第16回 AAMT長尾賞 受賞者講演

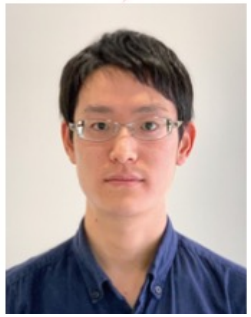
WMT-2020ニュース翻訳タスクに参加して：Team Tohoku-AIP-NTT at WMT-2020

Team Tohoku-AIP-NTT at WMT-2020

清野 舜[1,2] 伊藤 拓海[2,1] 今野 颯人[2,1] 森下 睦[3,2] 鈴木 潤[2,1]

[1]理研AIP [2]東北大学 [3]NTT コミュニケーション科学基礎研究所




代表発表 鈴木潤 (東北大学)

第16回 AAMT長尾賞 受賞のお礼

このたびは，WMT-2020ニュース翻訳タスクに参加した「Team Tohoku-AIP-NTT（チームメンバー：清野 舜, 伊藤 拓海, 今野 颯人, 森下 睦, 鈴木 潤）」が構築した機械翻訳システム，および，ニュース翻訳タスクにおける成績を高く評価していただき，第16回AAMT長尾賞に選出していただきましたこと，選考委員の方々をはじめ関係者の皆様に，この場をお借りして御礼申し上げます.

本日の話題

- 第16回 AAMT長尾賞受賞の経緯
  - WMT-2020 ニュース翻訳タスクへの参加，およびその結果
- 参加システムの概要
- いくつかの知見の共有
- コンペティションに参加することの意義
- まとめ

AAMT/Japio特許翻訳研究会主催の「特許情報シンポジウム」2021.2.26
に同様の内容を講演済み

- 第16回 AAMT長尾賞受賞の経緯

WMT概要

- 2006年に第1回WS開催
  - 以降毎年開催
- 第11回よりWSから国際会議に昇格
- 毎年併設で機械翻訳に関連するコンペティションを開催
  - 複数のタスクと言語毎に分けられたトラックが存在
  - 参加チームごとに参加タスクおよびトラックを選択しレギュレーションに従ってシステムを構築
  - システムは各タスクの評価基準に従い順位付け

機械翻訳の研究分野において最も有名かつ競争の激しいコンペティション

EMNLP 2020
FIFTH CONFERENCE ON
MACHINE TRANSLATION (WMT20)

November 19-20, 2020
Online

Home

[HOME] [SCHEDULE] [PAPERS] [VIRTUAL CONFERENCE] [RESULTS]
TRANSLATION TASKS: [NEWS] [LIFELONG LEARNING] [ROBUSTNESS] [SIMILAR LANGUAGES] [UNSUP AND VERY LOW RES] [BIOMEDICAL] [CHAT]
OTHER TASKS: [AUTOMATIC POST-EDITING] [METRICS] [QUALITY ESTIMATION] [PARALLEL CORPUS FILTERING]

NEW: *Draft* proceedings available from links above (updated 2020-12-22)

This conference builds on a series of annual workshops and conferences on statistical machine translation, going back to 2006:

- the NAACL-2006 Workshop on Statistical Machine Translation,
- the ACL-2007 Workshop on Statistical Machine Translation,
- the ACL-2008 Workshop on Statistical Machine Translation,
- the EACL-2009 Workshop on Statistical Machine Translation,
- the ACL-2010 Workshop on Statistical Machine Translation
- the EMNLP-2011 Workshop on Statistical Machine Translation,
- the NAACL-2012 Workshop on Statistical Machine Translation,
- the ACL-2013 Workshop on Statistical Machine Translation,
- the ACL-2014 Workshop on Statistical Machine Translation,
- the EMNLP-2015 Workshop on Statistical Machine Translation,
- the First Conference on Machine Translation (at ACL-2016),
- the Second Conference on Machine Translation (at EMNLP-2017),
- the Third Conference on Machine Translation (at EMNLP-2018).
- the Fourth Conference on Machine Translation (at ACL-2019).

https://www.statmt.org/wmt20/

コンペティションの流れ

- 日程

IMPORTANT DATES

Release of training data for shared tasks (by)	29 February, 2020
Test suite source texts must reach us	~~May 30~~ June 12, 2020
Test data released	~~June 8~~ June 22, 2020
Translation submission deadline	~~June 15~~ June 29, 2020 (Anywhere on Earth)
Translated test suites shipped back to test suites authors	~~July 2~~ July 6, 2020
Start of manual evaluation	TBD July 2020
End of manual evaluation	TBD July 2020

- 約3ヶ月強のシステム開発期間
- 約1週間の評価データの翻訳作成期間
  （標準的なコンペティションの開催形態）

- 評価はBLEUスコアによる自動評価（一次評価）および人手評価（最終評価）

WMTのニュース翻訳タスク

- 概要
  - 第1回から続くWMTを代表するタスク
  - WMTコンペティションを代表する伝統的なタスク
  - 毎年多くの参加チームで熾烈な争い
  - 機械翻訳の研究分野に多大な影響を与えてきた
    - データ，方法論，

- 主催者からの主な配布物

対訳コーパス

File	CS-EN	DE-EN	IU-EN	JA-EN	KM-EN	PL-EN	PS-EN	RU-EN	TA-EN	ZH-EN	FR-DE
Europarl v10	✓	✓				✓					✓
ParaCrawl v5.1	✓	✓		✓	✓	✓	✓	✓			✓
Common Crawl corpus	✓	✓						✓			✓
News Commentary v15	✓	✓		✓				✓		✓	✓
CzEng 2.0	✓										
Yandex Corpus								✓			

単言語コーパス

Corpus	CS	DE	EN	FR	IU	JA	KM	PL	PS	RU	TA	ZH
News crawl	✓	✓	✓	✓		✓		✓		✓	✓	✓
News discussions			✓	✓								
Europarl v10	✓	✓	✓	✓				✓				
News Commentary	✓	✓	✓	✓		✓				✓		✓
Common Crawl	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Wiki dumps			✓				✓		✓		✓	

昨年までの評価データ（開発用データ）

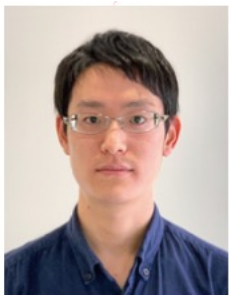
Year	CS-EN	DE-EN	IU-EN	JA-EN	KM-EN	PL-EN	PS-EN	RU-EN	T
2008	✓	✓							
2009	✓	✓							
2010	✓	✓							
2011	✓	✓							
2012	✓	✓						✓	
2013	✓	✓						✓	
2014	✓	✓						✓	
2015	✓	✓						✓	
2016	✓	✓						✓	
2017	✓	✓						✓	
2018	✓	✓						✓	
2019	✓	✓						✓	
2020 (dev)			✓	✓	✓	✓	✓		

