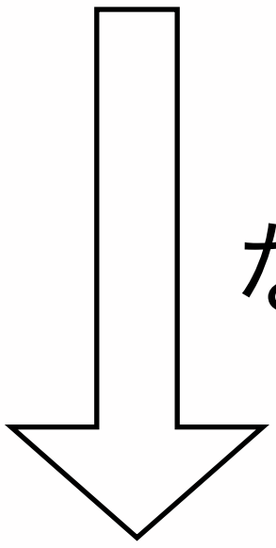
最適化

渡辺太郎

taro.watanabe at nict.go.jp

https://sites.google.com/site/alaginmt2014/

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 お い 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 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

どうしましょう?

f = 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	-2	-3	-4	-9
I need to master machine translation	-3	-4	-4	-11
machine translation want to study	-2	-5	-1	-8
I don't want to learn anything	-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

kbestを正しく並び替えるように重みを学習

重み付け

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

- より一般化:

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

- 複数の素性h(e,d, f)をlog-linearに組み合わせ

最適化 = 最適なwを学習

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

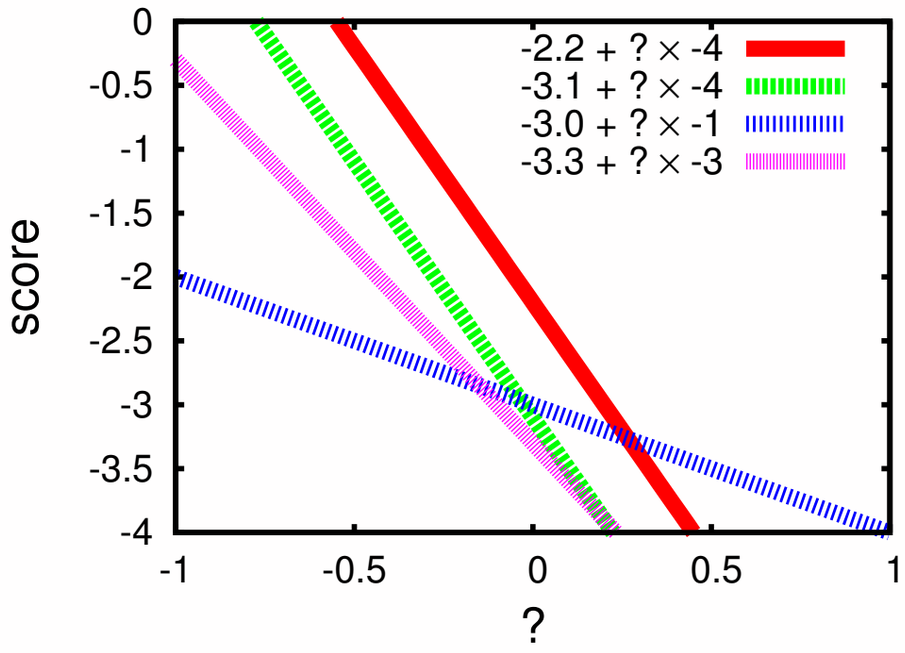
kyoto-dev.{ja,en}

一つの次元に着目

f = 機械翻訳について勉強したい

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

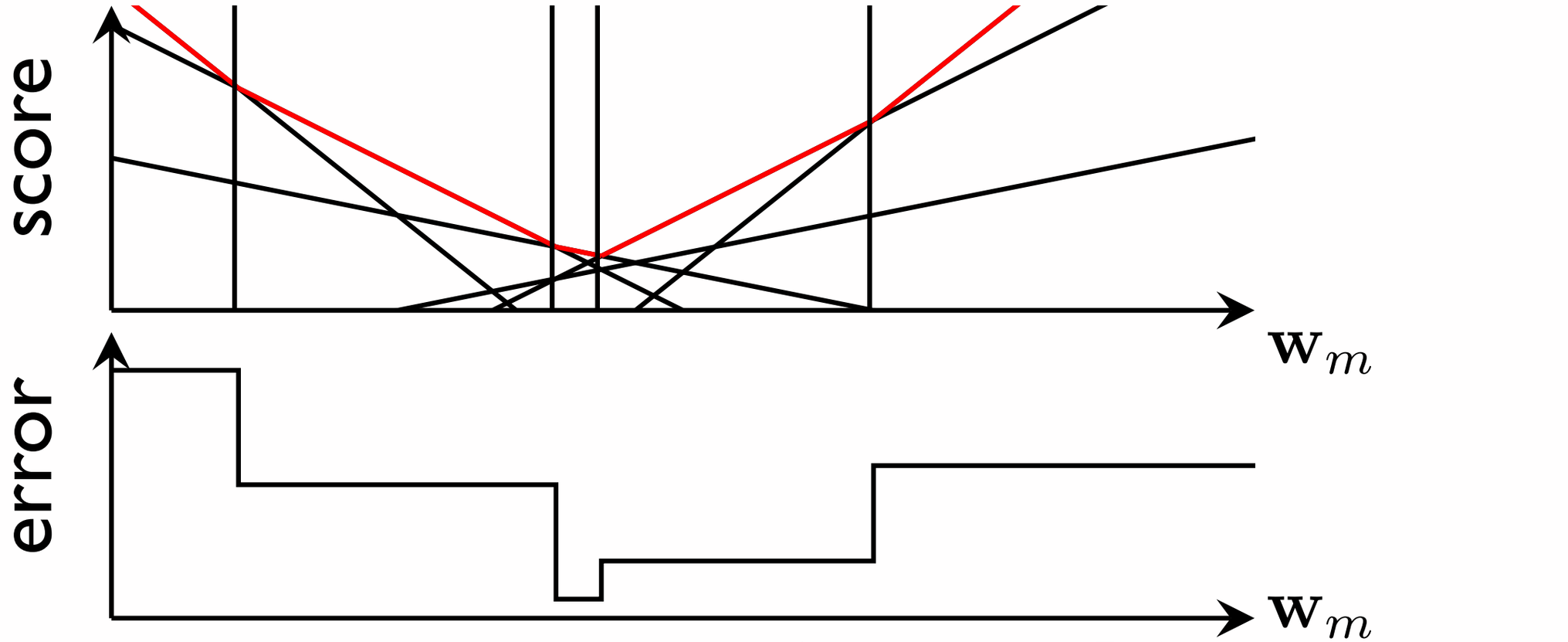
I want to study about machine translation	0.5×-2	0.4×-3	?×-4
I need to master machine translation	0.5×-3	0.4×-4	?×-4
machine translation want to study	0.5×-2	0.4×-5	?×-1
I don't want to learn anything	0.5×-5	0.4×-2	?×-3



