

< 韓国日本語学会学術大会, Sep. 19th, 2020 >

機械翻訳のしくみ，翻訳との違い

藤田 篤



情報通信研究機構

NICT, Japan

講演内容の概要

■ 機械翻訳(machine translation; MT)の進展

- 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

- 翻訳に必要なこと
- MTが実際に果たしていること
- MTを使う際の注意点

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 and 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 (日本語版) and English (英語版) versions, each available in 5-minute and 18-minute formats. The bottom window shows a grid of service tiles. On the left, there's a section for disaster response with 'DISAANA' and 'D-SUMM' systems. The main grid includes tiles for '日本標準時' (Japan Standard Time), '校正サービス' (Calibration Service), '総合テストベッド' (General Test Bed), 'CYDER' (Cyber Defense), 'SecHack365', '宇宙天気予報' (Space Weather Forecast), '災害時関連技術' (Disaster-related Technology), 'WISDOMX' (Web Question Answering System), 'VoiceTra' (Voice Translation), '翻訳バンク' (Translation Bank), 'ひまわり8号' (Himawari-8 Real-time Web), 'NICTERWEB', 'NOTICE', 'NICT SEEDs', 'ICTスタートアップ' (ICT Startup Support Center), '情報バリアフリー' (Information Barrier Free), 'Smart IoT Acceleration Forum', 'AI Japan R&D Network', '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, and 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

CYDER

実践的サイバー防衛演習

SecHack365

宇宙天気予報

校正サービス

総合テストベッド

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紹介ビデオ

5分版 18分版

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のため2014年に新設

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):

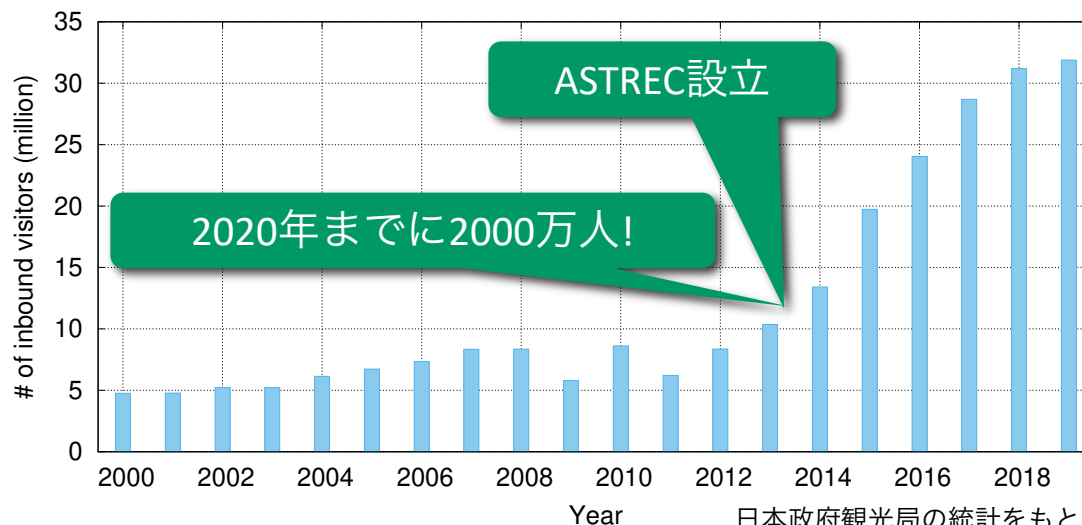
- Header: ASTREC logo (先進的音声翻訳研究開発推進センター), NICT logo, and navigation links: ホーム, 概要, 組織, NEWS, 採用情報, アクセス, English.
- Banner: さあ、世界中の人と話をしよう。言語の壁をなくす多言語音声翻訳アプリ VoiceTra. Includes a smartphone showing the app interface.
- Sidebar (Left):
 - OUTLINE: 世界の「言葉の壁」を打ち破り、グローバル実装を推進します。
 - NEWS:
 - メディア 2020/07/30 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 (button), ログインしたままにする (checkbox).
 - 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。

ASTREC設立の経緯

- 2013年9月
東京オリンピック・パラリンピックの開催決定
- 2014年4月
総務省 グローバルコミュニケーション計画
 - 世界の「言葉の壁」をなくす
 - 多言語音声翻訳技術の開発と社会実装
- 急増する訪日外国人 [日本政府観光局 17]



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」
Tumbler T-01
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」(ボイスビズ)
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

多言語音声翻訳アプリVoiceTra



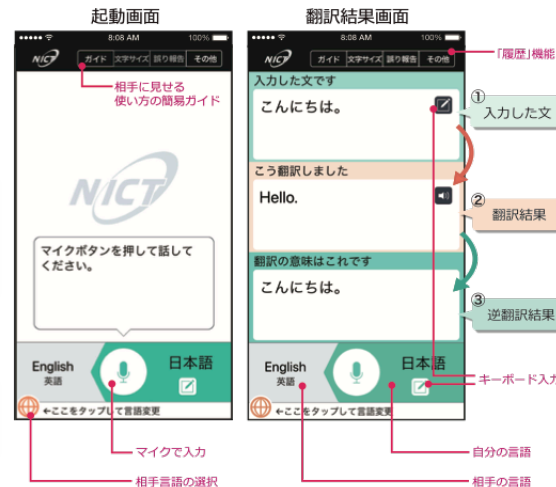
NICTの無料アプリ

- 31言語に対応
- 旅行会話が得意
- 500万ダウンロード

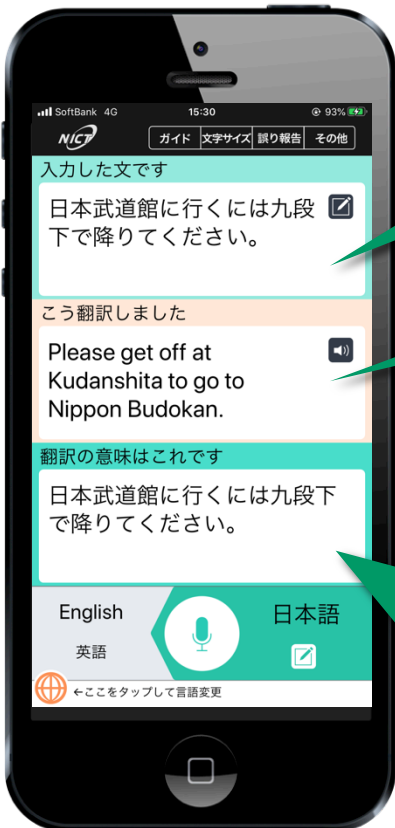
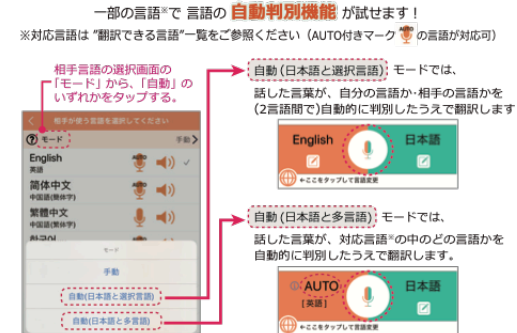
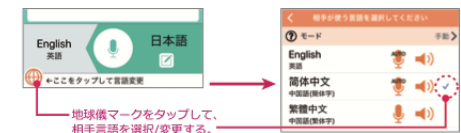
重点10言語 + 3言語を強化中

