英日機械翻訳のための対訳コーパスフィルタリングの検討

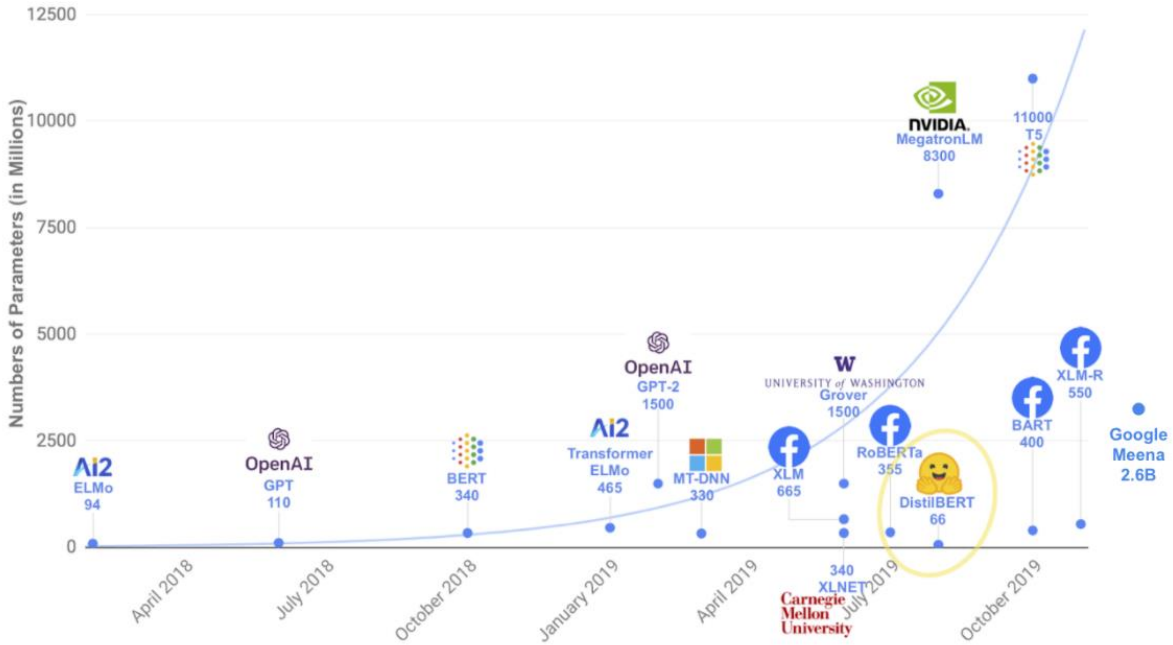
愛媛大学工学部工学科（学部3年）　本田志遠　正木亮太朗

愛媛大学大学院理工学研究科　梶原智之

背景

◆自然言語処理の扱うデータ量は指数関数的に増加している

◆機械翻訳でも訓練データを増やすことで、翻訳品質を向上させている



https://ja.stateofaiguides.com/20200914-future-of-nlp/

課題

大規模データに基づく機械翻訳の課題

- ◆翻訳器の訓練に時間がかかる
- ◆大量の計算資源を運用するコストがかかる
- **◆大規模な対訳データは自動収集される場合が多く、ノイズが含まれる**

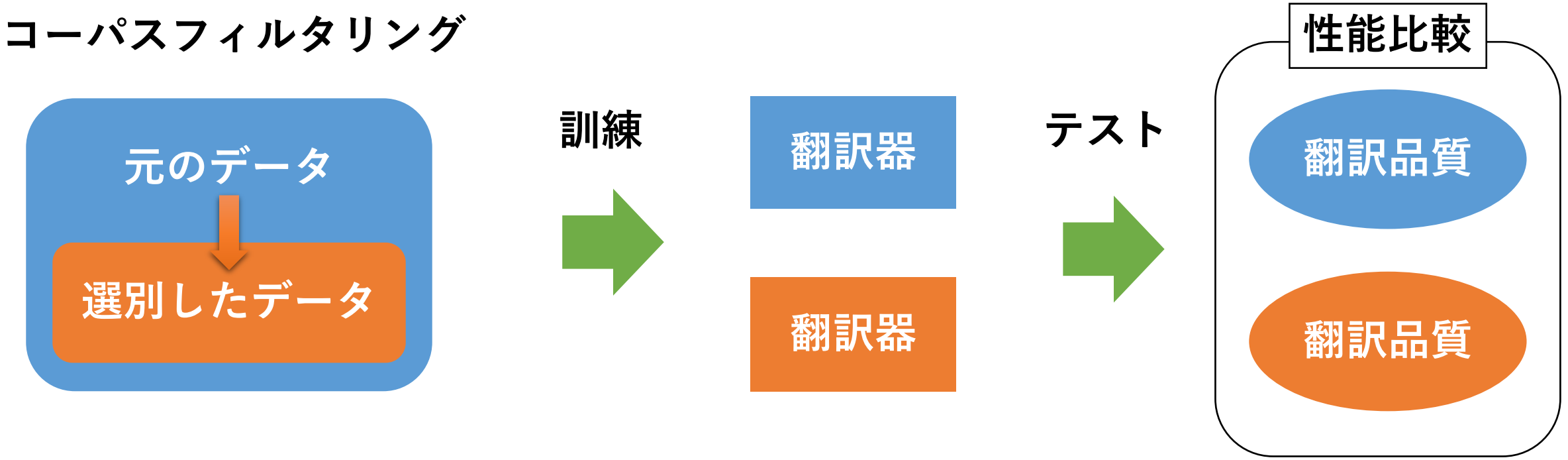
小規模な対訳データから高品質な翻訳器を訓練する

対訳データに含まれるノイズ例

日英対訳データで最も大きいJParaCrawlでの例

ノイズのタイプ	英語文	日本語文
A：非英語文・非日本語文	Ding Ye On, Lee Bong In	丁用根、李鳳仁
B：短すぎる文・長すぎる文	RA: Guy J	RA: Guy J ニュース
C：意味的に対応しない文対	You will always need to have the back up 7 computer I was using XP.	私はPhotoshop 7の2つのレイヤーを持っています。

本研究の流れ



目標：半分の規模の訓練データから高品質な翻訳器を得る

提案手法

提案手法

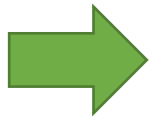
ノイズのタイプ	解決策
A：非英語文・非日本語文	1. 言語判定ツールで判断 2. 文字種の割合で判断
B：短すぎる文・長すぎる文	1. 文字数で判断 2. 単語数で判断
C：意味的に対応しない文対	1. 多言語文符号化器（mUSE）で判断 2. 多言語文符号化器（LaBSE）で判断