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

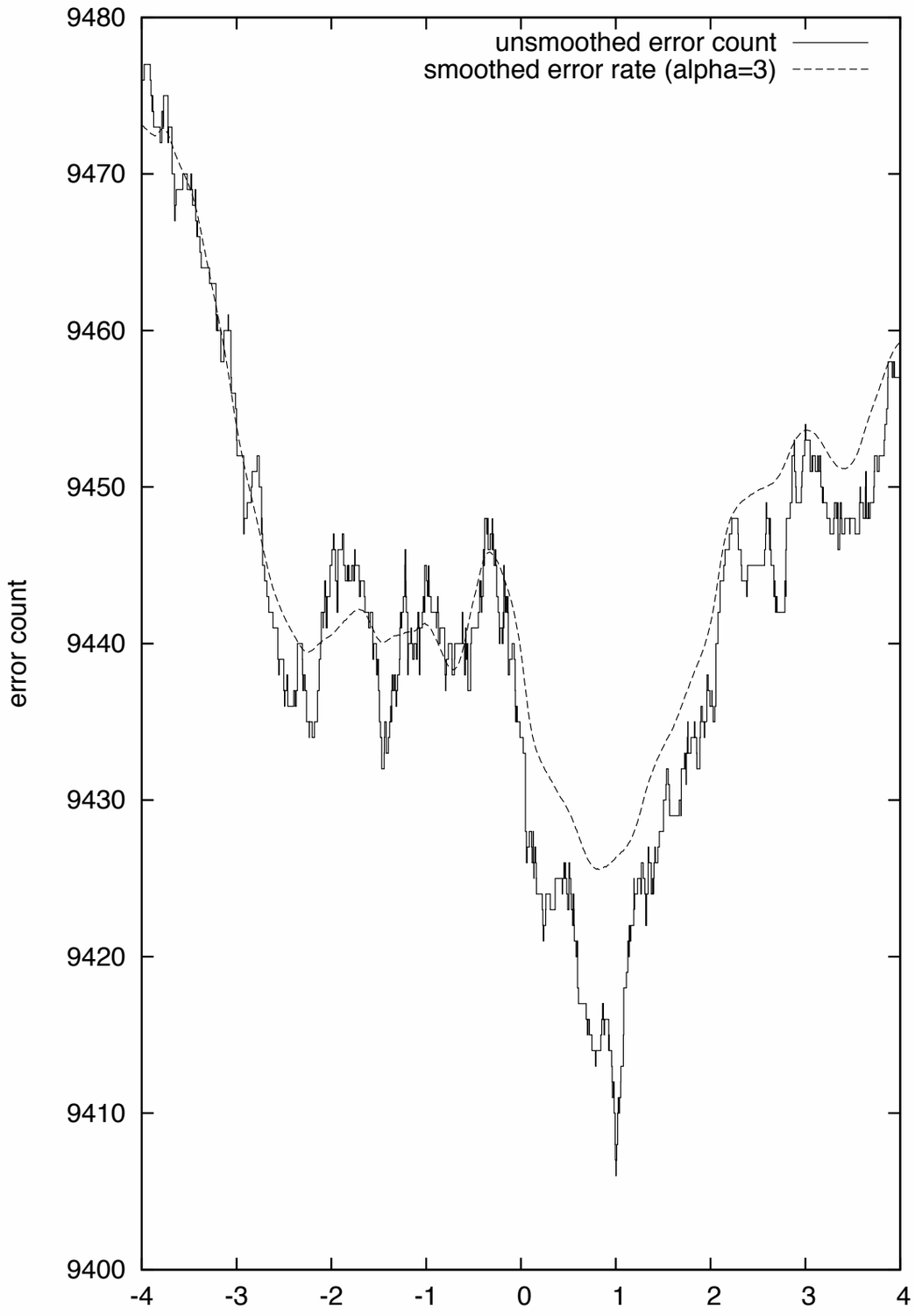
線分探索

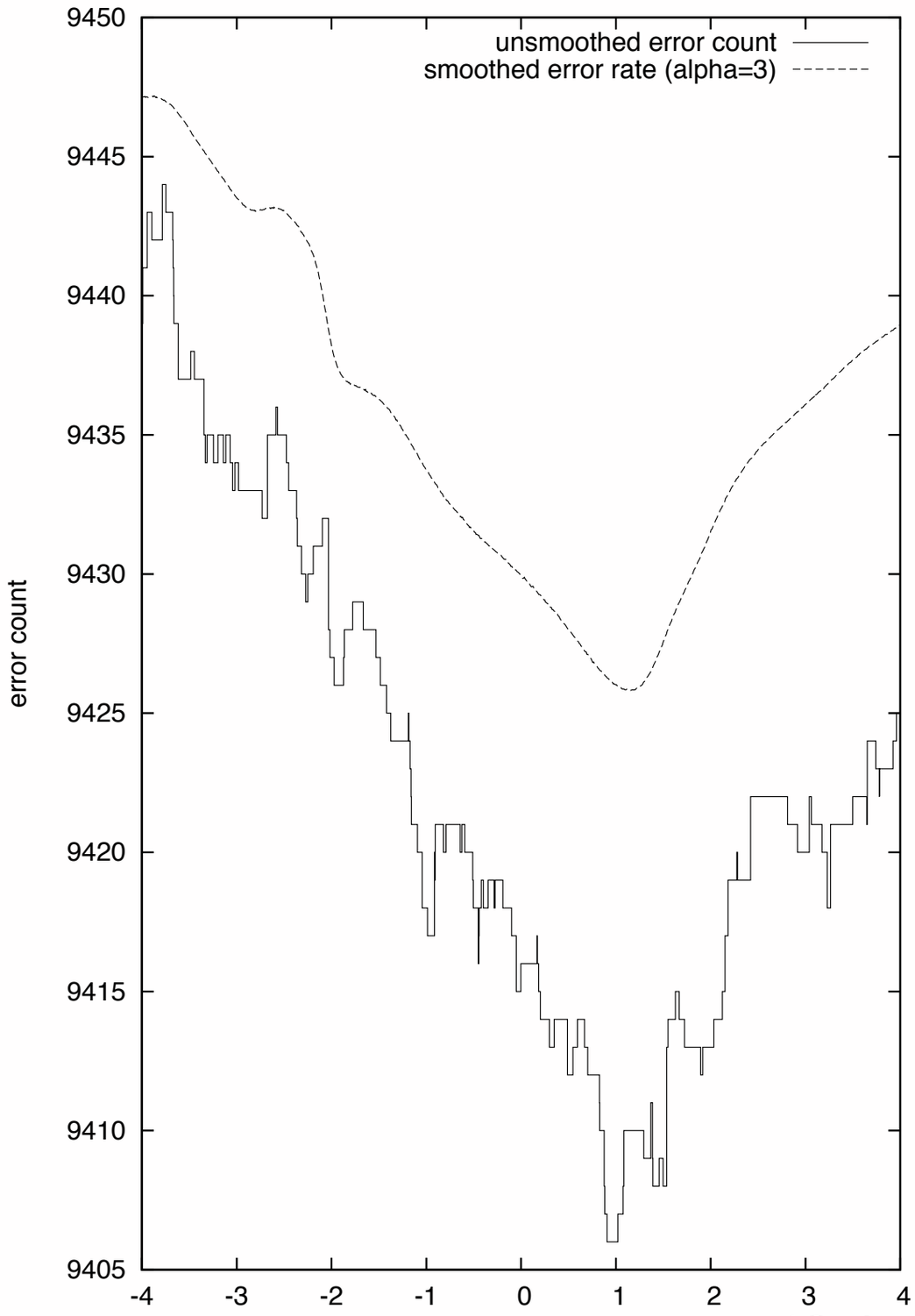
$$\hat{e} = \underset{e}{\operatorname{argmax}}\, \boldsymbol{w}_m^\top \cdot \underbrace{\boldsymbol{h}_m(\boldsymbol{f}^{(s)}, \boldsymbol{d}, \boldsymbol{e})}_{\text{傾き}} + \underbrace{\boldsymbol{w}_{m\_}^\top \cdot \boldsymbol{h}_{m\_}(\boldsymbol{f}^{(s)}, \boldsymbol{d}, \boldsymbol{e})}_{\text{切片}}$$



