

翻訳時に必要な情報を参照しない 機械翻訳が生じる問題

藤田 篤



情報通信研究機構

自己紹介: 藤田 篤 (ふじた あつし)



■ 国立研究所の職業研究者 (2014～)

- 情報通信研究機構 (NICT)
 - ユニバーサルコミュニケーション研究所
 - 先進的音声翻訳研究開発推進センター
 - ✓ 先進的翻訳技術研究室 主任研究員

■ 専門: 計算言語学, 自然言語処理

- とくに同義の言語表現の処理に関心

	同一の言語内	異なる言語間
<code>set <string> y = generate(x; 𐄂);</code>	言い換え生成	機械翻訳
<code>bool z = classify(x, y; 𐄂);</code>	言い換え認識 (同義知識獲得)	翻訳品質推定 (翻訳知識獲得)

■ 現在分担しているプロジェクト

- NICTの中長期計画: 2021-2026, 機械翻訳の性能向上
- 科研費基盤研究(S): 2019-2024, 翻訳規範とコンピテンスの外在化

今回の講演

■ ゴール: 機械翻訳(machine translation; MT)の課題の共有

- 翻訳とは異なり, 必要な情報を参照していないこと
- それによって生じる問題の分析方法
- ニュース記事の英日翻訳の誤り分析結果

■ 構成

- MTの概要
 - 情報通信研究機構(NICT)における研究開発
 - 近年用いられているMTのしくみ
 - MTが今後できそうなこと
- 翻訳とMTの違い
 - 翻訳に必要な情報
 - それらを欠くMTが生じる問題
 - MTを使う際の注意点
 - 存在感を増すMTとどう付き合っていくか?

NICTにおける機械翻訳技術の研究開発

国立研究開発法人 情報通信研究機構

情報通信技術の研究開発 + 情報通信分野の事業支援

The image displays two overlapping browser windows of the NICT website. The top window shows the main header with the NICT logo, navigation links (ENGLISH TOP, アクセス, お問い合わせ), and a search bar. Below the header is a large banner for 'NICT紹介ビデオ' (NICT Introduction Video) with options for Japanese (5分版 / 18分版) and English (5分版 / 18分版) versions. The bottom window shows a grid of service tiles including: 日本標準時 (Japan Standard Time), 日本標準時の業務 (Japan Standard Time Business), CYDER (実践的サイバー防衛演習), SecHack365, 宇宙天気予報 (Space Weather Forecast), 校正サービス (Correction Service), 総合テストベッド (General Test Bed), 災害時関連技術 (Disaster-related Technology), WISDOMX (Web質問応答システム), VoiceTra (多言語音声翻訳アプリ), 翻訳バンク (Translation Bank), ひまわり8号 (ひまわり8号 リアルタイムWeb), DISAANA (対災害 SNS 情報分析システム), D-SUMM (災害状況要約システム), NICTERWEB (Cybersecurity Laboratory), NOTICE (National Operation Towards IoT Clean Environment), NICT SEEDs (Smart IoT Acceleration Forum), AI Japan R&D Network, ICTスタートアップ支援センター, 情報バリアフリーのための情報提供サイト, YouTube NICTchannel, and @NICT_Publicity #NICT. The footer contains navigation links: NICTについて, 研究紹介, 研究成果, オープンイノベーション, 広報・出版.

ENGLISH TOP

アクセス お問い合わせ ENHANCED BY Google

NICT 国立研究開発法人 情報通信研究機構

NICTについて 研究紹介 研究成果 オープンイノベーション 公募・調達 広報・出版 採用情報 インターンシップ

NICT紹介ビデオ

日本語版 5分版 / 18分版
英語版 5分版 / 18分版

観る Watch

最新

日本標準時 日本標準時の業務 標準電波 JJY 時刻供給サービス

校正サービス 総合テストベッド

CYDER 実践的サイバー防衛演習

SecHack365

宇宙天気予報

災害時関連技術 (DISAANA / D-SUMM 等)

WISDOMX Web質問応答システム

VoiceTra 多言語音声翻訳アプリ

翻訳バンク

ひまわり8号 リアルタイムWeb

DISAANA 対災害 SNS 情報分析システム

D-SUMM 災害状況要約システム

NICTERWEB Cybersecurity Laboratory

NOTICE National Operation Towards IoT Clean Environment

NICT SEEDs Smart IoT Acceleration Forum

AI Japan R&D Network

ICTスタートアップ支援センター

情報バリアフリーのための情報提供サイト

YouTube NICTchannel

@NICT_Publicity #NICT

NICTについて 研究紹介 研究成果 オープンイノベーション 広報・出版

先進的音声翻訳研究開発推進センター

多言語音声翻訳技術の研究開発拠点 (2014年設立)

The image displays two overlapping screenshots of the ASTREC website. The top screenshot shows the homepage with a banner for 'VoiceTra' and a sidebar with 'OUTLINE' and 'NEWS'. The bottom screenshot shows a login page for 'みんなの自動翻訳@TexTra' with a login form and a sidebar with 'OUTLINE' and 'NEWS'.

Top Screenshot (Homepage):

- Header: ASTREC logo (先進的音声翻訳研究開発推進センター), NICT logo, and navigation links: ホーム, 概要, 組織, NEWS, 採用情報, アクセス, English.
- Banner: さあ、世界中の人と話をしよう。言語の壁をなくす多言語音声翻訳アプリ VoiceTra. Includes a smartphone showing the app interface.
- Sidebar (Left):
 - OUTLINE: 世界の「言葉の壁」を打ち破り、グローバル実装を推進します。
 - NEWS:
 - メディア 2020/07/30 メディア掲載情報を更新しました。
 - メディア 2020/02/26 NICTの多言語音声翻訳技術や隔田フェローへのインタビューに関する記事が、Newsweek日本版3月3日号およびWEB版に掲載されました。

Bottom Screenshot (Login Page):

- Header: ASTREC logo (先進的音声翻訳研究開発推進センター), NICT logo, and navigation links: ホーム, 概要, 組織, NEWS, 採用情報, アクセス, English.
- Main Content:
 - みんなの自動翻訳@TexTra®
 - 「みんなの自動翻訳@TexTra®」は、自動翻訳をみんなで育てるサイトです。
 - Login Form: ユーザー名, パスワード, ログイン, ログインしたままにする.
 - 自動翻訳を使ってみよう! 「みんなの自動翻訳@TexTra®」にユーザー登録し、サイト上で翻訳エディタを使用して自分で翻訳.
- Sidebar (Left):
 - OUTLINE: 世界の「言葉の壁」を打ち破り、グローバルで自由な交流の実現を目標として、多言語音声翻訳技術の研究開発と産学官の連携による社会実装を推進します。
 - NEWS:
 - メディア 2020/07/30 メディア掲載情報を更新しました。
 - メディア 2020/02/26 NICTの多言語音声翻訳技術や隔田フェローへのインタビューに関する記事が、Newsweek日本版3月3日号およびWEB版に掲載されました。

多言語音声翻訳アプリVoiceTra

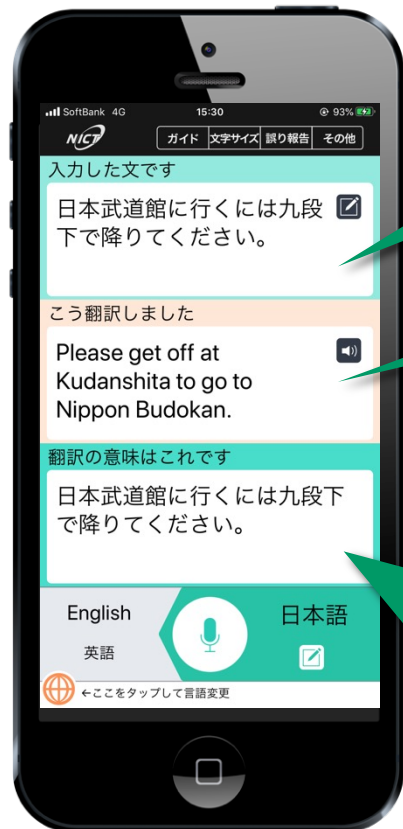


NICTの無料アプリ

- 31言語に対応
- 旅行会話が得意
- 700万ダウンロード

重点10言語 + 5言語を強化中

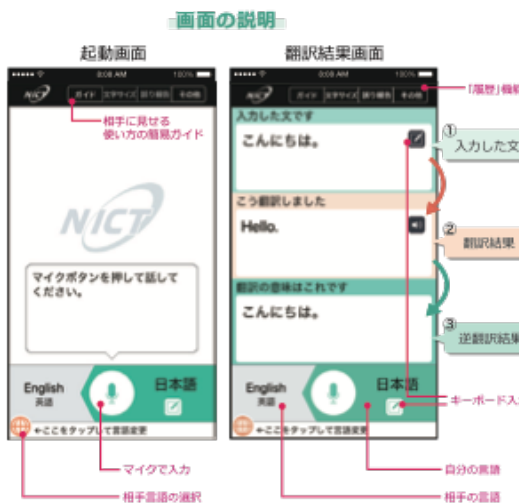
- 日, 英, 中, 韓, タイ, ベトナム, インドネシア, ミャンマー, スペイン, フランス
- フィリピン, ブラジルポルトガル, クメール, ネパール, モンゴル



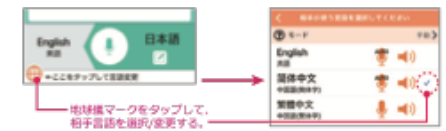
入力音声の
音声認識結果

翻訳結果

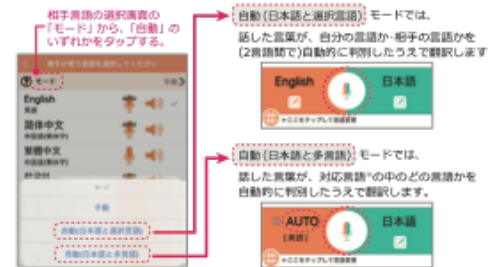
折り返し
翻訳結果
(翻訳結果を
自分の言語で
確認できる)



相手の言語の選択



一部の言語で言語の自動判別機能が試せます！
※対応言語は「翻訳できる言語」一覧をご参照ください（AUTO付きマークの言語が対応可）



翻訳できる言語 (31言語) 中国語・ポルトガル語の方を含む

— 日本語～モンゴル語を重点研究言語（*印は強化対象言語）としています —

🔊 音声で入力できる (19言語対応)

🗣️ 何語かを自動判別もできる* (10言語対応)

🔊 音声が出力される (18言語対応)

🔊 日本語
🔊 英語
🔊 中国語 (簡体字)
🔊 中国語 (繁体字)
🔊 韓国語
🔊 タイ語
🔊 フランス語
🔊 インドネシア語

🔊 ベトナム語
🔊 スペイン語
🔊 ミャンマー語
🔊 フィリピン語
🔊 ポルトガル語 (ブラジル)
🔊 クメール語*
🔊 ネパール語*
🔊 モンゴル語*

アラビア語
イタリア語
ウルドゥ語
オランダ語
シンハラ語
デンマーク語
ドイツ語
トルコ語

ハンガリー語
ヒンディ語
ポーランド語
ポルトガル語
マレー語
ラオ語
ロシア語

現在の多言語音声翻訳技術の社会展開例

観光

アプリ



「はなして翻訳」
NTTドコモ



「NEC翻訳」
日本電気



「mimi」音声翻訳
powered by NICT
Fairy Devices

「しゃべって翻訳 for A」
ATR-Trek



「native.heart」
ブリックス スマートカルチャーゲートウェイ



「どこでも翻訳」
フィート



「接客音声翻訳」
三菱地所(凸版印刷)

専用端末



「iil(イーリー)」
ログバー



「iil インバウンド」
ログバー



「POCKETALK」
ソースネクスト



「対面ホンヤク」
パナソニック



「Fairy I/O」
Fairy Devices

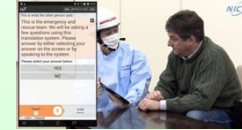
公的機関

警察



- ・「PIT/スマホと腕時計型端末」岡山県警
- ・「VoiceTra」道府県警(NICT)※実証実験中

消防



「救急ボイストラ」
消防庁

自治体



「VoiceBiz」(ボイスBiz)
凸版印刷

ビジネス

電話



自動翻訳電話
「PANインバウンド」
東風津梁

工場



「多言語業務向けシステム」
住友ゴム工業(ATR-Trek)

金融



「ハイブリッド式多言語
通訳サービスKOTOBAL」
コニカミノルタ

交通機関



「駅コンシェル」
日立ソリューションズ・
テクノロジー

鉄道



「鉄道向け多機能翻訳アプリ」
京急・NICT・ブリックス・日立G
(※実証実験時)

タクシー



「TOTTRA」
鳥取県ハイヤータクシー協会
Fairy Devices

プラット フォーム



「NICT VoiceTra
プラットフォーム」
NICT



「みらい翻訳プラット
フォーム音声翻訳API」
みらい翻訳



「mimi」Cloud API Service
Fairy Devices



「多言語音声APIサービス」
日本電気

医療機関

病院



「医療通訳タブレット
MELON」
コニカミノルタ

みんなの自動翻訳@TexTra

NICT運用の無料のテキスト翻訳サービス

- 各分野向けにチューンしたモデルが利用可能
 - 特許
 - 金融
 - 対話

みんなの自動翻訳@TexTra

Language: atsushi

質問・要望 リサイクル

自動翻訳

英語 → 日本語

汎用NT 【英語 - 日本語】 etc 2

翻訳

ヘルプ

Expressions that convey the same meaning using different linguistic forms in the same language are called paraphrases. Techniques for generating and recognizing paraphrases play an important role in many natural language processing systems, because "equivalence" is such a basic semantic relationship.