対案手法A－1：言語判定ツールで判断

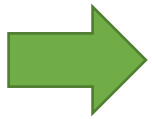
言語判定ツールであるlangdetectを用いて、

SRCが英語ではない or TGTが日本語ではない文対をノイズとする

今後も社会貢献できる開発に取り組みたいと思っています。



langdetect



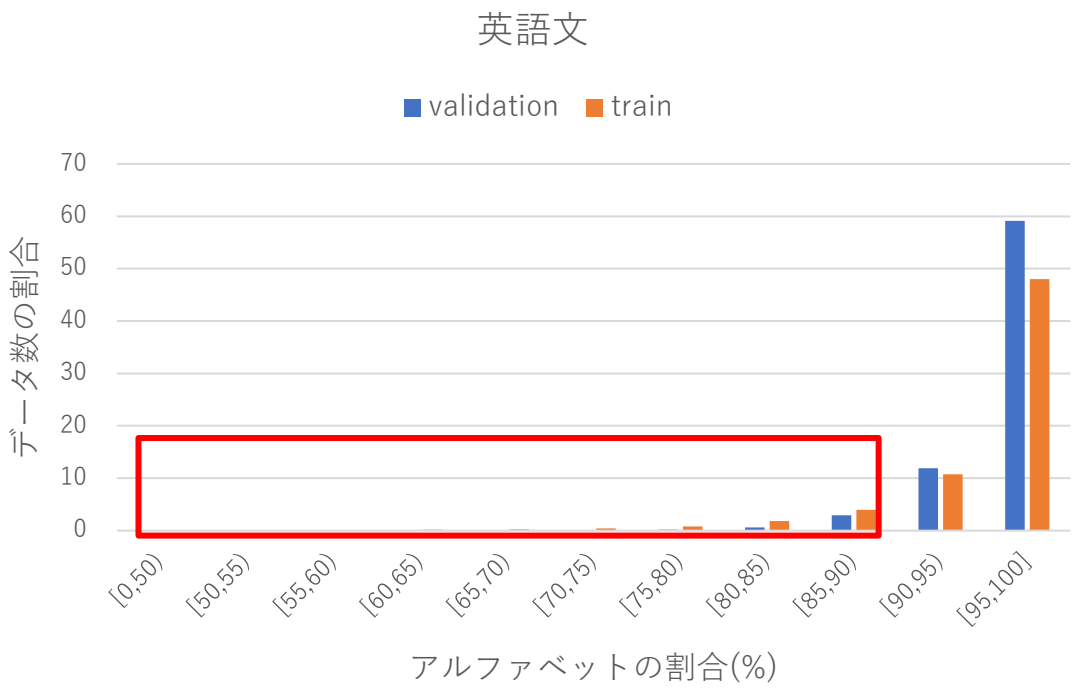
“ja”

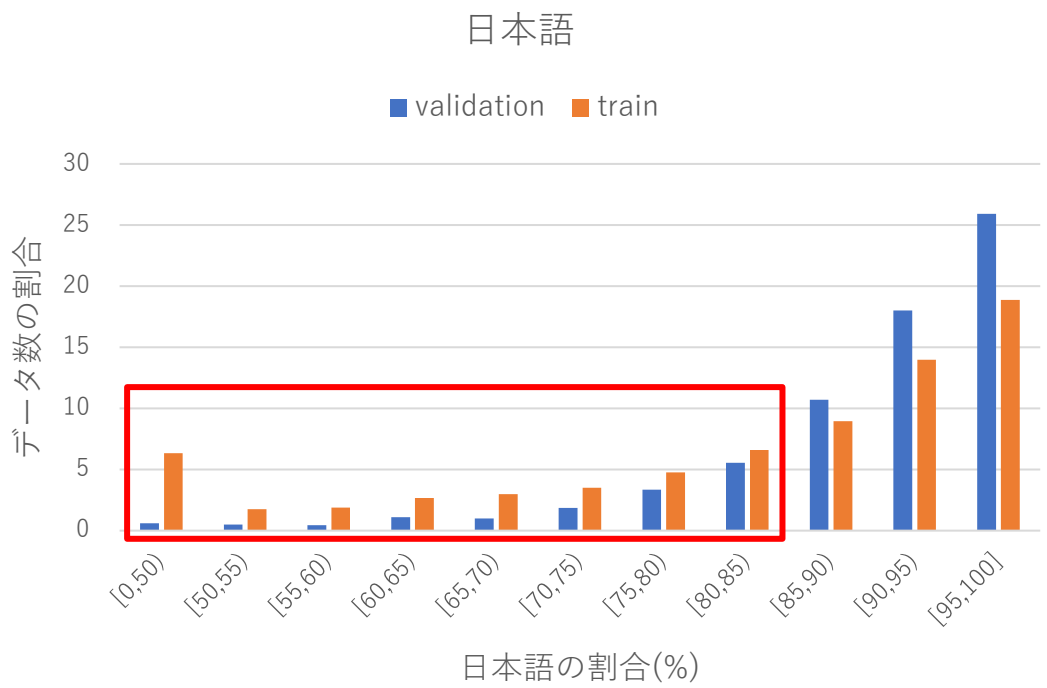
データ	正解率
手動構築した検証用データ	98.5%
自動収集した訓練用データ	95.7%

https://github.com/Mimino666/langdetect

提案手法A－2：文字種の割合で判断

英語文に含まれる**アルファベットの割合**、
日本語文に含まれる**日本語（ひらがな・カタカナ・漢字）の割合**から
検証用データよりも訓練用データの割合が高くなる部分をノイズとして判断





