

第9回 長尾賞 学生奨励賞 受賞者講演

学習時と推論時における入力データの特徴の違いを考慮した
ニューラル機械翻訳モデルの学習手法

東京工業大学情報理工学院 (現 NHK放送技術研究所)

美野 秀弥

目的

- 翻訳モデルの学習時と翻訳時のデータの特徴の違いによる翻訳精度低下の解決
- データの特徴の分類
 - 学習時と翻訳時のデータの特徴の違いを下記の4つ観点で分類し、それぞれの違いに対応したニューラル機械翻訳のタスクを扱う
 - (a) 翻訳対象
 - (b) 翻訳結果
 - (c) 翻訳対象と翻訳結果との関係
 - (d) 外部情報

データの特徴の分類項目

(a) 翻訳対象

- 入力データの語彙, 体裁, 内容の特徴
- 例: ニュース文や特許文書の入力

(b) 翻訳結果

- 出力データの語彙, 体裁, 内容の特徴
- 例: 平易に翻訳した出力, 専門語彙は固定訳を用いて翻訳した出力

(c) 翻訳対象と翻訳結果との関係

- 例: 翻訳対象のデータを文字数を制限して翻訳する (少ない or 多い) , 意訳する, 低品質な翻訳精度で翻訳する

(d) 外部情報

- 翻訳対象ではない入力データの情報 (翻訳の補助情報)
- 例: 対訳辞書, 文脈情報

取り組むタスク

■分類項目におけるデータの特徴の違いに対応した2つのニューラル機械翻訳タスクに焦点を当て、双方において翻訳精度向上のための手法を提案する

1. ドメインタグを用いたニューラル機械翻訳
✓ (a) 翻訳対象, (b) 翻訳結果, (c) 翻訳対象と翻訳結果との関係, に対応
2. 目的言語側の前文を用いた文脈情報考慮型ニューラル機械翻訳
✓ (d) 外部情報, に対応

1. ドメインタグを用いたニューラル機械翻訳

ドメインアダプテーションを用いたニューラル機械翻訳

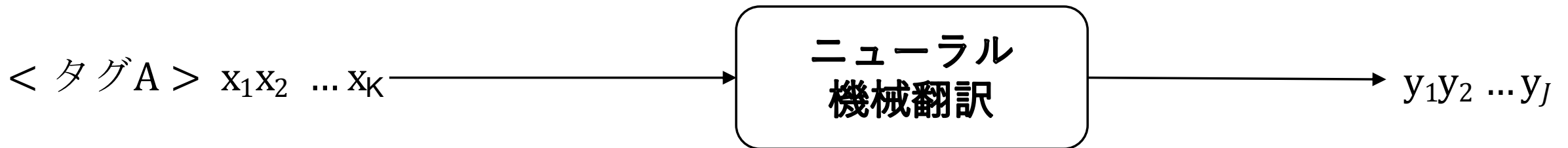
- 翻訳対象（インドメイン）と翻訳対象以外（アウトオブドメイン）の対訳データの両方を効果的に学習して翻訳精度を向上させる
 - 目的：インドメインのデータの特徴とは異なる特徴を持つアウトオブドメインのデータを学習に使うとインドメインのデータの翻訳精度が低下する課題の解決
 - 本論文では、学習時または翻訳時に用いる入出力データ全体の特徴をドメインと定義
- 本論文では、ドメインタグを用いたニューラル機械翻訳を扱う

ドメインタグを用いたニューラル機械翻訳

- 各データのドメインのタグ (<タグA>, <タグB>, ...) を原言語側のデータに付与して学習
- 翻訳時は翻訳対象のインドメインのタグを付与して翻訳

学習時：対訳データのドメインを表すタグをそれぞれ付与

翻訳時：翻訳対象のドメインのタグを付与



- どのようなタグを設定するかによって精度は変わる

本章で用いる日英ニュースコーパス [田中ら, 2021]

■内容等価コーパス（インドメインデータ）

- 原言語側と目的言語側の内容が等価になるように人手で翻訳した対訳データ

■自動アライメントコーパス（アウトオブドメインデータ）

- 事前に人手で翻訳された原言語側と目的言語側のデータ同士を自動で対応づけした対訳データ

■逆翻訳コーパス（アウトオブドメインデータ）

- 事前に用意した機械翻訳機で目的言語側のデータを原言語側に翻訳して作られた対訳データ

	過剰訳	訳抜け
日本語文	臨時休校の間、 <u>友人宅を</u> 行き来して遊んでいた孫は、授業再開が決まると「学校に行ける！」 <u>「給食何かな」と</u> 、登校を楽しみにしていたという。	
英語文	<u>According to the man,</u> his grandchild was looking forward to the resumption of classes while the school was closed.	

図：自動アライメントコーパスのデータ例。一部の単語やフレーズが翻訳されていない「訳抜け」や翻訳元のない情報が翻訳結果に含まれる「過剰訳」を含む。

本章で用いる日英ニュースコーパス [田中ら, 2021]

■各コーパスの特徴

コーパス	原言語側	目的言語側	原言語側と目的言語側のデータ間の関係
内容等価コーパス	元のニュース	内容が等価になるように翻訳	内容が等価である
自動アライメントコーパス	元のニュース	元のニュース	訳抜けや過剰訳などが含まれる
逆翻訳コーパス (内容等価)	内容等価コーパスに合わせた機械翻訳の出力	元のニュース	機械翻訳特有の訳抜け、過剰訳、誤訳が含まれる
逆翻訳コーパス (アライメント)	自動アライメントコーパスに合わせた機械翻訳の出力	元のニュース	機械翻訳特有の訳抜け、過剰訳、誤訳が含まれる

