Competence-based Curriculum Learning for Neural Machine Translation

Emmanouil Antonios Platanios, Otalia Stretcu,
Graham Neubig, Barnabas Poczos, Tom M. Mitchell.
In Proc. of NAACL, pp.1162-1172, 2019.

大阪大学 データビリティフロンティア機構
特任助教　梶原 智之
https://sites.google.com/site/moguranosenshi/

概要：Competence-based Curriculum Learning for NMT

- ニューラル機械翻訳は大きなニューラルネットワークを用いる
  - ➢ 訓練が遅い
  - ➢ 多くのヒューリスティクスに頼る
    e.g. 学習率のスケジューリング、大きなバッチサイズ、...
    ↑ **広範囲のハイパラ調整が必要なので望ましくない**

- 本研究では**カリキュラムラーニング**を提案
  - ✓ 訓練時間を大幅に短縮
  - ✓ ハイパラ調整からの解放
  - ✓ 翻訳品質を改善

- サンプルの**難易度**とモデルの**能力**に基づき、
  訓練中にどのサンプルをいつモデルに見せるかを工夫する
- RNNとSANの実験の結果、特にSANに有効で、
  訓練時間を70%短縮した上でBLEUを2.2ポイント改善

課題：ニューラル機械翻訳における訓練の難しさ

- 長い訓練時間
- 学習率やバッチサイズなどの多くのヒューリスティクス

→ **広範囲のハイパラチューニングに膨大なコストがかかる**

解法：Curriculum Learning

簡単な問題から学習を始め、徐々に難しい問題の学習へ進む

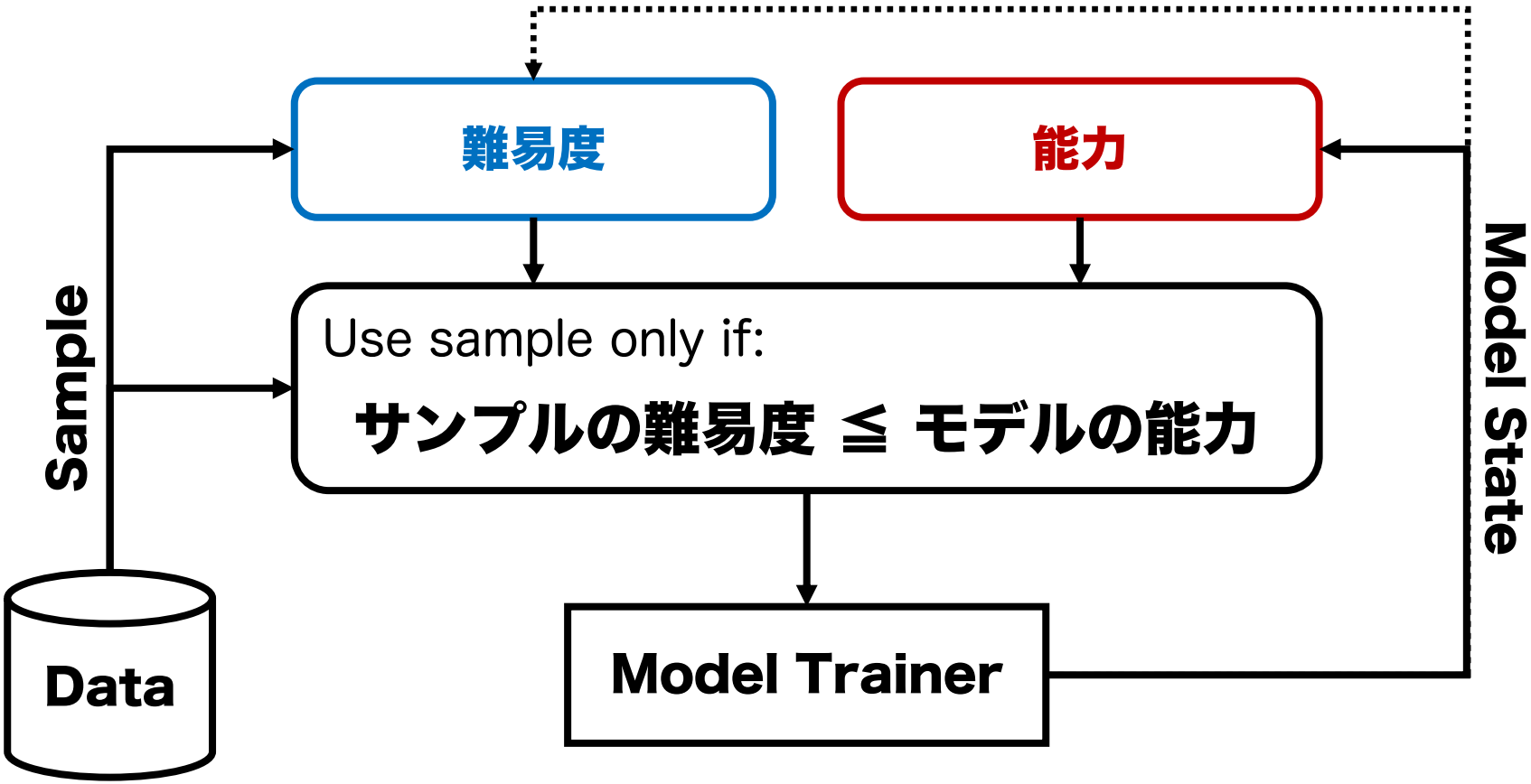
- ✓ 収束が早くなる
- ✓ より良い品質へ到達できる

Easy	Medium	Hard
Thank you.	Thank you very much.	Thank you for your helping me with my work.

Training Time

提案手法：Competence-based Curriculum Learning

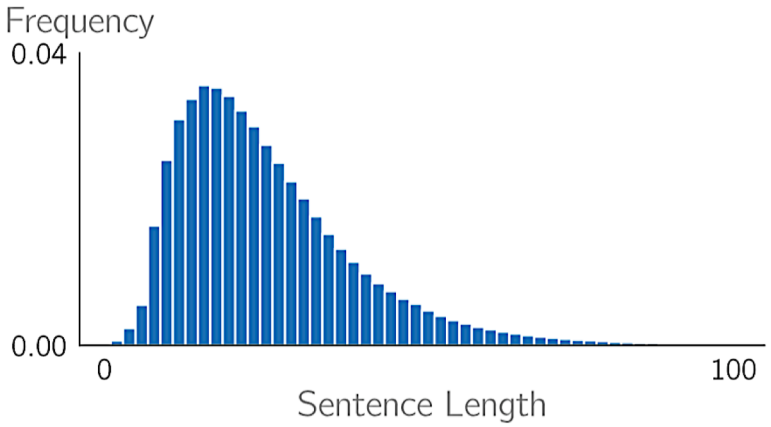
仮説：モデルの現在の能力に適した難易度の文を選んで訓練すると良い



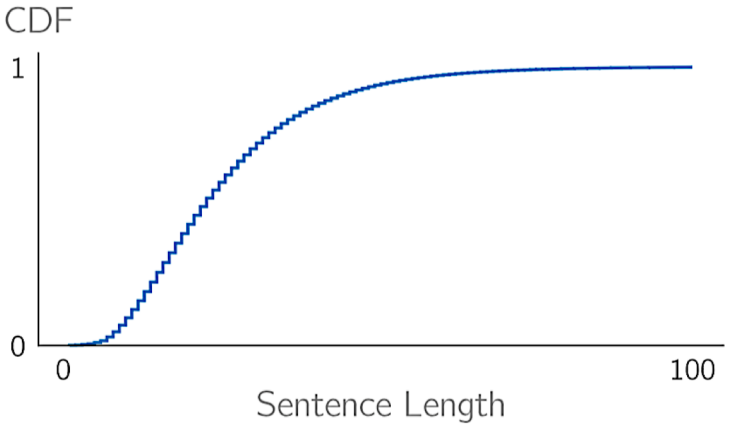
アルゴリズム (1/2)

1. 各サンプル $s_i \in \mathcal{D}$ について難易度 $d(s_i)$ を計算

Sentence	Length
Thank you very much!	4
Barack Obama loves ...	13
My name is ...	6
What did she say ...	123