- 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, ネパール, ブラジルポルトガル

画面の説明



相手言語の選択



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)

翻訳できる言語 (31言語)

中国語・ポルトガル語の方言を含む

--- 日本語 ~ ポルトガル語(ブラジル) を重点研究言語としています ---

🔊 音声で入力できる (18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 (17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)
- アラビア語
- イタリア語
- ウルドゥ語

- 🔊 オランダ語
- 🔊 クメール語
- 🔊 シンハラ語
- 🔊 デンマーク語
- 🔊 ドイツ語
- 🔊 トルコ語
- 🔊 ネパール語
- 🔊 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
- 🔊 モンゴル語
- 🔊 ラオ語
- 🔊 ロシア語

VoiceTraの音声翻訳のしくみ

(1)入力された音声をネットワーク上のサーバに送信

VoiceTra

音声入力

駅はどこですか

音声出力

Where is the station?

(2)翻訳された音声サーバから戻ってくる

多言語音声翻訳サーバ

音声認識

音声→テキスト

テキスト翻訳

e.g., 日本語テキスト→英語テキスト

音声合成

テキスト→音声

駅はどこですか

Where is the station?

統計・機械学習・
人工知能技術

統計・機械学習・
人工知能技術

統計・機械学習・
人工知能技術

日本語の音声データ・
テキストデータ・辞書

日本語と英語の
対訳辞書・対訳データ

英語の音声データ・
テキストデータ・辞書

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

自動翻訳

英語 ↔ 日本語

汎用NT 【英語 - 日本語】 etc 2

翻訳

The above equalizer circuit 30 can be made to output the negative-phase output signal by reversing the positive- and negative-phase input signals INP and INN, and in this case, the equalizer circuit outputs a signal having a waveform similarly compensating for the influence of noise contained in the negative-phase input signal INN.

翻訳不要記号を押入

☐ 文章を結合する

☒ 文章を区切る

比較 原文

The above equalizer circuit 30 can be made to output the negative-phase output signal by reversing the positive- and negative-phase input signals INP and INN, and in this case, the equalizer circuit outputs a signal having a waveform similarly compensating for the influence of noise contained in the negative-phase input signal INN.

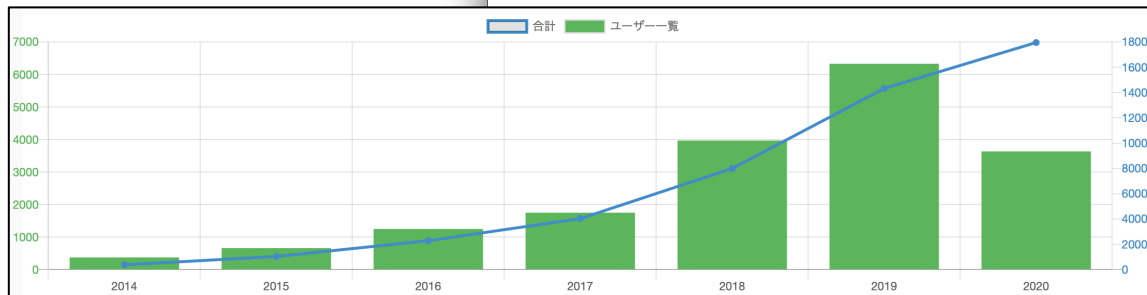
☒ 汎用NT 【英語 - 日本語】

上記等化回路30は、正相入力信号INPおよび逆相入力信号INNを反転させて逆相出力信号を出力させることができ、この場合、逆相入力信号INNに含まれるノイズの影響を同様に補償した波形の信号を出力する。

☒ 特許NT 【英語 - 日本語】

上記のイコライザ回路30は、正相入力信号INPと逆相入力信号INNとを反転させて逆相出力信号を出力させることができ、この場合、逆相入力信号INNに含まれるノイズの影響を同様に補償した波形の信号を出力する。

訳文コピー リセット



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

深層学習に必要な膨大な学習データ

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

統計処理に必要な大量な対訳データ

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

1万規則/言語対

人が書ける
ルールデータ



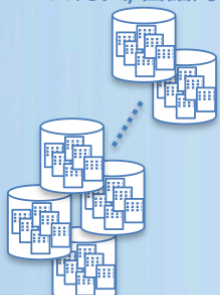
1986
音声翻訳の研究開始

実験室で実験

2000 社会還元加速プロジェクト

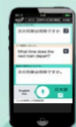
出典：
https://www.soumu.go.jp/main_content/000678485.pdf

100万文/言語対



派生技術
要約自動生成技術
（日仏翻訳）

ニューラル翻訳の実装



社会へ

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

2017



多様な翻訳装置の開発開始

2020

日常生活やビジネスを
支える翻訳

2025

文脈・話者の意図等を
補う翻訳

2030

シビアな交渉にも
使える通訳

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



文字で確認
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の性能評価

■ 性能評価 → 性能向上

- 色々な観点: 翻訳品質, 処理速度, 対応言語, 対応分野

■ 翻訳品質の評価尺度

- (全)自動評価: 信頼性^低, 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(見かけ上の)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?}, コスト^大
 - 正確さ(adequacy) / 流暢さ(fluency)
 - 後編集(post-edit)の分量: e.g., HTER [Snover+ 06]



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」

- 冬の時代の到来

■ 1990年頃以降: コーパスに基づくMTの研究が活発化

- コーパス: (人間が作った)テキストを集めたデータ

1984

用例ベースMT
(EBMT)

[Nagao 84]

