

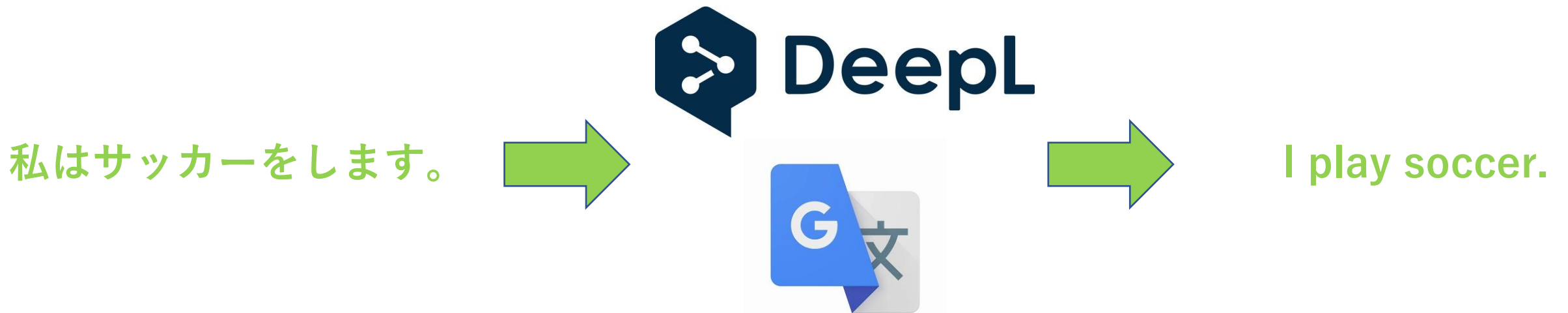
少量データからの 効率的な翻訳器の訓練

メンバー： 工学部工学科（学部4年） 正木亮太郎 本田志遠

