

第8回AAMTセミナー AAMT若手翻訳研究会
2024年3月22日

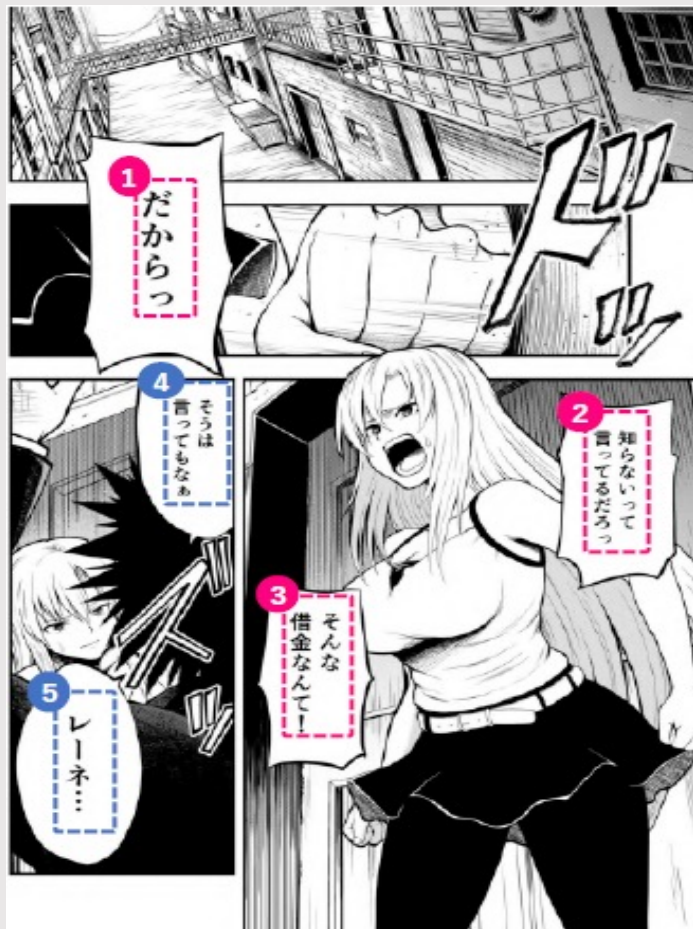
吹き出しよりも長い文脈情報を用いた 漫画の機械翻訳

戒能大翔¹, 杉原壮一郎¹, 梶原智之¹, 二宮崇¹, Tanner Joshua², 石渡祥之佑²

1. 愛媛大学

2. Mantra株式会社

漫画の翻訳の例



日本語文



英語文

©朽鷹みつき

背景

- 近年、非公式に翻訳された海賊版の漫画が海外で流通
 - ▶ 2021年の小学館の調査では正規版の約5倍の漫画が海賊版として海外で流通^[1]
 - ▶ 閲覧数が1億回を超える海賊版も存在^[1]
- リアルタイムな正規の翻訳出版が少ないことが要因の1つ



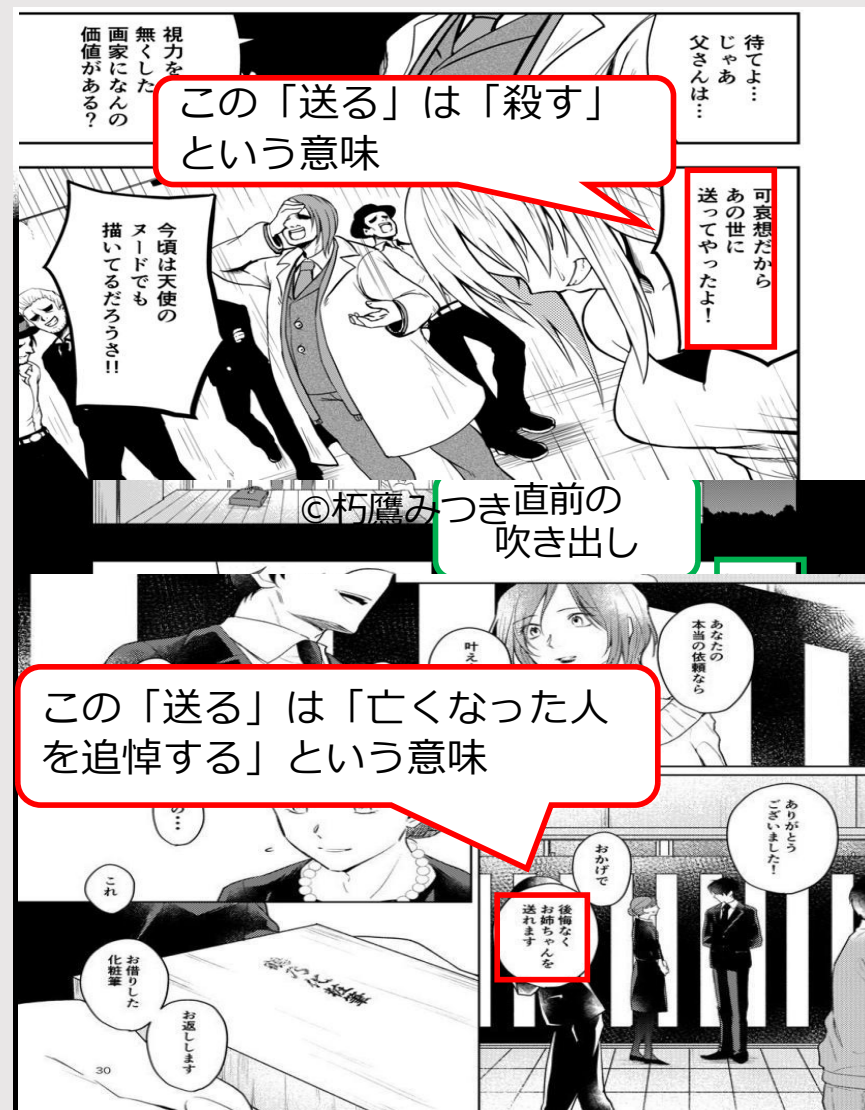
自然言語処理の技術による漫画の機械翻訳が期待されている^[2]