チーム構成 / 参加トラック

- チーム構成







理研 （学生研究員/客員研究員） NTT

（社会人博士） 東北大学 （社会人博士）

- 参加トラック
  - ① 英語 -> ドイツ語　② ドイツ語-> 英語
    ③ 英語 -> 日本語　④ 日本語 -> 英語
    の 4トラックに参加

WMT-2020 参加トラックの結果 (1/3)

- 自動評価結果（BLEUスコア）

独→英: 1位

Team	BLEU
Tohoku-AIP-NTT	43.8
Huoshan_Translate	43.5
OPPO	43.2
UEDIN	42.3
Online-B	41.9

英→独: 1位

Team	BLEU
Tohoku-AIP-NTT	38.8
Tencent_Translation	38.6
OPPO	38.6
Huoshan_Translate	38.2
eTranslation	37.9

日→英: 2位

Team	BLEU
NiuTrans	26.7
Tohoku-AIP-NTT	25.5
OPPO	24.8
NICT_Kyoto	22.8
eTranslation	22.2

英→日: 4位

Team	BLEU
NiuTrans	28.4
OPPO	27.3
ENMT	25.9
Tohoku-AIP-NTT	25.8
NICT_Kyoto	23.9

WMT-2020 参加トラックの結果 (2/3)

- 人手評価結果

人間，または，オンライン翻訳システム
(本タスクに参加していないシステム)

独 → 英: 同率1位

Ave.	Ave. z	System
82.6	0.228	VolcTrans
84.6	0.220	OPPO
82.2	0.186	HUMAN
81.5	0.179	Tohoku-AIP-NTT
81.3	0.179	Online-A
81.5	0.172	Online-G
79.8	0.171	PROMT-NMT
82.1	0.167	Online-B

英 → 独: 同率1位

Ave.	Ave. z	System
90.5	0.569	HUMAN-B
87.4	0.495	OPPO
88.6	0.468	Tohoku-AIP-NTT
85.7	0.446	HUMAN-A
84.5	0.416	Online-B
84.3	0.385	Tencent-Translation
84.6	0.326	VolcTrans
85.3	0.322	Online-A
82.5	0.312	eTranslation

日 → 英: 同率1位

Ave.	Ave. z	System
75.1	0.184	Tohoku-AIP-NTT
76.4	0.147	NiuTrans
74.1	0.088	OPPO
75.2	0.084	NICT-Kyoto
73.3	0.068	Online-B
70.9	0.026	Online-A
71.1	0.019	eTranslation
64.1	−0.208	zlabs-nlp
66.0	−0.220	Online-G
61.7	−0.240	Online-Z

英 → 日: 同率1位

Ave.	Ave. z	System
79.7	0.576	HUMAN
77.7	0.502	NiuTrans
76.1	0.496	Tohoku-AIP-NTT
75.8	0.496	OPPO
75.9	0.492	ENMT
71.8	0.375	NICT-Kyoto
71.3	0.349	Online-A
70.2	0.335	Online-B
63.9	0.159	zlabs-nlp
59.8	0.032	Online-Z

注) 人手評価に統計的有意差がない場合，同率一位という扱い

WMT-2020 参加トラックの結果 (3/3)

- 他の論文中の評価でも好成績

[Mathur+, WMT'20] Results of the WMT20 Metrics Shared Task

Evaluation	OPPO	TOHOKU
avg metric (HUMAN-A ref)	8.85	**8.95**
avg metric (Human-B ref)	10.15	**10.26**
WMT	**84.6**	81.5
z-score	**0.220**	0.179
prof. linguist	81.0	**81.7**
z-score	-0.005	**0.010**

Table 13: WMT 2020 German→English comparing the reference-based ratings acquired with crowd workers/researcher (WMT) against source-based ratings acquired with professional linguists.

Evaluation	OPPO	TOHOKU	HUMAN-A
avg metric (Human-B ref)	10.05	**10.09**	9.14
avg metric (Human-P ref)	11.93	12.07	**15.74**
WMT	87.39	**88.62**	85.10
z-score	**0.495**	0.468	0.379
prof. linguist	73.66	74.70	**84.09**
z-score	-0.051	-0.037	**0.088**

Table 14: WMT 2020 English→German comparing the source-based ratings acquired with crowd workers/researcher (WMT) against source-based ratings acquired with professional linguists.

The results in Table 13 show that professional linguists in fact prefer the output of TOHOKU as predicted by all automatic metrics.

[Avramidis+, WMT'20] Fine-grained linguistic evaluation for state-of-the-art Machine Translation

category	items	Tohoku	VolcTrans
Ambiguity	81	**82.7**	**77.8**
Composition	49	**98.0**	**98.0**
Coordination & ellipsis	78	**89.7**	**91.0**
False friends	36	**72.2**	**80.6**
Function word	72	**86.1**	**80.6**
LDD & interrogatives	174	**89.1**	**86.2**
MWE	80	**80.0**	**75.0**
Named entitiy & terminology	89	**92.1**	**84.3**
Negation	20	**100.0**	**100.0**
Non-verbal agreement	61	**91.8**	**88.5**
Punctuation	60	**96.7**	**98.3**
Subordination	180	**90.6**	**88.3**
Verb tense/aspect/mood	4447	**84.6**	**85.3**
Verb valency	87	**79.3**	**81.6**
micro-average	5514	**85.3**	**85.4**
macro-average	5514	**88.1**	86.8
BLEU		43.8	43.5

複単語表現，固有表現，機能語，動詞の時制の取り扱い，など

Tohoku achieves the best category macro-averaged accuracy of 88.1%, whereas it is sharing the first position with VolcTrans based on their micro-averaged accuracy (85.3-85.4%).