ドメインタグを用いた先行研究

■コーパスタグ

- 各対訳コーパスに1つのタグを付与する手法 [Kobus et al., 2017]
 - ✓ジャンルの異なる対訳コーパスに異なるドメインタグを付与

■ノイズタグ

- 機械翻訳を用いて構築したコーパスに別個のタグを付与する手法 [Vaibhav et al., 2019]

■2つのタグ

- 各対訳コーパスに2つのタグを付与する手法 [Berard et al., 2019]
 - ✓コーパスタグとノイズタグの両方を付与

ドメインタグを用いた従来手法の課題

■従来手法ではコーパス間の特徴を表現しきれていない

ドメインタグ	内容等価 コーパス	自動アライメント コーパス	逆翻訳 コーパス (内容等価)	逆翻訳 コーパス (アライメント)
コーパスタグ (Kobus et al. 2017)	<CE>	<AA>	<CE>	<AA>
ノイズタグ (Vaibhav et al. 2019)	<CE>	<AA>	<BT-CE>	<BT-AA>
2つのタグ (Berard et al. 2019a)	<CE>	<AA>	<BT> <CE>	<BT> <AA>

- 例：「逆翻訳コーパス（内容等価）」と「自動アライメントコーパス」は目的言語側のデータはともに「元のニュース」を用いているが従来手法ではこの情報が落ちてしまう

提案手法

- 原言語側のデータの特徴，目的言語側のデータの特徴，原言語側と目的言語側間の関係，に着目したタグを付与する

ドメインタグ	内容等価 コーパス	自動アライメント コーパス	逆翻訳 コーパス (内容等価)	逆翻訳 コーパス (アライメント)
コーパスタグ (Kobus et al. 2017)	<CE>	<AA>	<CE>	<AA>
ノイズタグ (Vaibhav et al. 2019)	<CE>	<AA>	<BT-CE>	<BT-AA>
2つのタグ (Berard et al. 2019a)	<CE>	<AA>	<BT> <CE>	<BT> <AA>
複数タグ (提案手法)	<NS-S> <CE-T> <NO-BT> <NO-AN>	<NS-S> <NS-T> <NO-BT> <AN>	<CE-S> <NS-T> <BT> <NO-AN>	<NS-S> <NS-T> <BT> <AN>

評価実験

■ 日英ニュースコーパスを用いた日英機械翻訳実験で評価

- テストセットは内容等価コーパスからあらかじめ重複しないように抽出した対訳データ（2,000文）を利用

■ 3種類の実験を実施

1. ドメインタグの効果

- ✓ ドメインタグを用いずに特徴の異なるコーパスを追加して比較

2. 提案手法の効果

- ✓ ドメインタグを用いた従来手法との比較
- ✓ さらにドメインタグを最大に増やした場合（7つのタグ）との比較

3. インドメインのデータ数による効果の違い

- ✓ インドメインデータを100K, 50K, 10K, 5K, 1Kと減らして比較

■ 評価尺度

- BLEU [Papineni et al., 2002]

実験結果 1 : ドメインタグの効果

- 特徴の異なるデータを追加してもドメインタグを用いない場合は翻訳精度は向上しない

学習に用いたコーパス	タグ	データ数	BLEU
内容等価コーパス	無	0.22M	20.93
内容等価コーパス + 自動アライメントコーパス	無	0.51M	20.68
内容等価コーパス + 自動アライメントコーパス + 逆翻訳コーパス (内容等価) + 逆翻訳コーパス (アライメント)	無	1.57M	20.36
自動アライメントコーパス	無	0.29M	10.30

実験結果 2 : 提案手法の効果

- 提案手法の翻訳精度が最も高くなった
- 2種類のタグと提案手法（4種類のタグを利用）のBLEUの差は+0.31であり、付与するドメインタグによって向上の程度は異なる

ドメインタグ	タグの数	データ量	BLEU	Δ
タグ無し	-	1.57M	20.36	-
コーパスタグ	1	1.57M	22.41	+2.05
ノイズタグ	1	1.57M	24.19	+3.83
2つのタグ	2	1.57M	24.25	+3.89
提案手法	4	1.57M	24.56	+4.20
7つのタグ	7	1.57M	24.26	+3.90

Δ : タグ無しの結果との差分

まとめ

■複数タグを用いたドメインタグを用いた手法を提案

- 日英ニュースコーパスを用いた日英機械翻訳実験において提案手法の効果を
確認

■今後の課題

- 翻訳精度の向上に寄与するドメインタグの選定方法の検討
- 他データセットでの実験

2. 目的言語側の前文を用いた文脈情報考慮型 ニューラル機械翻訳

文脈情報考慮型ニューラル機械翻訳

■ 翻訳対象の周辺の文などを文脈情報として活用して翻訳精度を向上させる

- 目的：1文ごとに学習，翻訳すると翻訳精度が低下する場合がある課題の解決

- 翻訳元日本語 1文だけでは情報が不足しておりの確な英語に翻訳できない
日英翻訳例

- ✓ 翻訳元日本語文

京都府内で感染が確認されたのは、これで14人になります。

- ✓ 正解英語訳 (感染者の性別の情報が必要)

She is the 14th person to have been confirmed to be infected in Kyoto Prefecture.