- 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

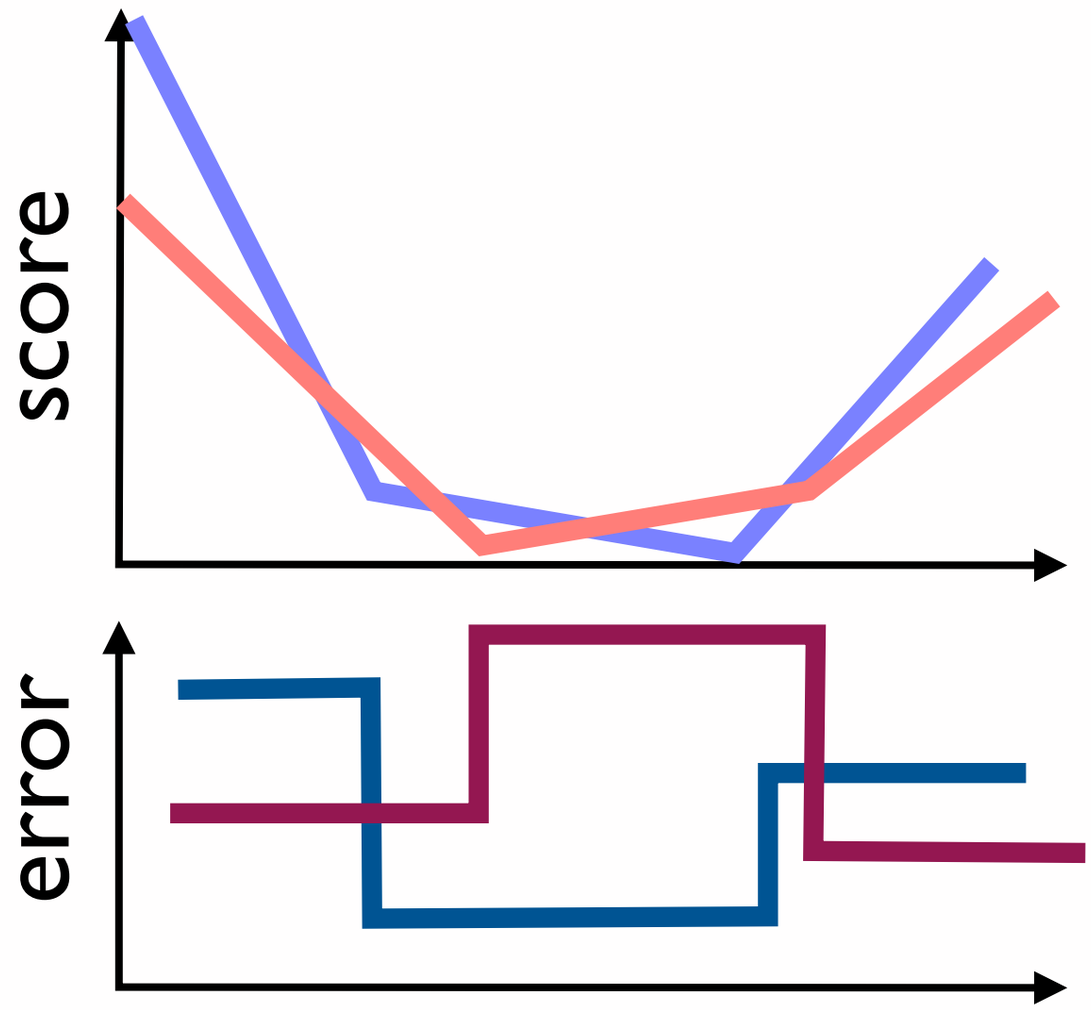
エラー曲線





(Och, 2003)

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

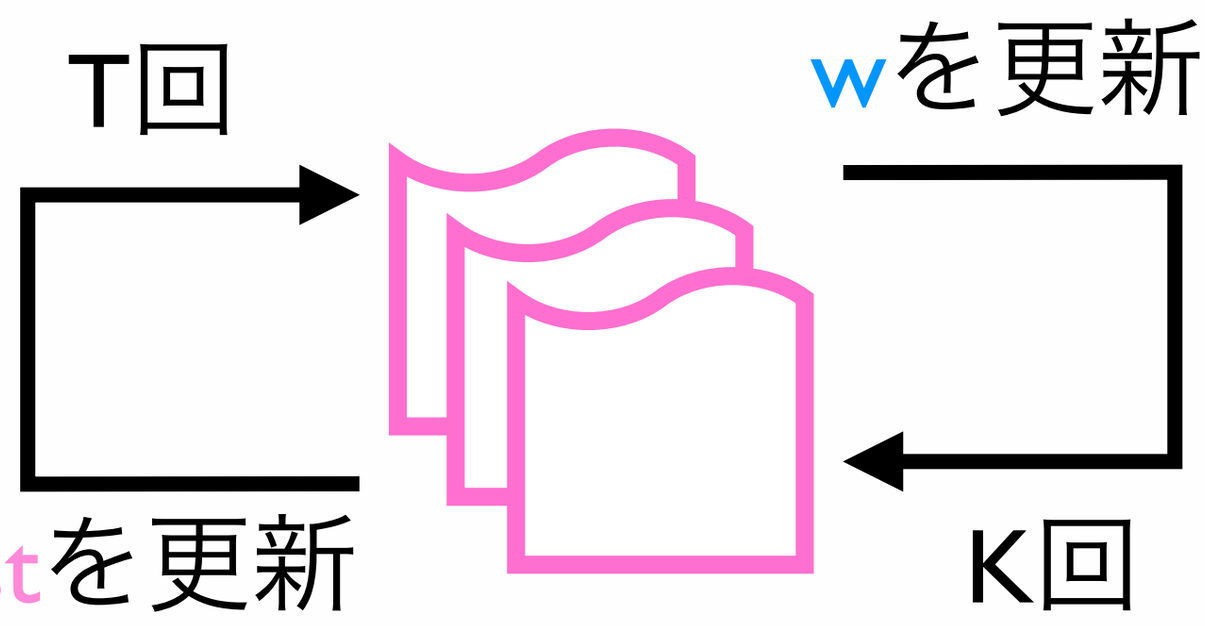
MERT

$$\hat{\boldsymbol{w}} = \arg\min_{\boldsymbol{w}} \ell_{\text{error}}\left(\left\{\arg\max_{\boldsymbol{e}} \boldsymbol{w}^{\top} \cdot \boldsymbol{h}(\boldsymbol{f}^{(s)}, \boldsymbol{e})\right\}_{s=1}^{S}, \left\{\boldsymbol{e}^{(s)}\right\}_{s=1}^{S}\right)$$