翻訳不要記号を挿入

オプション

- ☐ 文章を結合する
- ☒ 文章を区切る
- ☐ 辞書引き不要

比較

原文

Expressions that convey the same meaning using different linguistic forms in the same language are called paraphrases.

汎用NT 【英語 - 日本語】 同じ言語で異なる言語形式を使用して同じ意味を伝える表現は、言い換えと呼ばれます。

特許NT 【英語 - 日本語】 同じ言語で異なる言語形式を使用して同じ意味を伝える表現は、パラフレーズと呼ばれる。

比較

原文

Techniques for generating and recognizing paraphrases play an important role in many natural language processing systems, because "equivalence" is such a basic semantic relationship.

汎用NT 【英語 - 日本語】 言い換えを生成し認識する技術は、多くの自然言語処理システムにおいて重要な役割を果たす。

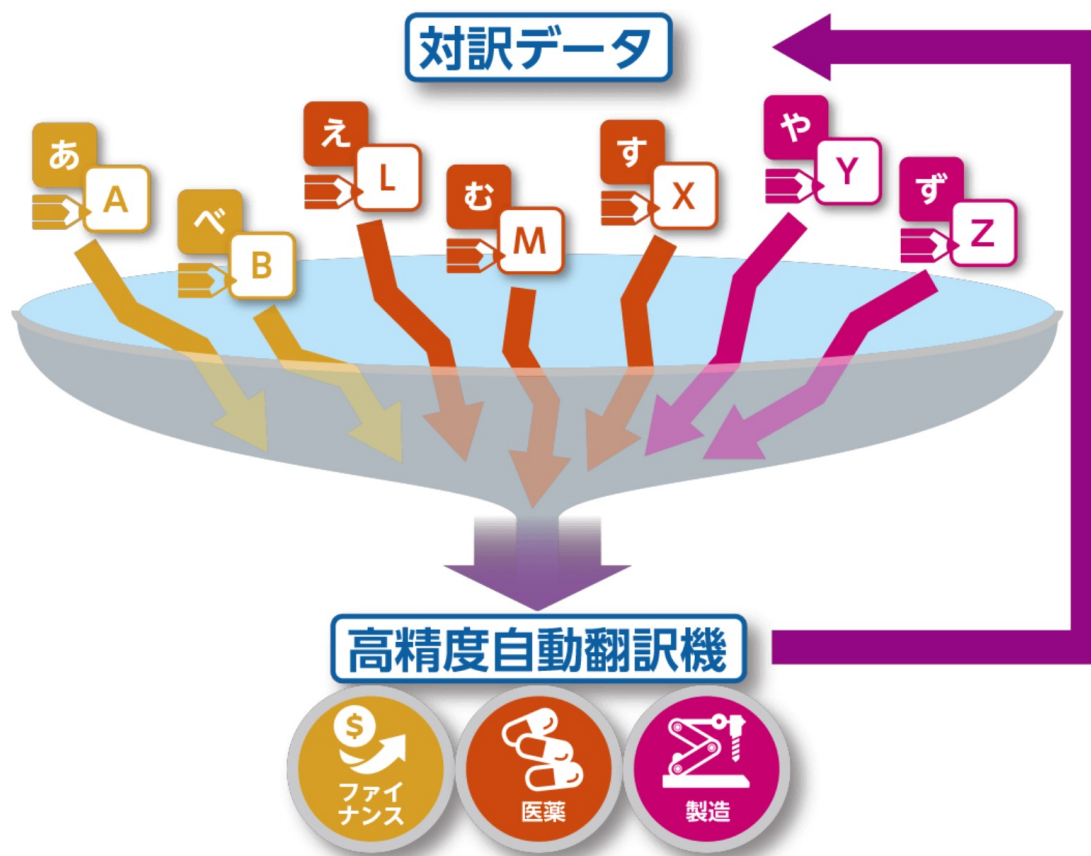
特許NT 【英語 - 日本語】 パラフレーズを生成および認識するための技術は、多くの自然言語処理システムにおいて重要な役割を果たす。



翻訳バンク@NICT

■ 2017年9月に運用開始

- 当面の目標は1億文対の集積
- 製薬，自動車，金融，ビジネス分野などで成果



グローバルコミュニケーション計画2025

多言語翻訳技術の進展

- 我が国の多言語翻訳技術は、30年の研究開発を経て実験室から社会へ、これからは日常生活からビジネスへ
- 「グローバルコミュニケーション計画」に基づく取組により、AI技術も導入し、翻訳精度を向上（12言語で実用レベルを達成）
- NICTから民間への技術移転も進め、多様なサービスが実用化・普及

言葉の壁がない世界

多言語翻訳技術の研究開発とパラダイムシフト

ニューラルネットベース（大規模コーパス+深層学習）

- ・ 不特定話者が対象
- ・ 雑音下でも高精度な音声認識
- ・ 違和感のない自然な文章に翻訳
- ・ 違和感のない自然な音声合成

コーパスベース（大規模コーパス+機械学習）

- ・ 不特定話者が対象
- ・ 丁寧な発話が必要
- ・ 一定の条件下であれば屋外でも音声認識可能
- ・ 生活会話の翻訳

ルールベース（If-Then-Else）

- ・ 特定話者が対象
- ・ 文節発声・定型文の発話
- ・ 静かな室内での発話
- ・ 会議室予約程度が可能

1986
音声翻訳の研究開始

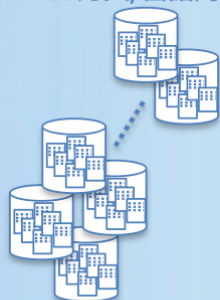


2008
実験室で実験

2000
社会還元加速プロジェクト

出典：
https://www.soumu.go.jp/main_content/000678485.pdf

100万文/言語対



深層学習に必要な膨大な学習データ

10万文/言語対



統計処理に必要な大量な対訳データ

ニューラル翻訳の実装

2017
社会へ

2015
多様な翻訳装置の開発開始

2014
「グローバルコミュニケーション計画」開始

2010
世界初のネットワーク型音声翻訳スマホアプリ VoiceTra

2009
全国5観光地域での大規模実証実験
→大量の実利用ログで精度向上

派生技術
要約自動生成技術（日仏翻訳）

2030

シビアな交渉にも使える通訳

2025

文脈・話者の意図等を補う通訳

2020

日常生活やビジネスを支える翻訳



グローバルコミュニケーション計画
2025（第2期：2020～2025）

グローバルコミュニケーション計画
（第1期：2014～2019）



多様な翻訳端末・アプリが登場



音声認識技術による聴覚障がい者支援サービス
自治体・郵便局窓口での翻訳サービス



（病院）
（鉄道会社）
医療・鉄道等の分野での翻訳サービスの利活用

民間企業への技術移転を促進する多言語音声翻訳プラットフォームの構築



NICTの重点対応言語（実用レベル）	
1 日本語	主に訪日外国人対応（当初目標）
2 英語	
3 中国語	
4 韓国語	
5 タイ語	
6 インドネシア語	主に在留外国人対応
7 ベトナム語	
8 ミャンマー語	
9 フランス語	
10 スペイン語	
11 ブラジルポルトガル語	
12 フィリピン語	

機械翻訳(MT)のしくみ

アルゴリズム × データ (× ハードウェア)

MTのアルゴリズムの変遷

■ MT黎明期

- 1947年: 数学者Weaverから数学者Wienerへの手紙
「暗号解読の要領で機械的に翻訳できるのではないか」
- 1950年代には研究が活発化
 - 規則に基づくMT: 人間が数万件の規則と辞書を整備

■ 1964年: ALPACレポート「実用化には程遠い」

- 言語処理の基礎研究に重点, MTにとっては冬の時代の到来

■ 1990年頃以降: 対訳コーパスに基づくMTの研究が活発化

- 対訳コーパス: 人間による翻訳テキストのデータ

1984

1988

2003

2014

2017

2021

用例ベースMT
(EBMT)

[Nagao 84]

統計的MT
(SMT)

[Brown+ 88]

句に基づくSMT
(PBSMT)

[Koehn+ 03]

ニューラルMT
(NMT)

[Sutskever+ 14]

Transformer
(NMTの一種)

[Vaswani+ 17]

対訳コーパス中の対訳文書の例

Italy have defeated Portugal 31-5 in Pool C of the 2007 Rugby World Cup at Parc des Princes, Paris, France.

Andrea Masi opened the scoring in the fourth minute with a try for Italy.

Despite controlling the game for much of the first half, Italy could not score any other tries before the interval but David Bortolussi kicked three penalties to extend their lead.

Portugal never gave up and David Penalva scored a try in the 33rd minute, providing their only points of the match.

Italy led 16-5 at half time but were matched by Portugal for much of the second half.

However Bortolussi scored his fourth penalty of the match, followed by tries from Mauro Bergamasco and a second from Andrea Masi to wrap up the win for Italy.

Currently third in Pool C with eight points, Italy face a tough match against second placed Scotland on 29 September.

New Zealand lead the group with ten points, ahead of Scotland on points difference.

Portugal are bottom of the group with no points, behind Romania with one.

フランスのパリ、パルク・デ・フランスで行われた2007年ラグビーワールドカップのプールCで、イタリアは31対5でポルトガルを下した。

アンドレア・マーシが開始4分後のトライでイタリアにとって最初の得点を入れた。

イタリアは試合前半のほとんどをコントロールしながら、その後インターバルまでのトライでは得点できなかったが、デビッド・ボルトルッシがペナルティキックで3点を入れリードを広げた。

ポルトガルはあきらめることなく、デビッド・ペナルバが33分でトライを決め、試合中唯一の得点をあげた。

イタリアはハーフタイムの時点で16対5とリードしていたが、後半ではポルトガルと互角になった。だがボルトルッシがこの試合で4つめとなるペナルティキックを成功させ、さらにマウロ・ベルガマスコのトライとアンドレア・マーシの2回目のトライでイタリアの勝利が確定的になった。