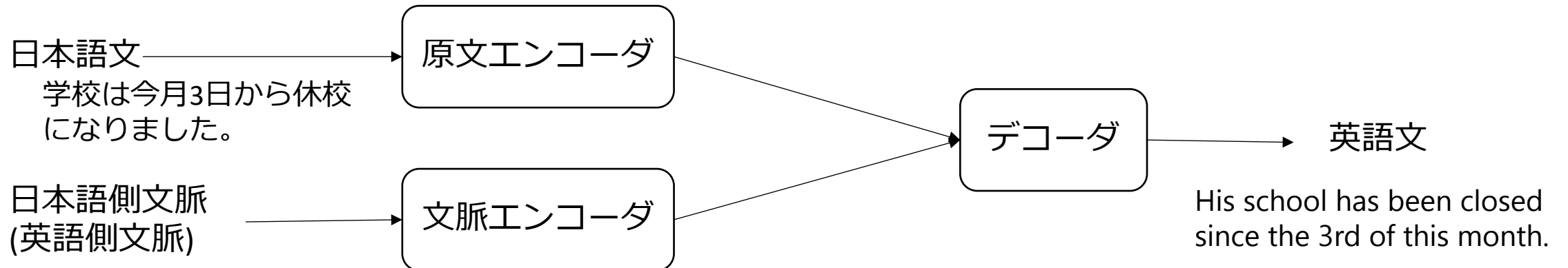
- ✓ 文脈情報 (日本語側の前文)

京都市では区役所の窓口の消毒作業を行うとともに、**女性**と一緒に働いていた16人の職員の健康状態について経過観察することになっています。

文脈情報考慮型ニューラル機械翻訳システム

[Bawden et al., 2018]

- 対訳データ (例：日本語+文脈, 英語)を学習、翻訳
 - 原文エンコーダ：翻訳する日本語文を受け取り、文ベクトルを生成
 - 文脈エンコーダ：文脈情報（翻訳対象の周辺の文）を受け取り、文脈情報のベクトルを生成
 - デコーダ：日本語文ベクトルと文脈文ベクトルを受け取り、英語文を1トークンずつ生成



神奈川県の高校 1年生の男子学生が...
(A male student in the first year of junior
high school in Kanagawa Prefecture...)

文脈情報考慮型ニューラル機械翻訳の先行研究で用いられる 文脈情報とその課題

■文脈情報として用いられるデータ

- 原言語側の前文 [Agrawal et al., 2018; Bawden et al., 2018]
- 目的言語側の前文 [Bawden et al., 2018]

■先行研究により、文脈情報に原言語側の前文を用いると翻訳精度は向上するが、目的言語側の前文を用いると翻訳精度は低下するという課題がある

目的言語側の前文を用いることが有効な例

■原言語側がない情報が目的言語側に含まれる翻訳

●ゼロ代名詞を含む例（日英翻訳）

	文脈情報（前文）	翻訳対象
原言語側	昼食にはサラダを食べる。	牛乳も飲みます
目的言語側	His friend eats a salad for lunch.	→原言語側前文を用いた翻訳結果： I also drink a milk. →目的言語側前文を用いた翻訳結果： His friend also drinks a milk.

●文末表現を含む例（英日翻訳）

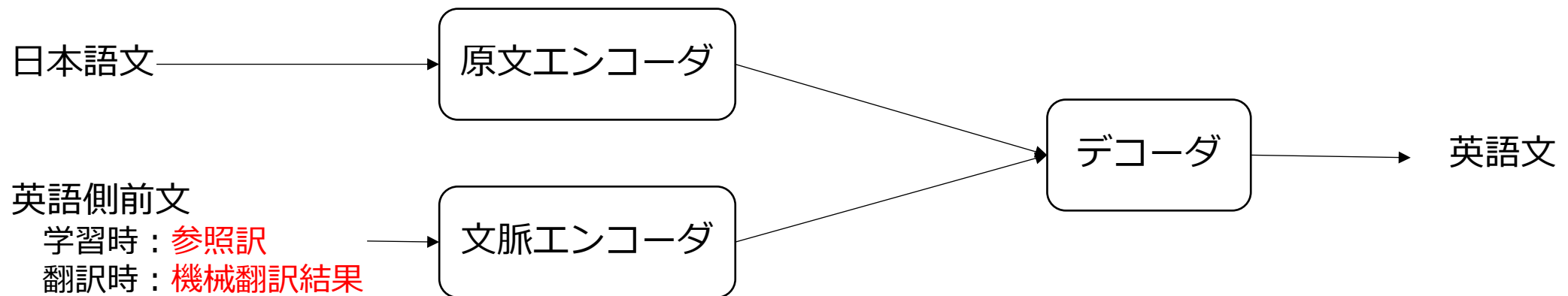
	文脈情報（前文）	翻訳対象
原言語側	I eat a salad for lunch.	I also drink a milk.
目的言語側	私は昼食にサラダを 食べる 。	→原言語側前文を用いた翻訳結果： 私は牛乳も飲み ます 。 →目的言語側前文を用いた翻訳結果： 私は牛乳も 飲む 。

目的

- 先行研究では目的言語側の文脈情報を用いると翻訳精度が低下した原因を推定
- 目的言語側の文脈を用いて翻訳精度を向上させる手法を提案する

従来手法の目的言語側の文脈を用いる機械翻訳システムの課題の原因

- 対訳データ (例：日本語+目的言語側の前文, 英語)を学習、翻訳
 - 文脈エンコーダ：英語側前文を受け取り、文脈文ベクトルを生成
 - ✓ 学習時：対訳データ内の前文の参照訳を利用
 - ✓ 翻訳時：前文の機械翻訳結果を利用