AAMT長尾賞受賞

- 受賞理由

機械翻訳の分野で権威の高い国際会議WMT2020のニュース記事翻訳タスクにおいて、参加した4言語対トラック(英独、独英、日英、英日)の

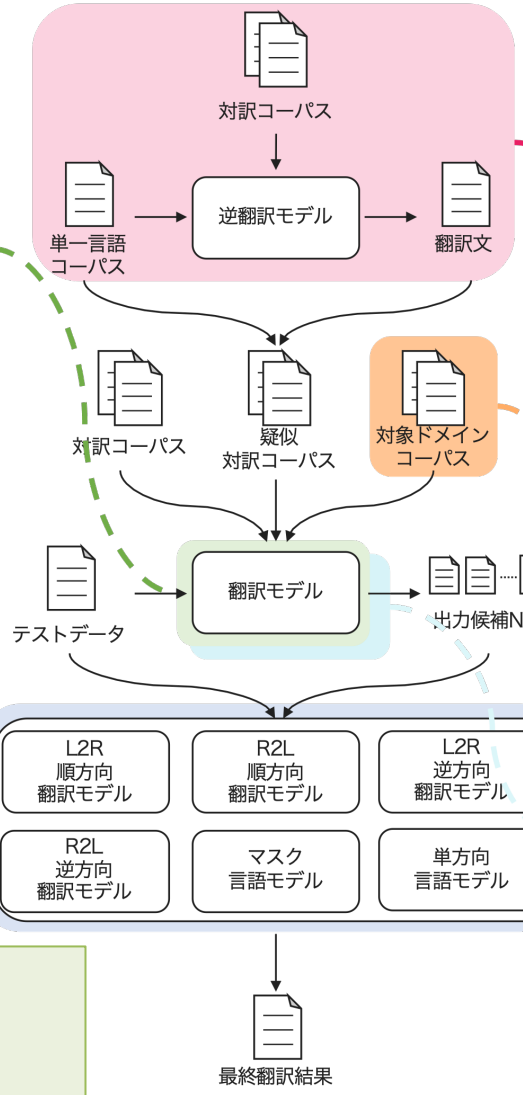
① 人手評価において全て(同率を含む)一位を獲得し、

② 英独および独英のトラックでは自動評価でも一位を獲得した。さらに、

③ 言語学者による言語現象別の翻訳品質評価においても他のシステムを大幅に上回る最良の評価を得るなど、非常に高品質な機械翻訳システムを実現し、国際的に高く評価されている。これらの成果は、機械翻訳システムの実用化に向けた日本の研究開発力の国際的プレゼンス向上に大きく貢献し、また得られた多くの知見を研究開発者に共有する活動を通じて機械翻訳システムの開発と実用化の促進にも大きく貢献している。大学、研究機関、企業の産官学連携による優れた成果であり、まさにAAMT長尾賞にふさわし。

- システム概要

コンペテイション参加システムの概略



①
対訳コーパスだけではデータが足りない
単一言語コーパスを使ってデータを増やしたい
⇛**逆翻訳によるデータ拡張**

②
翻訳モデル（Transformer）はハイパーパラメータに対して非常にセンシティブ
ハイパラの知見も日進月歩
⇛**ハイパーパラメータの調整**

③
テストデータは新聞記事ドメインなので、
新聞記事のデータに適応させたい
⇛**ファインチューニング**

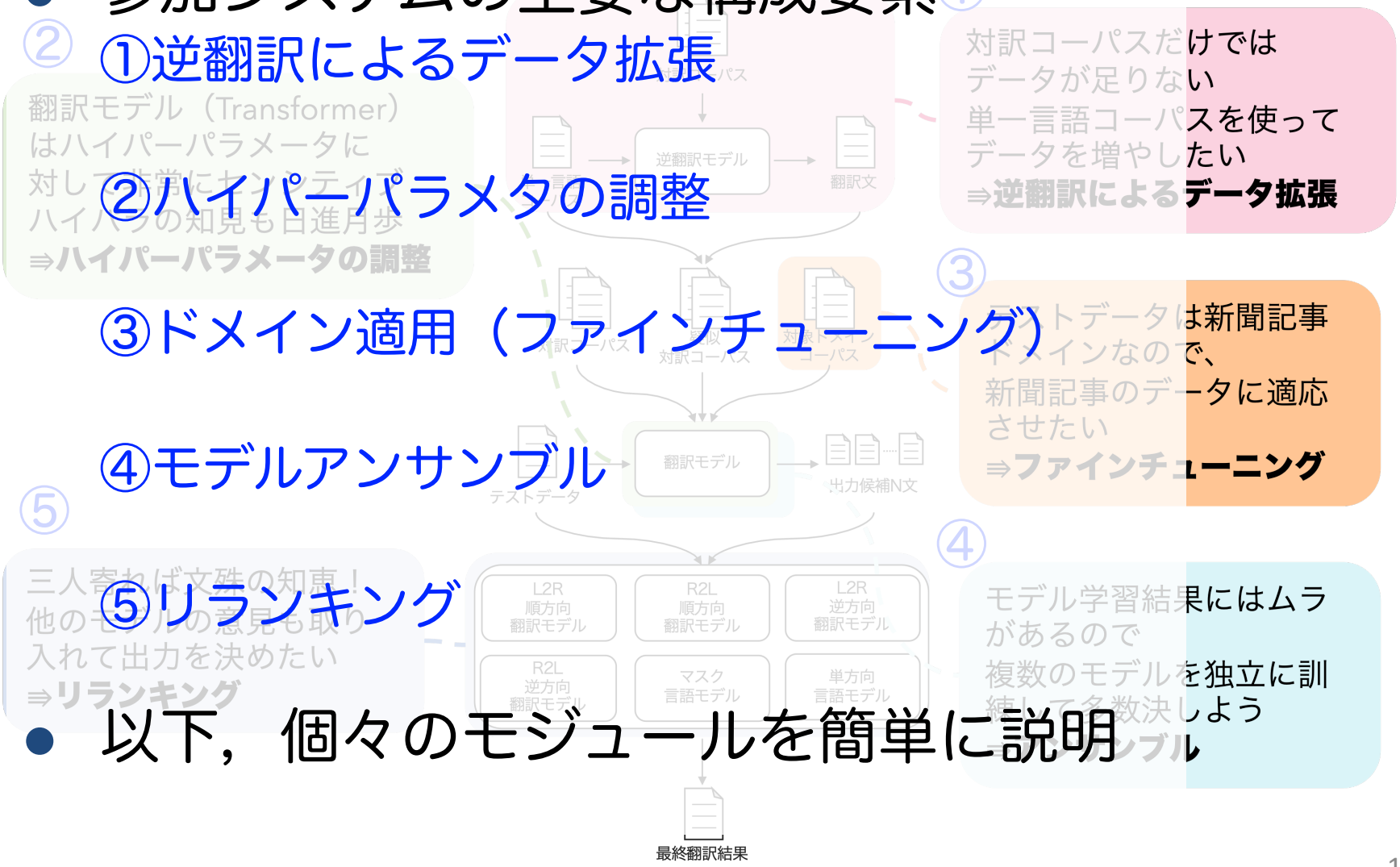
④
モデル学習結果にはムラがあるので
複数のモデルを独立に訓練して多数決しよう
⇛**アンサンブル**