指導教員： 理工学研究科（助教） 梶原智之

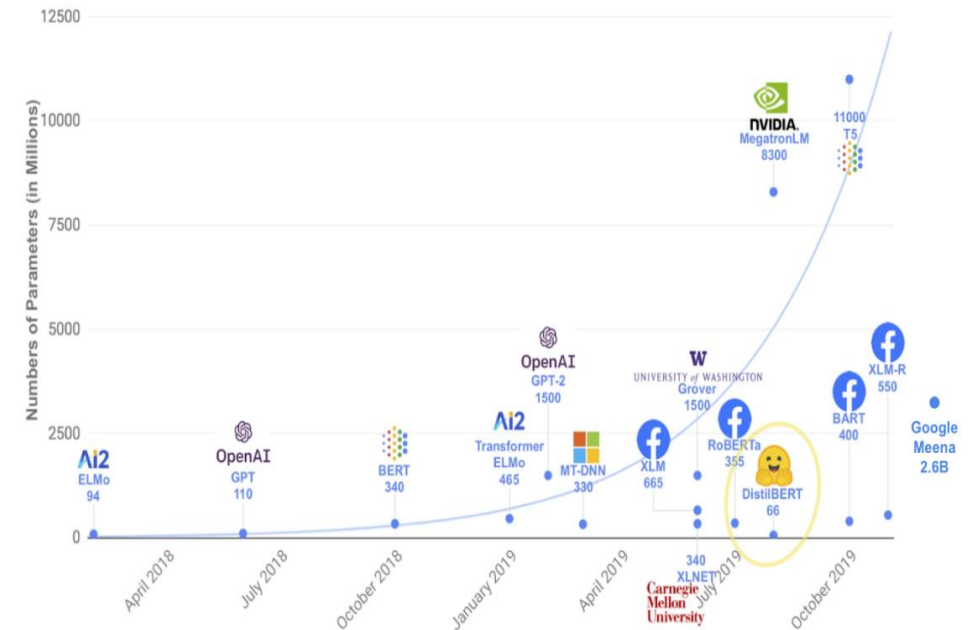
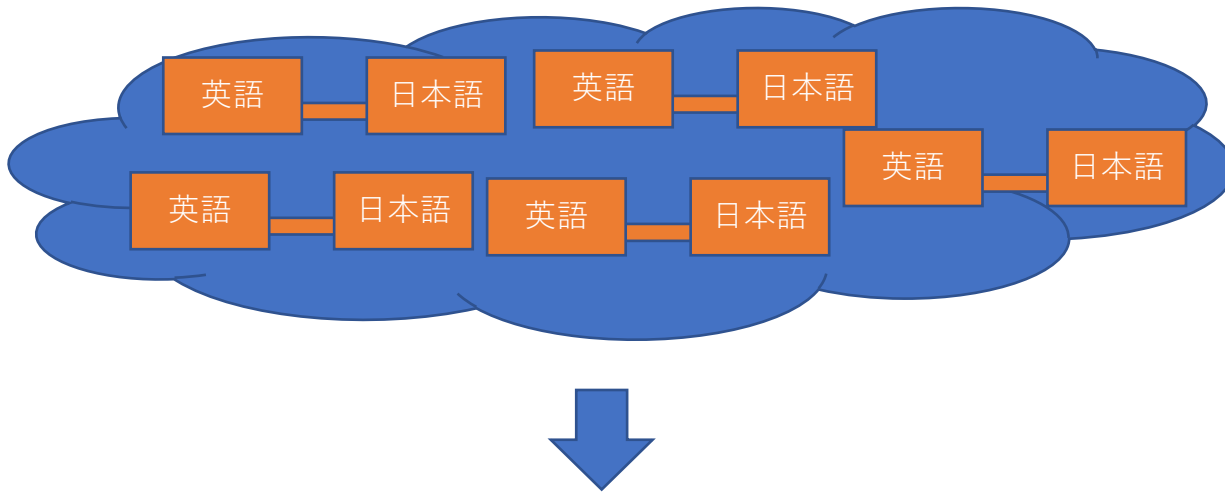
機械翻訳とは