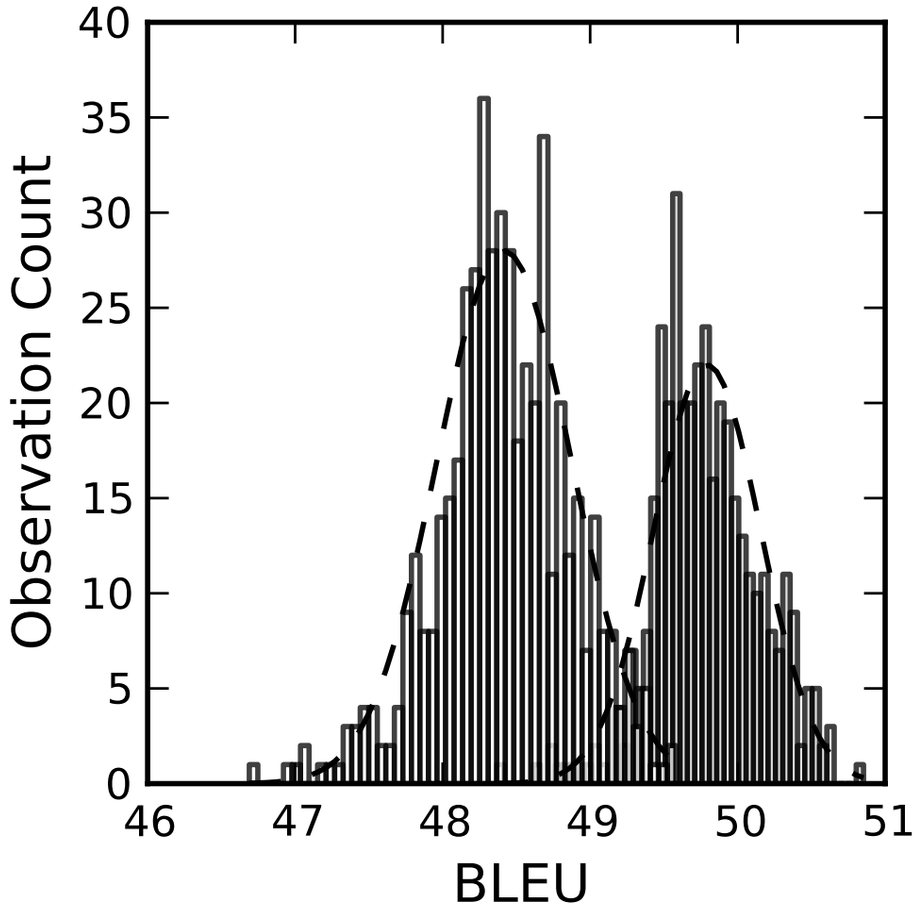
- MERT(Minimum Error Rate Training) (Och, 2003)
  - $\ell_{\text{error}}(.)$に対して、様々なエラー関数を使用可能(例えば、1 - BLEU)
  - wを更新するたびに、argmaxを計算: kbest近似