提案手法B：文字数または単語数で判断

閾値を超えて**文字数・単語数が多いまたは少ない文対を訓練用データから徐外**

文字数

今後も社会貢献できる開発に取り組みたいと思っています。⇒ 27個

SentencePieceによる単語分割

今後も 社会貢献 できる 開発 に取り組み たいと思っています 。⇒ 7個

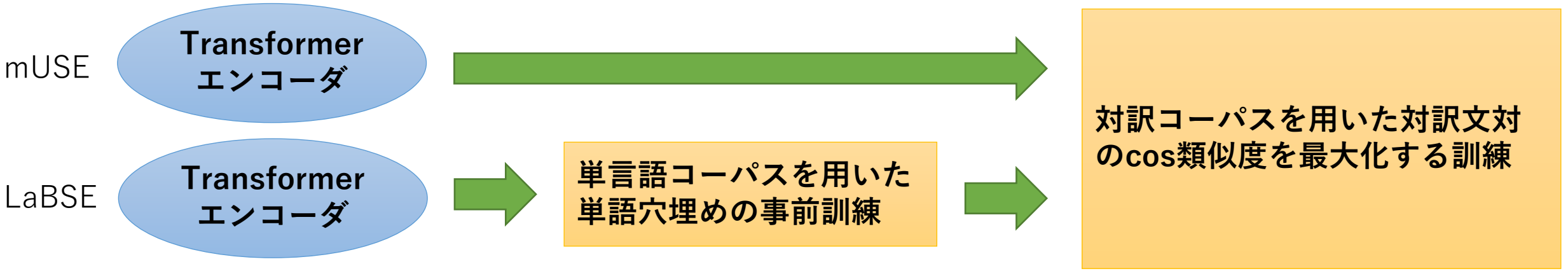
提案手法C：多言語文符号化器で判断（1/2）

◆mUSE

- Transformerに基づく多言語符号化器（日英を含む16言語）
- 対訳文対のcos類似度が高くなるように文ベクトルを学習

◆LaBSE

- Transformerに基づく多言語符号化器（日英を含む109言語）
- 単語穴埋め事前訓練 → 対訳文対のcos類似度を最大化する再訓練



提案手法C：多言語文符号化器で判断（2/2）

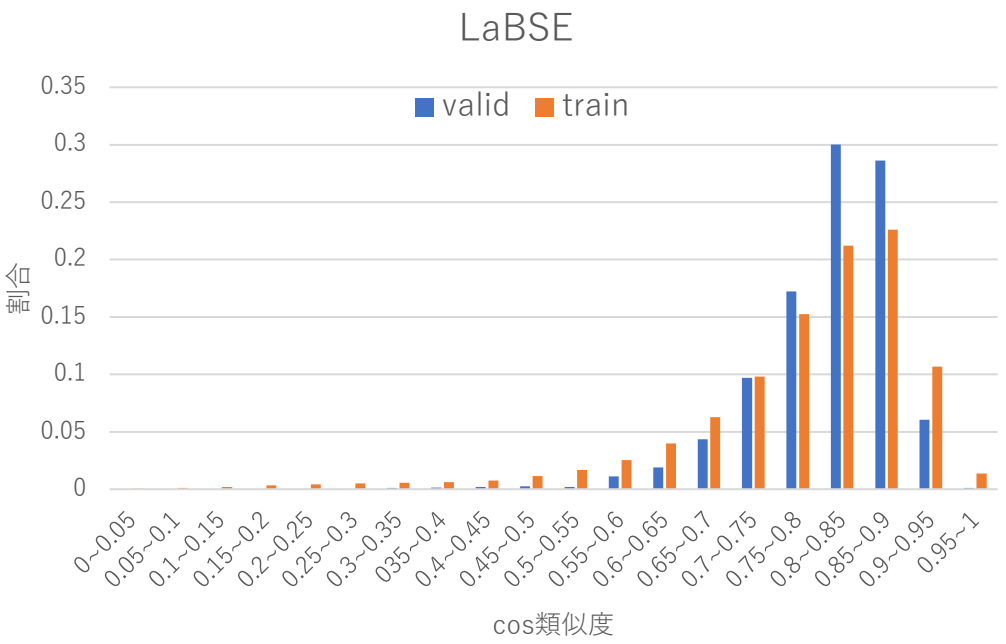
◆**文ベクトル間のcos類似度**からノイズを判断

◆ノイズを削除した上で、cos類似度が高いものを優先的に訓練に用いる

◆cos類似度が高すぎる文対は定型表現が多い

⇒cos類似度が高すぎる文対もノイズとして判断

cos類似度	英語文	日本語文
0.97	Device Management System Services -	デバイス管理システムサービス -