[1] <https://prtimes.jp/main/html/rd/p/000000004.000059295.html>

[2] Ryota Hinami et al. Towards Fully Automated Manga Translation, In Proc. of AAAI.2021

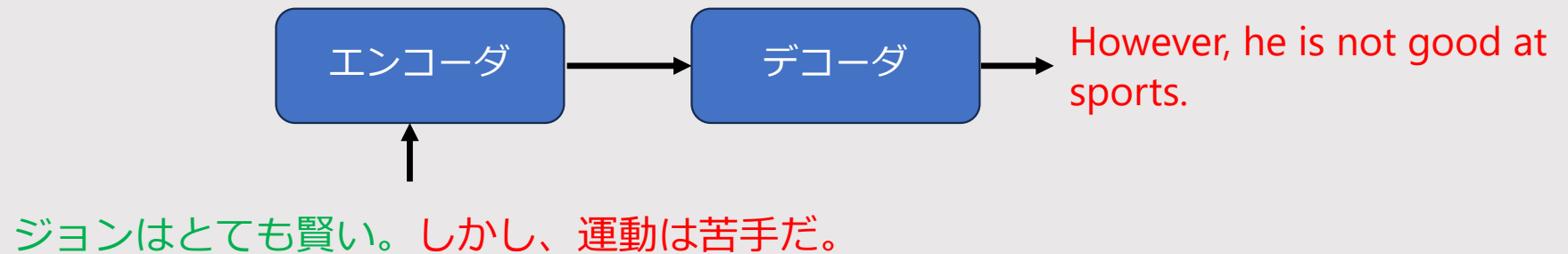
漫画の機械翻訳の課題

- 他のコマや吹き出しを文脈として考慮する必要がある
 - ▶ 漫画の吹き出しは単体では翻訳に必要な情報を十分に持たない
- 漫画ごとに適切なスタイルで翻訳する必要がある
 - ▶ 作品や作者、ジャンルごとに特徴が異なる