現在8ポイントとプールCでは3位のイタリアは、9月29日に2位のスコットランドとの厳しい戦いに臨む。

ニュージーランドが10ポイントでグループをリードし、勝ち点差でスコットランドに勝っている。

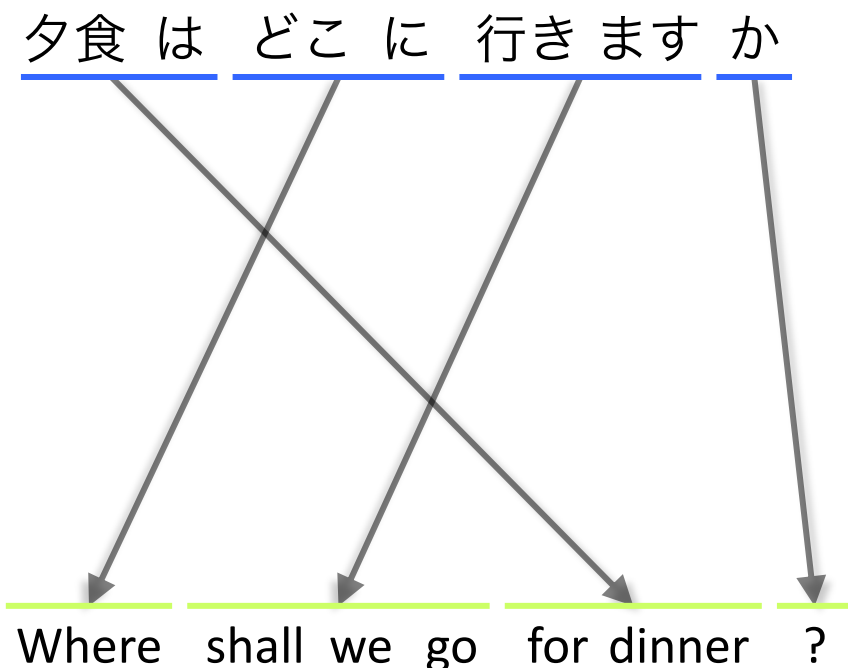
ポルトガルはポイントなしでグループ最下位、1ポイントでルーマニアの下位にいる。

このALT [Riza+ 16] の例は文書単位の翻訳だがMT研究では文単位の翻訳もよく使われる

PBSMTとNMT

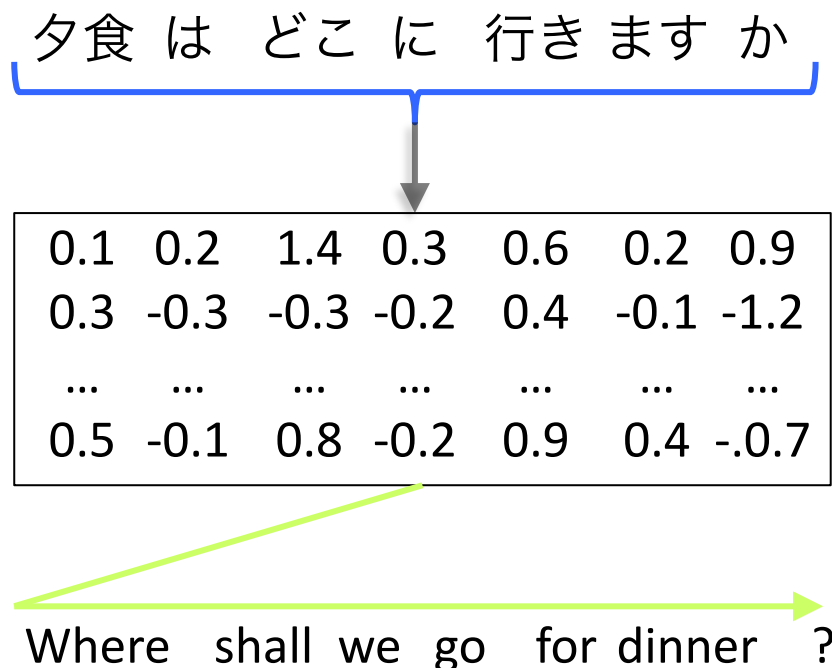
■ PBSMT: 句に基づくSMT

- 対訳データ中に頻出する
句の対応を陽に捉える
- 組み合わせの確率的探索
 - 語句単位の訳語, 並び替え, 流暢さなど



■ NMT: ニューラルMT

- 原文を数値データに変換
 - e.g., ベクトルや行列
- 語の系列を生成
 - e.g., 文頭から1語ずつ



NMTの基本構造

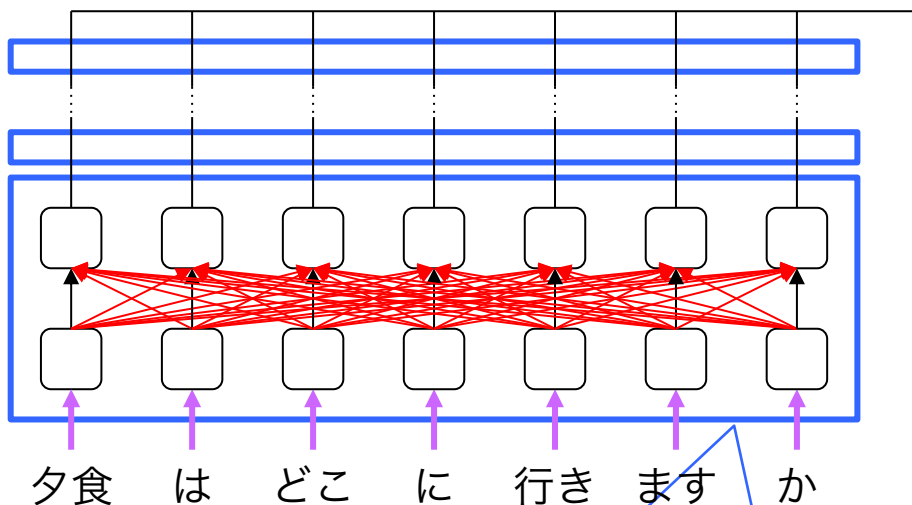
■ 語の系列 → 語の系列

- 要素: 語の分散表現, エンコーダ, デコーダ, 注意機構
- e.g., Transformer [Vaswani+ 17]

語の分散表現 (embedding)

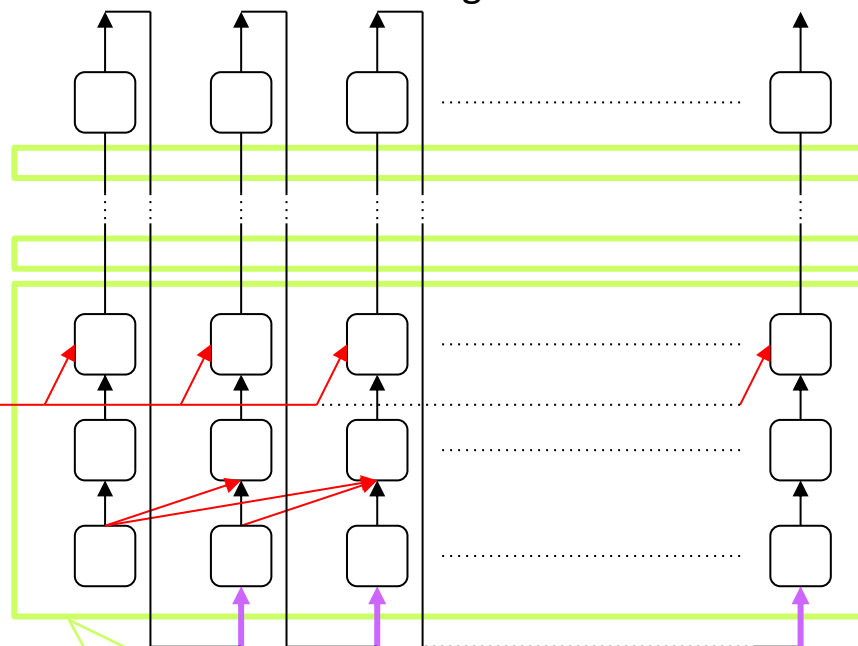
夕食 $\leftrightarrow$ (0.1, -1.4, 2.9, ..., -0.7)

k次元ベクトル



エンコーダ (encoder)
原文を数値データに変換

where shall we go for dinner ? <EOS>



デコーダ (decoder)
数値データから語の系列を生成

解析/生成のトライアングル

■ [Vauquois 68]



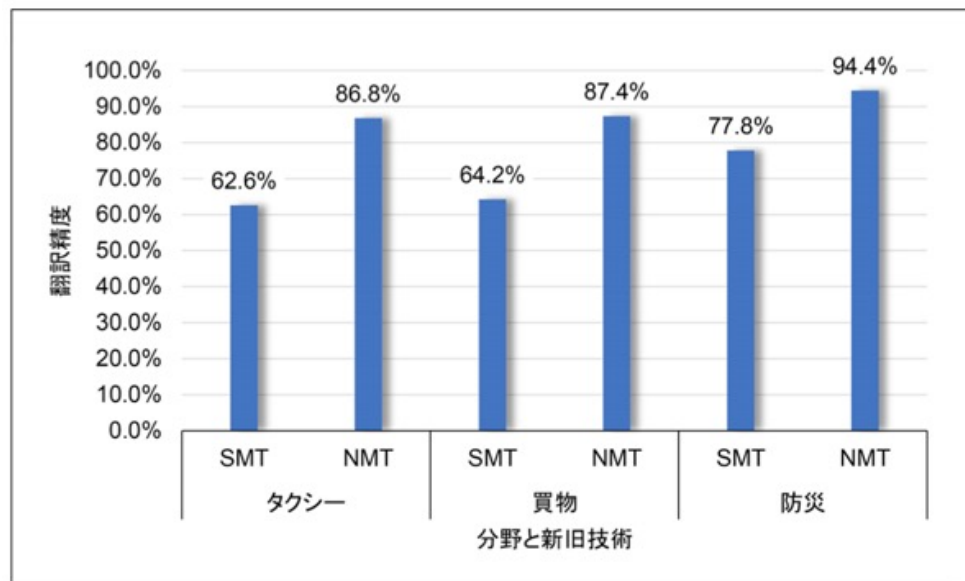
MTの性能評価

- 性能評価は性能向上を目指すうえで不可欠
 - 色々な観点: 翻訳品質, 処理速度, 対応言語, 対応分野
- 翻訳品質の評価
 - (全)自動評価: 信頼性中?, 客観性高, コスト低 (MTの研究開発を促進)
 - 人間が作成した参照訳との表層的な類似度
 - e.g., BLEU [Papineni+ 02], METEOR [Banerjee+ 05], RIBES [Isozaki+ 10]
 - 人間が関わる評価: 信頼性高?, 客観性低?, コスト高
 - 正確さ(adequacy) / 流暢さ(fluency), それに類するもの [Goto+ 13]
 - 後編集(post-edit)の分量: e.g., HTER [Snover+ 06]
 - 翻訳誤りの分布や分量 [藤田 20][Freitag+ 21]



PBSMT→NMTで翻訳品質が大きく向上

■ e.g., NICTの英日NMT [NICT 17]



■ 言語や分野によっては
「人間並み」!?

- “Human parity”
[Hassan+ 18][Barrault+ 19]
- というわけでもない
[藤田 20][Läubli 20][Freitag 20]



With hundreds of languages used by people on our platforms and thousands more spoken around the world, developing powerful and flexible machine translation systems has long been a research focus for Facebook. Today we are proud to announce that Facebook AI models achieved first place in several language tasks included in this year's annual news translation competition, hosted by the [Fourth Conference on Machine Translation](#) (also known as WMT). Our models outperformed all other entrants' models in the four tasks we participated in, including English to German, the most competitive task in the contest, with entries drawn from a wide range of high-performing research teams. For this language direction, our translations have been declared superhuman by the WMT organizers, meaning that human evaluators preferred them over translations done by human experts.

Our models used large-scale sampled back-translation, noisy channel modeling, and data-cleaning techniques to achieve the highest performance for translating from English to German, German to English, English to Russian, and Russian to English. These models, along with our work on [cross-lingual pretraining](#) and [self-supervised learning for other modalities](#), will enable us to break down language barriers and to build better content understanding systems to keep people safe.

Microsoft | Research Research areas ▾ Researcher tools Programs & Events ▾ Careers People More ▾ All Microsoft ▾

Achieving Human Parity on Automatic Chinese to English News Translation

Hany Hassan Awadalla, Anthony Aue, Chang Chen, Vishal Chowdhary, Jonathan Clark, Christian Federmann, Xuedong Huang, Marcin Junczys-Dowmunt, Will Lewis, Mu Li, Shujie Liu, Tie-Yan Liu, Renqian Luo, Anul Menezes, Tao Qin, Frank Seide, Xu Tan, Fei Tian, Lijun Wu, Shuangzhi Wu, Yingce Xia, Dongdong Zhang, Zhirui Zhang, Ming Zhou
March 2018
arXiv:1803.05567
[View Publication](#)

[Download BibTex](#)

Machine translation has made rapid advances in recent years. Millions of people are using it today in online translation systems and mobile applications in order to communicate across language barriers. The question naturally arises whether such systems can approach or achieve parity with human translations. In this paper, we first address the problem of how to define and accurately measure human parity in translation. We then describe Microsoft's machine translation system and measure the quality of its translations on the widely used WMT 2017 news translation task from Chinese to English. We find that our latest neural machine translation system has reached a new state-of-the-art, and that the translation quality is at human parity when compared to professional human translations. We also find that it significantly exceeds the quality of crowd-sourced non-professional translations.