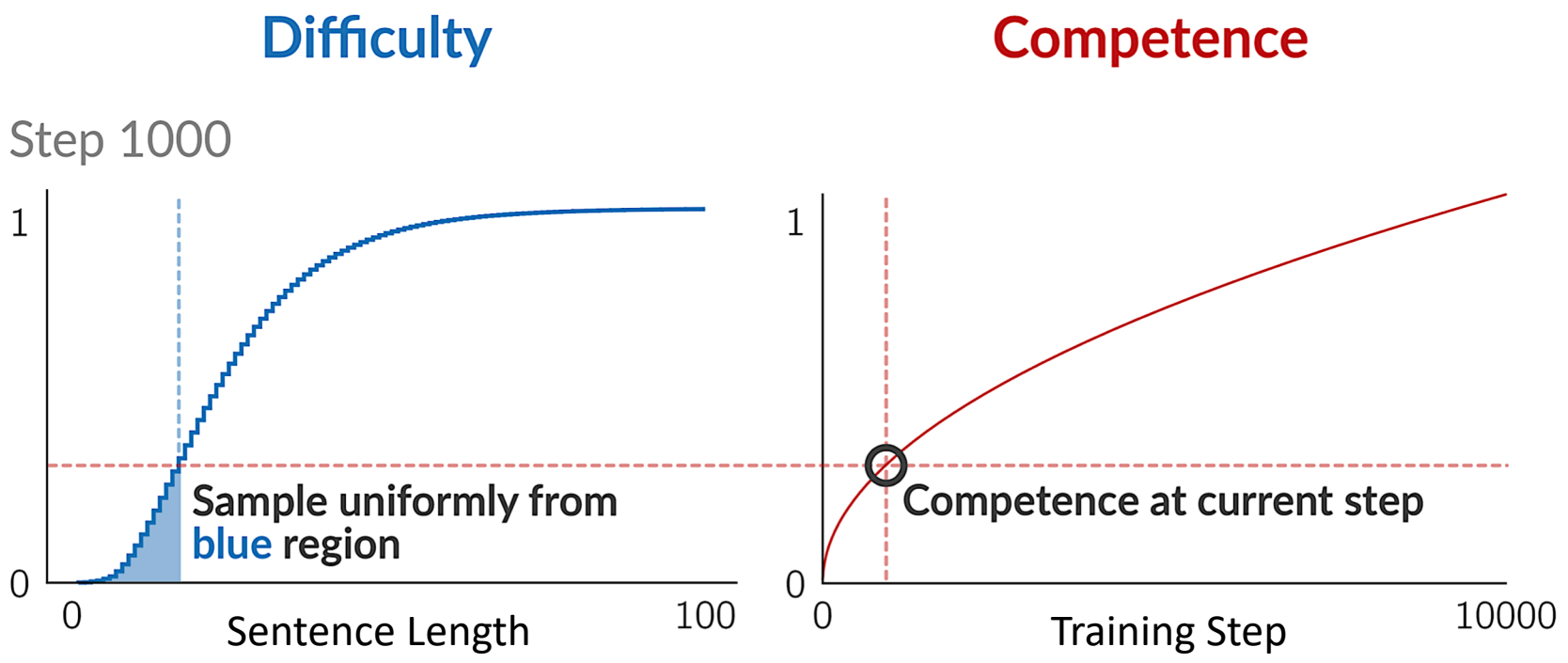
2. 難易度の累積密度関数 $\bar{d}(s_i) \in [0, 1]$ を計算



Sentence	Difficulty
Thank you very much!	0.01
Barack Obama loves ...	0.15
My name is ...	0.03
What did she say ...	0.95

アルゴリズム (2/2)

1. 各サンプル $s_i \in \mathcal{D}$ について難易度 $d(s_i)$ を計算
2. 難易度の累積密度関数 $\bar{d}(s_i) \in [0, 1]$ を計算
3. **For** 訓練ステップ $t = 1$, ... **do**
   i. モデルの能力 $c(t)$ を計算
   ii. $\bar{d}(s_i) \leq c(t)$ を満たす s_i をサンプリングして B_t を構成
   iii. バッチ B_t を用いてモデルを訓練



提案手法の詳細

訓練データを $\mathcal{D} = \{s_i\}_{i=1}^{M}$ 、文を $s_i = \{w_0^i, \ldots, w_{N_i}^i\}$ 、
訓練ステップ数を T 、モデルの能力の初期値を c_0 として、

