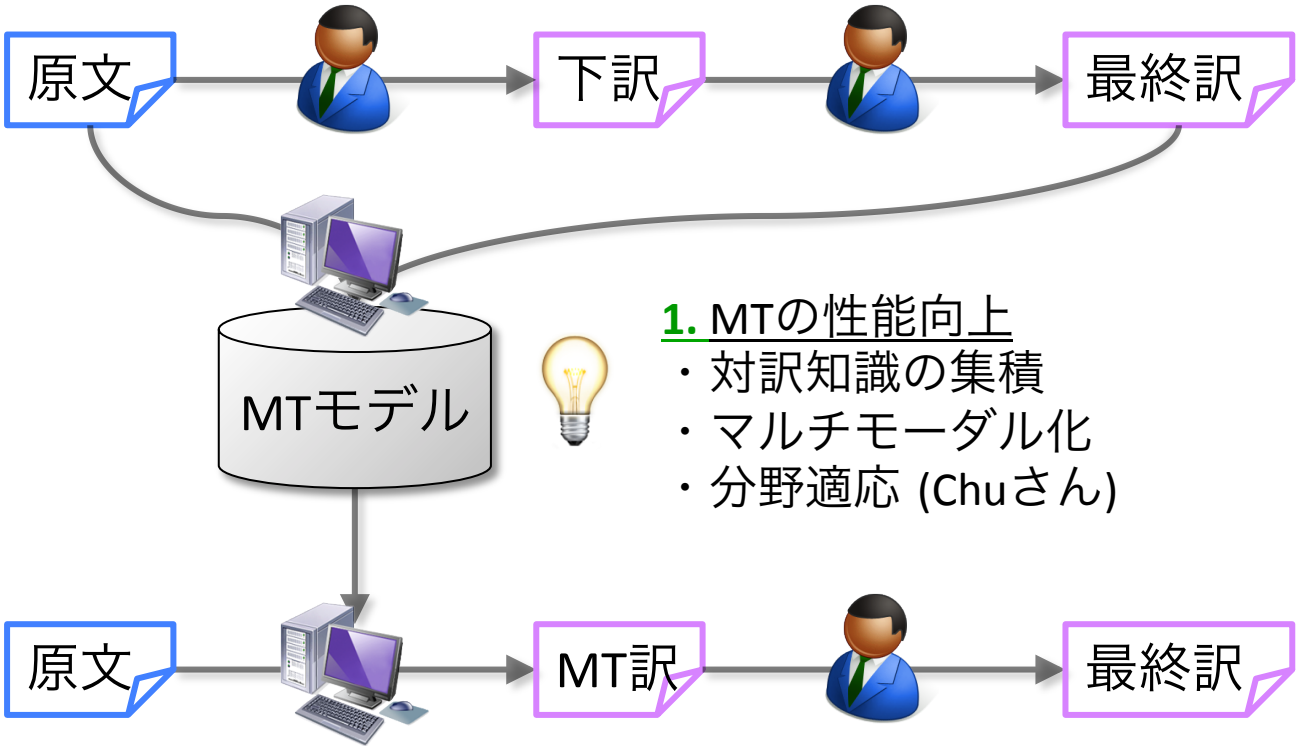
機械翻訳(MT)が
できそうなこと
できないこと

藤田 篤
情報通信研究機構 (NICT)
先進的音声翻訳研究開発推進センター (ASTREC)
atsushi.fujita@nict.go.jp / fujita@paraphrasing.org