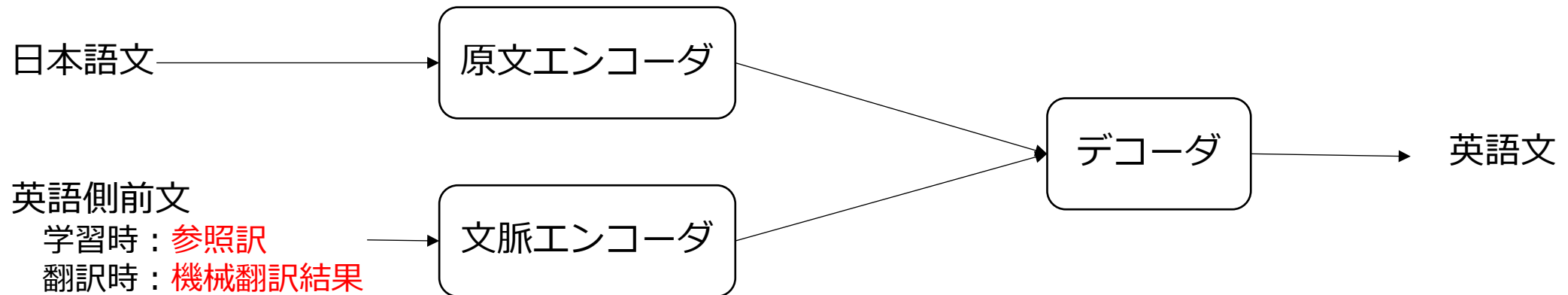
従来手法の目的言語側の文脈を用いる機械翻訳システムの課題の原因

■ 対訳データ (例：日本語+目的言語側の前文, 英語)を学習、翻訳

● 文脈エンコーダ：英語側前文を受け取り、文脈文ベクトルを生成

- ✓ 学習時：正確な訳である**参照訳**を利用
- ✓ 翻訳時：ノイズが含まれる**機械翻訳結果**を利用

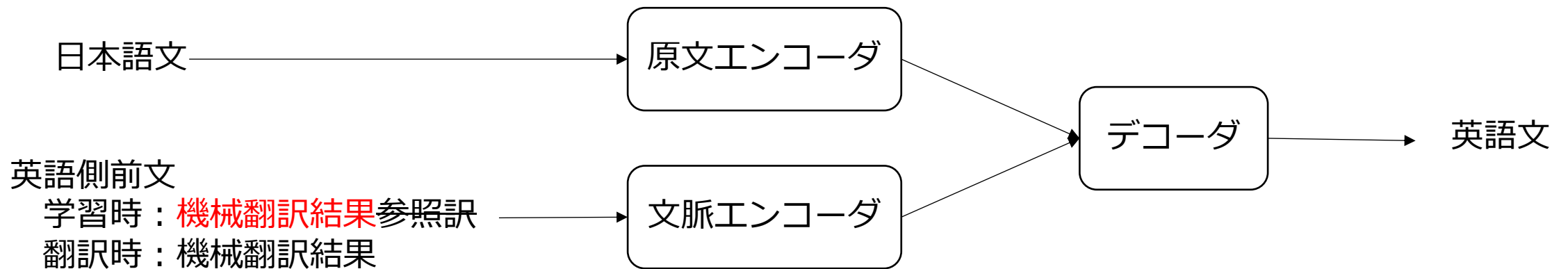
原因：学習時と翻訳時に用いる文脈情報のデータの特徴が異なっている



目的言語側の文脈情報を用いることで先行研究で翻訳精度が低下した原因が外部情報（目的言語側の前文）のデータの違いを考慮していないことにあるとし、データの違いを考慮した手法を提案する

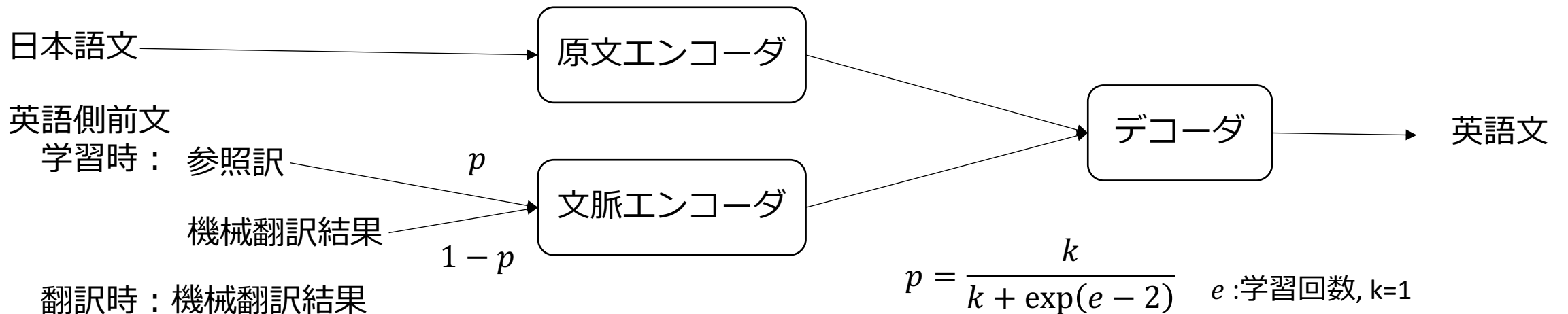
提案手法 1：機械翻訳結果を利用

- 学習時と翻訳時ともに目的言語側前文の機械翻訳結果を入力
 - 原文エンコーダ、デコーダは既存手法と同様
 - 学習時と翻訳時に用いる文脈情報の違いによるデータの特徴の違いが解消
 - 課題：機械翻訳結果に含まれるノイズが学習に悪影響を及ぼす可能性がある