0.96	Selecting a MIDI Remote Control Device	MIDI リモートコントロールデバイスの選択
0.96	Automation System Safety Guide: PDF file	自動化システム安全ガイド：PDF ファイル



提案手法の詳細

手法	詳細
手法A-1（langdetect）	英語文で英語と判定され、かつ日本語文で日本語と判定された文対のみ用いる
手法A-2（文字種）	英語文に含まれるアルファベットの割合が90％以上の文かつ 日本語文に含まれる日本語の割合が80％以上の文のみ用いる
手法B-1（文字数）	文字数が英語文では30以上200未満の文かつ日本語では15以上100未満の文対のみ用いる
手法B-2（単語数）	分割数が英語文・日本語文ともに10以上55未満の文のみを用いる
手法C-1（mUSE）	cos類似度が0.4以上0.7未満の文対を用いる
手法C-2（LaBSE）	cos類似度が0.7以上0.9未満の文対を用いる

評価実験

実験設定

データ

◆訓練用データ
- JParaCrawl [Morishita+ 2020]
- コーパス：Web
- サイズ：1億文対

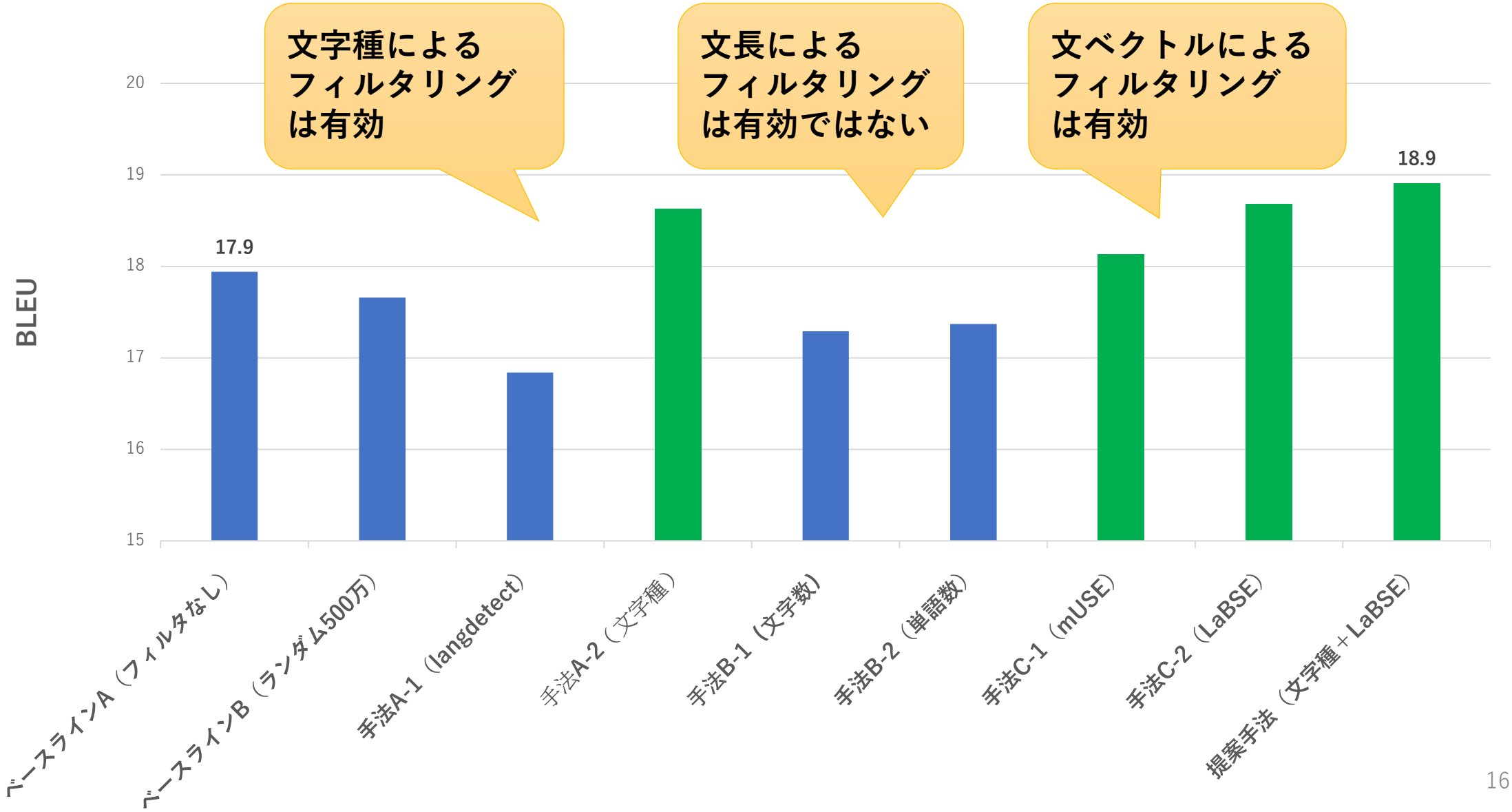
◆評価用データ
- WMT20[Barrault+ 2020]
- コーパス：News
- サイズ
  validation：1,998
  test：1,000