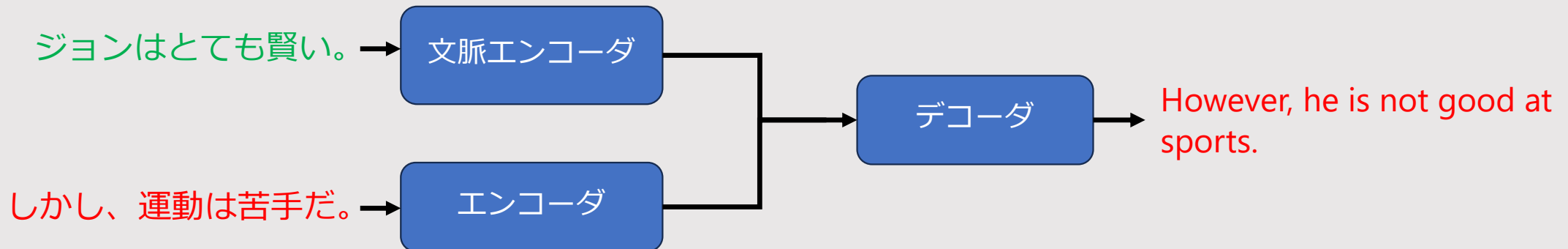


関連研究：文脈を考慮する機械翻訳

- 結合ベースの手法

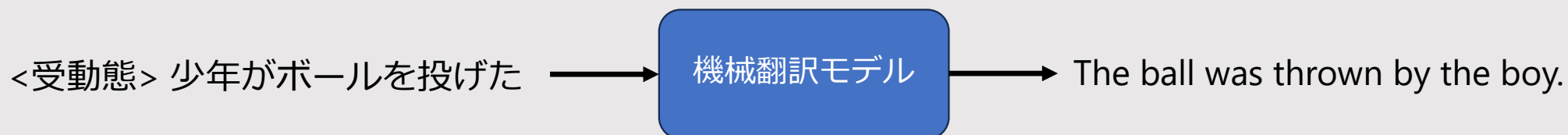


- マルチエンコーダを用いる手法



関連研究：機械翻訳のスタイル制御

- 機械翻訳などの系列変換モデルの出力文のスタイルは訓練データのテキストのスタイルに依存
- 状況に応じて出力文のスタイルを制御する研究が行われている
- 多くの研究が特殊トークンを文に付与してスタイルを制御する手法を採用^{[4][5]}



[3] Rico Sennrich et al. Controlling Politeness in Neural Machine Translation via Side Constraints, In Proc of NAACL. 2016

[4] Hayahide Yamagishi. Improving Japanese-to-English Neural Machine Translation, In Proc of IJCNLP, 2017

先行研究：他の吹き出しを文脈とする手法^[2]

- 直前の吹き出しを考慮して翻訳
病院の方から<SEP>あまり見ない方がいいと言われて
- コマ単位での翻訳
・・・いえ
<SEP> 病院の方から
<SEP> あまり見ない方がいいと言われて
- 課題：コマを超える文脈を考慮できない

他のコマや漫画の作品や作者、ジャンルなどの作品情報(属性)といったコマを超えた文脈の考慮が必要

翻訳対象のコマ

直前の吹き出し

病院の方から

あまり見ない方がいいと言われて

・・・いえ

翻訳対象の吹き出し

ではしようねえ

ご覧になりましたか?

お姉さんの事故後のお顔

©吉良いと

[2] Ryota Hinami et al. Towards Fully Automated Manga Translation, In Proc. of AAAI.2021

提案手法

直前のコマ

直前のコマを考慮する翻訳

結合ベースの文脈考慮型機械翻訳を応用し、
直前のコマを考慮したコマ単位の翻訳を行う

翻訳例

十年後 -夏 <コマ内> 相川仁 <コマ内> お前には
一流の大学を目指す学力がある <コマ外> **なのになぜ
三流大の推薦なんかを狙う?** <コマ内> **実に夢がない**



why get recommendation letters for a third
tiered college? <コマ内> don't you have aspirations?



漫画の属性を考慮する翻訳

- 特殊トークンを文に付与してスタイルを制御する手法を応用し、漫画の属性を付与して翻訳
- 漫画の**作品・作者・出版社・掲載誌・ジャンル**の5つの属性を考慮
(Wikipediaやマンガペディアから収集)

翻訳対象	属性情報	入力
「あさがお」は	作品	<バレーンドリーム> 「あさがお」は
	作者	<朽鷹みつき> 「あさがお」は
	ジャンル	<学園> 「あさがお」は



実験

実験設定

- データセット
 - ▶ 漫画の機械翻訳のためのデータセットであるManga Corpus^[2]
 - ▶ 検証用データおよび評価用データとして50作品を無作為に抽出し、各作品の最新巻から10ページずつ無作為抽出した
 - ▶ その他のページは訓練データとして使用
- データの前処理
 - ▶ 日本語にはKytea^[6]、英語にはMoses^[7]を用いて単語分割
 - ▶ 単語分割後、Byte Pair Encoding^[8]でサブワード分割
(語彙サイズは日英合わせて32,000)
- 翻訳モデル
 - ▶ Transformer (big)^[9]

[2] Ryota Hinami et al. Towards Fully Automated Manga Translation, In Proc. of AAAI.2021