1988

統計的MT
(SMT)

[Brown+ 88]

2003

句に基づくSMT
(PBSMT)

[Koehn+ 03]

2014

ニューラルMT
(NMT)

[Sutskever+ 14]

2017

Transformer
(NMTの一種)

[Vaswani+ 17]

NMTの基本構造

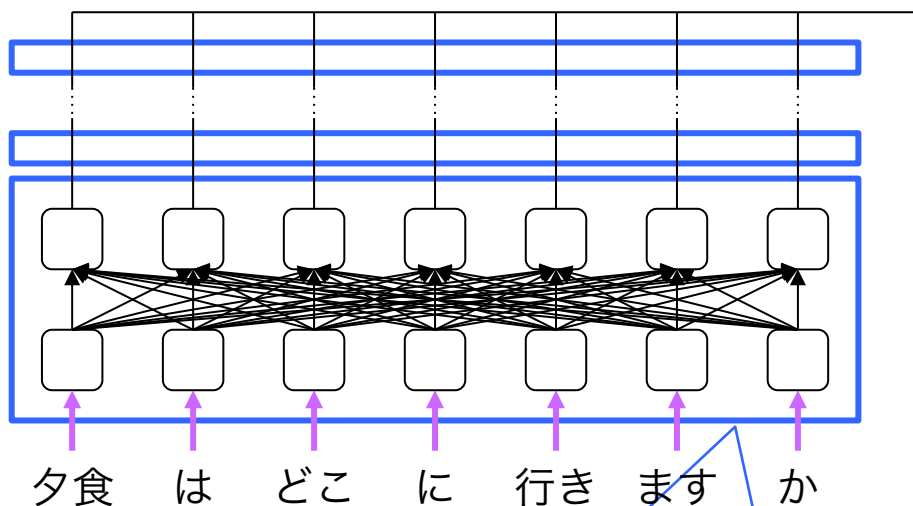
■ 語の系列 → 語の系列

- 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]

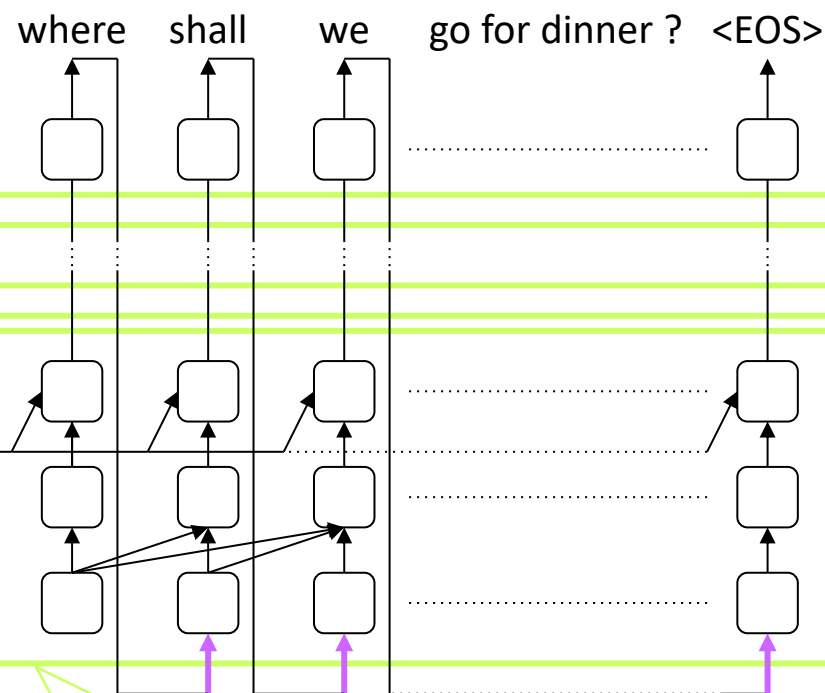
語の分散表現 (embedding)