提案手法 2: 参照訳と機械翻訳結果を利用 (提案手法 1 の改善)

- 学習時に目的言語側前文の参照訳と機械翻訳結果の両方を確率値 p でサンプリングして入力
 - カリキュラムラーニングの手法を参考に、最初の学習では「参照訳」を中心に学習して徐々に「機械翻訳結果」を多く学習していく
 - ✓ 最初に「参照訳」を用いることで、(ノイズがないため) 学習が促進される
 - ✓ 段階的に「機械翻訳結果」の割合を増やすことで、(翻訳時と同じ特徴を有する) 機械翻訳結果に含まれるノイズに対応した学習ができる



評価実験

■日英・英日機械翻訳実験

データの種類	学習データ文数	評価用テストセット文数
ニュース	220,180	2,000
TEDトーク	194,170	1,285

■目的言語側の文脈情報

- 計算機コストを抑えるために前の1文のみを用いる

■比較手法

- 文レベル機械翻訳システム：文脈情報を用いずに対訳データのみで学習
- 文脈情報考慮型機械翻訳システム：英語側の前文参照訳と対訳データで学習

■評価尺度

- BLEU [Papineni et al., 2002]

日英・英日機械翻訳実験結果

機械翻訳システム	学習時 文脈情報	翻訳時 文脈情報	ニュース		TEDトーク	
			日英	英日	日英	英日
文レベル機械翻訳システム	-	-	24.23	41.99	8.48	15.69
文脈情報考慮型機械翻訳システム	参照訳	機械翻訳	24.80	42.40	8.98	16.15
提案手法 1	機械翻訳	機械翻訳	24.31	42.45	9.15	16.27
提案手法 2	参照訳 + 機械翻訳	機械翻訳	24.86	42.79	9.66	16.37

■ 具体例（日英）

（英語側の前文）

Well, you can do it for yourself.

（翻訳対象日本語文）

そして信頼性を高めることができます。

正解英語訳

You can make your own sense of trust.

文レベル機械翻訳システム

And then we can increase trust .

提案手法 2

You can improve your confidence.

まとめ

- 既存手法の文脈情報考慮型機械翻訳システムの課題を示し、文脈情報として目的言語側の前文を利用したニューラル機械翻訳の新たな学習手法を提案した
 - 日英・英日機械翻訳の評価実験により、提案手法の効果を確認した
- 今後の課題
 - データを変えた時の最適なハイパーパラメータを確認
 - 文脈情報として複数文を用いた場合の効果を確認

全体のまとめ

全体のまとめ

- 学習時と翻訳時のデータの特徴の違いを4つの観点で分類した
- 各観点において特徴の異なるデータを用いたことで翻訳精度が低下した2つのニューラル機械翻訳のタスクに取り組んだ
 - ドメインタグを用いたニューラル機械翻訳
 - 目的言語側の前文を用いた文脈情報考慮型ニューラル機械翻訳
- 双方のタスクにおいて翻訳精度を向上させる手法を提案して効果を確認した
 - 複数の特徴タグを用いたニューラル機械翻訳
 - 目的言語側の前文として参照訳と機械翻訳結果の両方を用いた文脈情報考慮型ニューラル機械翻訳

関連研究と提案手法におけるデータの特徴の分類項目の対応

手法	データの特徴の分類項目			
	(a) 翻訳対象	(b) 翻訳結果	(c) 翻訳対象と 翻訳結果との関係	(d) 外部情報
ドメイン適応を用いたニューラル機械翻訳				
コーパスタグ (Kobus et al. 2017)	✓	(✓)	(✓)	-
ノイズタグ (Vaibhav et al. 2019)			✓	-
2つのタグ (Berard et al. 2019a)	✓		✓	-
提案手法	✓	✓	✓	-
目的言語側の前文を用いた文脈情報考慮型ニューラル機械翻訳				
従来手法 (Bawden et al. 2018)	-	-	-	
提案手法	-	-	-	✓

今後の課題

- 本論文で扱った2つのニューラル機械翻訳以外のタスクにおいて、学習時と翻訳時とのデータの違いにより翻訳精度が低下しているかどうか検証する
- ドメインが不明なコーパスに対応できる手法を検討する
 - 例えばウェブから収集したデータなどは特徴が違うデータが混在しており、それらを区別することができない
 - クラスタリングなどの手法を用いて分類する必要があり、有効なデータの分類手法の検討が必要となる

謝辞

本研究を進め、論文をまとめるにあたり、ご指導とご助言を賜りました多くの方々に対し、感謝の言葉を述べさせていただきます。

主指導教官であった徳永健伸教授には、本研究だけでなく、自然言語処理全般の研究に関して、日頃より多くのご指導とご助言を賜りました。深く感謝いたします。

本研究成果の一部は国立研究開発法人情報通信研究機構の委託研究「多言語音声翻訳高度化のためのディープラーニング技術の研究開発」(課題197)、「自動翻訳の精度向上のための「マルチモーダル情報の外部制御可能なモデリング」の研究開発」(課題225)により得られたものです。

補足資料

取り扱う 2 つのタスクの関係性

■ タスク自体は別個の課題を扱ったタスク

- ドメインタグを用いたニューラル機械翻訳
 - ✓ 特徴の異なる対訳コーパスを用いて翻訳精度を向上させる
 - ✓ 対訳データが少ない場合に有効
- 文脈情報考慮型ニューラル機械翻訳
 - ✓ 文脈情報を用いて翻訳精度を向上させる
 - ✓ 1文を越えた情報や知識が必要な場合に有効