⑤
三人寄れば文殊の知恵！
他のモデルの意見も取り入れて出力を決めたい
⇛**リランキング**

なんかとても複雑なシステム？

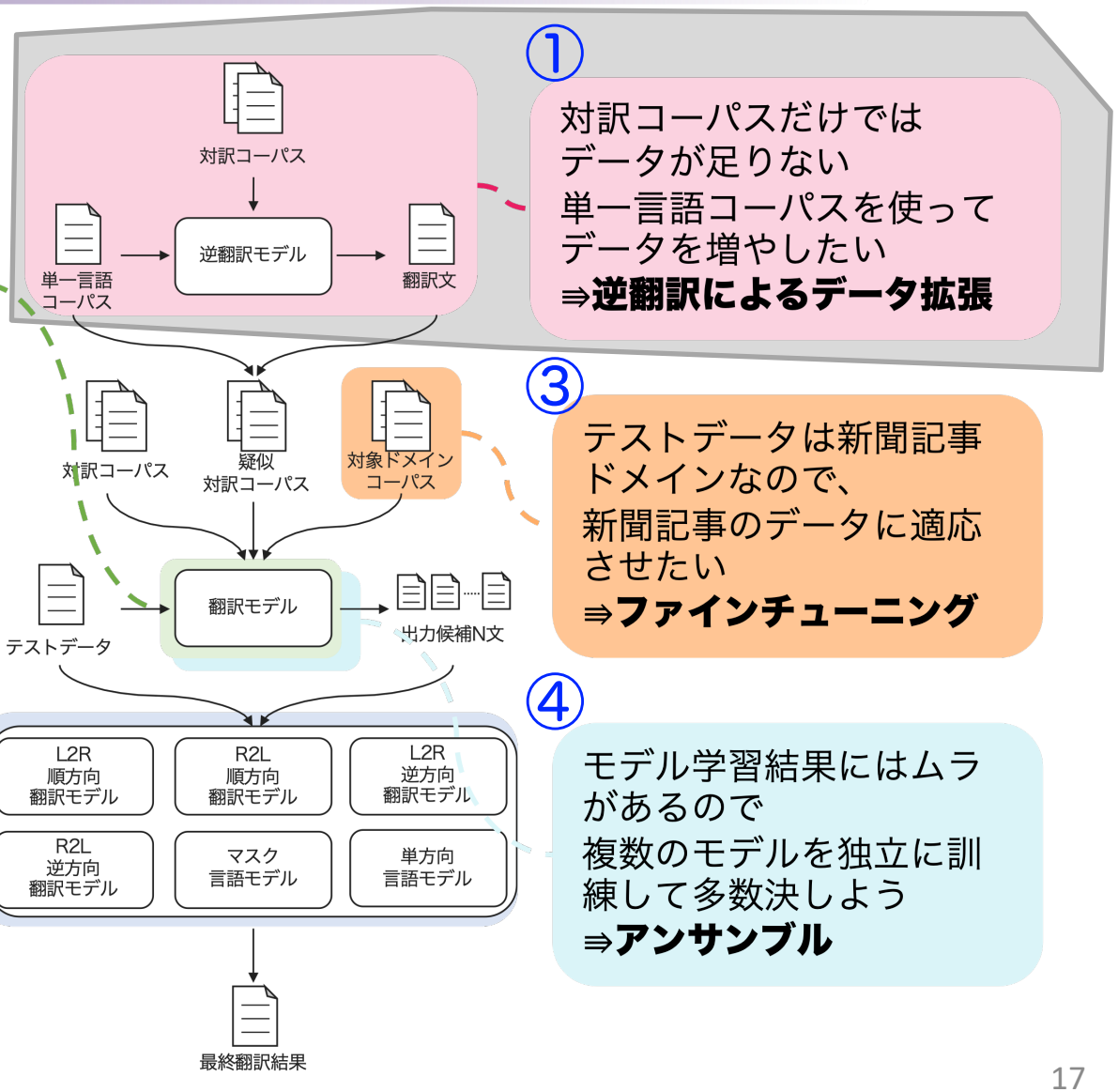
コンペティション参加システムの概略

- 参加システムの主要な構成要素
  - ①逆翻訳によるデータ拡張
  - ②ハイパーパラメタの調整
  - ③ドメイン適用（ファインチューニング）
  - ④モデルアンサンブル
  - ⑤リランキング
- 以下，個々のモジュールを簡単に説明



[再掲]

コンペティション参加システムの概略



①逆翻訳によるデータ拡張

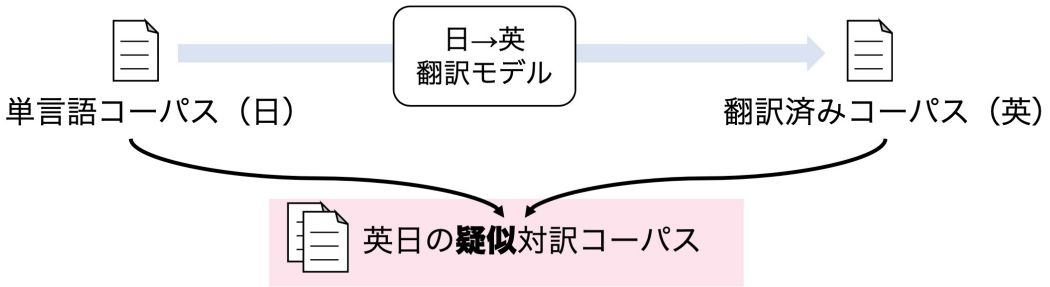
[Sennrich+2016]