モデル

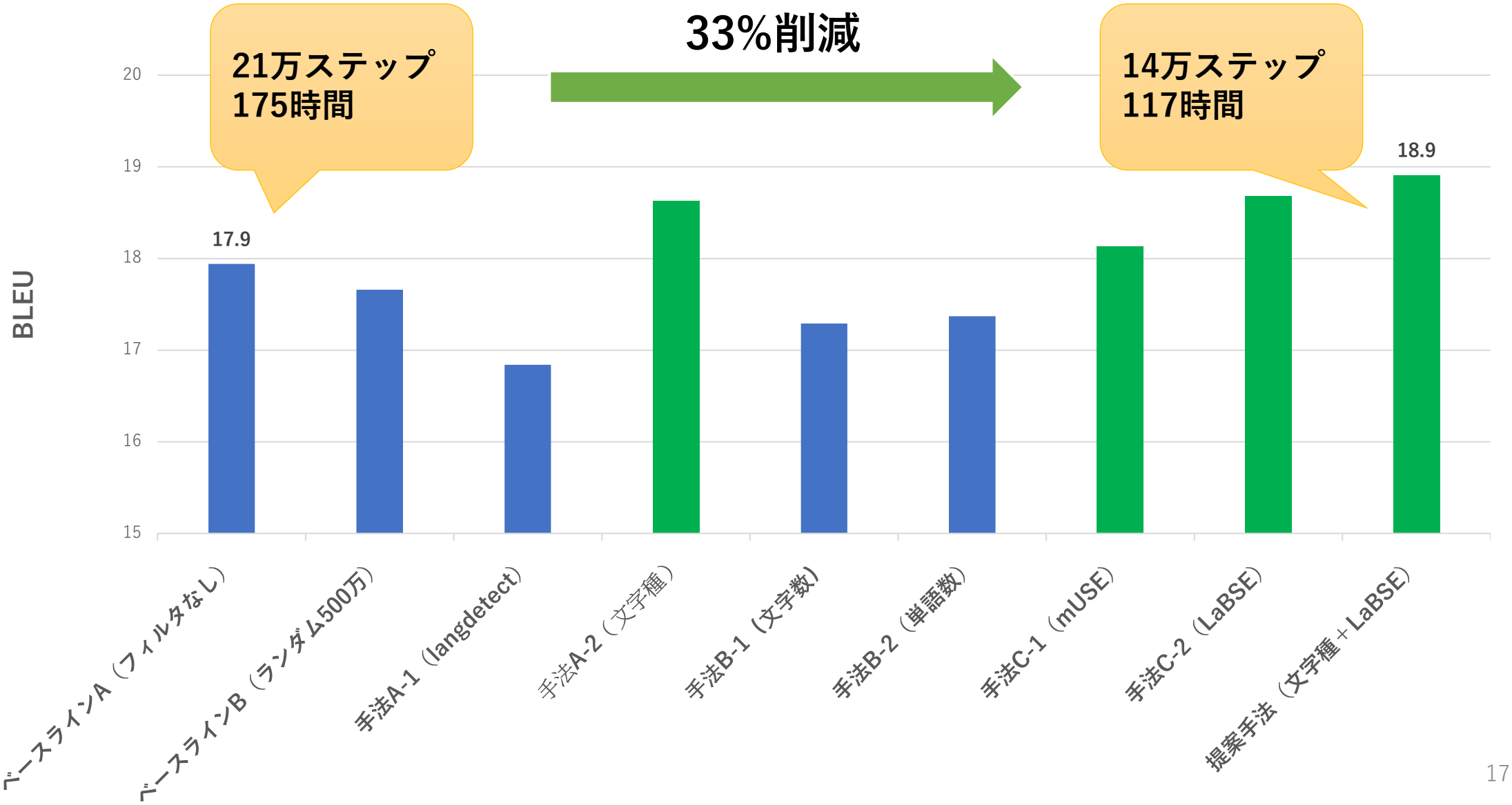
◆モデル：Transformer
◆Layerの数： 6
◆Headの数： 8
◆次元数：512次元
◆バッチサイズ：4096
◆訓練データの規模
ベースライン：1000万文対
提案手法：ノイズを除外した中から無作為抽出した500万文対
◆翻訳の方向：英語→日本語

[Morishita+ 2020] Morishita et al. (LREC-2020) JParaCrawl: A Large Scale Web-Based English-Japanese Parallel Corpus
[Barrault+ 2020] Barrault et al. (WMT-2020) Findings of the 2020 Conference on Machine Translation (WMT20)

実験結果：翻訳品質


文字種による
フィルタリング
は有効
文長による
フィルタリング
は有効ではない
文ベクトルによる
フィルタリング
は有効
BLEU
20
19
18
17
16
15
17.9
18.9
ベースラインA（フィルタなし）
ベースラインB（ランダム500万）