[View Publication](#)

Groups

[Machine Translation](#)

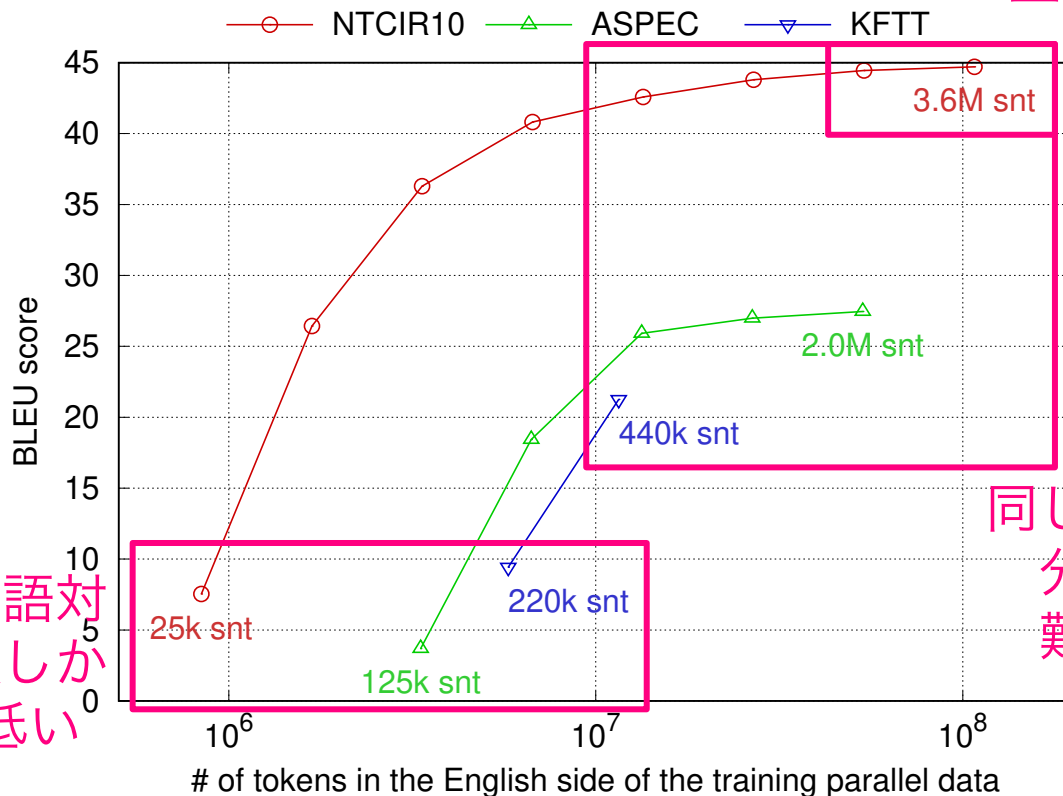
Research Areas

[Artificial Intelligence](#)

NMTの訓練には大規模な対訳が不可欠

- 学習させる対訳データが多いほど高性能
 - より多様な表現・事例をカバーできる
 - より正確に学習できる → 予測能力が上がる
 - e.g., 日英NMTの学習曲線

データが潤沢にあり
実用に耐えるのは
一部の言語・分野



多くの分野、言語対
では対訳が少数しか
なく、性能は低い

同じ翻訳方向でも
分野によって
難度は異なる

対訳コーパスの入手方法

■ 人手で翻訳

- 15～20字の日本語文の翻訳 = 100～400円
- クラウドソーシング
- 自分で翻訳 (主に評価用データ)

■ 既訳の収集

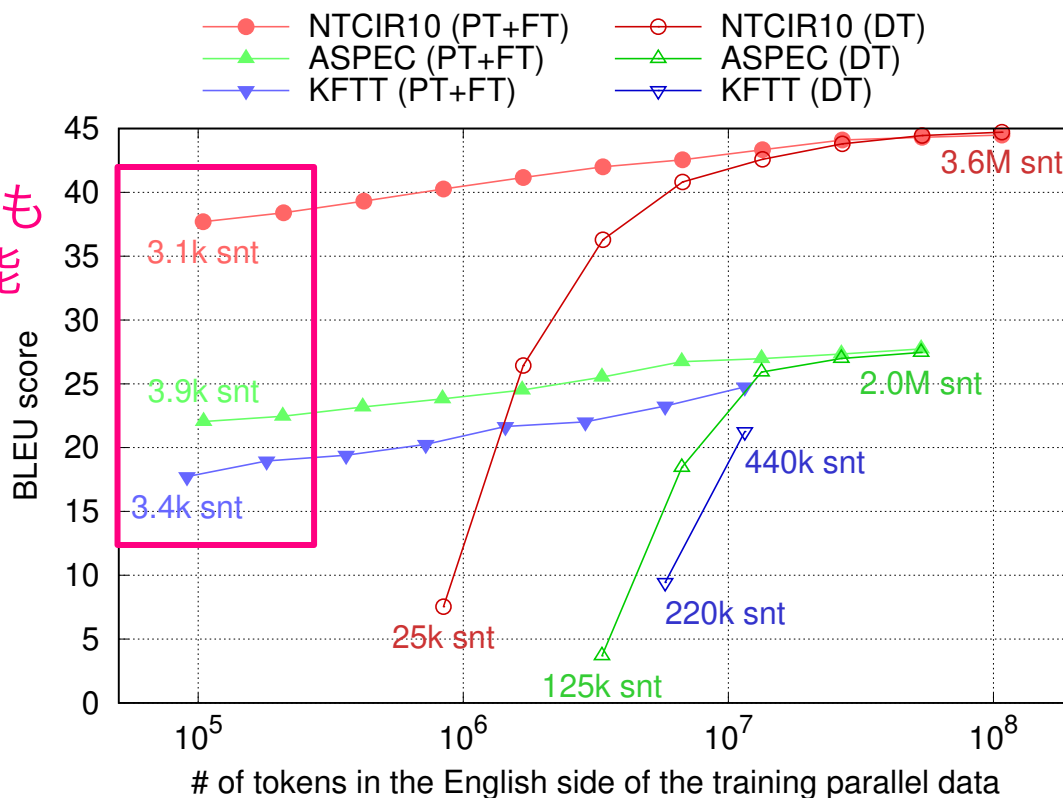
- 欧州議会文書: 200万文強 [Koehn 05]
- 国際出願特許: 日英3.5億文 [JPO & NICT 15]
- 科学技術論文の要旨: 300万文 [Nakazawa+ 16]
- OPUS [Tiedemann 12]
- 翻訳バンク [NICT 17]
- Webからのクロール [Pryzant+ 17][Morishita+ 20]

他の種類のデータの利活用 (1/3)

■ 他分野からの転移学習 [Chu+ 17]

- 事前学習: 他分野の大規模な対訳コーパスから各言語の理解・生成と言語間の翻訳関係を学習
- パラメタ調整: 特定対象分野の対訳コーパスで微調整
- e.g., JParaCrawl (10M文対) [Morishita+ 20] による事前学習の効果

小規模な対訳でも
それなりの性能

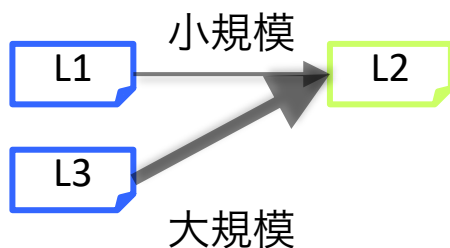


他の種類のデータの利活用 (2/3)

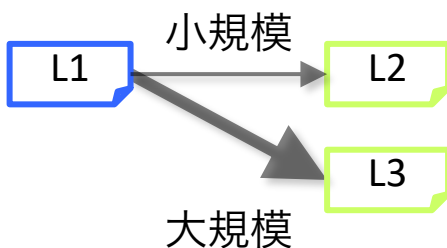
■ 他の言語対からの転移学習 [Zoph+ 16][Johnson+ 17]

- 事前学習: 他の言語対の大規模な対訳コーパスから各言語の理解・生成と言語間の翻訳関係を学習
- パラメタ調整: 特定対象言語対の対訳コーパスで微調整
- e.g., 仏-英 → 越-英 (デコーダ共有)

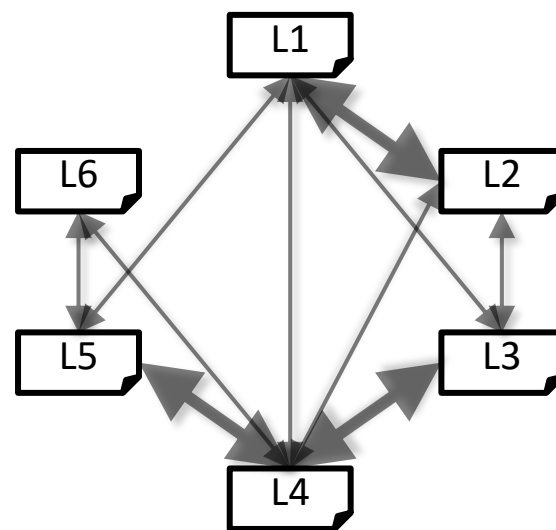
デコーダ共有



エンコーダ共有



Multi-to-multi

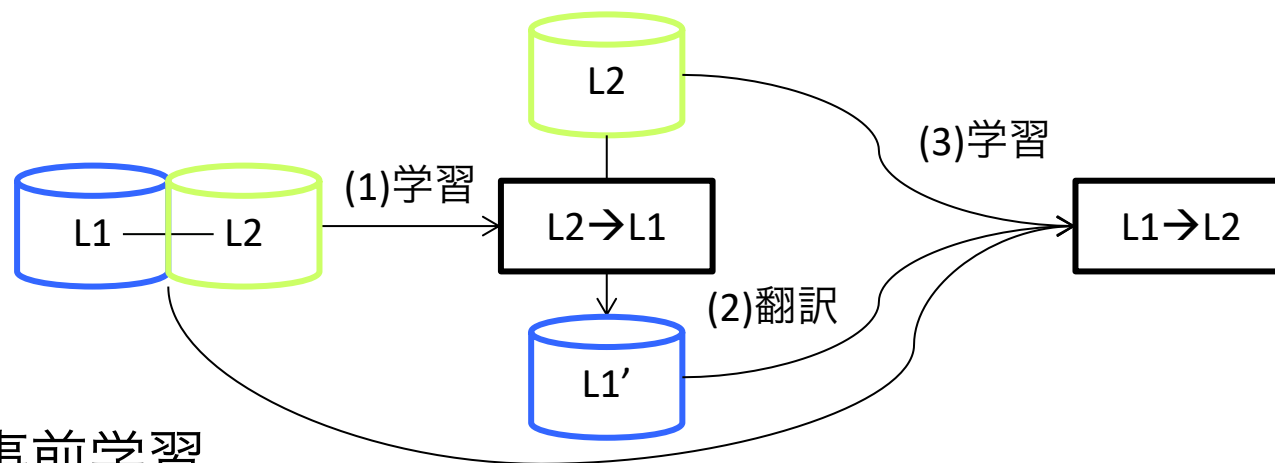
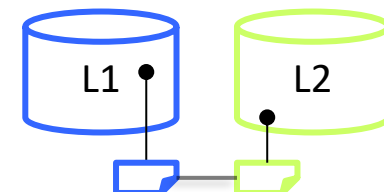


L1→L6やL2→L5はZero-shot

他の種類のデータの活用 (3/3)

■ 単言語データ

- 疑似対訳の獲得
 - 対訳らしい文対の抽出 [Marie+ 17a; 19a]
 - MTシステムによる疑似対訳文対の生成
 - e.g., 逆翻訳: $L2 \rightarrow L1'$ を $L1 \rightarrow L2$ の訓練に利用 [Sennrich+ 16a]
 - e.g., 単言語データについても疑似文を生成・利用 [Marie+ 21]



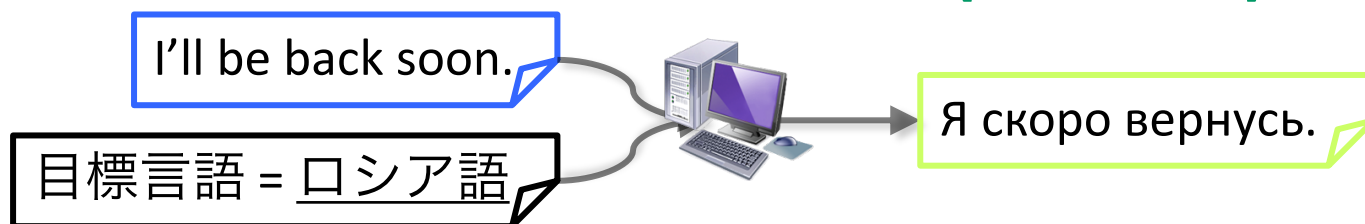
■ 様々な事前学習

- 語の分散表現などのパラメタを学習して援用
 - BERT [Devlin+ 19], XLM [Conneau+ 19], MASS [Song+ 19], BART [Lewis+ 20], mBART [Liu+ 20]

NMTができそうなこと (1/2)

■ どんな情報でも陽に与えられれば利用できる

- 専門分野やテキストタイプへの適応 [Luong+ 15][Chu+ 17]
- 起点言語や目標言語の指定による多言語翻訳 [Johnson+ 16]



- 文脈の情報に基づく曖昧性解消 [Tiedemann+ 17][Voita+ 18]