- 対象とする翻訳方向とは逆向きの翻訳モデルにより目的言語の文から原言語文を生成
  => 擬似データとして翻訳モデルの学習に活用

手順例

① 逆翻訳モデルの訓練

② 日本語単言語コーパスを翻訳し、疑似データを生成

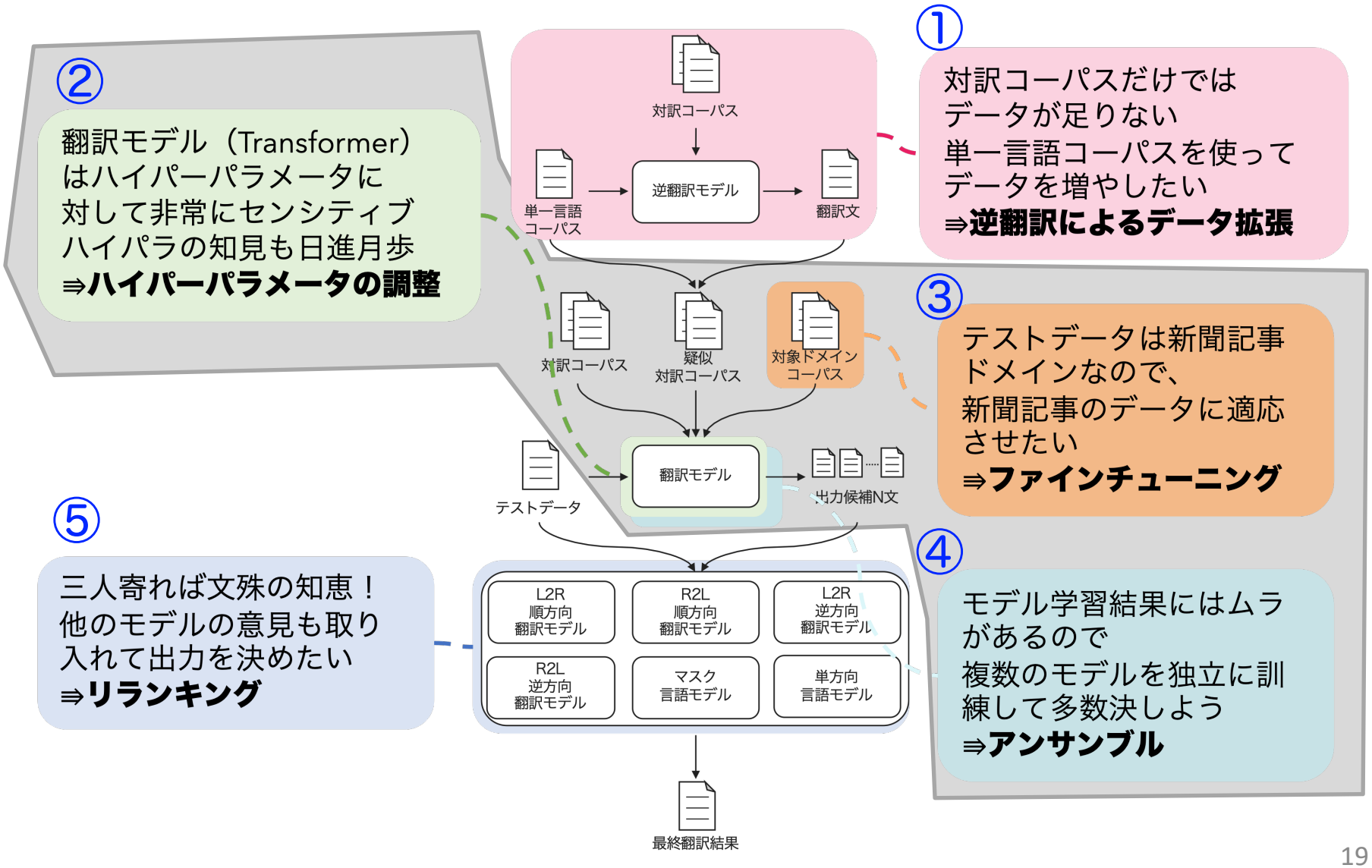


③ 疑似データを用いた学習

単純なトリックとしては，
日 => 英，英 => 日
独 => 英，英 => 独
のタスクに参加しているので，逆方向の翻訳モデルをそのまま使えば良い
（特別な処理ではない）

[再掲]

コンペティション参加システムの概略



②ハイパーパラメタの調整

- ベースモデル: Transformer [Vaswani+2017]
  - 近年のデファクトスタンダードなモデル

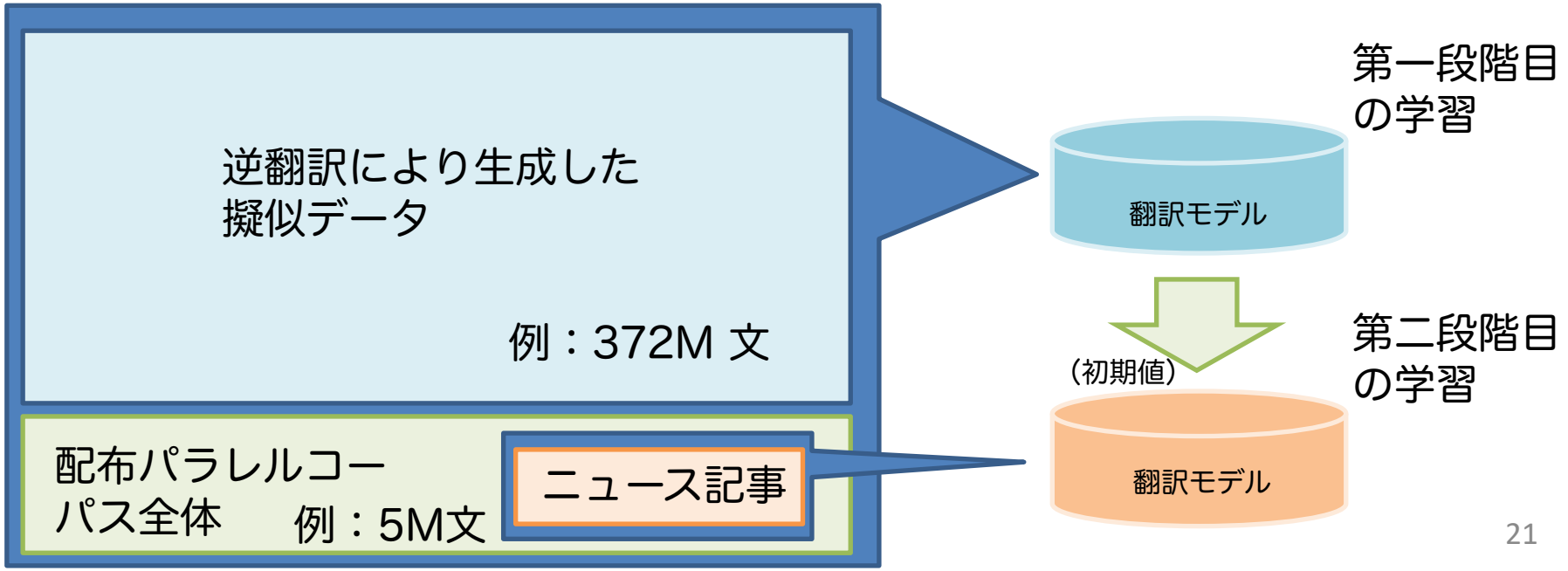
- 選択したハイパーパラメタ
  - FF層の次元数/layer数の増加
    - 次元数を倍，層の数 6 => 9
  - Larger batch size [Ott+2018]
    - 通常 4,000 トークン => 512,000
  - 学習率増加 [Ott+2018]
    - Adamのstep size 0.0005 => 0.001
  - チェックポイント平均法
    - 各10epoch毎に直近10モデルで平均化
    - BLEUスコアが改善すると報告 [Popel+2018]
  - layer-normalization
    - Pre-normを採用
    - 多層Transformerの学習が安定するとの報告 [Xiong+2018]