■ 「データの特徴の違いによる翻訳精度の低下」という共通の課題を有している

手法	データの特徴の分類項目			
	(a) 翻訳対象	(b) 翻訳結果	(c) 翻訳対象と翻訳結果との関係	(d) 外部情報
ドメインタグを用いたニューラル機械翻訳				
コーパスタグ (Kobus et al. 2017)	✓	(✓)	(✓)	-
ノイズタグ (Vaibhav et al. 2019)			✓	-
2つのタグ (Berard et al. 2019a)	✓		✓	-
提案手法	✓	✓	✓	-
目的言語側の前文を用いた文脈情報考慮型ニューラル機械翻訳				
従来手法 (Bawden et al. 2018)	-	-	-	
提案手法	-	-	-	✓

ドメインタグを用いた先行研究と提案手法におけるデータの特徴の分類項目の対応

- 提案手法は分類項目全てを考慮したタグ体系となっている

手法	データの特徴の分類項目			
	(a) 翻訳対象	(b) 翻訳結果	(c) 翻訳対象と翻訳結果との関係	(d) 外部情報
ドメイン適応を用いたニューラル機械翻訳				
コーパスタグ (Kobus et al. 2017)	✓	(✓)	(✓)	-
ノイズタグ (Vaibhav et al. 2019)			✓	-
2つのタグ (Berard et al. 2019a)	✓		✓	-
提案手法	✓	✓	✓	-

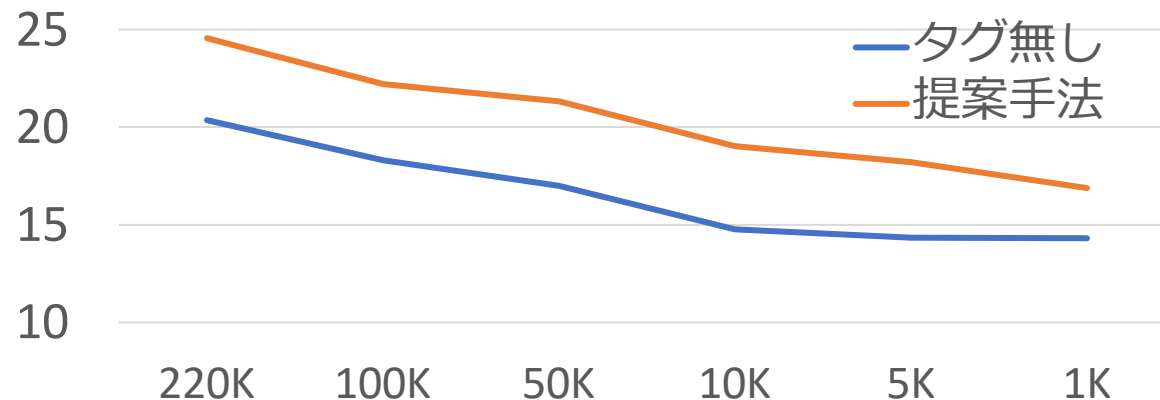
本章で用いる日英ニュースコーパス [田中ら, 2021]

■統計情報

コーパス	データ数	日本語側		英語側	
		文数	文字数	文数	単語数
内容等価コーパス	220,180	220,180	10,427,732	235,407	6,052,647
自動アライメントコーパス	286,247	286,247	17,521,620	286,247	9,090,966
逆翻訳コーパス (内容等価)	533,581	533,581	22,798,841	533,581	14,348,304
逆翻訳コーパス (アライメント)	533,581	533,581	21,861,647	533,581	14,348,304

実験結果 3 : インドメインのデータ数による効果の変化

- インドメインのデータ数が減少しても提案手法の翻訳精度は向上
- インドメインのデータ数が1Kの場合の提案手法の効果は下がった



ドメインタグ	インドメインのデータ数					
	220K	100K	50K	10K	5K	1K
タグ無し	20.36	18.30	17.01	14.75	14.34	14.31
提案手法	24.56	22.20	21.33	19.03	18.22	16.88
Δ	+4.20	+3.90	+4.32	+4.28	+3.88	+2.57
合計データ数	1.57M	1.45M	1.40M	1.36M	1.35M	1.35M

目的言語側の前文を用いる機械翻訳システムの従来手法と提案手法におけるデータの特徴の分類項目の対応

- 提案手法は(d) 外部情報の特徴の違いを考慮した手法となっている

手法	データの特徴の分類項目			
	(a) 翻訳対象	(b) 翻訳結果	(c) 翻訳対象と翻訳結果との関係	(d) 外部情報
目的言語側の前文を用いた文脈情報考慮型ニューラル機械翻訳				
従来手法	-	-	-	
(Bawden et al. 2018)	-	-	-	
提案手法	-	-	-	✓

分析 1 : 人手評価

■TEDトークコーパスの日英機械翻訳タスクのテストセットからランダムに300文を抽出して人手評価を実施した

- 日本語側に主語の省略があるかどうか確認
- 主語の省略がある場合は英語側で正しく主語を補完して翻訳しているか確認

機械翻訳システム	主語の省略なし	主語の省略あり	
		正解訳数	誤訳数
文レベル機械翻訳システム	201	49	50
提案手法 2 : 正解訳と機械翻訳結果を利用	201	63	36

TED トークコーパス 日英機械翻訳 タスク 実例 1

翻訳対象の日本語文	名前はありませんが彼の話はできます
目的言語側の前文（機械翻訳結果）	I'm going to talk about a boy in a village.
参照訳	I don't know his name, but I do know his story.
文レベル機械翻訳システム	I don't know the name, but I can tell you his story.
提案手法 2 : 正解訳と機械翻訳結果を利用	I don't know his name, but I can tell you his story.