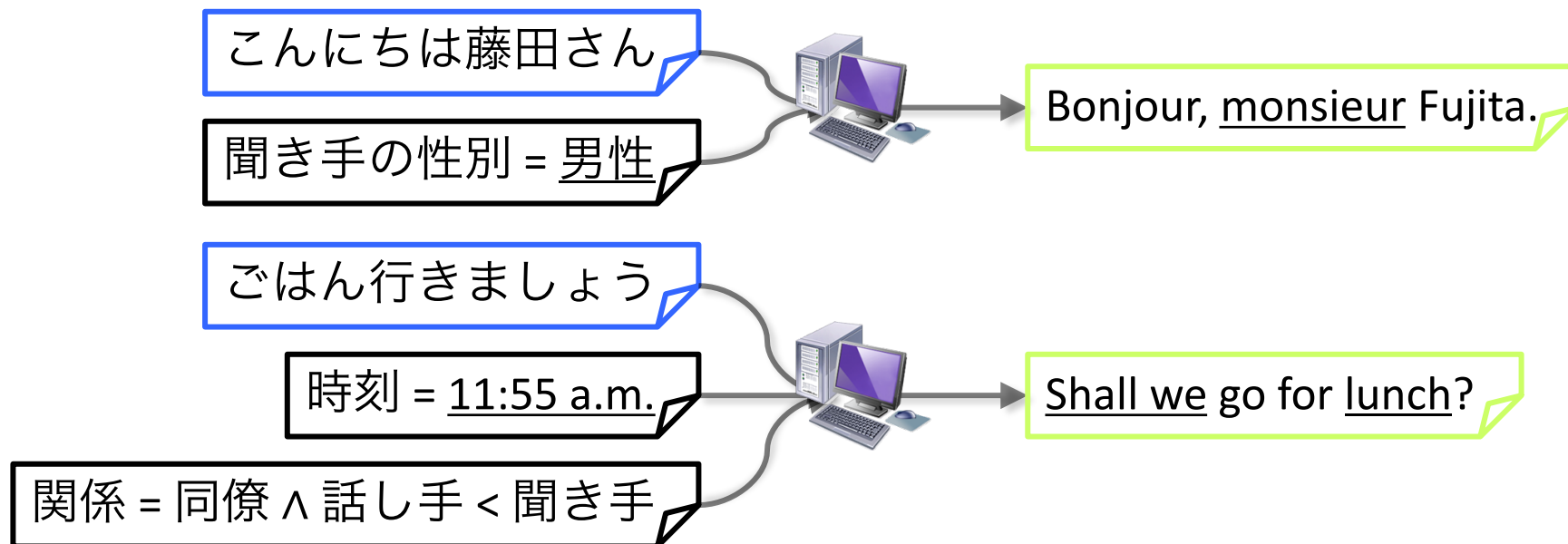
- 丁寧さやスタイルの制御 [Sennrich+ 16][Takeno+ 17]



NMTができそうなこと (2/2)

■ どんな情報でも陽に与えられれば利用できる

- テキストや文書の外にある情報の参照
 - 著者/話し手/読者/聞き手の属性
 - 参与者間の社会的関係/親密さ
 - 発話の文脈: e.g., 時間, 位置情報



Q. どんな情報が性能向上に寄与するか?

Q. データ作成も含めて、どこまで実現可能か?

MTと翻訳の違い

翻訳プロセスの確認 (1/2)

■ 翻訳して、と文書を与えられたら、どうする? [影浦 17]

目的は?

結果物の用途は?

誰が読む?

スコープ

分野は?

文書タイプは?

いつ書かれた?

文書属性

...

So the following statement should have been taken into account, among others:

States Parties particularly condemn racial segregation and apartheid and undertake to prevent, prohibit and eradicate all practices of this nature in territories under their jurisdiction.

Justice Baltasar Garzón decided to indict Augusto Pinochet in 1998, he was ...

国際条約からの引用

目標言語を公用語とする国はどこか?

その国はこの条約に批准しているか?

公式の翻訳は存在するか?

人名

既訳は存在するか?

翻訳プロセスの確認 (2/2)

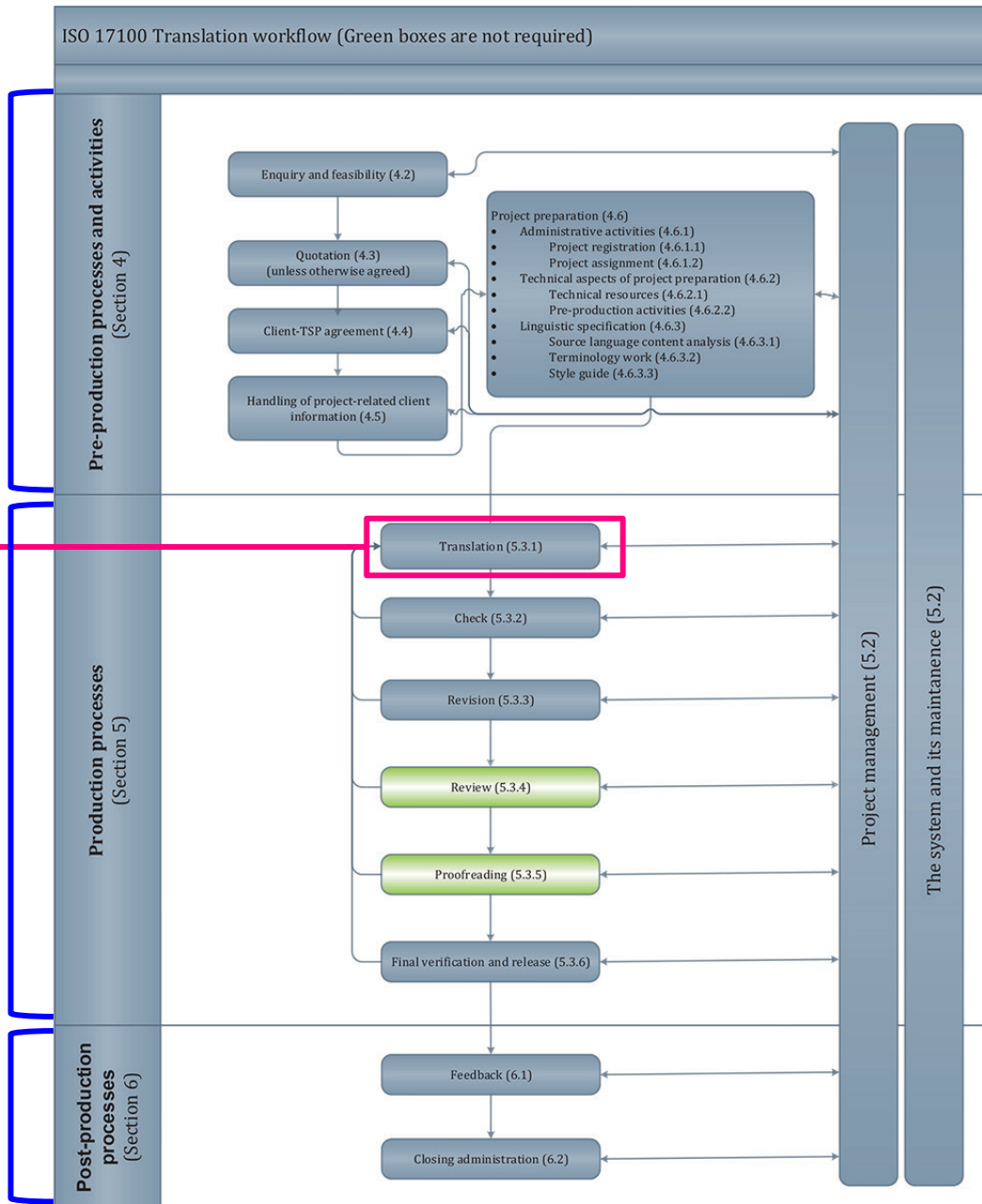
cf. ISO 17100 [ISO 15]

制作前のプロセス および活動

MTが担うのは
せいぜい狭義の翻訳
(上段からの入力なし)

制作プロセス

制作後のプロセス



翻訳とは何か?

■ 起点言語文書が担う役割を目標言語で果たすこと

- 文書を扱う
 - 所与の文字列のみでもその字義的な意味のみでもない
 - 文書の内部構造
 - 構成要素の役割
 - それらの相互作用
- 文書の外側にある仕様も要考慮
 - スコパス [Vermeer 89]: 文書の役割や翻訳の目的
 - ノーム: 目標言語における規範
 - レジスタ: 文書が所属する言語仕様域

Watch your fingers.

翻訳 vs. MTができそうなこと

■ 翻訳

■ MTができそうなこと = $X \text{文} Y \text{訳} + \alpha$

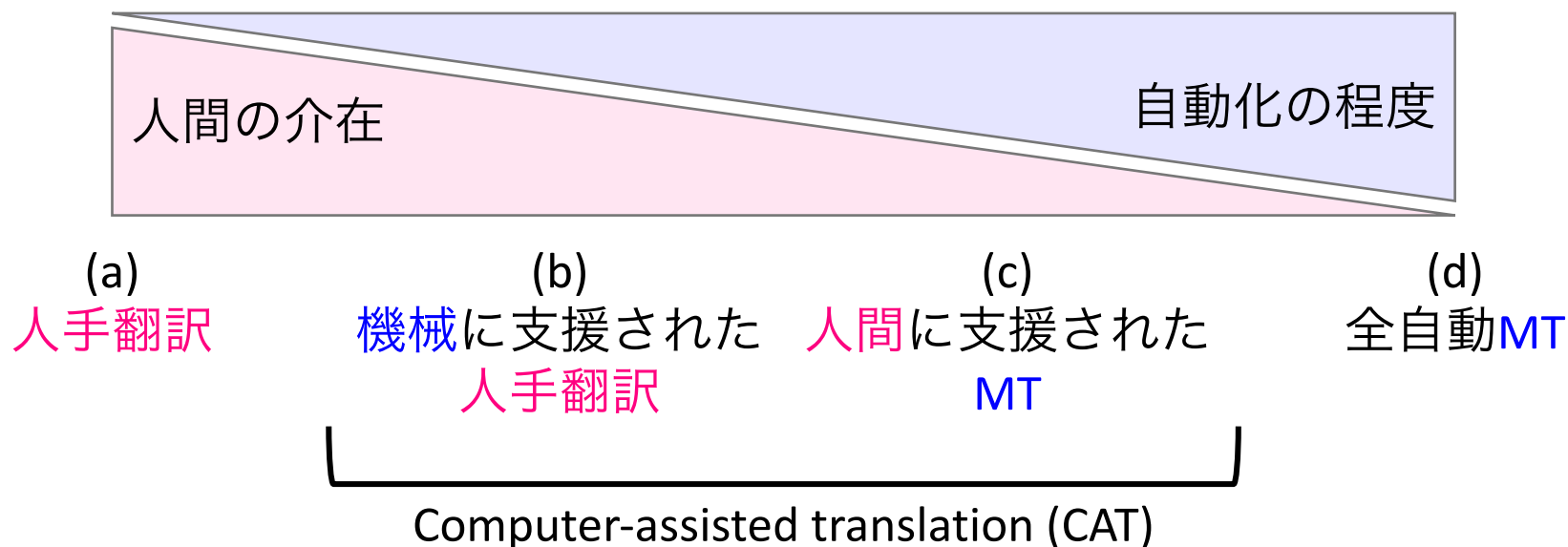
- 起点言語の文字列から目標言語の文字列への変換に過ぎない
 - 「意味」を理解しているとは限らない(しているかもしれない)
- 翻訳の仕様は訓練データが間接的に定めていると仮定している
 - 当該文字列を含む文書の社会的・歴史的 position付けや役割
 - 翻訳の目的や結果物の消費者
- 文単位という制約はなくなりつつある
 - 文書内文脈の参照 [Voita+ 18;19][Lopes+ 20]
 - 他のモーダルの参照 [Barrault+ 18]
- どんな情報でも陽に与えられれば利用できる

MTの品質の不安定性

■ MTのみで完璧な翻訳は不可能 [Bar-Hillel 51]