- 自然言語処理の応用事例の一つ
- ある言語の文を別の言語に自動的に翻訳
- 人間が翻訳した文に近づけることを目標とする
- 機械翻訳が用いられているものとしてDeepL、Google翻訳がある



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/>

課題と目的

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

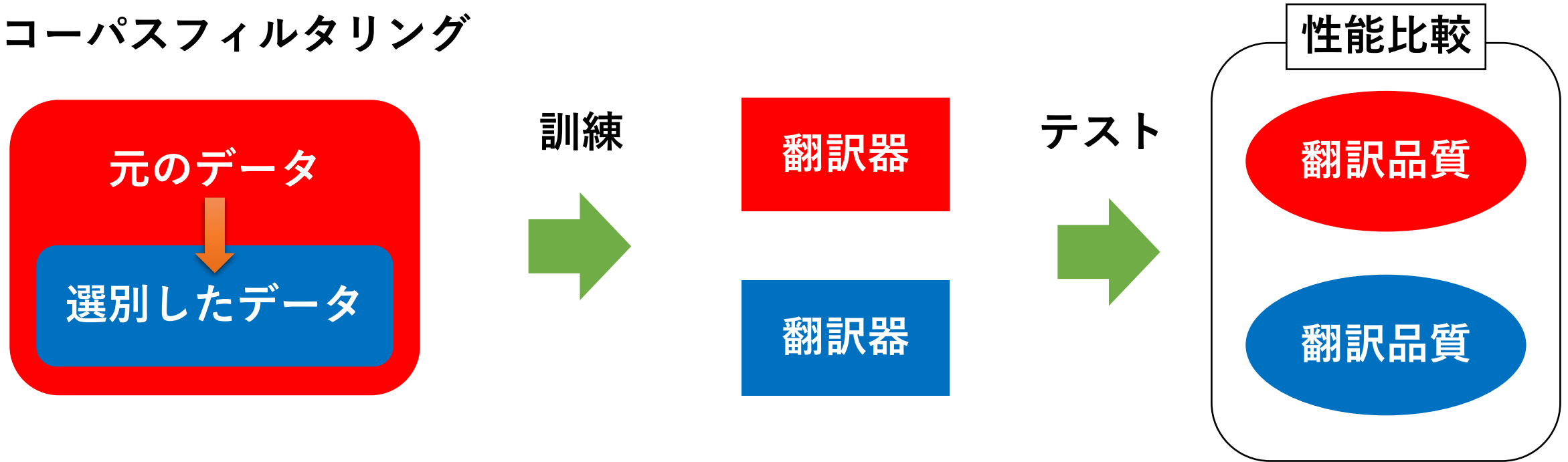
対訳データに含まれるノイズの例

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

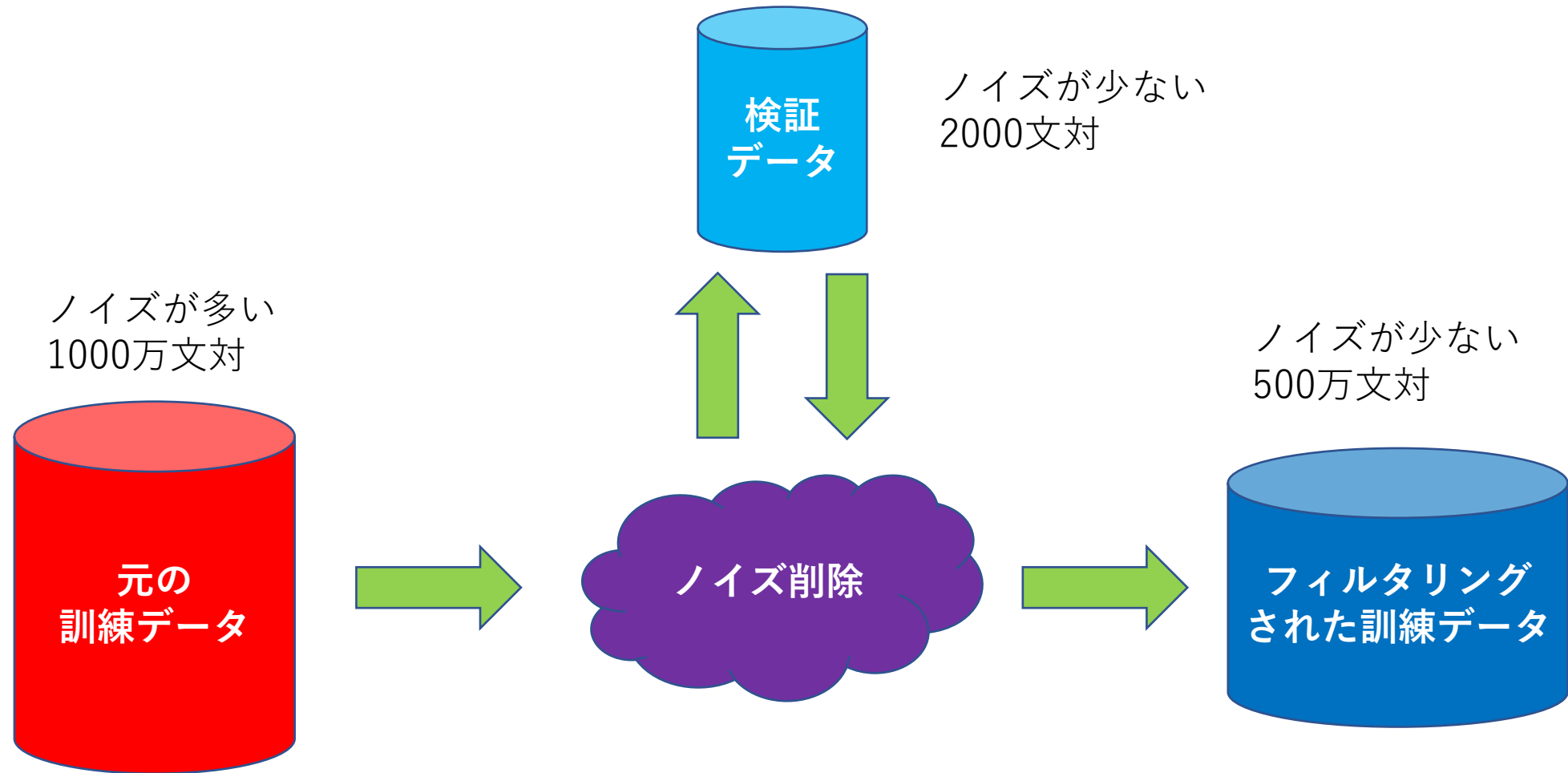