- 文レベル機械翻訳システムでは“the name”となっているが、提案手法では“his name”と人称所有格を正しく翻訳できている

TEDトークコーパス日英機械翻訳タスク実例 2

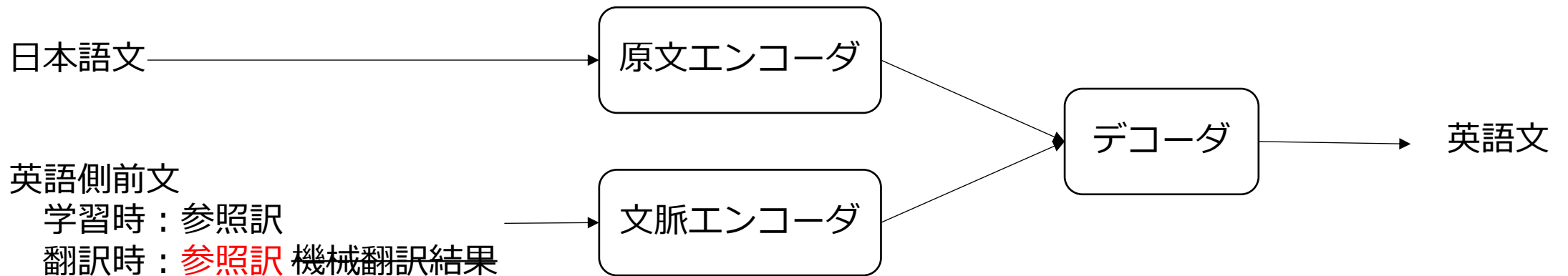
(提案手法が誤った例)

翻訳対象の日本語文	ヒールでダンスできれば当然ヒールで歩けますからね
目的言語側の前文（機械翻訳結果）	Yeah, on Broadway. It's crazy. My parents couldn't afford physical therapy, so they sent me to dancing school.
目的言語側の2つ前の文（機械翻訳結果）	I was in Broadway, and I was dancing.
参照訳	I learned how to dance in heels, which means I can walk in heels.
文レベル機械翻訳システム	Because, of course, if you can dance on your heels, you can walk on your feet.
提案手法 2 : 正解訳と機械翻訳結果を利用	They could dance with their feet, because they could walk with their legs.

- 文レベル機械翻訳システム, 提案手法とともに主語や所有格の補完が不適切であった.
- 前文が長く意味を把握しきれずに誤訳した可能性がある.
- 2つ前の文を使うことで正しく翻訳できる可能性がある.

分析 2 : データの特徴の違いの影響 (学習時と翻訳時ともに参照訳を利用)

- 目的言語側前文の参照訳を文脈エンコーダに入力して学習
 - 原文エンコーダ、デコーダは既存手法と同様
 - 学習時と翻訳時に用いる文脈情報の違いによるデータの特徴の違いが解消される
 - 実際には翻訳時に参照訳を使える状況にはないため、現実的な設定ではない
- 翻訳時に目的言語側前文に機械翻訳を用いた場合と参照訳を用いた場合との間の翻訳精度を比較する



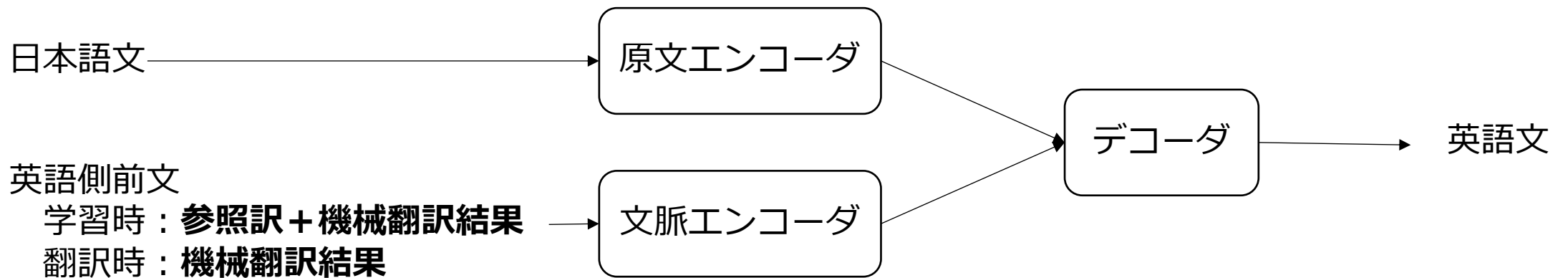
分析2の実験結果

機械翻訳システム	学習時 文脈情報	翻訳時 文脈情報	ニュース		TEDトーク	
			日英	英日	日英	英日
文レベル機械翻訳システム	-	-	24.23	41.99	8.48	15.69
文脈情報考慮型機械翻訳システム	参照訳	機械翻訳	24.80	42.40	8.98	16.15
文脈情報考慮型機械翻訳システム	参照訳	参照訳	25.09	42.50	9.51	16.39

- 学習時に機械翻訳を用いた場合と比較して参照訳を用いた場合の翻訳精度は向上した
- この翻訳精度の差がデータの特徴の違いによる影響だと考えられる

分析3：カリキュラムラーニングの効果 (参照訳と機械翻訳結果の両方を利用)