- 現時点でも，X文Y訳としての品質もまだ担保できない
- 人間が適切に使いこなす必要あり

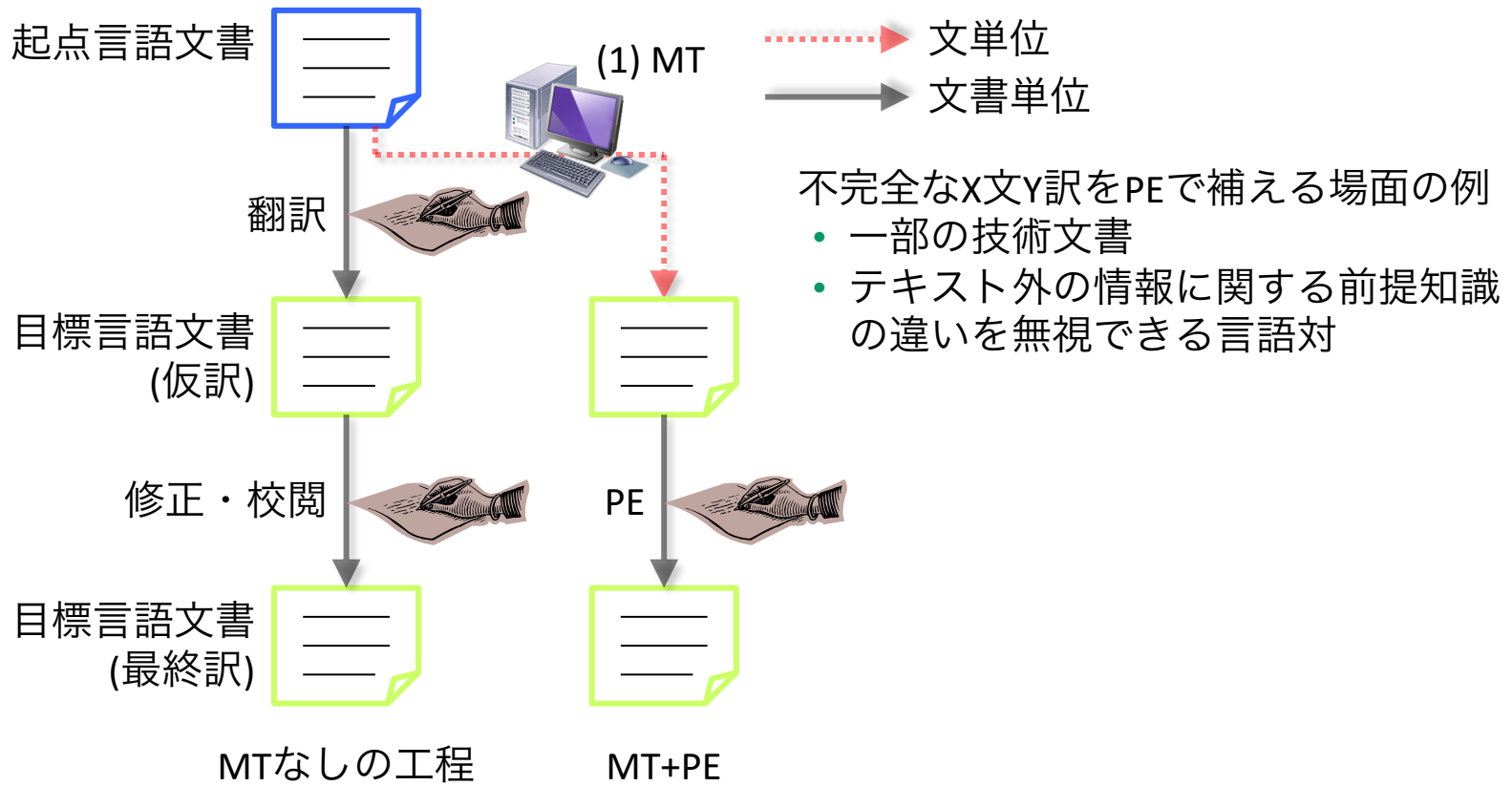
■ 4つの戦略 [Hutchins+ 92]



MTと翻訳の間の壁を乗り越える戦略の例

■ 後編集(post-edit; PE): MTの結果を人間が手直しする

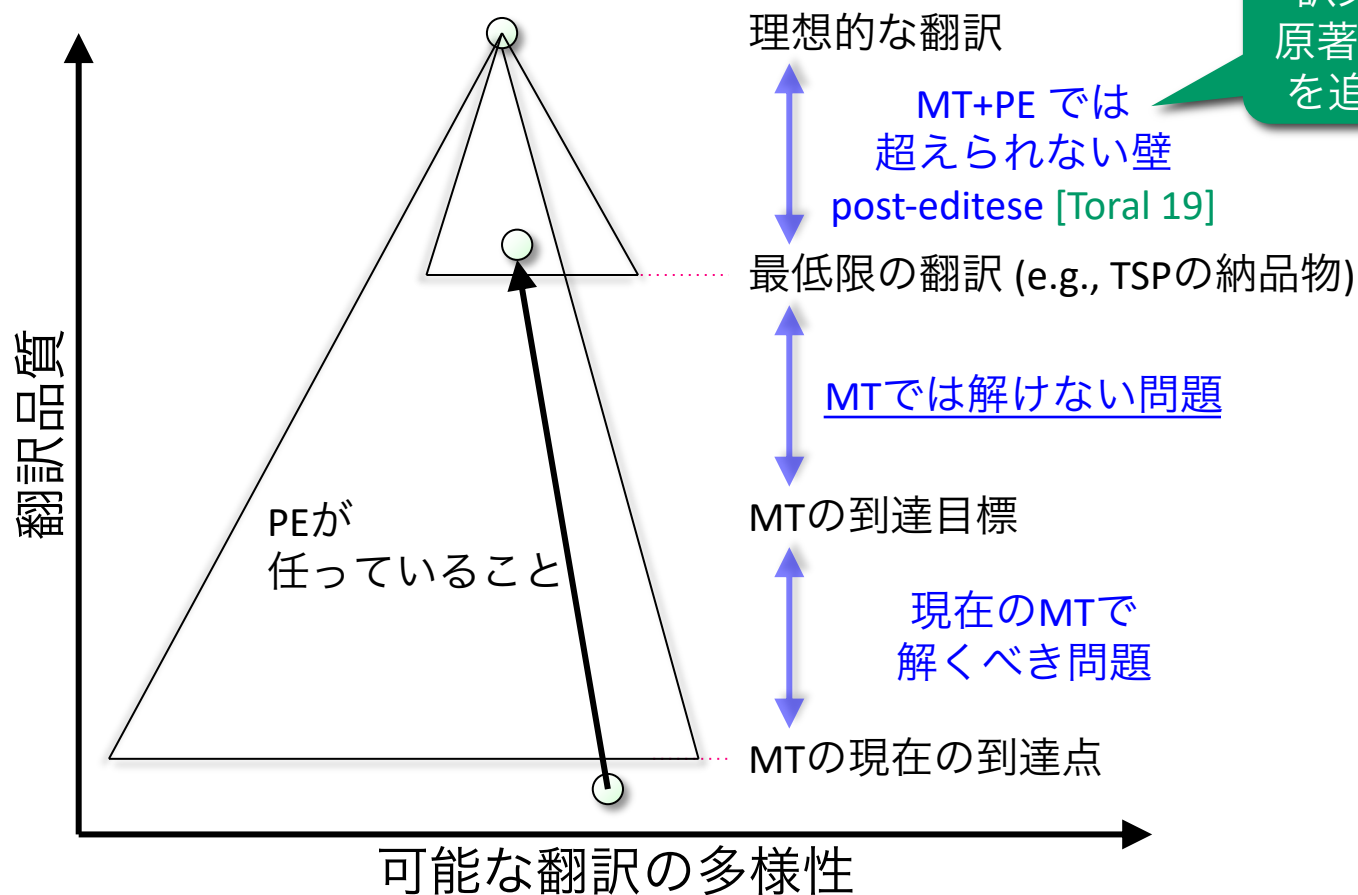
- ISO 18587 [ISO 17]
- 「尻拭い」と言って忌避される場合も



PEが担っていること

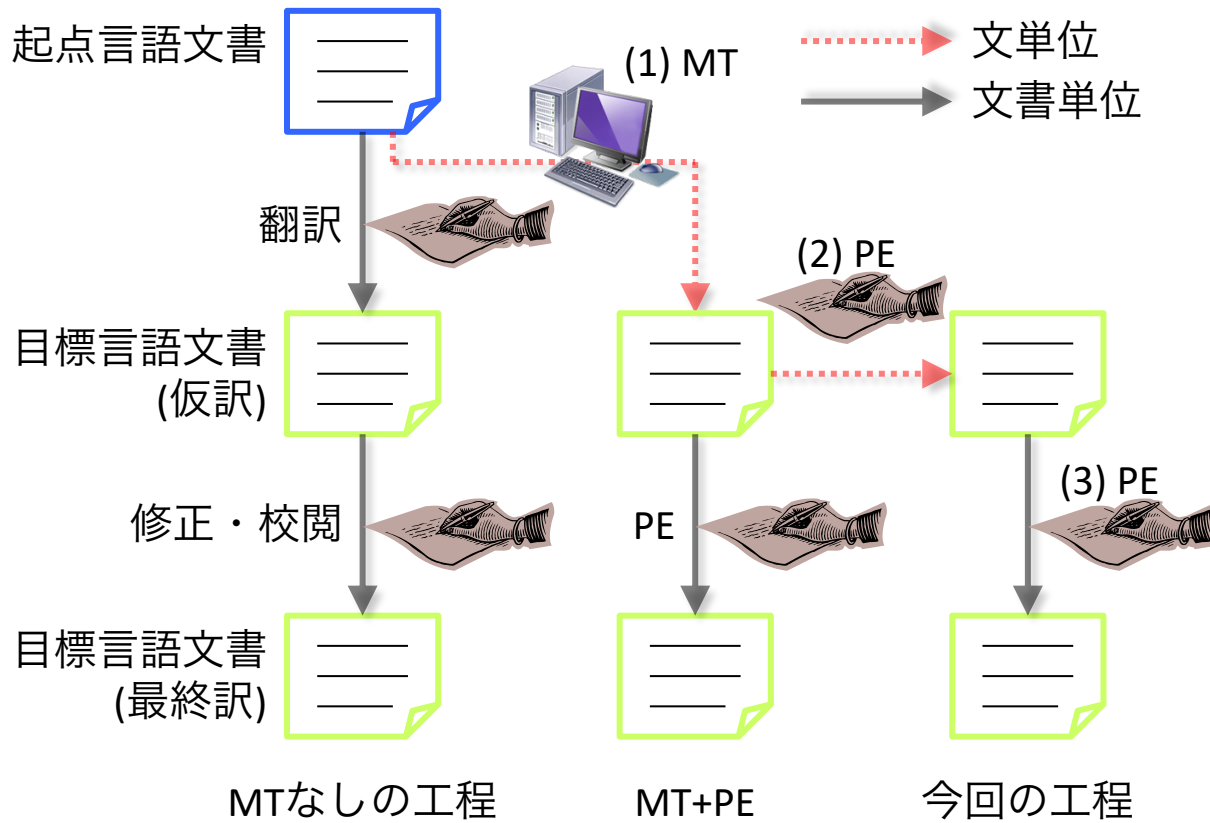
■ 2種類の問題をまとめて解決

- 現在のMTの問題設定(X文Y訳)で残っている問題
- 必要な情報を参照しないために生じている, MTでは解けない問題



2種類の問題を分離して分析する方法 (1/2)

- PEを2段階に分割した工程で修正事例を収集 [Scarton+ 15]
 - 文単位PE: 当該文のみを参照して必要最小限の編集 [藤田 20]
 - 文書単位PE: 必要な情報をすべて参照しつつ, 必要最小限の編集



2種類の問題を分離して分析する方法 (2/2)

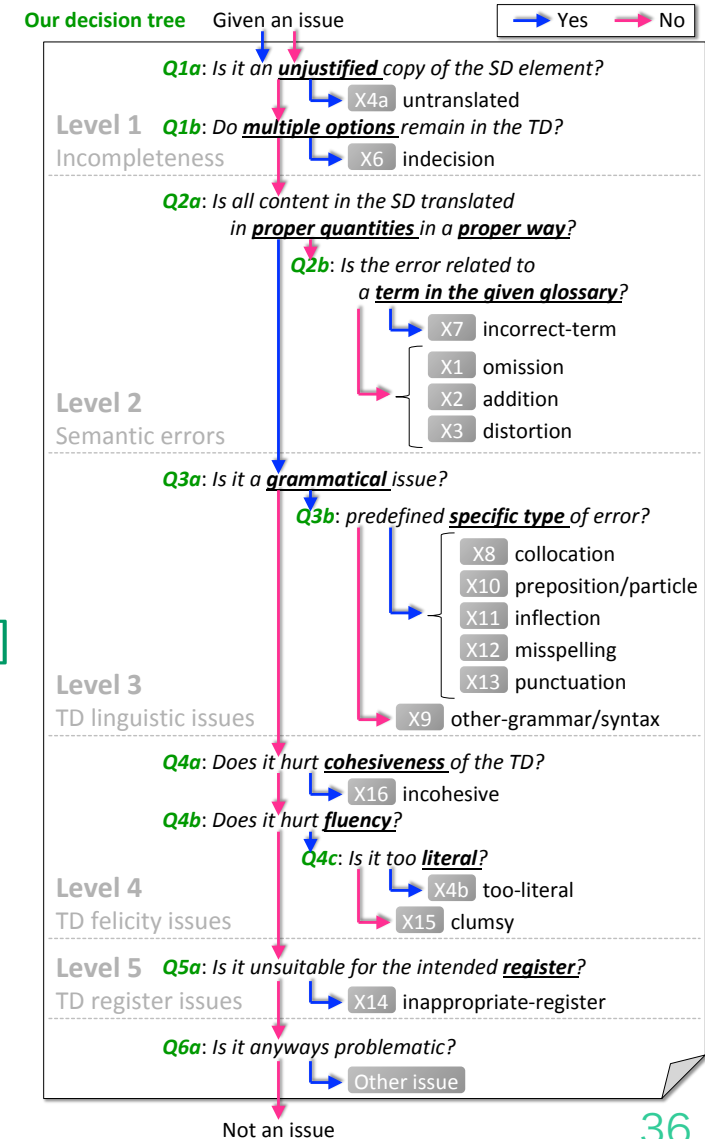
■ 文書単位PEの修正内容の分析

- PEの際に参照された情報の分類

PEに必要な 参照情報		非言語的な情報や 文書の外側の情報	
		不要	必要
当該文書内の 言語的な情報	不要	(a)	(c)
	必要	(b)	(d)