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

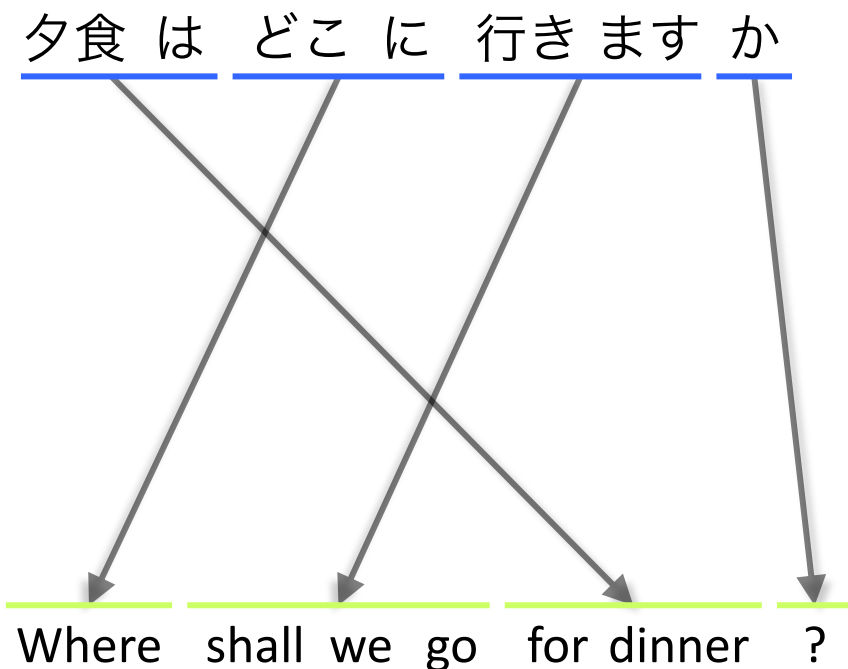


デコーダ (decoder)
数値データから語の系列を生成

PBSMTとNMT

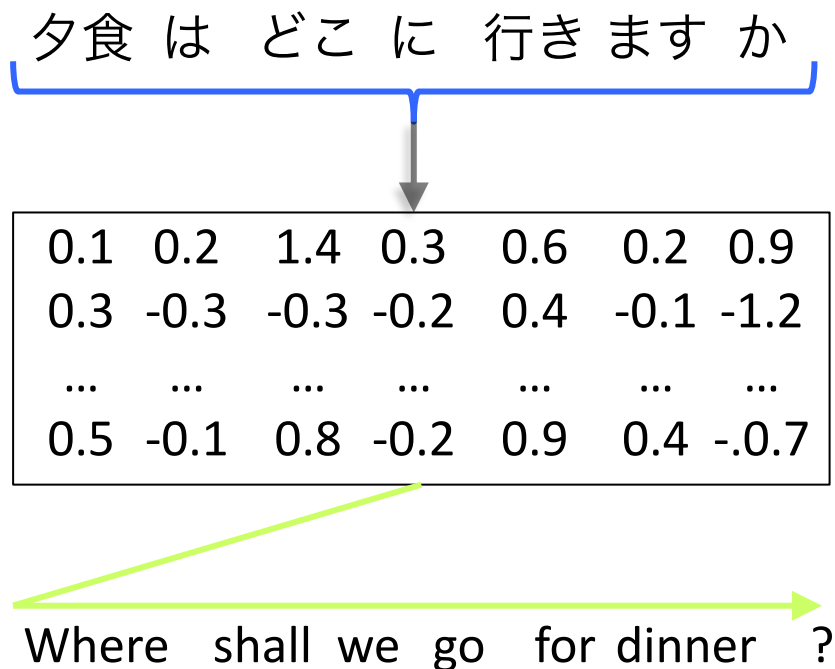
■ PBSMT: 句に基づくSMT

- 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



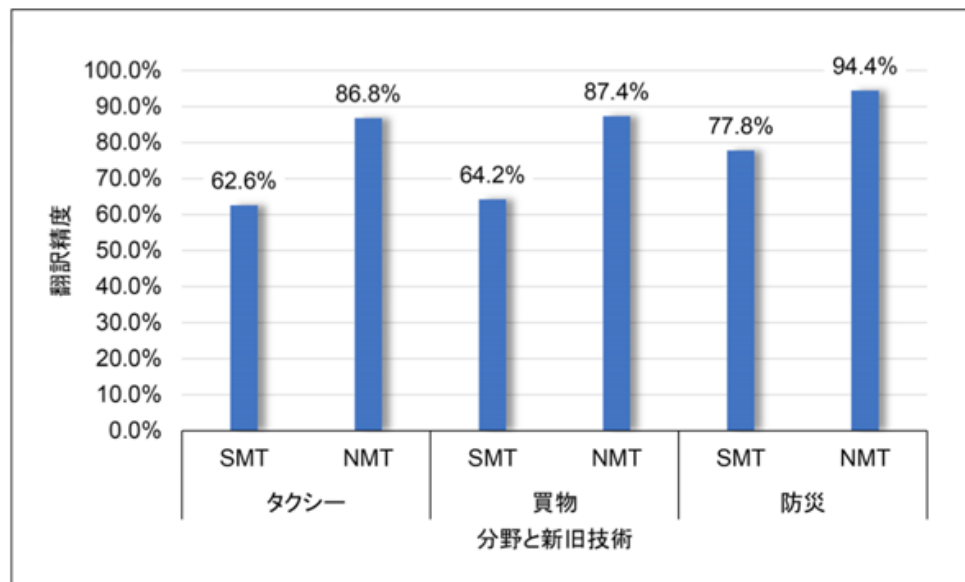
■ NMT: ニューラルMT

- 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



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」!?
[Barrault+ 19]



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, Arul 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](#)

対訳データ

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。

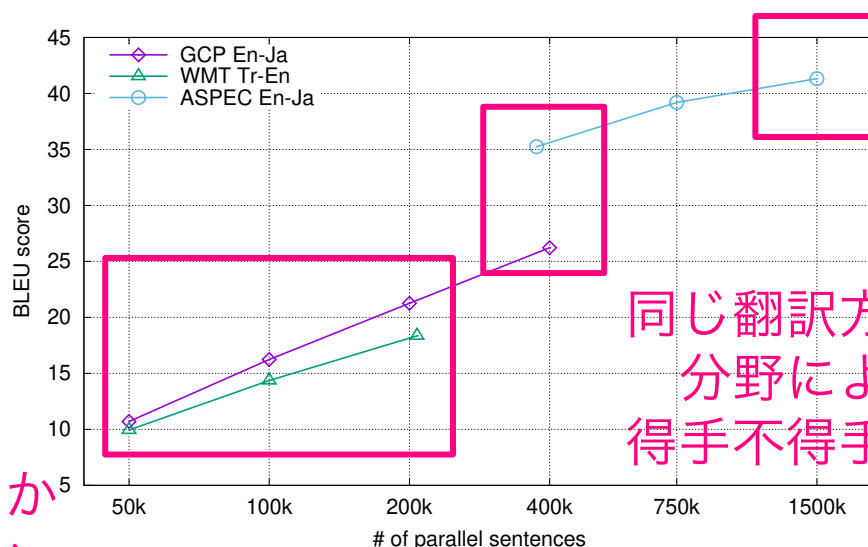
この例 [Riza+ 16] は文書単位の翻訳だが
MT研究では文単位の翻訳もよく使われる

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., 我々のNMTの実験結果の例 [Dabre+ 19]



数百万文あれば
利用できるレベル
(一部の言語・分野)

同じ翻訳方向でも
分野によって
得手不得手がある

対訳が数万文対しかない
と性能は低い
(多くの分野,
少資源言語対)

対訳コーパスの拡充

■ 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)