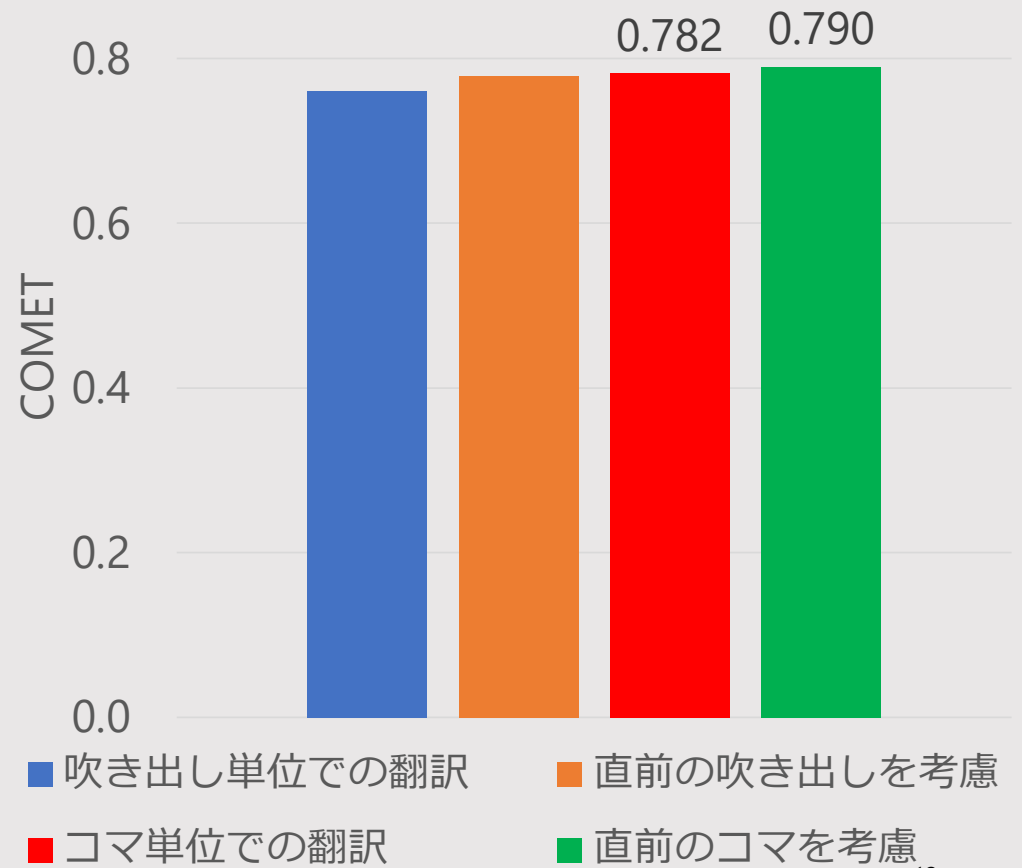
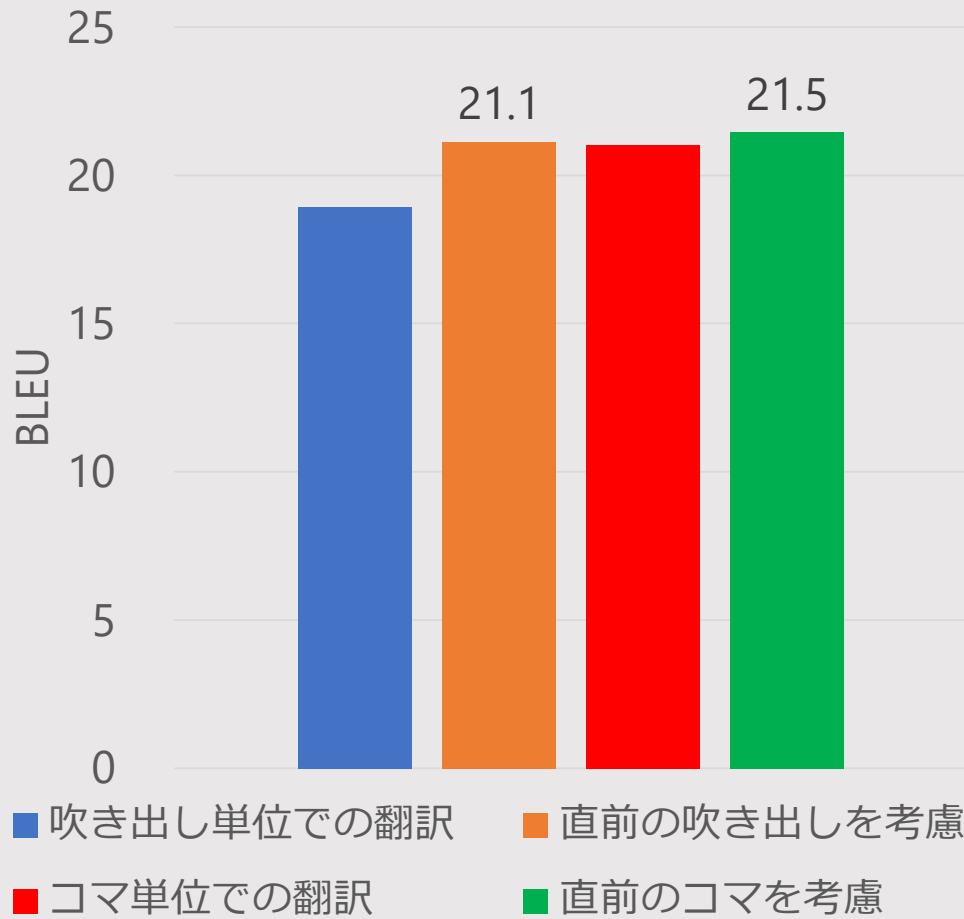
[5] Graham Neubig et al. Pointwise Prediction for Robust, Adaptable Japanese Morphological Analysis, In Proc of ACL. 2011

[6] Philipp Koehn et al. Moses: Open Source Toolkit for Statistical Machine Translation, In Proc of ACL. 2007