翻訳に対する自然言語処理のアプローチ

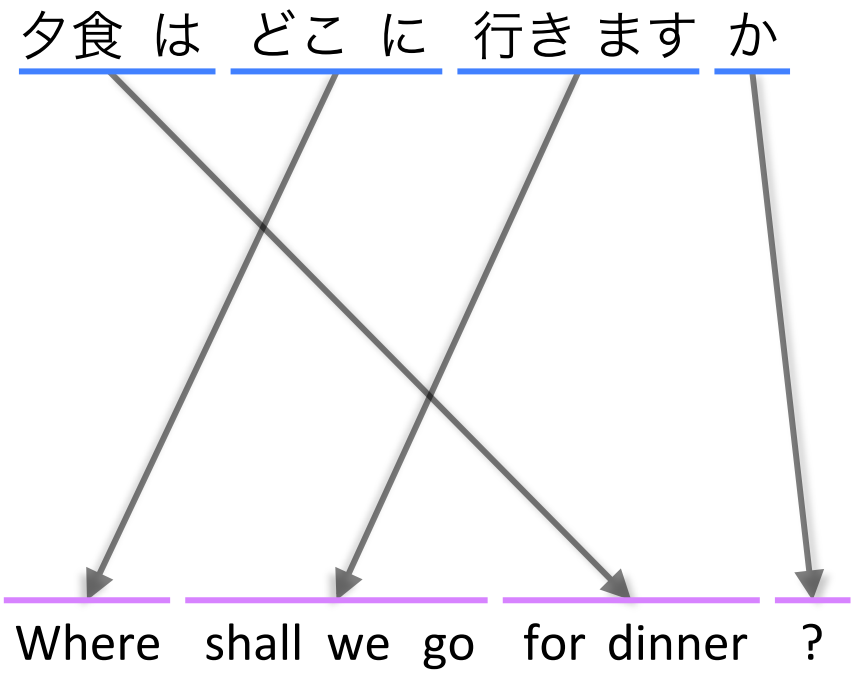


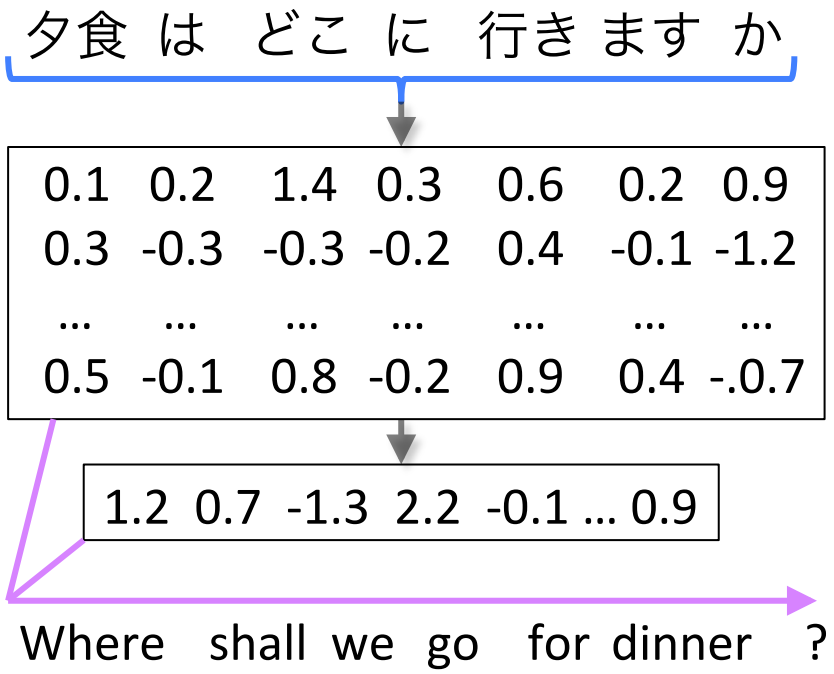
1. MTの性能向上
- 対訳知識の集積
- マルチモーダル化
- 分野適応 (Chuさん)

2. 完璧ではないMTの活用
- ポストエディットとその自動化
- 品質推定
- プリエディット (宮田さん)

SMTとNMT

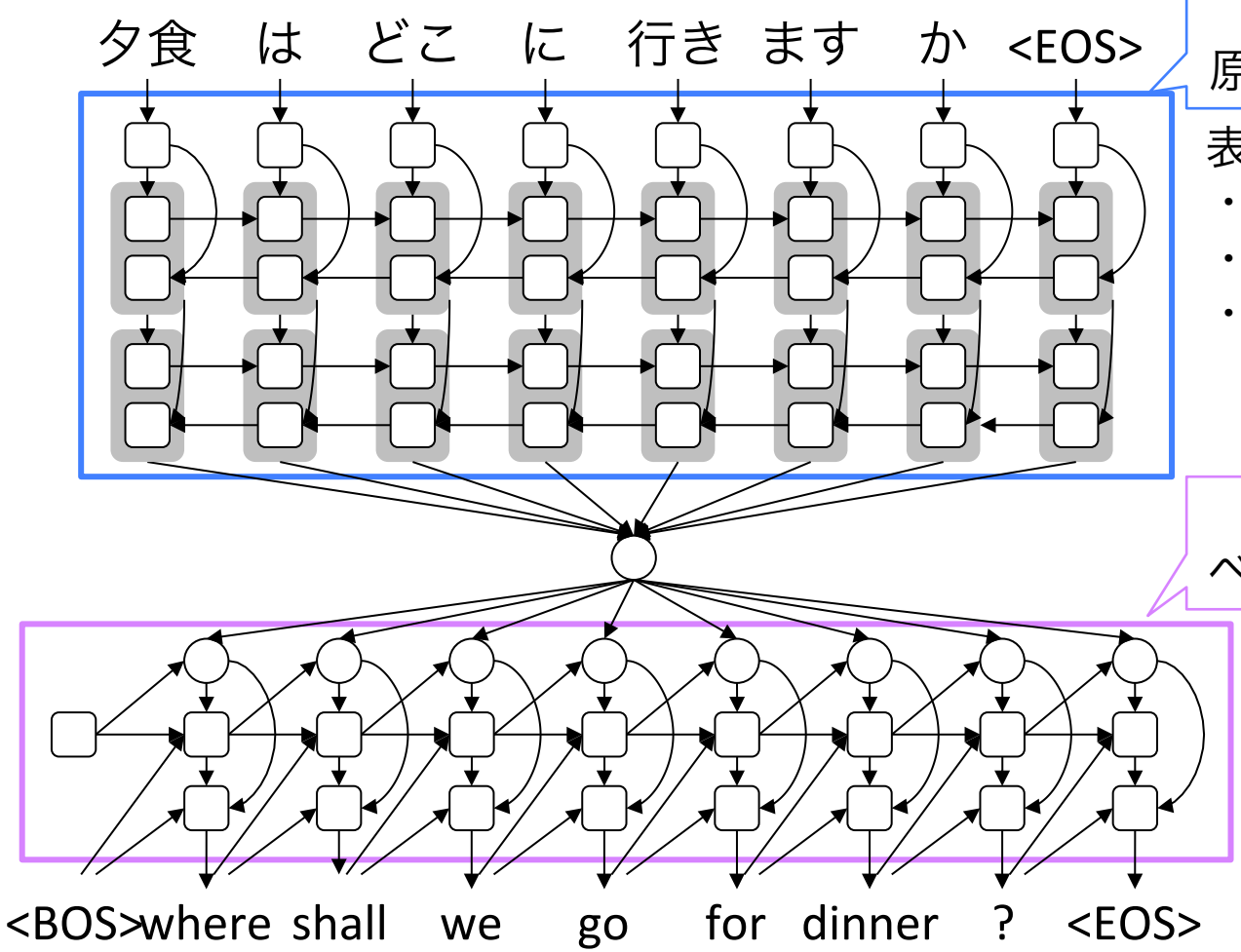
- SMT: 統計的機械翻訳
  - 対応関係を陽に捉える
  - 確率的コピペ
- NMT: ニューラル機械翻訳
  - 原文を数値データに変換
    - ベクトル，行列，テンソル
  - 語の系列を生成





