shai Shalev-Shwartz, and Yoram Singer. 2006. Online passive-aggressive algorithms. *Journal of Machine Learning Research*, 7:551-585 March.
- Mark Hopkins and Jonathan May. 2011. Tuning as ranking. In *Proc. of EMNLP 2011*, pages 1352-1362, Edinburgh Scotland, UK., July.

参考文献

- Franz Josef Och and Hermann Ney. 2002. Discriminative training and maximum entropy models for statistical machine translation. In *Proc. of ACL 2002*, pages 295-302 Philadelphia, Pennsylvania, USA, July.
- Franz Josef Och. 2003. Minimum error rate training in statistical machine translation. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 160-167, Sapporo, Japan, July.
- Adam Pauls, John Denero, and Dan Klein. 2009. Consensus training for consensus decoding in machine translation In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 1418-1427, Singapore August.
- Antti-Veikko Rosti, Bing Zhang, Spyros Matsoukas, and Richard Schwartz. 2010. Bbn system description for wmt10 system combination task. In *Proc. of SMT-MetricsMATR 2010* pages 321-326, Uppsala, Sweden, July.
- Antti-Veikko Rosti, Bing Zhang, Spyros Matsoukas, and Richard Schwartz. 2011. Expected bleu training for graphs Bbn system description for wmt11 system combination task In *Proc. of SMT 2011*, pages 159-165, Edinburgh, Scotland July.
- David A. Smith and Jason Eisner. 2006. Minimum risk annealing for training log-linear models. In *Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions*, pages 787-794, Sydney, Australia, July.
- Taro Watanabe, Jun Suzuki, Hajime Tsukada, and Hidek Isozaki. 2007. Online Large-Margin Training for Statistical Machine Translation. In *Proc. of EMNLP-CoNLL 2007*, pages 764-773, Prague, Czech Republic, June.