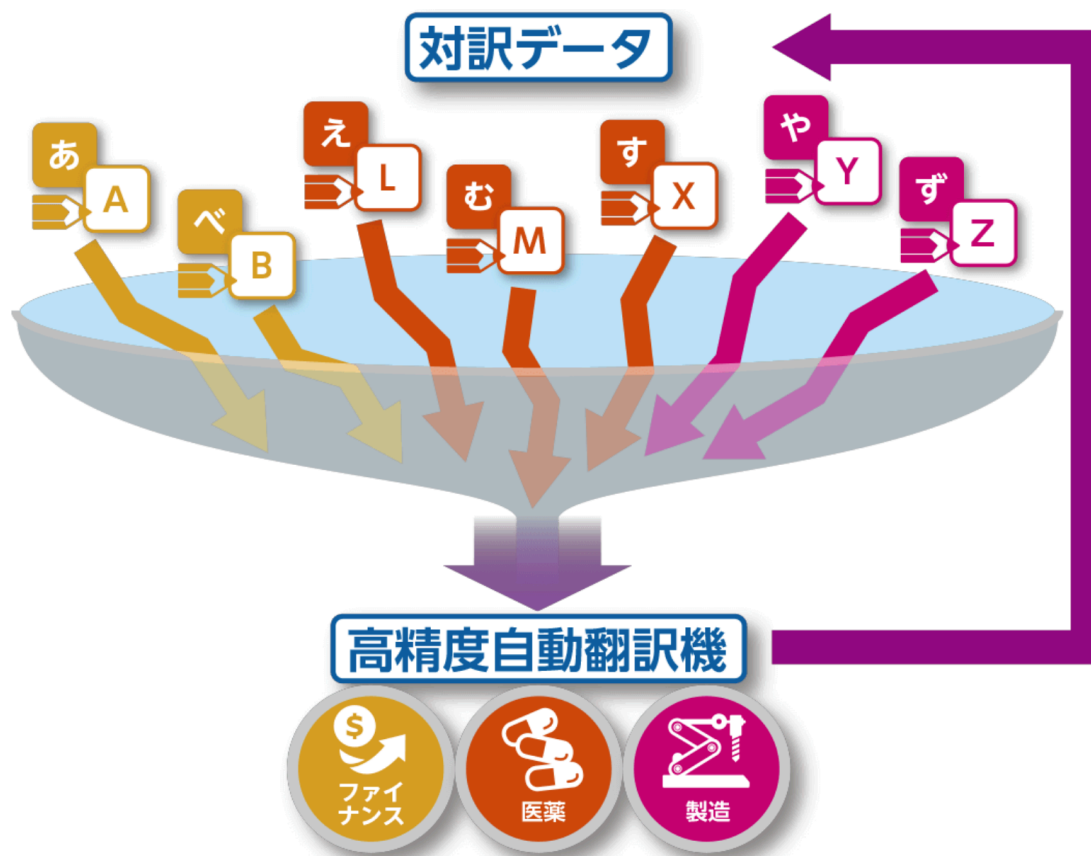
■ 既訳の収集

- 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 (Google 他)

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，ビジネス分野で成果

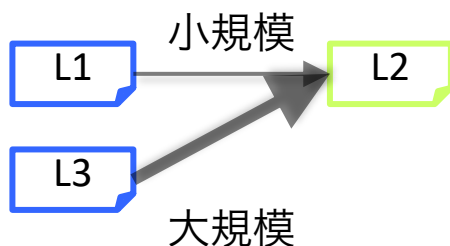


他の種類のデータの利活用 (1/2)

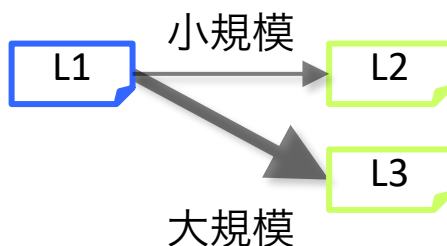
■ 他言語対・他分野の対訳

- 他言語対からの転移学習 [Zoph+ 16][Johnson+ 17]
 - e.g., 仏-英 → 越-英
- 他分野からの転移学習 [Chu+ 17]
 - e.g., 科学技術文書 → TEDトーク

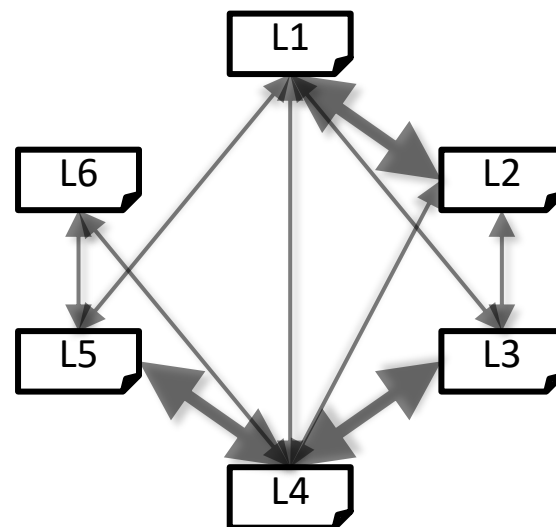
デコーダ共有



エンコーダ共有



Multi-to-multi



L1→L6やL2→L5はZero-shot

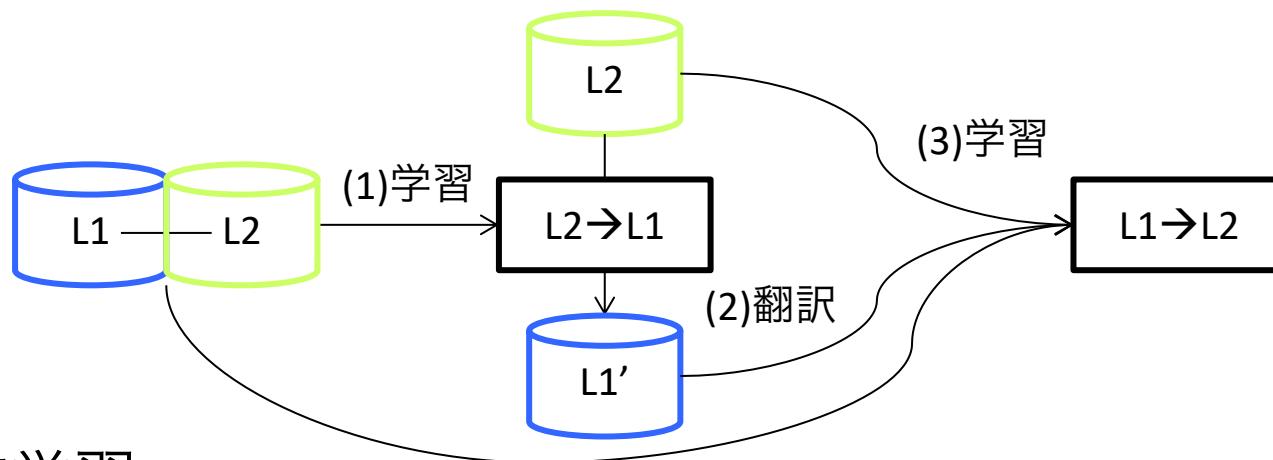
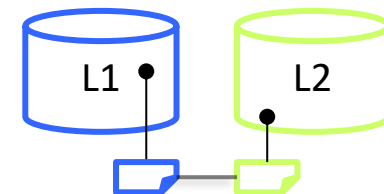
他の種類のデータの活用 (2/2)

■ 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]



- 事前学習

- 語の分散表現やエンコーダ内部の変換行列を学習
- e.g., BERT [Devlin+ 19], XLM [Conneau+ 19], MASS [Song+ 19]