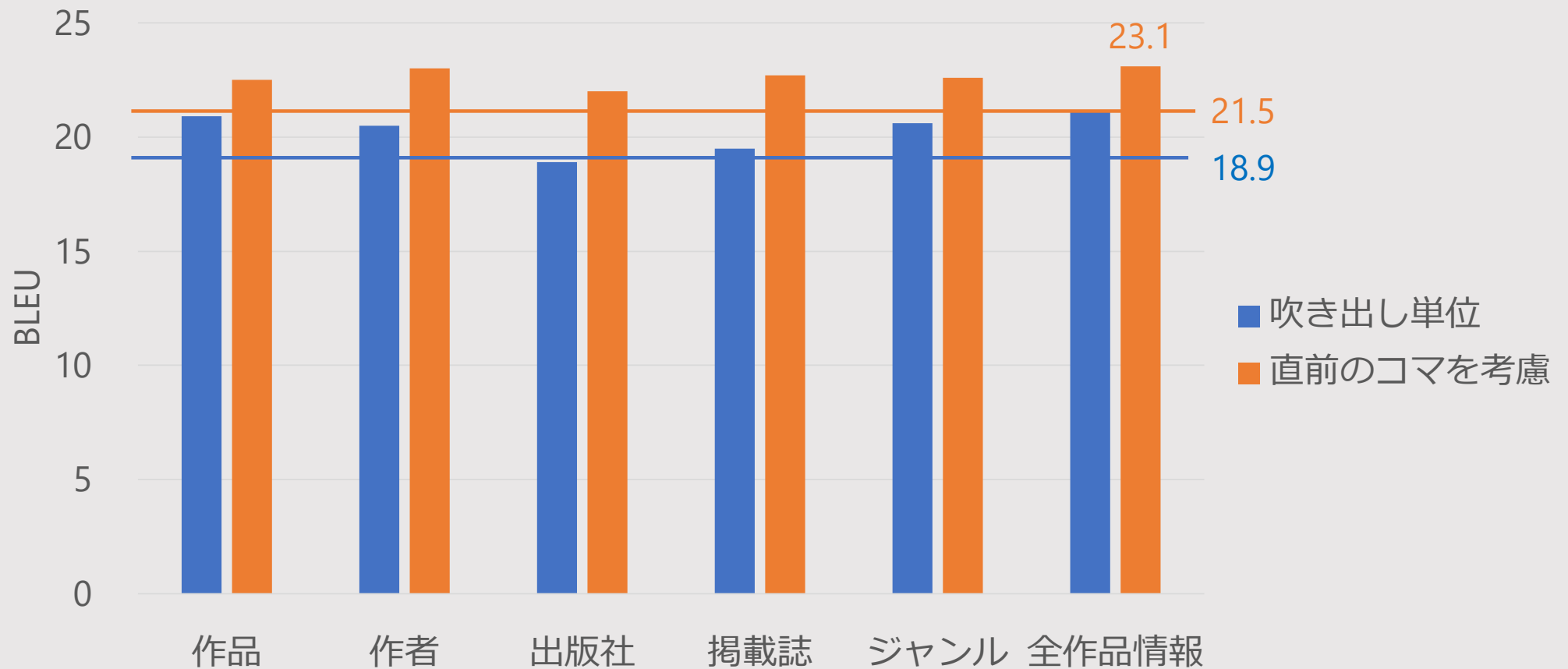
[7] Rico Sennrich et al. Neural Machine Translation of Rare Words with Subword Units, In Proc of ACL. 2016