- 目的言語側前文の参照訳と機械翻訳結果の両方を一定の割合で文脈エンコーダに入力して学習
 - 原文エンコーダ、デコーダは既存手法と同様
 - 学習時に用いる参照訳と機械翻訳結果は半々で利用（割合は固定）
 - 翻訳時は機械翻訳結果を利用
- 参照訳と機械翻訳結果のデータの割合を制御して学習するカリキュラムラーニングと比較



分析3の実験結果

機械翻訳システム	学習時 文脈情報	翻訳時 文脈情報	ニュース		TEDトーク	
			日英	英日	日英	英日
文レベル機械翻訳システム	-	-	24.23	41.99	8.48	15.69
文脈情報考慮型機械翻訳システム	参照訳+ 機械翻訳 (固定)	機械翻訳	24.30	42.44	9.05	16.28
文脈情報考慮型機械翻訳システム	参照訳+ 機械翻訳 (変化)	機械翻訳	24.86	42.79	9.66	16.37

- 学習時に用いる文脈情報のデータの割合を制御することで翻訳精度が向上
- この翻訳精度の差がカリキュラムラーニングによる影響であると考えられる

目的言語側の前文を用いた文脈情報考慮型ニューラル 機械翻訳の翻訳実験結果

機械翻訳システム	学習時 文脈情報	翻訳時 文脈情報	ニュース		TEDトーク	
			日英	英日	日英	英日
文レベル機械翻訳システム	-	-	24.23	41.99	8.48	15.69
文脈情報考慮型機械翻訳システム（従来手法）	参照訳	機械翻訳	24.80	42.40	8.98	16.15
提案手法 1	機械翻訳	機械翻訳	24.31	42.45	9.15	16.27
提案手法 2	参照訳 + 機械翻訳 （変化）	機械翻訳	24.86	42.79	9.66	16.37
分析 3	参照訳 + 機械翻訳 （固定）	機械翻訳	24.30	42.44	9.05	16.28
分析 2	参照訳	参照訳	25.09	42.50	9.51	16.39

貢献

■全体の貢献

- 学習時と翻訳時におけるデータの特徴を4つの観点で分類
 - ✓体系的に分類したことで、データの特徴の違いを詳細に区別して下記の2つのニューラル機械翻訳の改善手法を提案できた

■ドメインタグを用いたニューラル機械翻訳

- 従来手法が対訳データ全体で特徴を分類したのに対し、提案手法ではデータの特徴の違いを原言語側の特徴、目的言語側の特徴、原言語側と目的言語側のデータ間の特徴に分類して学習することで翻訳精度を向上させた

■目的言語側の前文を用いた文脈情報考慮型ニューラル機械翻訳

- 翻訳精度が低下した従来手法の原因が学習時と翻訳時における目的言語側の文脈情報のデータの特徴の違いにあることを明らかにし、学習時の文脈情報を学習の進行に応じて参照訳から機械翻訳結果へ段階的に切り替えることで翻訳精度を向上させた

参考文献 (1/3)

[Bahdanau et al.] Bahdanau, D., Cho, K., and Bengio, Y. (2015). "Neural Machine Translation by Jointly Learning to Align and Translate." In 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings.

[Bawden et al.] Bawden, R., Sennrich, R., Birch, A., and Haddow, B. (2018). "Evaluating Discourse Phenomena in Neural Machine Translation." In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), pp. 1304–1313, New Orleans, Louisiana. Association for Computational Linguistics.

[Berard et al.] Berard, A., Calapodescu, I., and Roux, C. (2019b). "Naver Labs Europe's Systems for the WMT19 Machine Translation Robustness Task." In Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1), pp. 526–532, Florence, Italy. Association for Computational Linguistics.

[Cho et al.] Cho, K., van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., and Bengio, Y. (2014). "Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation." In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 1724–1734, Doha, Qatar. Association for Computational Linguistics.

参考文献(2/3)

[Kingma and Ba] Kingma, D. P. and Ba, J. (2015). "Adam: A method for stochastic optimization." In International Conference on Learning Representations (ICLR).

[Kobus et al.] Kobus, C., Crego, J., and Senellart, J. (2017). "Domain Control for Neural Machine Translation." In Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017, pp. 372–378, Varna, Bulgaria. INCOMA Ltd.

[Papineni et al.] Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). "Bleu: a Method for Automatic Evaluation of Machine Translation." In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, pp. 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

[Robbins and Monro] Robbins, H. and Monro, S. (1951). "A stochastic approximation method." Annals of Mathematical Statistics, 22, pp. 400–407.

参考文献(3/3)

[Sutskever et al.] Sutskever, I., Vinyals, O., and Le, Q. V. (2014). "Sequence to Sequence Learning with Neural Networks." In Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2, NIPS'14, pp. 3104–3112, Cambridge, MA, USA. MIT Press.

[田中英輝他] 田中英輝, 中澤敏明, 美野秀弥, 伊藤均, 後藤功雄, 山田一郎, 川上貴之, 大嶋聖一, 朝賀英裕(2021). 時事通信社ニュースの日英対訳コーパスの構築-第3報. 言語処理学会第27回年次大会, pp. 228–231.

[Vaibhav et al.] Vaibhav, V., Singh, S., Stewart, C., and Neubig, G. (2019). "Improving Robustness of Machine Translation with Synthetic Noise." In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pp. 1916–1920, Minneapolis, Minnesota. Association for Computational Linguistics.

[Vaswani et al.] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. u., and Polosukhin, I. (2017). "Attention is All you Need." In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R. (Eds.), Advances in Neural Information Processing Systems 30, pp. 5998–6008. Curran Associates, Inc.