サンプルの難易度

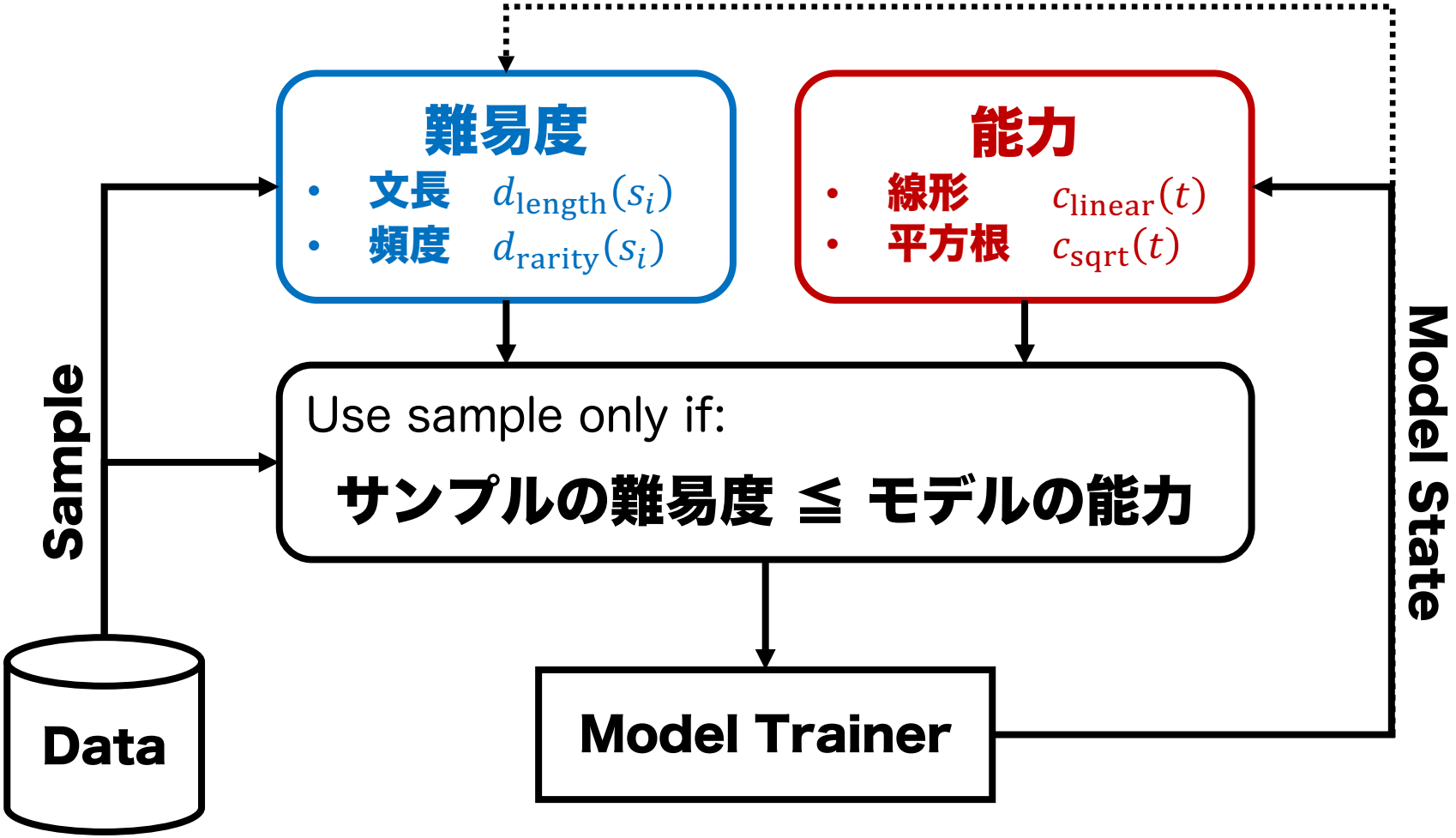
- 文長　$d_{\text{length}}(s_i) \coloneqq N_i$
- 頻度　$d_{\text{rarity}}(s_i) \coloneqq -\sum_{k=1}^{N_i} \log p(w_k^i)$

モデルの能力

- 線形　$c_{\text{linear}}(t) \coloneqq \min\left(1, t\frac{1-c_0}{T} + c_0\right)$
- 平方根　$c_{\text{sqrt}}(t) \coloneqq \min\left(1, \sqrt{t\frac{1-c_0^2}{T} + c_0^2}\right)$

提案手法まとめ：Competence-based Curriculum Learning

仮説：モデルの現在の能力に適した難易度の文を選んで訓練すると良い



実験設定

データセット

	訓練	検証	評価
IWSLT-15: En→Vi	13 万文対	768 文対	1,268 文対
IWSLT-16: Fr→En	22 万文対	1,080 文対	1,133 文対
WMT-16: En→De	450 万文対	3,003 文対	2,999 文対