コーパスフィルタリング



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

コーパスフィルタリングの全体像



提案手法A-1：言語判定ツールで判断

ノイズA(非英語・非日本語)に対する方法

言語判定ツールであるlangdetectを用いて、

入力文が英語ではない or 出力文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です。



langdetect



“ja”

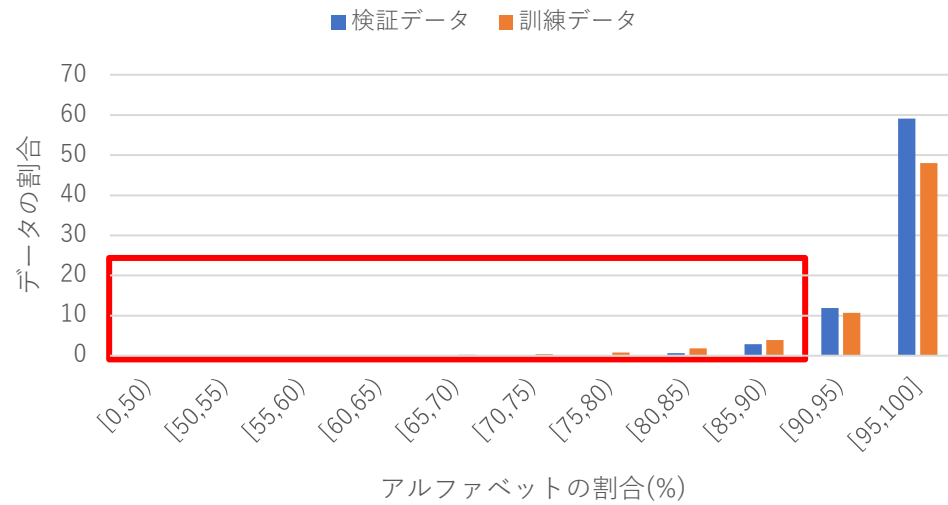
データ	正解文対数
手動構築した検証データ	1968/1998
自動収集した訓練データ	9,687,143/10,000,213