MTができそうなこと = X 文 Y 訳 + α

■ X 文 Y 訳

- テキストからテキストへの変換に過ぎない
 - テキストの「意味」を理解しているとも限らない
 - テキストの社会的・歴史的 position付けは理解できない

■ + α

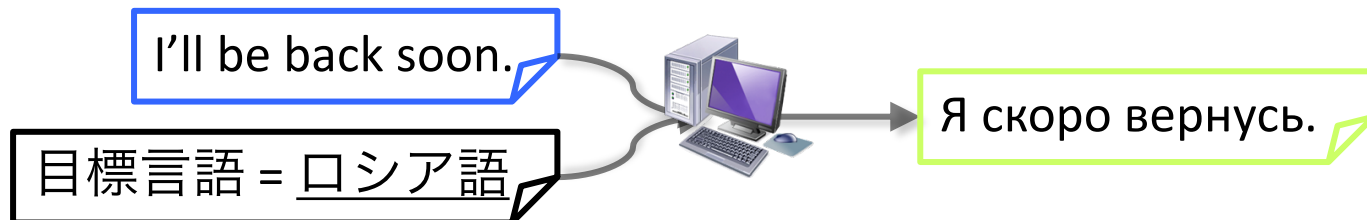
- どんな情報でも数値データに変換して利用できる
 - 翻訳対象テキストと同様、陽に与える必要はある

MTができそうなこと (1/2)

■ どんな情報でも数値データに変換して利用できる

- テキスト内の情報

- 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]



- 専門分野やテキストタイプへの適応 [Luong+ 15][Chu+ 17]

MTができそうなこと (2/2)

■ どんな情報でも数値データに変換して利用できる

- テキストの外界の情報

- 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

Tu bois du thé

こんにちは藤田さん

聞き手の性別 = 男性

Bonjour, monsieur Fujita.

ごはん行きましょう

時刻 = 11:55 a.m.

Shall we go for lunch?

関係 = 同僚 ∧ 話し手 < 聞き手

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

翻訳プロセスの確認

[影浦 17]の問い

■ 翻訳して、と文書を与えられたら、どうする？

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
 - 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

■ cf. MT

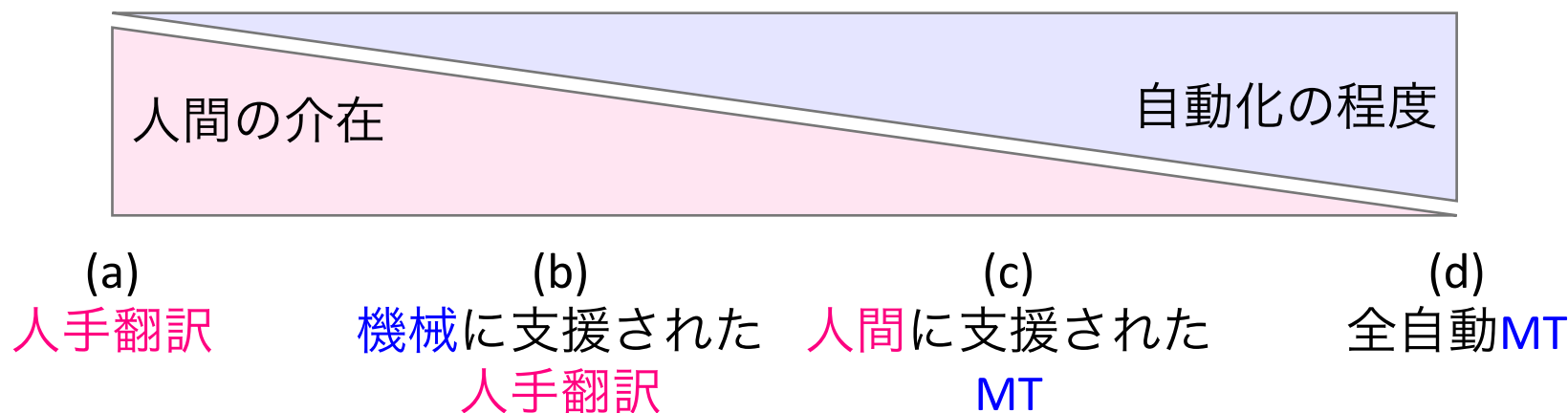
- 定式化のために単純化/無視 → X文Y訳

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]

- x文y訳としての品質の担保も、現時点では不可能
- 人間が適切に使いこなす必要あり

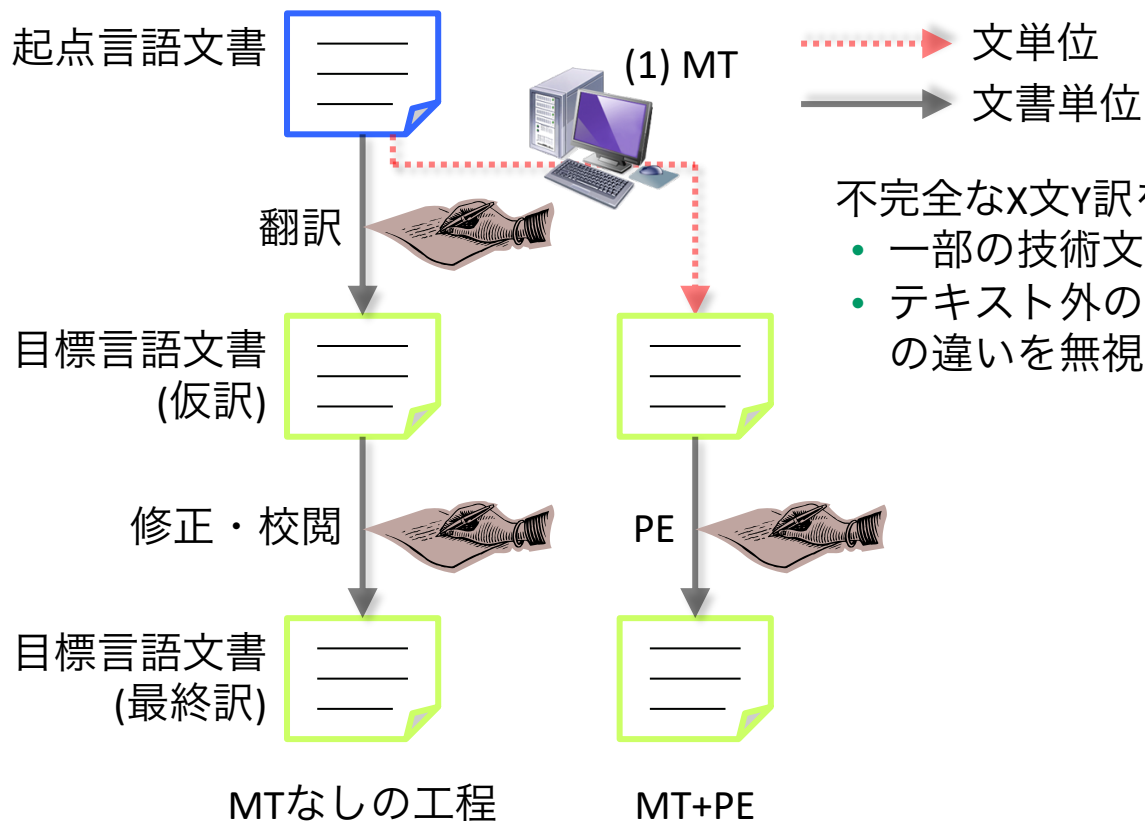
■ 4つの戦略 [Hutchins+ 92]