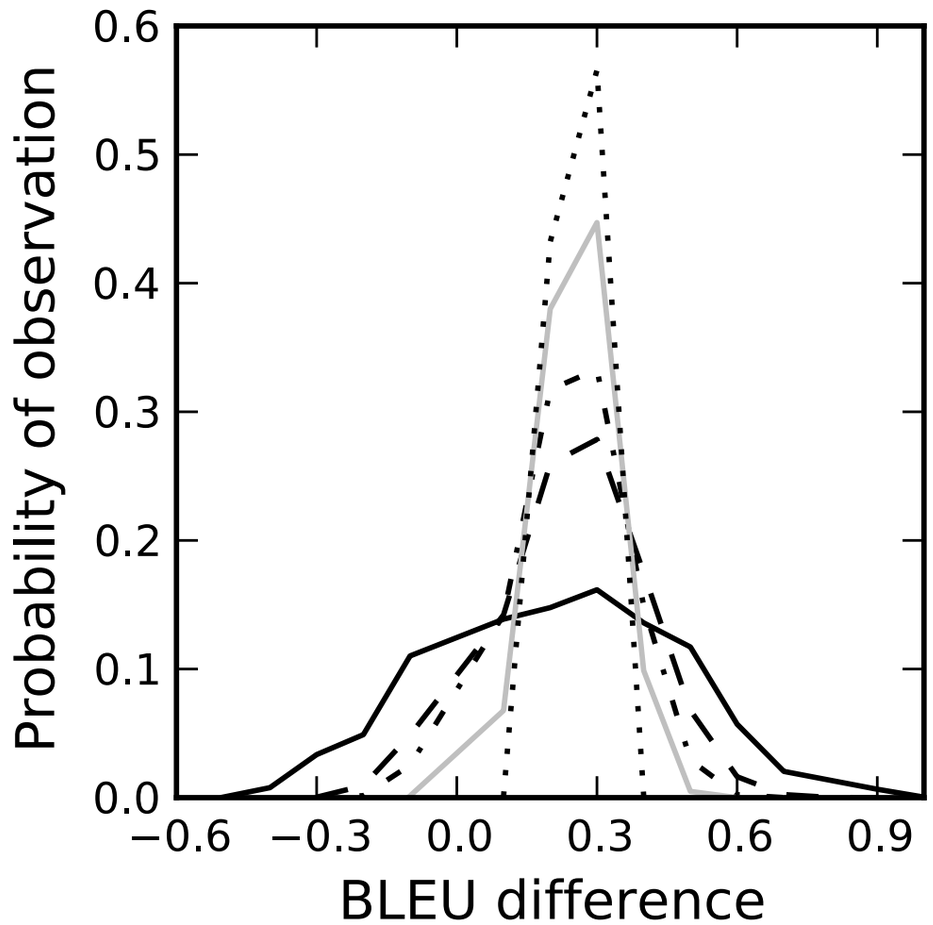
おおきなループ

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



さらに、大きなループ

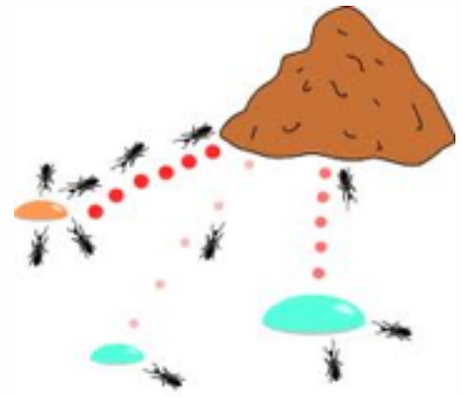




(Clark et al., 2011)

- MERTは、パラメータの初期値に大きく依存
- ランダムな初期値から複数MERT

最適化



$$\hat{\boldsymbol{w}} = \arg\min_{\boldsymbol{w}\in\mathcal{W}} \mathbb{E}_{Pr(F,E)}[\ell(F, E; \boldsymbol{w})]$$

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

- 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** $t = 1...T$ **do**
4: $\boldsymbol{w}$でデコード、kbestリストを生成
5: kbestリストをCへ結合
6: $\boldsymbol{w}^{(t+1)} = \arg\min_{\boldsymbol{w}} \ell(F, E, C; \boldsymbol{w}) + \lambda\Omega(\boldsymbol{w})$
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

$$Pr(e_1 \text{or } e_2 | e_1 \text{or } e_2 \text{or } e_3 \text{or } e_4)$$
$$> Pr(e_3 \text{or } e_4 | e_1 \text{or } e_2 \text{or } e_3 \text{or } e_4)$$

- 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

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

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

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

- オラクル翻訳($\mathbf{o}^{(s)}$)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

$$\text{error}(\boldsymbol{e}_1) < \text{error}(\boldsymbol{e}_3) \iff \boldsymbol{w}^\top \boldsymbol{h}(\boldsymbol{f}, \boldsymbol{e}_1) > \boldsymbol{w}^\top \boldsymbol{h}(\boldsymbol{f}, \boldsymbol{e}_3)$$

- 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

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

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

- 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

$$\mathbb{E}\left[\begin{array}{l}\mathrm{error}(e_1)\textbf{or}\\ \mathrm{error}(e_2)\textbf{or}\\ \mathrm{error}(e_3)\textbf{or}\\ \mathrm{error}(e_4)\end{array}\right]$$

- 翻訳エラーの期待値を最小化

期待値

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

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

(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

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

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.
- 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.
- Mark Hopkins and Jonathan May. 2011. Tuning as ranking. In *Proc. of EMNLP 2011*, pages 1352-1362, Edinburgh Scotland, UK., July.
- 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.

参考文献

- 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 COL ING/ACL 2006 Main Conference Poster Sessions*, pages 787-794, Sydney, Australia, July.