提案手法A-2：文字種の割合で判断

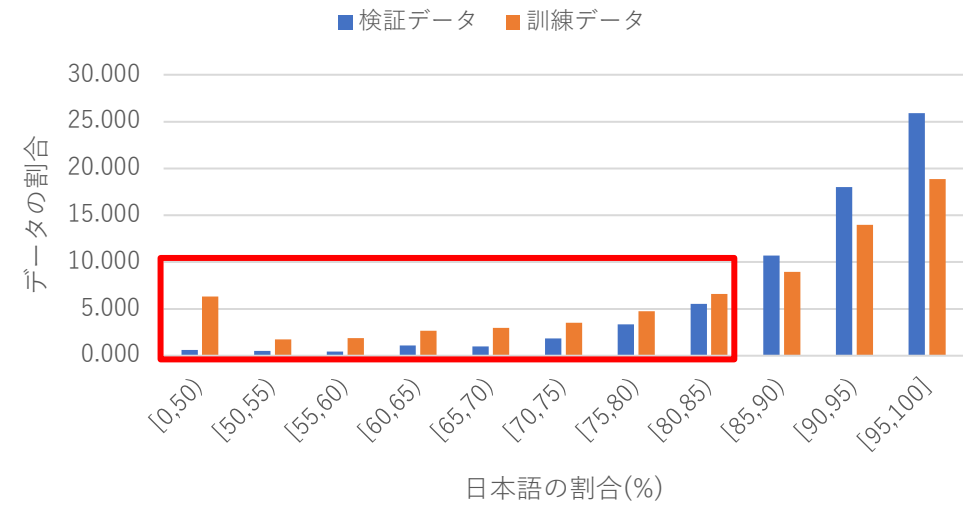
ノイズA(非英語・非日本語)に対する方法

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
検証データよりも訓練データの割合が高くなる部分をノイズとして判断

英語文



日本語文



提案手法B：文字数または単語数で判断

ノイズB(長すぎる文・短すぎる文)に対する方法

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除外
文字数

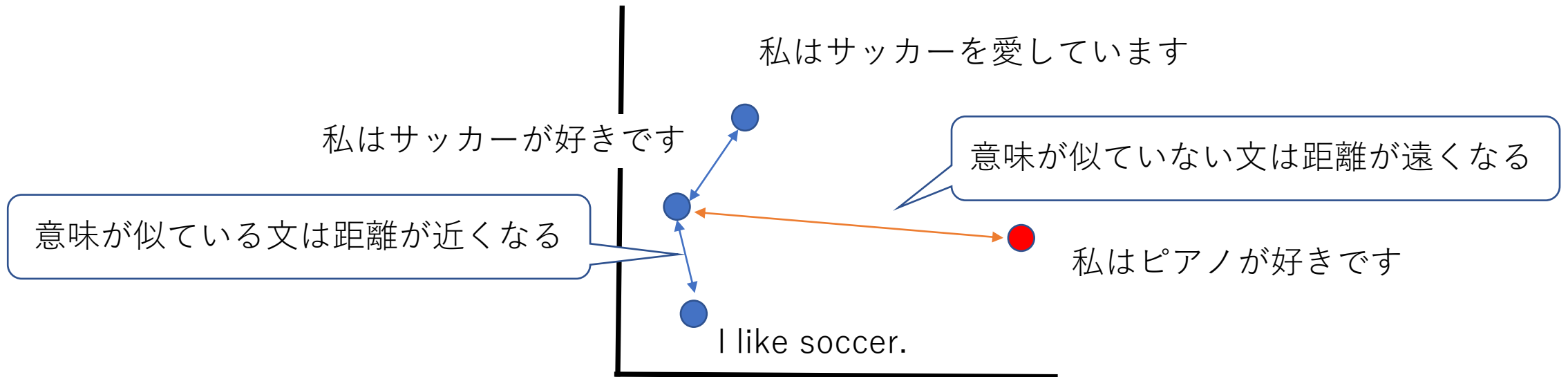
今後も社会貢献できる開発に取り組みたいと思っています。⇒ 27個

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：文ベクトルで判断

- ノイズC(意味的に対応しない文対)に対する方法
- 文符号化器を用いて文をベクトル空間上の点として表現
- 意味が近い文同士はベクトルが近く、似ていない文同士は遠くなる