様々な論文に書かれている知見を集約

③ドメイン適用（ファインチューニング）

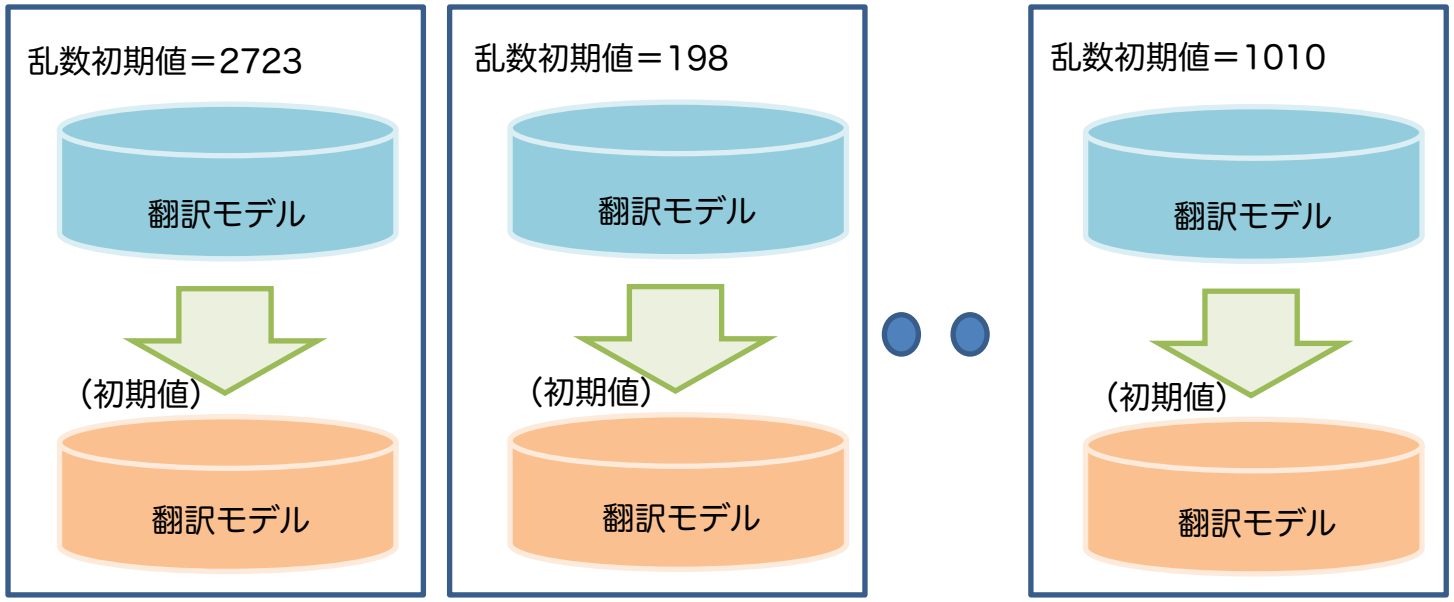
- 参加タスクはニュース翻訳 => ニュース記事による学習データのみを使ってベースモデルを再学習

- 基本的にドメイン適用の際の処理と同じ



④モデルアンサンブル

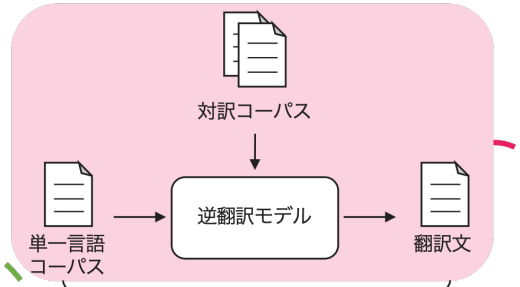
- 乱数シードの違う複数のモデルを学習

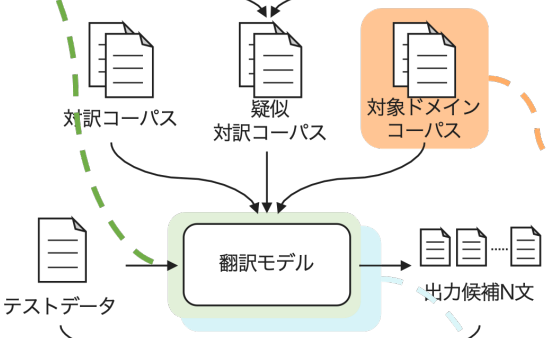


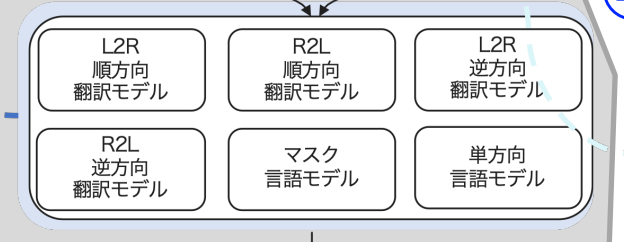
- 翻訳時の各トークンを予測する際に用意した全てのモデルの予測結果を用いて最終的に選択すべきトークン列を予測

[再掲]

コンペティション参加システムの概略







①

対訳コーパスだけではデータが足りない
単一言語コーパスを使ってデータを増やしたい
⇛**逆翻訳によるデータ拡張**

②

翻訳モデル（Transformer）はハイパーパラメータに対して非常にセンシティブ
ハイパラの知見も日進月歩
⇛**ハイパーパラメータの調整**

③

テストデータは新聞記事ドメインなので、
新聞記事のデータに適応させたい
⇛**ファインチューニング**

④

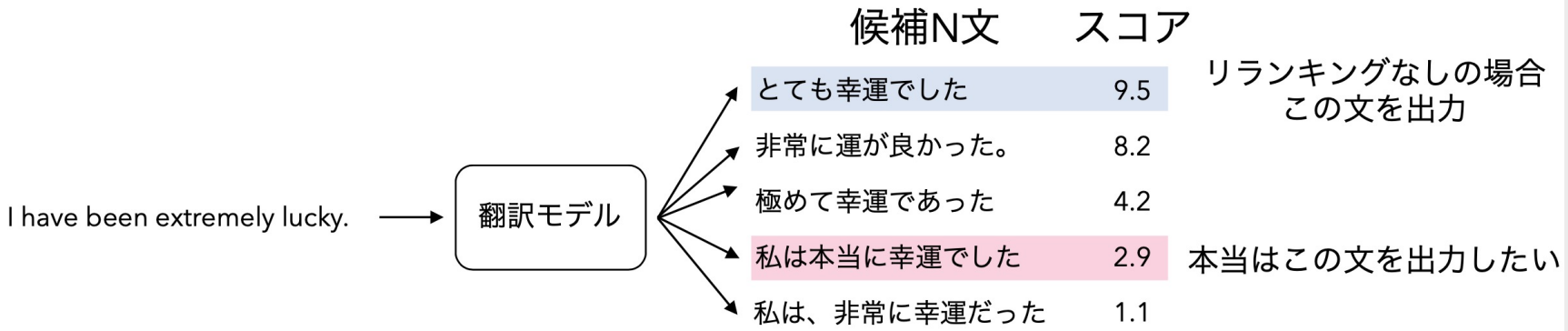
モデル学習結果にはムラがあるので
複数のモデルを独立に訓練して多数決しよう
⇛**アンサンブル**

⑤

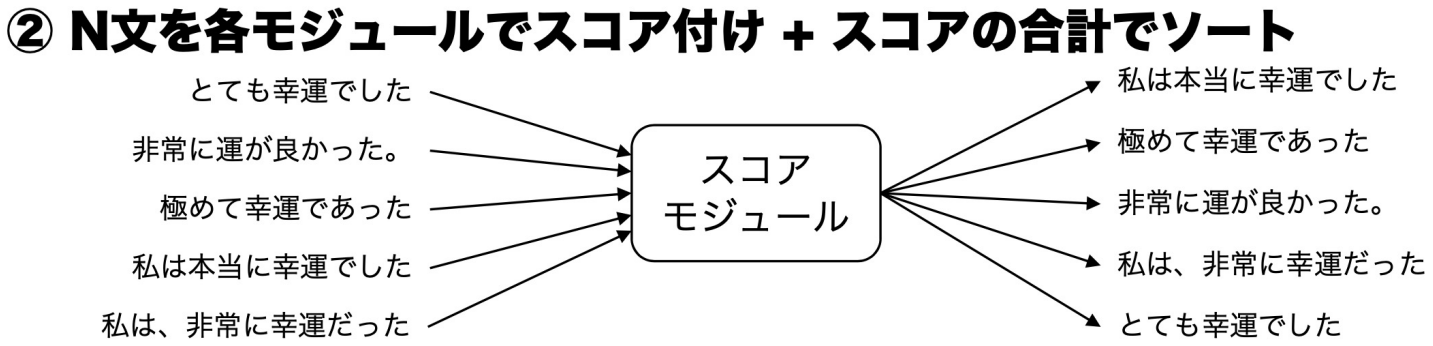
三人寄れば文殊の知恵！
他のモデルの意見も取り入れて出力を決めたい
⇛**リランキング**