[8] Vaswani et al. Attention is All You Need, In Proc. of NIPS.2017

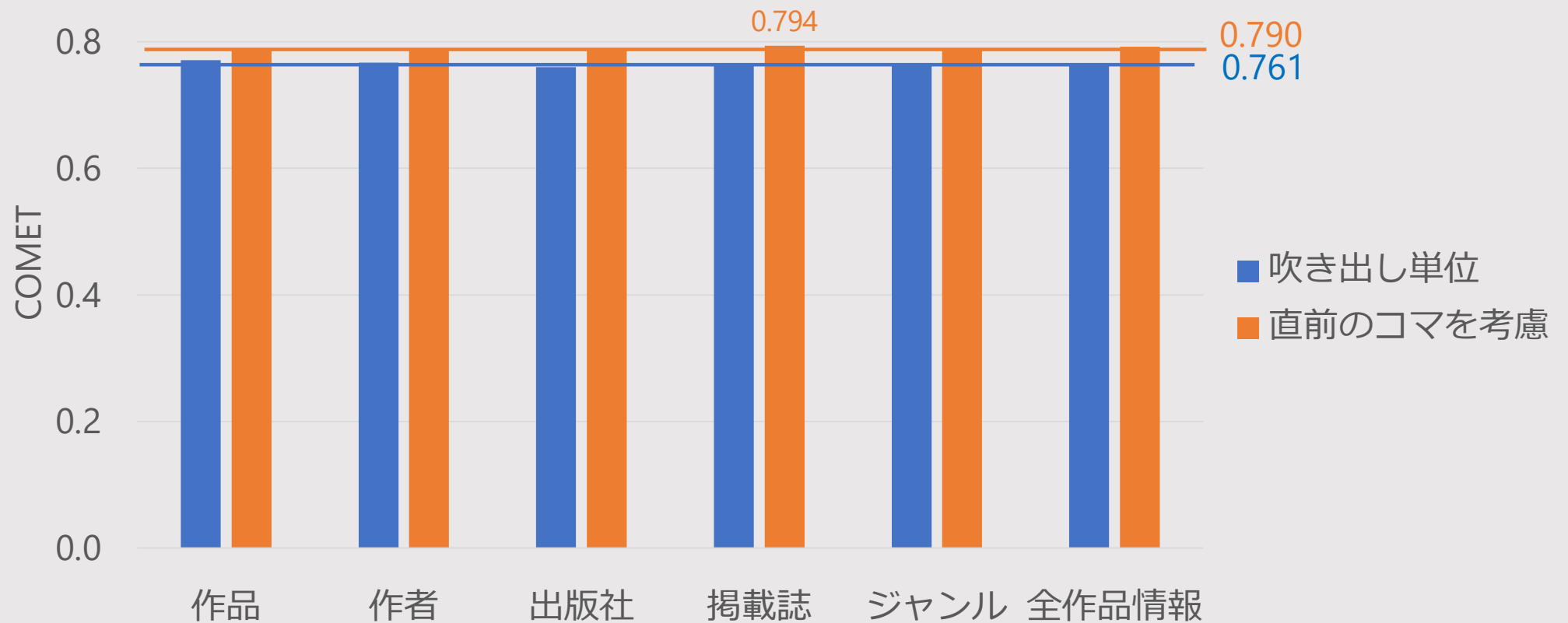
実験結果：直前のコマを考慮することで性能が向上



実験結果：漫画の属性情報を考慮することでBLEUが向上



実験結果：漫画の属性情報を考慮することでCOMETも向上



直前のコマを考慮するとコマ内がない主語を補う

	入力	出力
吹き出し単位での翻訳	実に夢がない	I don't have a dream.
直前の吹き出しを考慮	なのになぜ~狙う? <SEP> 実に夢がない	I don't have any dreams.
コマ単位での翻訳	なのになぜ~狙う? <SEP> 実に夢がない	I don't have any dreams.
直前のコマを考慮	十年後一夏 <コマ内> 相川仁 <コマ内> お前には~学力がある <コマ外>なのになぜ~狙う? <コマ外> 実に夢がない	You really don't have a dream.
正解文	実に夢がない	Don't you have aspirations?