提案手法のまとめ

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

評価実験

実験設定

◆データ

- ◆日本語と英語の文対
- ◆1000万文対→500万文対

◆モデル

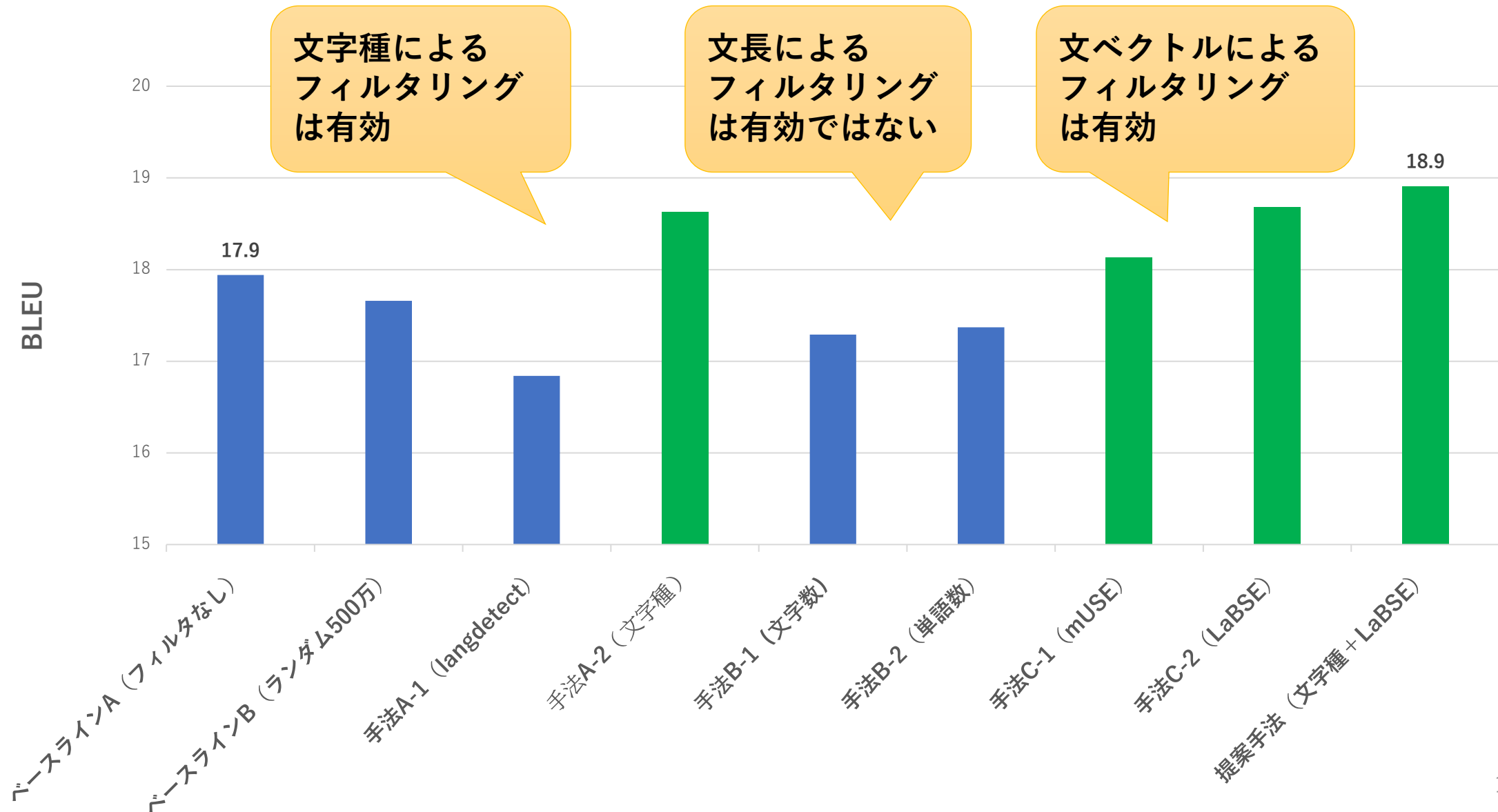
- ◆Transformer翻訳器
- ◆google翻訳などと同じディープラーニング