(c) 人間に支援されたMT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]



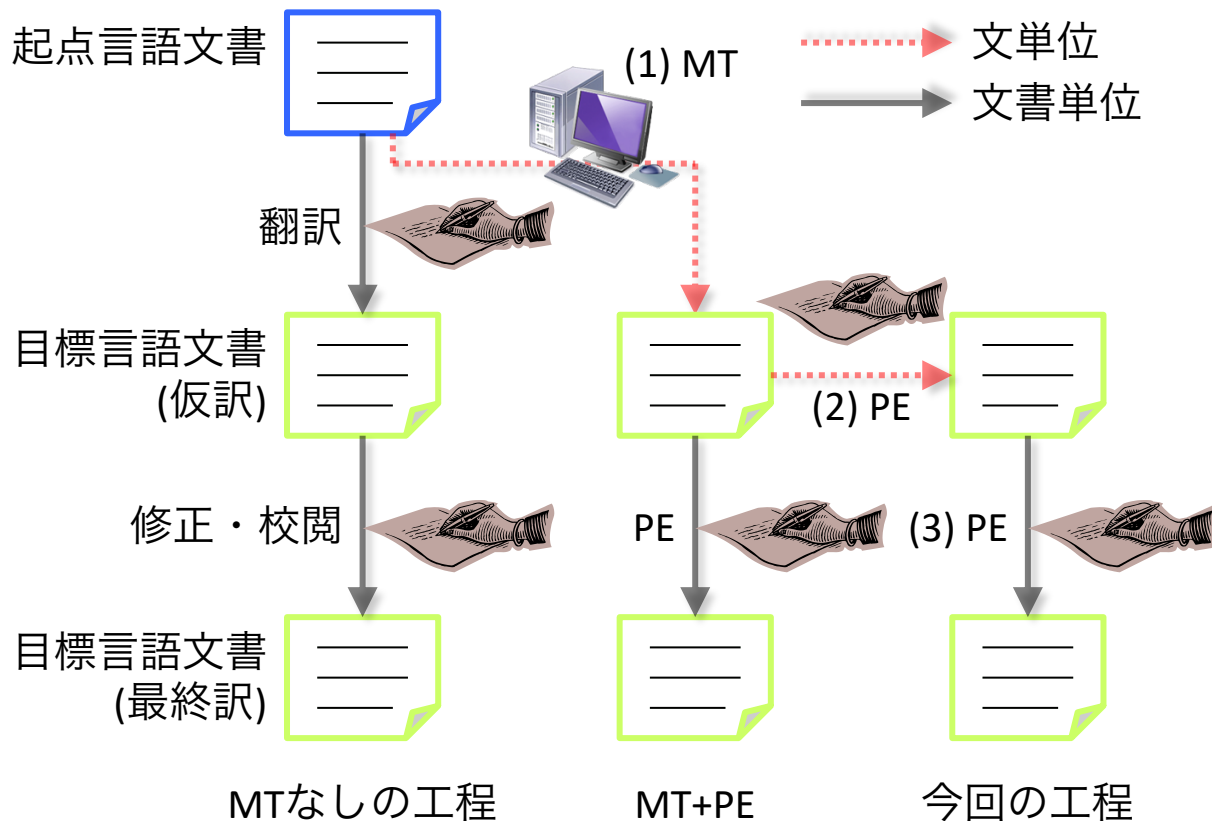
不完全なX文Y訳をPEで補える場面の例

- 一部の技術文書
- テキスト外の情報に関する前提知識の違いを無視できる言語対

(d) 全自動MTでどこまでできそうか?

■ PEを2段階に分けて課題を分析 [藤田 20]

- ニュース記事の英日翻訳



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

工程(1) 文単位の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。

文単位のPEの結果 (今のMTの目標)

工程(2) 文単位の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。

文書単位のPEの結果

工程(3) 文書単位の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。

修正事例の列挙・分解・分類

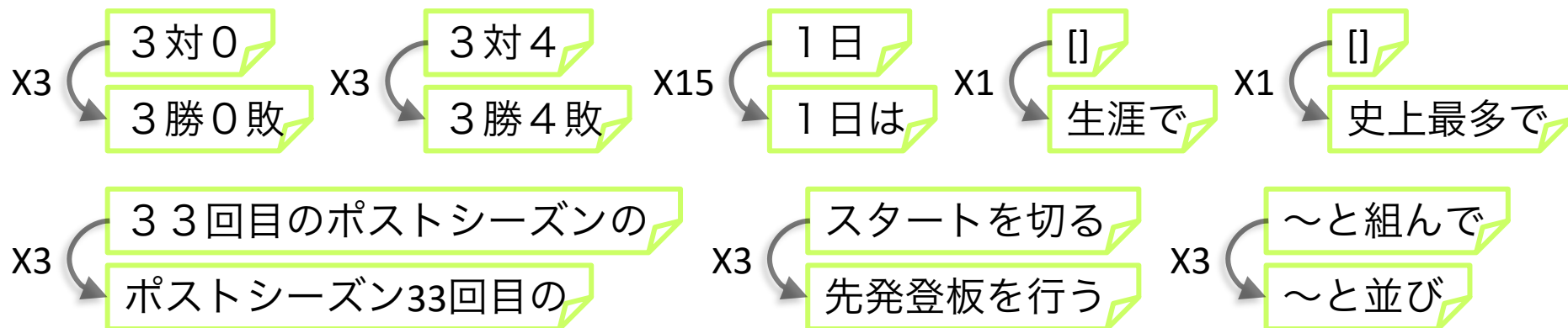
文単位のPEの結果(今のMTの目標)