文単位PEの内容は(a)に該当

- 人間翻訳向け校閲カテゴリ体系 [豊島+ 16] を用いた分類



翻訳時に必要な情報を参照しない
機械翻訳が生じる問題
～ニュース記事の英日翻訳の事例研究～

[藤田 20] の発表内容+α

ニュース記事の英日翻訳の事例研究

■ Asian Language Treebank [Riza+ 16]

- 起点言語: それなりに分かる英語
- 目標言語: 母語=日本語
- その他の要求仕様
 - 原文が文書である
 - 訓練データがある: できるだけ高品質なMTから始めるため
 - 参照訳がある: PEの作業量の抑制のため

用途	文書数	文数	トークン数	
			英語	日本語
訓練データ	1,698	18,088	472.1k	610.8k
開発データ	98	1,000	25.6k	33.0k
評価データ	97	1,018	26.3k	33.4k

(1) 文単位MT訳の生成

■ 英日NMTの学習

- 3フェーズ [Chu+ 17]
 1. 事前学習 (NICT内製TexTraデータ)
 2. 混合パラメタ調整 (TexTraデータ + ALT訓練データ)
 3. 非混合パラメタ調整 (ALT訓練データ)
- ツールやパラメタ設定などの詳細は [藤田 20] を参照されたい

■ ALTの評価データを翻訳

- ビーム幅10のビームサーチ
- 長さに関する正規化重みは開発データを使って最適化

(2) 文単位PE

■ 全1,018文をシャッフル

■ 文ごとに必要最低限の修正を施す

- 「最低限」: まったく独立に作成された模範訳よりはMT訳に似ている
 - 編集回数(MT訳 → 修正訳) ≤ 編集回数(MT訳 → 模範訳)
- 編集回数の定量化: TER [Snover+ 06]

原文 The tremor, lasting about 2 minutes, sparked panic.

MT訳 約 2分間 続く 振動 は パニック を 引き起こし た 。

修正訳 揺れ は 約 2分間 続き 、 パニック を 引き起こし た 。

↑ ↑ ↑ ↑ ↑ ↑ ↑ ↑ ↑ ↑ ↑ ↑

Ins Ins OK OK Sub Sub Del OK OK OK OK OK

■ 結果

- 95% (970/1,018) の文にPEが必要 → MTはまだまだ
 - cf. human parity [Hassan+ 18]

(3) 文書単位PE

- シャッフルしてあった97文書の文を整列
- 文書ごとに必要最低限の修正を施す
 - 文ごとのときと同じ制約
 - 結束性, 一貫性などに注意
- 結果
 - 31% (320/1,018) の文にPEが必要 → 文単位の翻訳の困難さ
 - PEの必要性はMTの品質の自動評価尺度では捉えられない

訳出物	BLUE計算時の参照訳	
	ALTの参照訳	文書単位PEの結果
ALTのみを用いた場合のMT訳	14.6	15.9
フェーズ1終了時のMT訳	29.0	40.3
フェーズ2終了時のMT訳	35.8	57.6
フェーズ3終了時のMT訳	36.0	58.6
文単位PEの結果	36.8	95.0
文書単位PEの結果	36.8	100.0

Post-edited
[Total 19]

PE事例 (ID=139760)

原文

John Paul's six-day tour was hugely popular, attracting hundreds of thousands of people from the Catholic religion. It became the first ever UK visit from the Pope, which will make Pope Benedict's visit the second papal one.

MT訳

(1) 文単位MT訳の生成

ジョン・ポールの6日間のツアーは非常に人気があって、カトリックの宗教から何十万もの人々を引きつけた。それはローマ教皇からの初めての英国訪問となり、ローマ教皇ベネディクトの第二の教皇訪問となる。

文単位PEの結果 (今のMTの目標)

(2) 文単位PE

ジョン・ポールの6日間のツアーは非常に人気があり、カトリックを信仰する何十万もの人々を引きつけた。それはローマ教皇の初めての英国訪問となり、ローマ教皇ベネディクトの訪問は2度目の教皇訪問となる。

文書単位PEの結果

(3) 文書単位PE

法王ヨハネ・パウロの6日間の訪問は非常に人気があり、カトリックを信仰する何十万もの人々を引きつけた。この時がローマ法王の初めての英国訪問となり、法王ベネディクトの訪問は2度目の法王訪問となる。

修正事例の分析

■ MTの限界点と最低限の翻訳の差を知りたい

- (3) 文書単位PEの内容を分析する

■ 手順

1. 修正事例の列挙: PE前後の表現対の抽出・分解 [宮田+ 17]

- 原文の対応箇所も見ながら

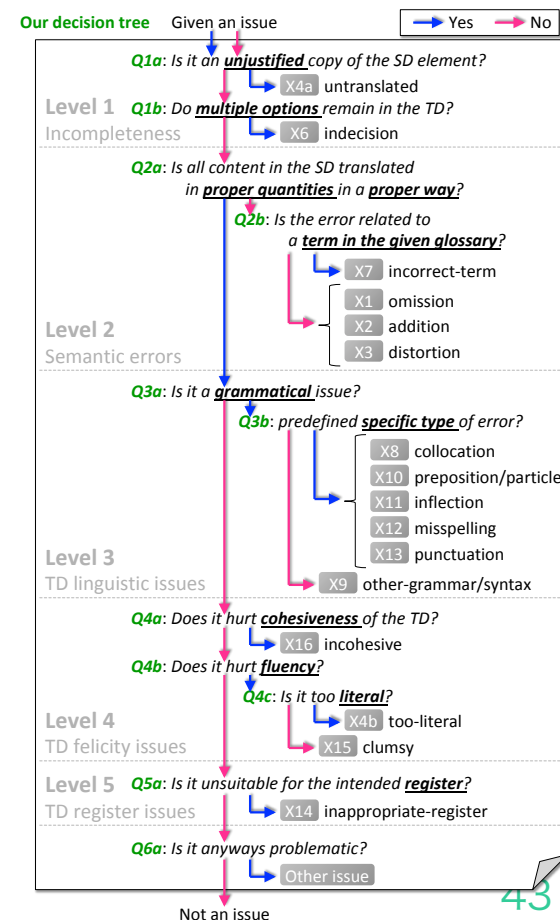
2. 個々の修正事例に対するアノテーション

- 参照すべき情報の有無とその種類

PEに必要な 参照情報		非言語的な情報や 文書の外側の情報	
		不要	必要
当該文書内の 言語的な情報	不要	(a)	(c)
	必要	(b)	(d)

- MNH-TTの校閲カテゴリ体系 [豊島+ 16]

- 人間の英日翻訳向けに設計された体系



修正事例の列挙・分解・分類

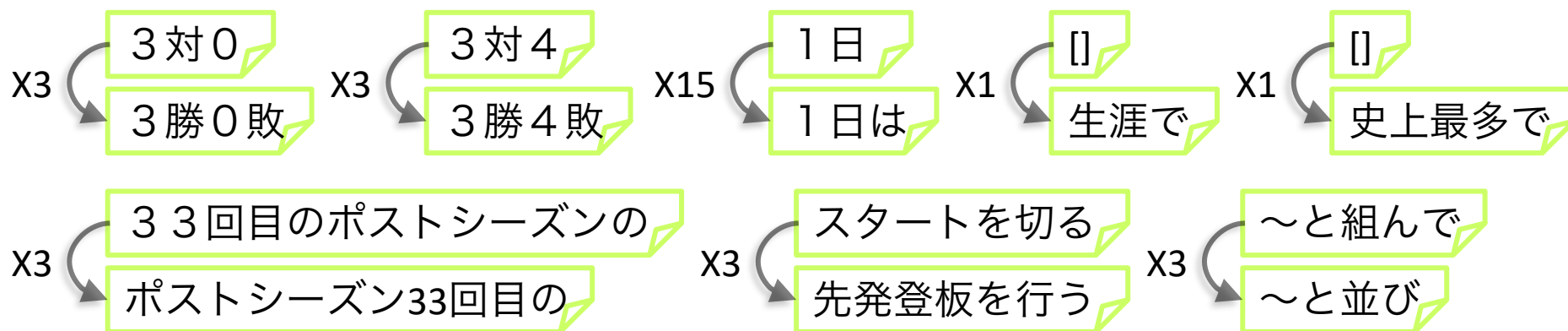
文単位PEの結果(今のMTの目標)

クレメンス（ワールドシリーズ7回出場で3対0、防御率1.90）は、少なくとも1日ペティット（ワールドシリーズ10回出場で3対4、3.90）と組んで、土曜日に33回目のポストシーズンのスタートを切る。

文書単位PEの結果

(3): 文書単位PE

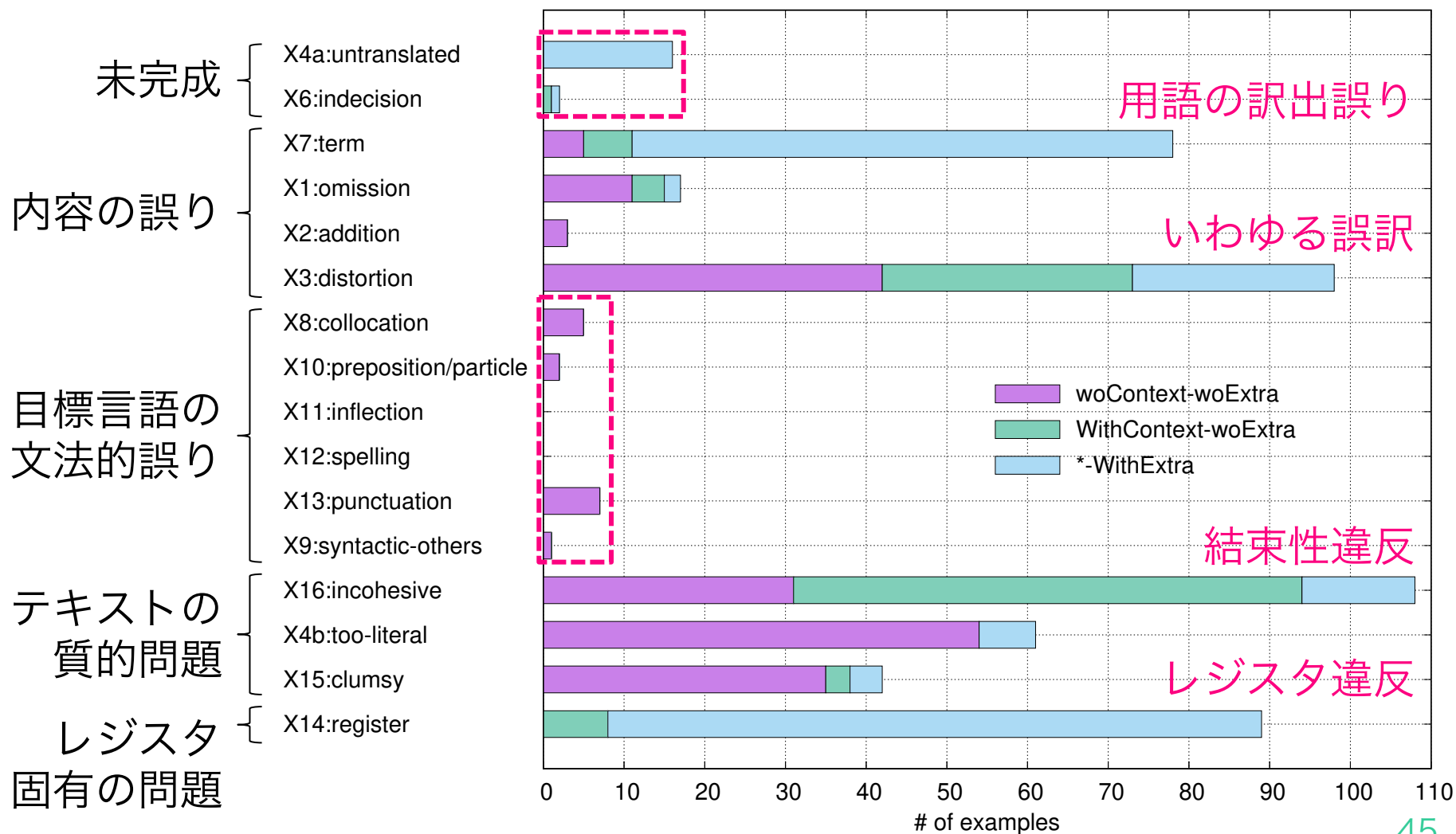
クレメンス（ワールドシリーズ7回出場で3勝0敗、防御率1.90）は、少なくとも1日はペティット（ワールドシリーズ10回出場で3勝4敗、3.90）と史上最多で並び、土曜日に生涯でポストシーズン33回目の先発登板を行う。