◆評価方法

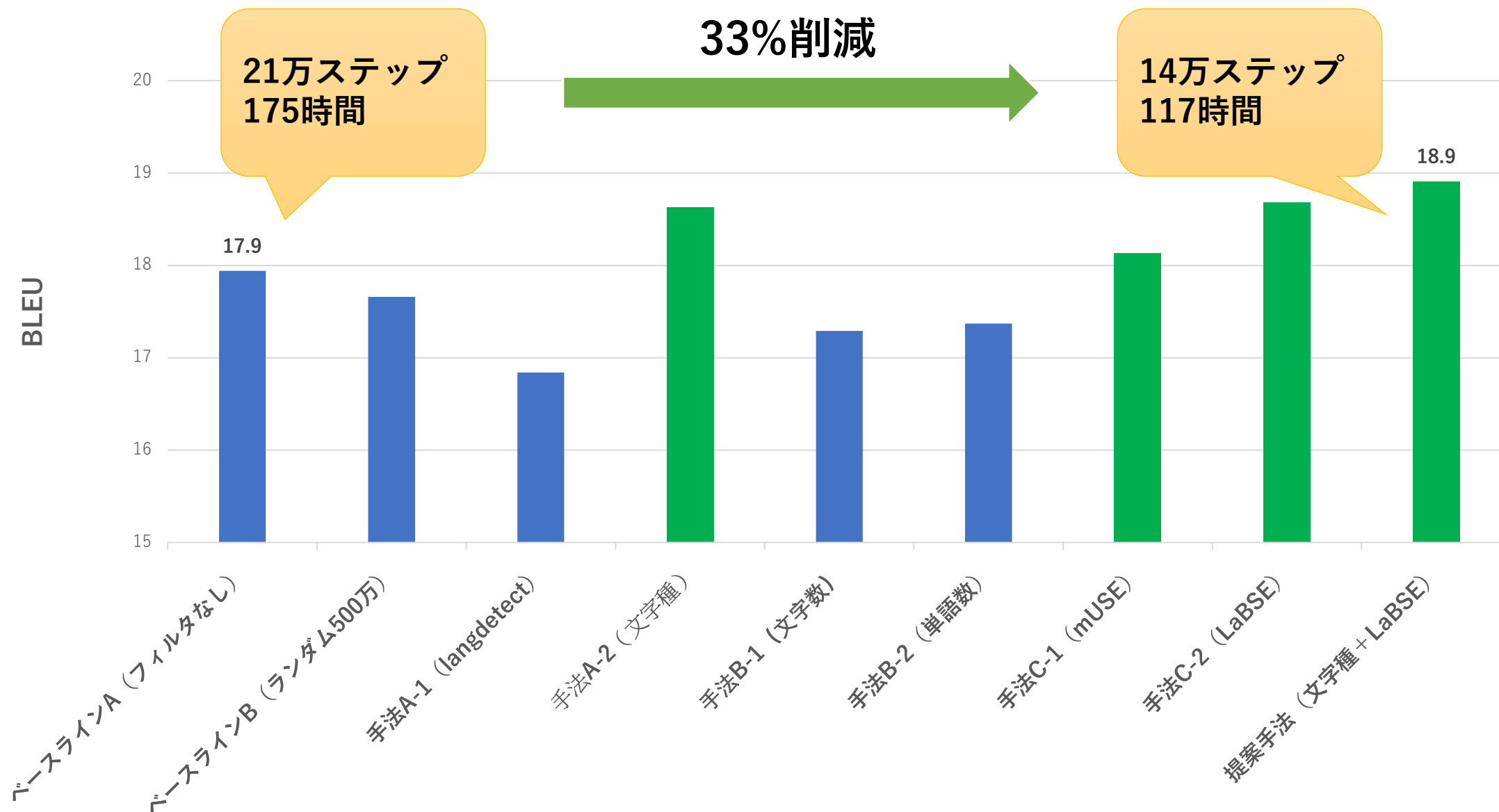
- ◆BLEU
- ◆0~100点までの翻訳品質
- ◆実際には40点くらいまでしか出ない
- ◆google翻訳の場合25点



実験結果：翻訳品質



実験結果：訓練時間



まとめ：少量データからの効率的な翻訳器の訓練

- ◆**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 → 有効ではない
 - 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

学会で発表し学生奨励賞を受賞

本田志遠, 正木亮太郎, 梶原智之. 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. 情報処理学会第84回全国大会, pp.793-794, March 2022.