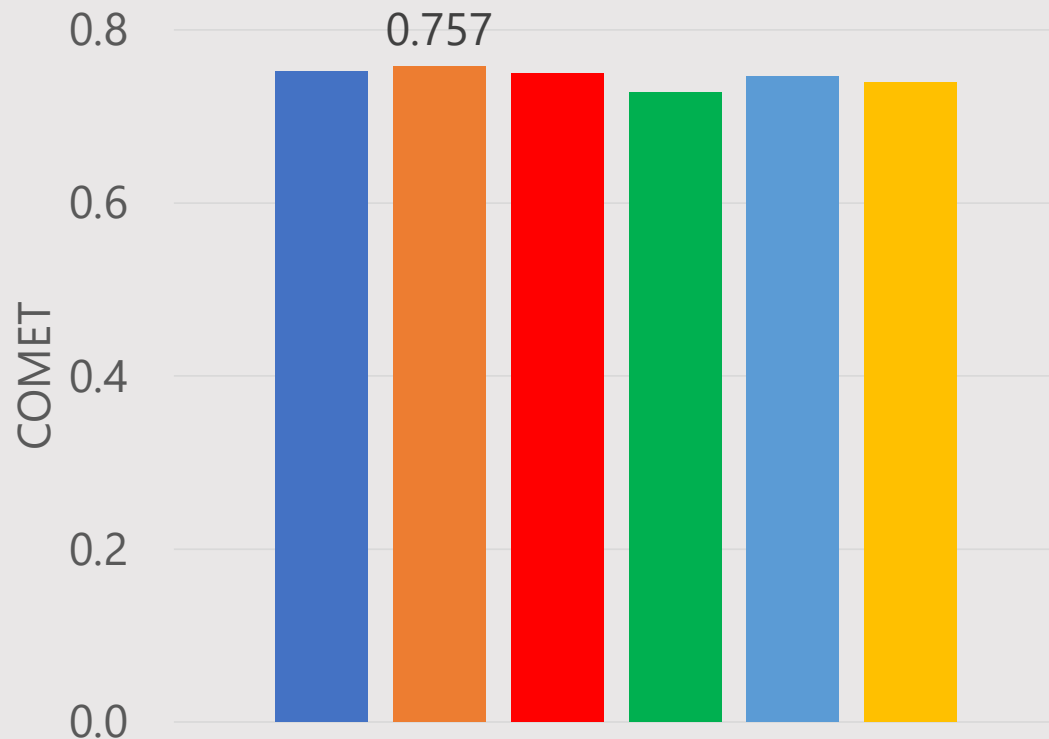
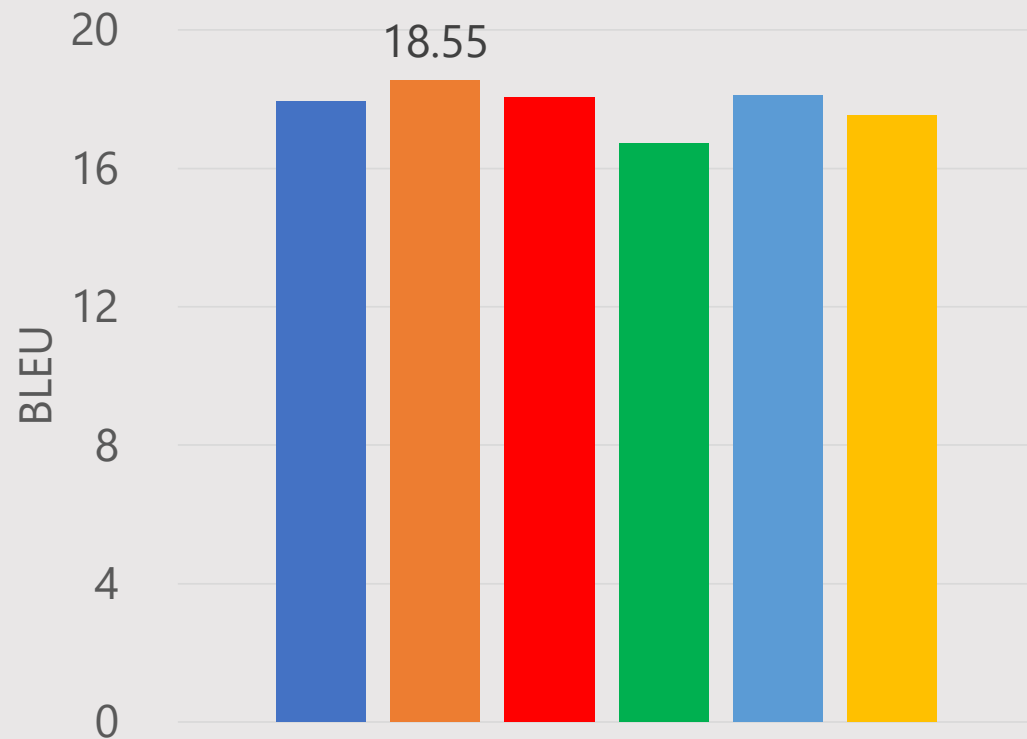
漫画の属性情報を考慮することで訳し分けが可能

「翔ぶ」の翻訳が変化

入力	ジャンル	出力
最初から翔ばないほうがいい	<少年漫画>	You shouldn't have jumped from the start.
	<野球>	You shouldn't have hit from the start.
	<ロボット>	You shouldn't have taken flight in the first place.
	<青年漫画>	It's better if you didn't fly from the start.