⑤リランキング

- 複数の候補を生成し別のスコア決定モジュールにより生成結果を再評価



- 本システムでは複数のモデルが出す候補をまとめてリランキング



[再掲]
コンペティション参加システムの概略

TOHOKU UNIVERSITY

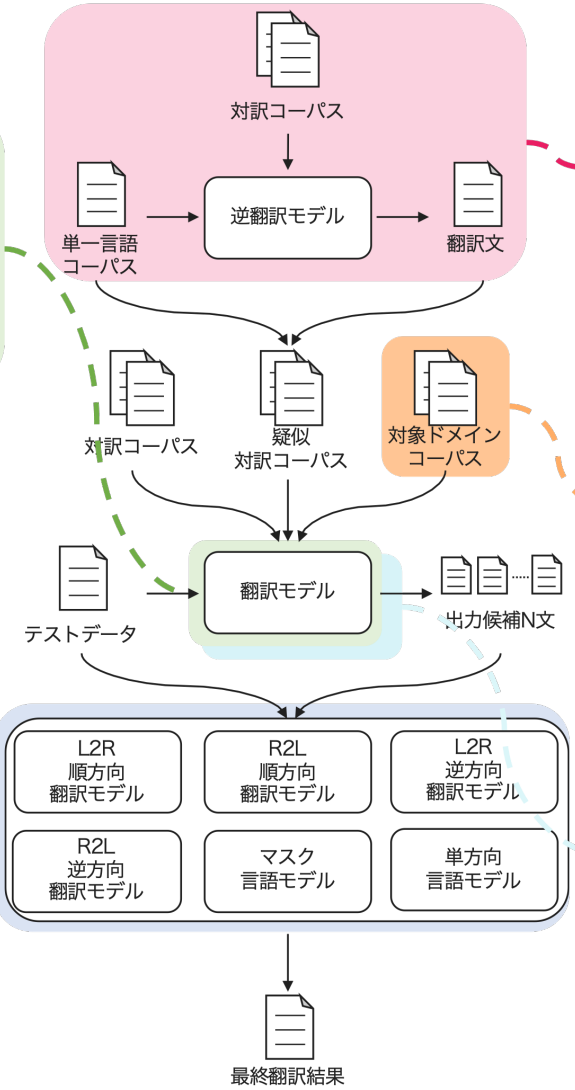
①
対訳コーパスだけではデータが足りない
単一言語コーパスを使ってデータを増やしたい
⇛**逆翻訳によるデータ拡張**

②
翻訳モデル（Transformer）はハイパーパラメータに対して非常にセンシティブ
ハイパラの知見も日進月歩
⇛**ハイパーパラメータの調整**

③
テストデータは新聞記事ドメインなので、
新聞記事のデータに適応させたい
⇛**ファインチューニング**

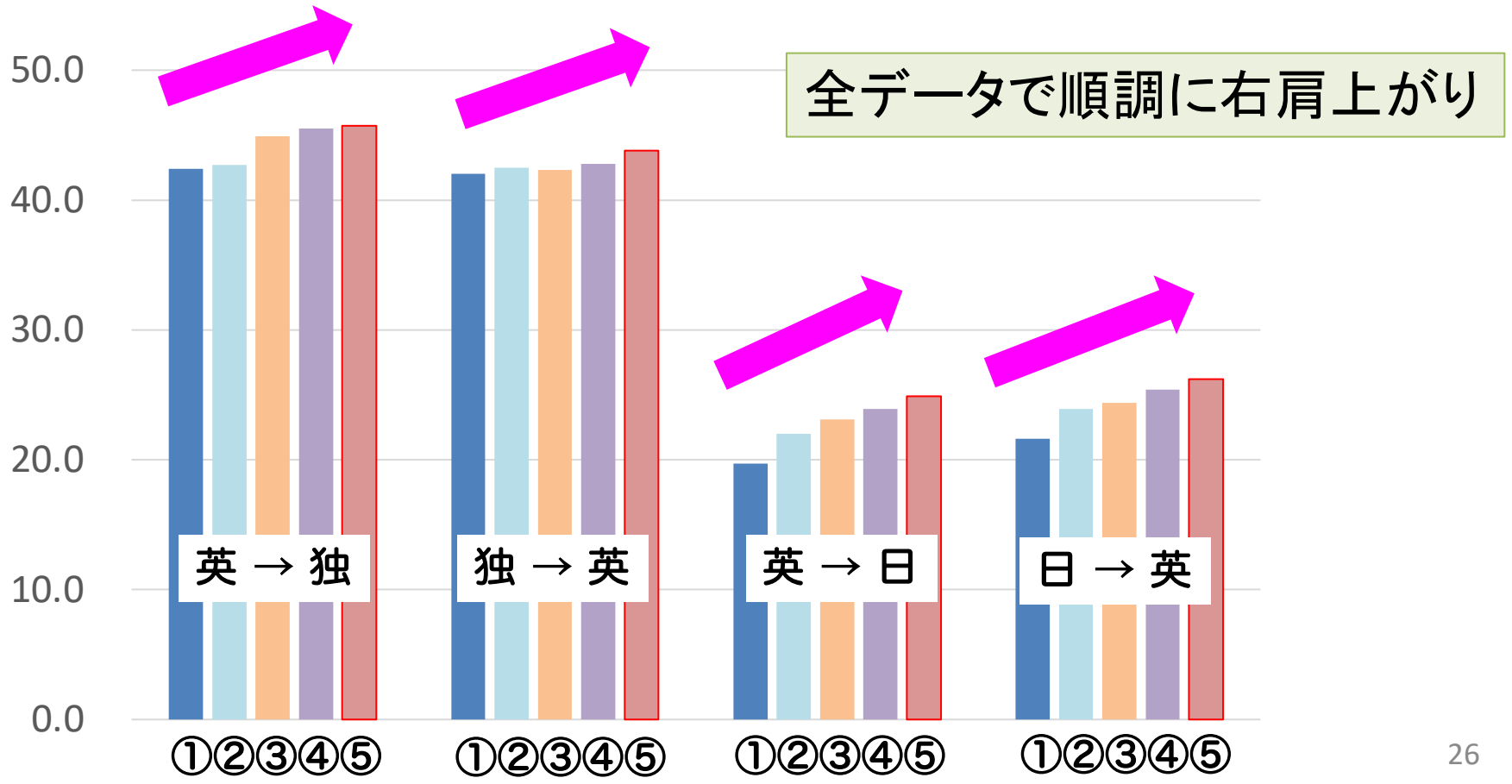
④
モデル学習結果にはムラがあるので
複数のモデルを独立に訓練して多数決しよう
⇛**アンサンブル**