モデル

- RNN: Bahdanau+ (2015) の2層モデル（WMTでは4層）
- SAN: Vaswani+ (2017) のBASEモデル

ハイパーパラメータ

- Initial Competence: $c_0 = 0.01$
- Curriculum Length: T →→→→→

※ カリキュラムなしで訓練したときのBLEUの90%を達成するのに必要な訓練ステップ数

	RNN	SAN
IWSLT	5,000	20,000
WMT	20,000	50,000

実験結果：En→Viでは最高性能を更新しつつ70%高速化を達成

		RNN					SAN				
		Base	文長 線形	文長 平方根	頻度 線形	頻度 平方根	Base	文長 線形	文長 平方根	頻度 線形	頻度 平方根
BLEU	En→Vi	26.27	26.57	**27.23**	26.72	26.87	28.06	29.14	29.57	29.03	**29.81**
	Fr→En	31.15	31.88	**31.92**	31.39	31.57	34.05	34.98	35.47	35.30	**35.83**
	En→De	26.53	**26.55**	**26.54**	**26.62**	**26.62**	27.95	28.71	29.28	29.93	**30.16**
Time	En→Vi	1.00	0.64	0.61	0.71	**0.57**	1.00	0.44	0.33	0.35	**0.31**
	Fr→En	1.00	**1.00**	0.93	**1.10**	**0.73**	1.00	0.49	0.44	0.42	**0.39**
	En→De	1.00	0.86	0.89	**1.00**	**0.83**	1.00	0.58	**0.55**	**0.55**	**0.55**

- SAN：提案手法が一貫して翻訳品質と訓練時間の両方を改善
  - BLEU：最大2.2ポイントの改善（En→De）
  - Time：最大70%の改善（En→Vi）
- RNN：効果は限定的（RNNはSANよりも訓練が容易なので）
- 難易度：文長か頻度かは場合による
- 能力：線形か平方根かは多くの場合に平方根能力モデルが有効

実験結果：学習率スケジュールのヒューリスティクスが不要

SANで高い性能を得るためには学習率スケジュールの調整が重要

- 例えば、Vaswani+ (2017)では、以下が提案されている

$$\mathrm{lr}(t) \coloneqq d_{\mathrm{embedding}}^{-0.5} \min\left(t^{-0.5}, t \cdot T_{\mathrm{warmup}}^{-1.5}\right)$$

- IWSLT-15 (En→Vi)　28.06 → 29.77 (Time = 1.00)
- IWSLT-16 (Fr→En)　34.05 → 34.88 (Time = 1.00)

提案手法（頻度×平方根）では、この調整から解放される

- IWSLT-15 (En→Vi)　28.06 → 29.81 (Time = 0.31)
- IWSLT-16 (Fr→En)　34.05 → 35.83 (Time = 0.39)

関連研究

- Zhang+ (IJCNLP-2017) Boosting Neural Machine Translation
  - 難しいサンプルだけで訓練する
  - 速くはならないが強くなる

  → **本研究では強くなるし速くもなる**

- Active Learning
  - Haffari+ (NAACL-2009) Active Learning for PBSMT　など
  - 訓練データにおける低頻度なn-gramを補う
  - 言及されてないけど最近だと Fadaee+ (EMNLP-2018) も？

  → **本研究における頻度に基づく難易度指標と関係がある**

- SANの訓練のヒント
  - 学習率スケジューリングは慎重に調整せよ（Shazeer+ 2018）
  - バッチサイズは大きめに設定せよ (Popel+ 2018)

  → **本研究では、学習率スケジューリングの調整なしで、小さなバッチサイズでも、最高性能に到達できる**

まとめ：Competence-based Curriculum Learning for NMT

- 手法：モデルの現在の**能力**に適した**難易度**の文を選んで訓練する
- 利点：
  - ✓ **訓練時間を大幅に短縮**　最大 70 %
  - ✓ **翻訳品質を改善**　最大 2.2 BLEU
  - ✓ **ハイパラ調整からの解放**　学習率スケジューリングなど
- 結果：性能がより訓練方法に敏感なSANモデルに対して特に有効
- 実装：https://github.com/eaplatanios/symphony-mt