手法A-1（langdetect）
手法A-2（文字種）
手法B-1（文字数）
手法B-2（単語数）
手法C-1（mUSE）
手法C-2（LaBSE）
提案手法（文字種＋LaBSE）

実験結果：訓練時間


21万ステップ
175時間
33%削減
14万ステップ
117時間
BLEU
20
19
18
17
16
15
17.9
18.9
ベースラインA（フィルタなし）
ベースラインB（ランダム500万）
手法A-1（langdetect）
手法A-2（文字種）
手法B-1（文字数）
手法B-2（単語数）
手法C-1（mUSE）
手法C-2（LaBSE）
提案手法（文字種＋LaBSE）

まとめ

◆**背景**：自動収集された大規模データにはノイズが含まれ、訓練効率が悪い

◆**提案**：英日機械翻訳のためのパラレルコーパスフィルタリングの手法を提案

- 文字種に基づく手法　→　実装によっては有効
- 文長に基づく手法　→　有向ではない
- 文ベクトルに基づく手法　→　有効

◆**結果**：英日機械翻訳の訓練データを1000万→500万に削減する設定

- 翻訳品質：BLEUスコアを1ポイント改善
- 訓練時間：33%削減

◆**今後**：日英方向での検証や、他のドメインの評価用データを用いた検証