⑤
三人寄れば文殊の知恵！
他のモデルの意見も取り入れて出力を決めたい
⇛**リランキング**



各モジュールの効果

- WMT-2019の評価データでの評価



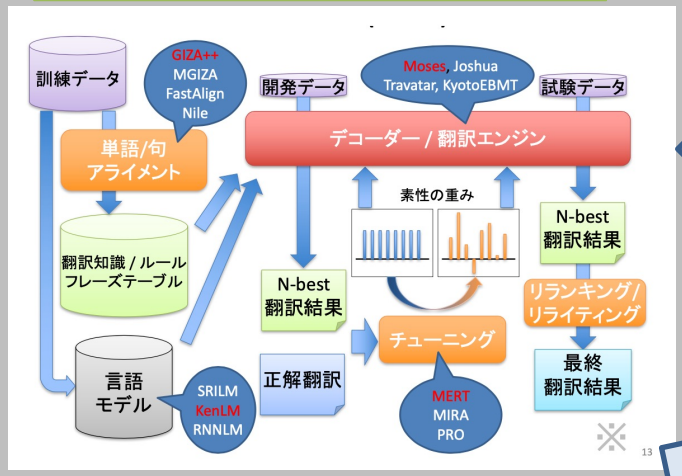
- いくつかの知見
  - ニューラル翻訳なのに複雑なモデル
  - うまくいかなかったモジュールの例
  - 計算リソース

- いくつかの知見
  - ニューラル翻訳なのに複雑なモデル

システム構成が複雑

統計翻訳 SMT

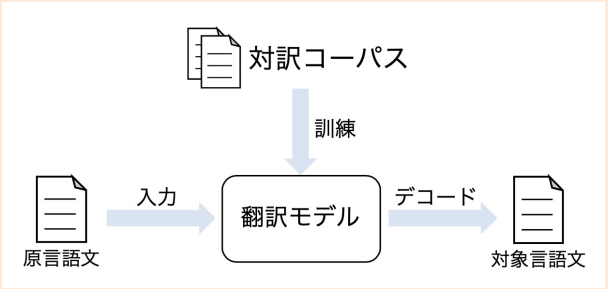
複雑！　メンテが大変！



一般論

ニューラル翻訳（通常版）

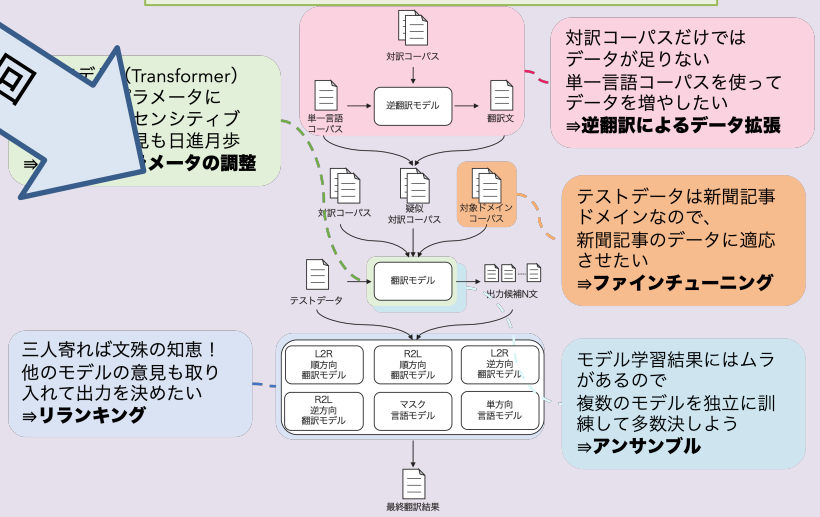
単純！　管理が楽！



今回

ニューラル翻訳（高性能版）

複雑！　メンテが大変！



※ゼロから始めるニューラルネットワーク機械翻訳
https://www.slideshare.net/ToshiakiNakazawa/nlp2017-nmt-tutorial
より図を引用

NMTといえど，性能を追求すると結局全体構成は複雑になり管理コストは高くなる

- いくつかの知見
  - うまくいかなかったモジュールの例

試行錯誤

- 各モジュールを組み合わせた結果は順調に右肩上がり
  - ただし，それは性能が向上しなかったモジュールを除外して評価しているから
  - 実際には最終的に採用したモジュール以外にも様々な試行錯誤をして最終形に到達

- 試行錯誤の例：データ選択
  - 逆翻訳のデータや対訳コーパスでも品質の良くない対訳が存在 => 取り除けば性能は上がるはず
  - しかし，残念ながら何度やっても良くならず断念

- いくつかの知見
  - 計算リソース

計算リソース

- 参加システムは合計11モデルで構成
  - 逆翻訳用の翻訳モデル
  - 順方向（left-to-right）デコーダモデル x4
  - 逆方向（right-to-left）デコーダモデル x4
  - リランキング用マスク型言語モデル
  - リランキング用単方向言語モデル
- 一つのモデルを学習するのに必要なリソース
  - V100 32GB 16日間相当
  - 4タスク全体だと x42 になる
  - システム作りの試行錯誤も合わせると膨大なリソースがどうしても必要になってしまう

- コンペティション参加のススメ

コンペティションに参加することの意義

- 世界の技術レベルを知ることができる
- 自分たちの技術レベルを知ることができる
- 普段の研究/開発では見えてこない難しさや改善点などが見えてくる
  - 論文の闇が見えてくることも．．．
- 最先端研究に必要な技能獲得や実験環境の獲得

最先端研究に必要な技能獲得や実験環境の獲得

- 1位（最高性能）を獲得することが「最終目的」ではない
  - その過程で得られる知見やノウハウ・経験が超重要
  - 大規模な実験計画への慣れ
- 最先端の研究への敷居を下げる
  - 例えば一位をとれば，それがすぐに最先端研究のベースラインシステムに

[余談]

- それでも有名コンペティションで一位になれば無名でも注目される?
  - 特に若手研究者には自分の技術を高める，自分を売り出していく場としてうまくコンペティションを活用して欲しい

- 全体のまとめ

まとめ: Take home messages

- 第16回 AAMT長尾賞受賞への経緯
  - WMT-2020 ニュース翻訳タスクへの参加，およびその結果
  - システム概要説明といくつかの知見共有

- コンペティション参加のススメ
  - コンペティションでのシステム構築には研究の実験管理などの技術と培う良い訓練の場
  - 勝つことは目標，だが最終ゴールではない
    - => とはいえ良い成績を獲得すれば有名になれる？
  - その過程で得られる知見やノウハウ・経験が超重要