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。

文書単位のPEの結果

工程(3): 文書単位のPE

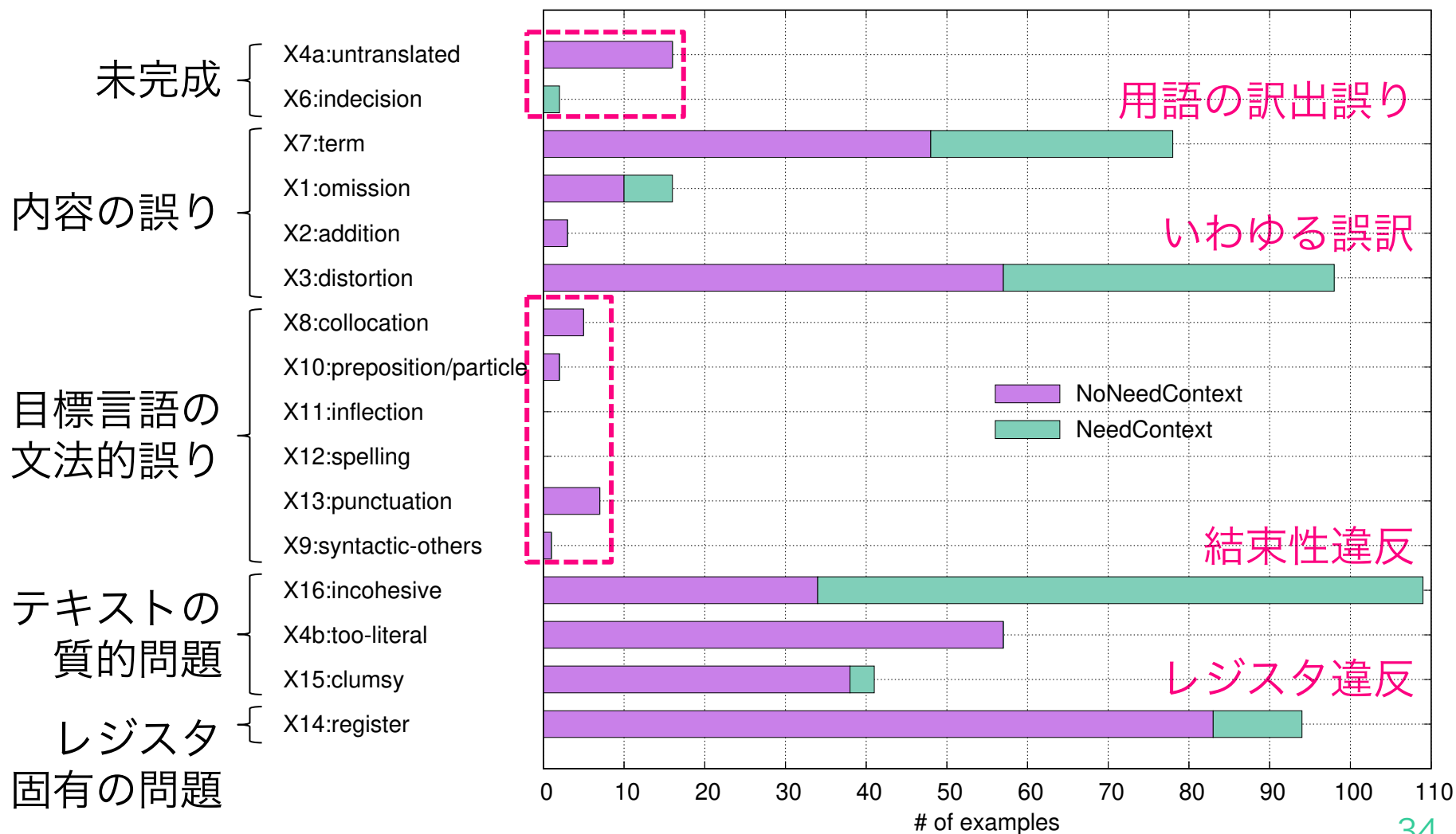
クレメンス（ワールドシリーズ7回出場で3勝0敗、防御率1.90）は、少なくとも1日はペティット（ワールドシリーズ10回出場で3勝4敗、3.90）と史上最多で並び、土曜日に生涯でポストシーズン33回目の先発登板を行う。



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]



文単位の翻訳における課題

■ X文Y訳を超えて実現すべきこと

- 必要な外部知識を要参照

- 一般常識
- 対象分野の専門的知識や表現
- 用語集: 既訳の踏襲, 一貫性
 - 固有名: 地名, 人名, 組織名など
 - 専門用語

- 翻訳の仕様を要考慮

- スコプス: 文書の役割や翻訳の目的
 - e.g., 想定読者, 文書の用途
- レジスタの規範, 案件ごとのスタイル等の指定
 - e.g., 固有名の翻字の要不要, メディアの空間的制約, 使用できる語や漢字, 括弧の使い方

Clemens ... 3-0 in seven starts

クレメンスは7度先発して3勝0敗

John Paul

ヨハネ・パウロ

machine translation

機械翻訳

文書内の他の文を参照する必要性

■ 修正事例中の32% (168/529)

- 【X16 結束性違反】 が最多の75件
 - 明示化: 指示詞や人称代名詞等の参照先の名詞句の明示
 - 省略: 自明な格要素の削除
 - 主題化/非主題化: 「は」と「が」の使い分け
 - 訳出の統一: 固有名は【X7 用語の訳出誤り】に分類
- 【X3 歪曲】 文脈依存の曖昧性を適切に解消する必要あり

Amnesty International has **also urged** the government of Pakistan to ensure the protection of journalists and media workers in the country.

X Amnestyはまた...

✓ Amnestyもまた...

Since completion of the surgery, the patient has been on immuno-suppressant drugs to prevent rejection of the new **limbs**.

X 足

✓ 手

翻訳におけるMTの使い方

MTを使う際の注意

■ 現在のMT

- 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)
- 翻訳誤りによって深刻な問題を生じうる可能性

■ 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

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)
 - 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
- MT結果であることを明記して注意を促す

まとめ

■ 機械翻訳(machine translation; MT)の進展

- 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

- 翻訳に必要なこと
- MTが実際に果たしていること
- MTを使う際の注意点