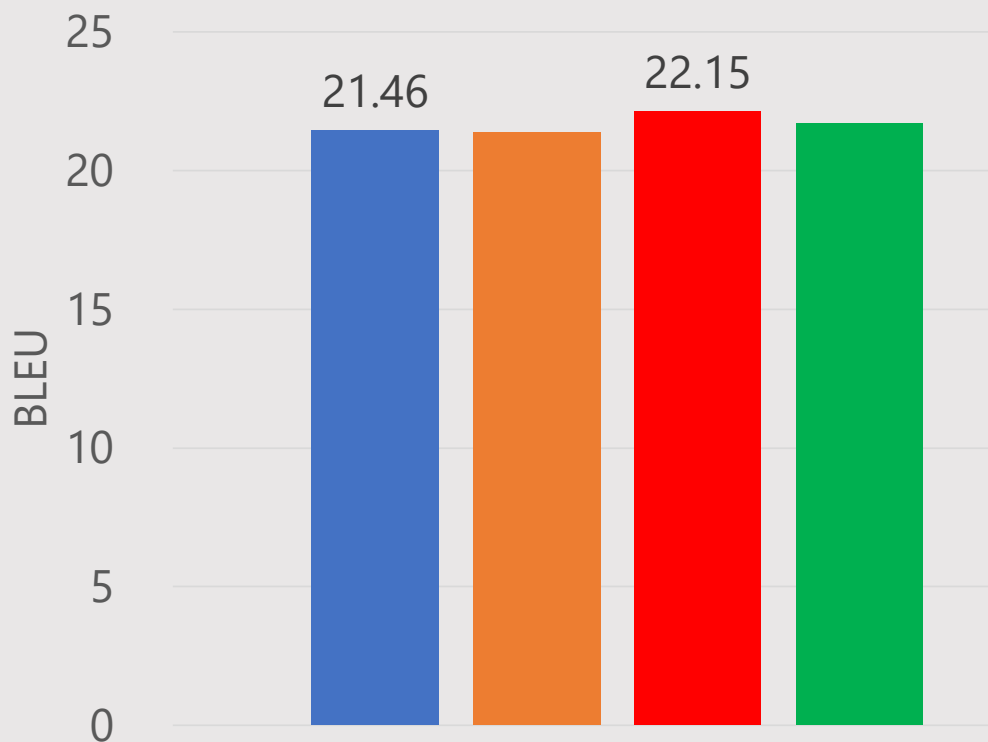
「~しないほうがいい」の翻訳が変化

直前のコマのみを用いるのが最も有効

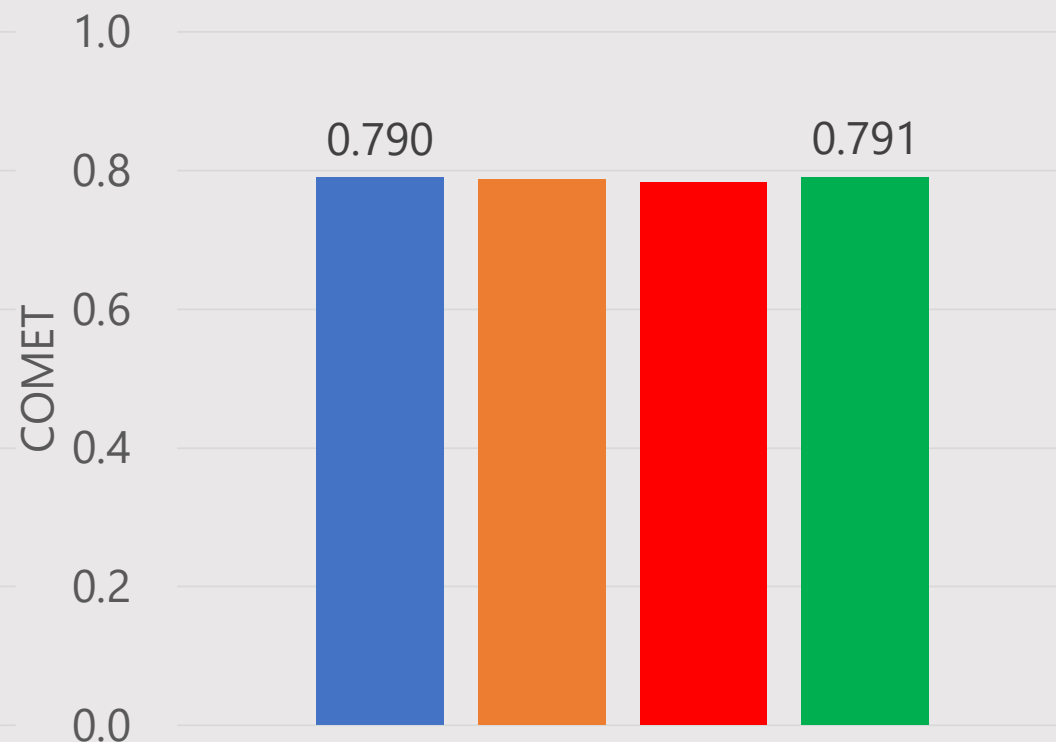


■ コマ単位での翻訳 ■ 直前のコマを考慮 ■ 前2コマまでを考慮 ■ コマ単位での翻訳 ■ 直前のコマを考慮 ■ 前2コマまでを考慮
■ 前3コマまでを考慮 ■ 前4コマまでを考慮 ■ 前5コマまでを考慮 ■ 前3コマまでを考慮 ■ 前4コマまでを考慮 ■ 前5コマまでを考慮

コマや吹き出しの順番を変えると性能が低下



- 直前のコマを考慮
- 直前のコマを考慮(コマの順番を入れ替え)
- 直前のコマを考慮(コマ内の吹き出しの順番を入れ替え)
- 次のコマを考慮



- 直前のコマを考慮
- 直前のコマを考慮(コマの順番を入れ替え)
- 直前のコマを考慮(コマ内の吹き出しの順番を入れ替え)
- 次のコマを考慮

まとめ

- 漫画の機械翻訳の品質改善のための長い文脈情報を扱う手法を提案した
 1. 直前のコマを考慮するコマ単位での翻訳
 2. 漫画の属性情報を考慮する翻訳
- 漫画を日本語から英語へ翻訳する実験の結果、2つの手法が漫画翻訳の品質を改善することを確認
- 2つの提案手法を組み合わせることで最高性能を達成した