例示のために対象部分のみを示しているが
実際には文脈込みの判断であることに注意

修正事例の分布 (合計529件)

- 文書中の他の文を参照する必要の有無で細分類
 - 人間の翻訳の分析用に作られた分類体系 [豊島+ 16]



必要だが参照できていない情報 (1/4)

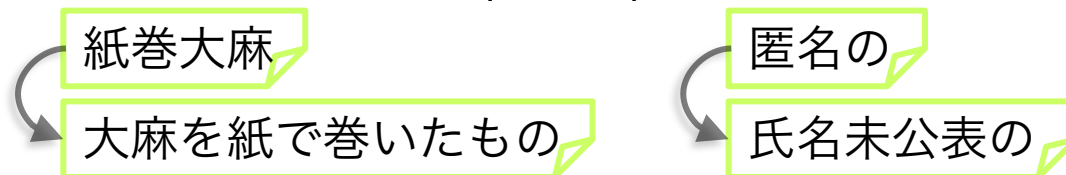
■ 翻訳に関する詳細な仕様 (121事例)

● 【X14 レジスタ違反】 (81事例)

■ 表記規則の非遵守: 語彙, 文字, 社名, 括弧, 文末表現



■ 情報レベルが未制御 (3事例)



● 【X4a 未翻訳】 (16事例)

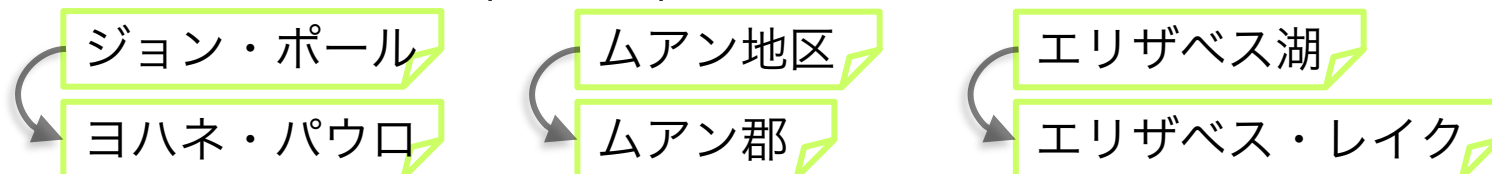
■ アルファベットのまま放置されている固有表現



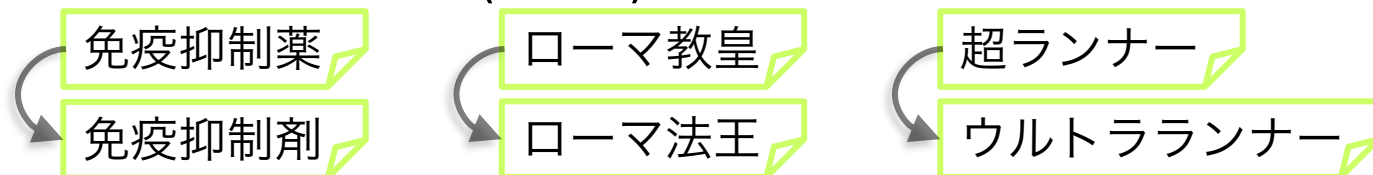
必要だが参照できていない情報 (2/4)

■ 用語集 (80事例)

- 【X7 用語の訳出誤り】 (64事例)
 - 用語の認定や既訳の遵守には用語集が不可欠
 - 固有表現の訳出誤り (56事例)



- 専門用語の訳出誤り (8事例)



- 【X4a 未翻訳】 (16事例, 前出)

必要だが参照できていない情報 (3/4)

■ 分野固有の知識 (31事例)

- 【X3 誤訳】 (18事例)

Clemens ... 3-0 in seven starts

クレメン스는7回出場で3対0

クレメン스는7度先発して3勝0敗

track work will recommence

保線工事が再開される

トラックでの調教が再開される

- 【X16 結束性違反】 (5事例)

At least seven people were killed by the action.

少なくとも7人がその戦闘によって死亡した

少なくとも7人がその軍事行動によって死亡した

"I am going to be so pumped up to get out there."

「そこに出るのはとても興奮するだろう。」

「マウンドに上がるのはとても興奮するだろう。」

必要だが参照できていない情報 (4/4)

■ 事実確認のための外部資料 (8事例)

- e.g., 表現の曖昧性解消
- 今回の観察事例は少ないが、決め打ちが偶然正解した場合もある
- 【X3 誤訳】 (7事例)

List MP Phil Twyford has not put his name forward.

国会議員フィル・トワイフォードは、自身の名前を提示しなかった。

国会議員フィル・トワイフォードは、立候補していない。

- 【X6 曖昧性未解消】 (1事例)

David Miliband, his brother and a fellow UK politician, was ...

彼の兄弟で英国の政治家仲間であるデイビッド・ミリバンドは ...

彼の兄で英国の政治家仲間であるデイビッド・ミリバンドは ...

参照すべき情報が欠けることで生じる問題

■ 欠けていた情報

- 翻訳に関する詳細な仕様
- 用語集
- 分野固有の知識
- 事実確認のための資料

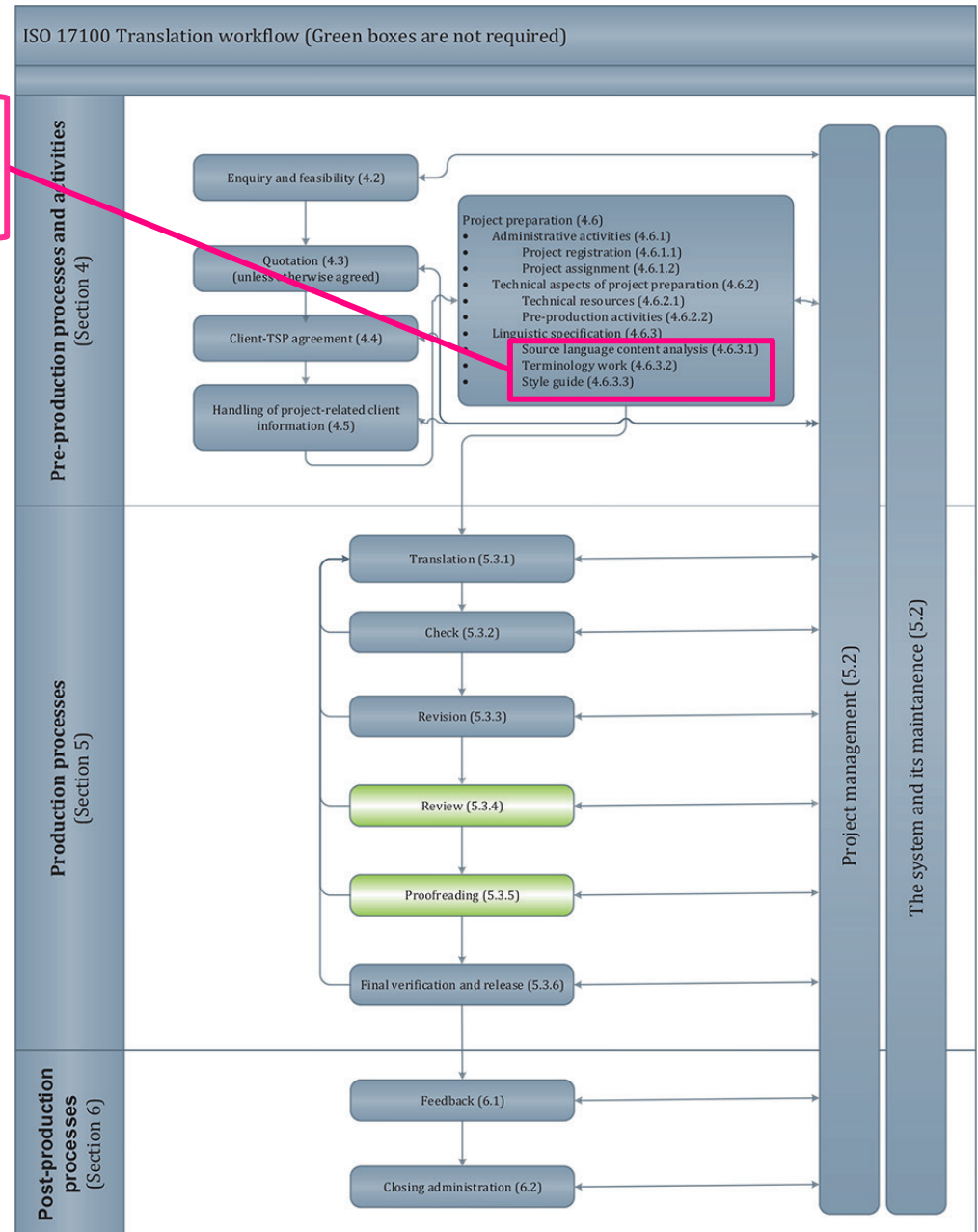
後者2つは翻訳者任せ?

■ 問題 (頻度の上位4つ)

- レジスタ違反 (未翻訳)
- 用語の訳出誤り
- いわゆる誤訳
- 結束性違反

■ 一部はNMTでも扱える

- 現状では間接的
- 陽に与えてちゃんと解くことを要検討



いくつかの提言とまとめ

MTの研究者や開発者に向けての提言

■ 何を目指すか?

- 全自動? 人間との共存?
- X文Y訳にしても問題は山積
 - 文単位PEで95%が修正された
 - 文書単位PEの約6割(312/529)が言語的情報のみを参照
- 参照・遵守していない情報を陽に扱う方向に進むか?
 - 知識, 翻訳の仕様, 用語集
 - 一部は制約付きデコーディング [Post+ 18] に期待
 - 一部はデータによって間接的に表そうとするに留まっている
 - 文脈情報
 - e.g., ゼロ照応, 指示詞, 上位-下位, オブジェクト-属性

■ 文書レベルの自動評価技術を要確立

- 訳出の一貫性, 結束性など

MTユーザに向けての提言

- 現在のMTは必ずしも「翻訳」をしない
 - 翻訳時に参照すべき情報の多くを見ていない
 - X文Y訳に過ぎない(うまくいったとしても)
- そのようなMT (サービス)を利用する際の注意
 - 使おうとしているMTサービスの利用規約を要確認
 - 何を参照しているか?
 - 翻訳誤りの責任は誰が負うのか?
 - 使い方をよく考える
 - 誰が消費するのか?
 - 翻訳の目的や仕様
 - さもなくば、深刻な翻訳過誤を生じうる

翻訳リテラシ
MTリテラシ

堺筋線

Sakai Muscle line

高塚川周辺に避難勧告が出ました

Foi emitido **aviso para se refugiar** proximidades do rio Takatsuka.

MTの使い方の例

■ 情報収集(assimilation): 消費者は自分

- 流暢でなくても意味が通じればOK
- 内容が正確でない可能性を常に疑う

■ 情報発信(dissemination): 消費者は他人

- 正確かつ流暢でないと発信者の信頼を損なう
- MT結果を人間が仕上げる
 - e.g., 後編集(PE), 複数人による検査
- 起点言語文書をMTできるように工夫する(i.e., 前編集) [Miyata+ 21]
 - e.g., 極端に長い文を避ける
 - e.g., 主語や述語, 構文を明瞭にする
- 折り返し翻訳を見て訳質を確認する [VoiceTra]
- MT結果であることを明記して注意を促す

まとめ

■ ゴール: 機械翻訳(machine translation; MT)の課題の共有

- 翻訳とは異なり, 必要な情報を参照していないこと
- それによって生じる問題の分析方法
- ニュース記事の英日翻訳の誤り分析結果

■ 構成

- MTの概要
 - 情報通信研究機構(NICT)における研究開発
 - 近年用いられているMTのしくみ
 - MTが今後できそうなこと
- 翻訳とMTの違い
 - 翻訳に必要な情報
 - それらを欠くMTが生じる問題
 - MTを使う際の注意点
 - 存在感を増すMTとどう付き合っていくか?