ニューラル機械翻訳 (NMT)

- 注意機構付きseq-to-seqモデル [Bahdanau+, 14]
  - 注意機構: 原文中の個々の語をチラ見する仕組み



エンコーダ (encoder)
原文をベクトルに変換 (記憶)

表現力と記憶力の向上
・双方向エンコード
・多層化 (ディープ化)
・サブワードの利用

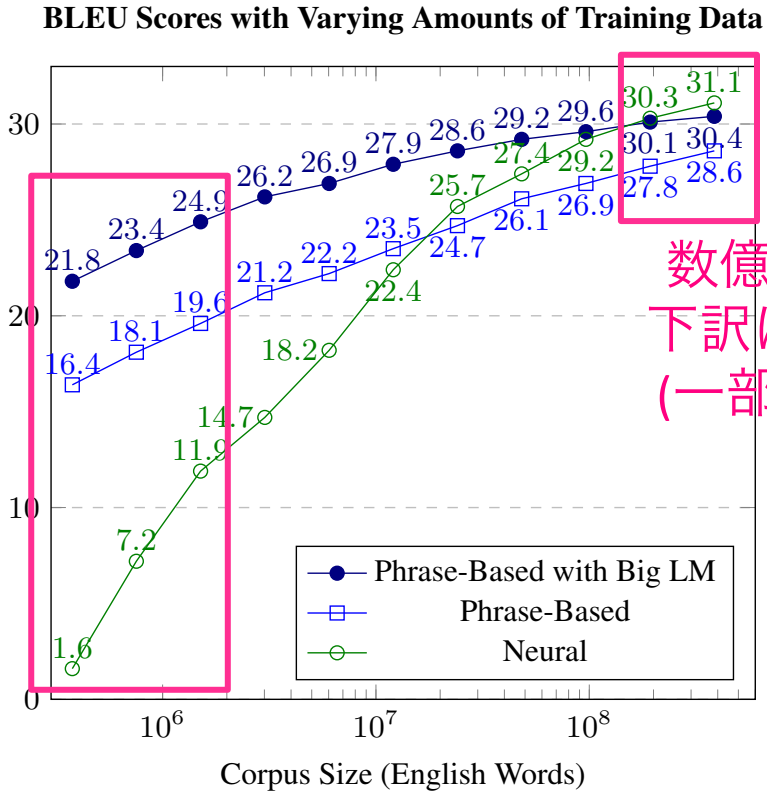
デコーダ (decoder)
ベクトルから語の系列を生成

最近: 回帰性からの脱却
・Transformer [Vaswani+, 17]
“Attention is All You Need”
・Non-autoregressive [Gu+, 17]

ハードウェア × アルゴリズム × データ

- 対訳コーパスが大きいほど高性能
  - より多様な表現・事例をカバーできる
  - 統計量をより正確に推定できる
  - NMTには不可欠: 小さいとSMTに完敗
    - e.g., En→Esニュース翻訳のBLEUスコア [Koehn+, 17]

対訳が数万文対だと
NMT << SMT
(多くの分野,
少資源言語対)



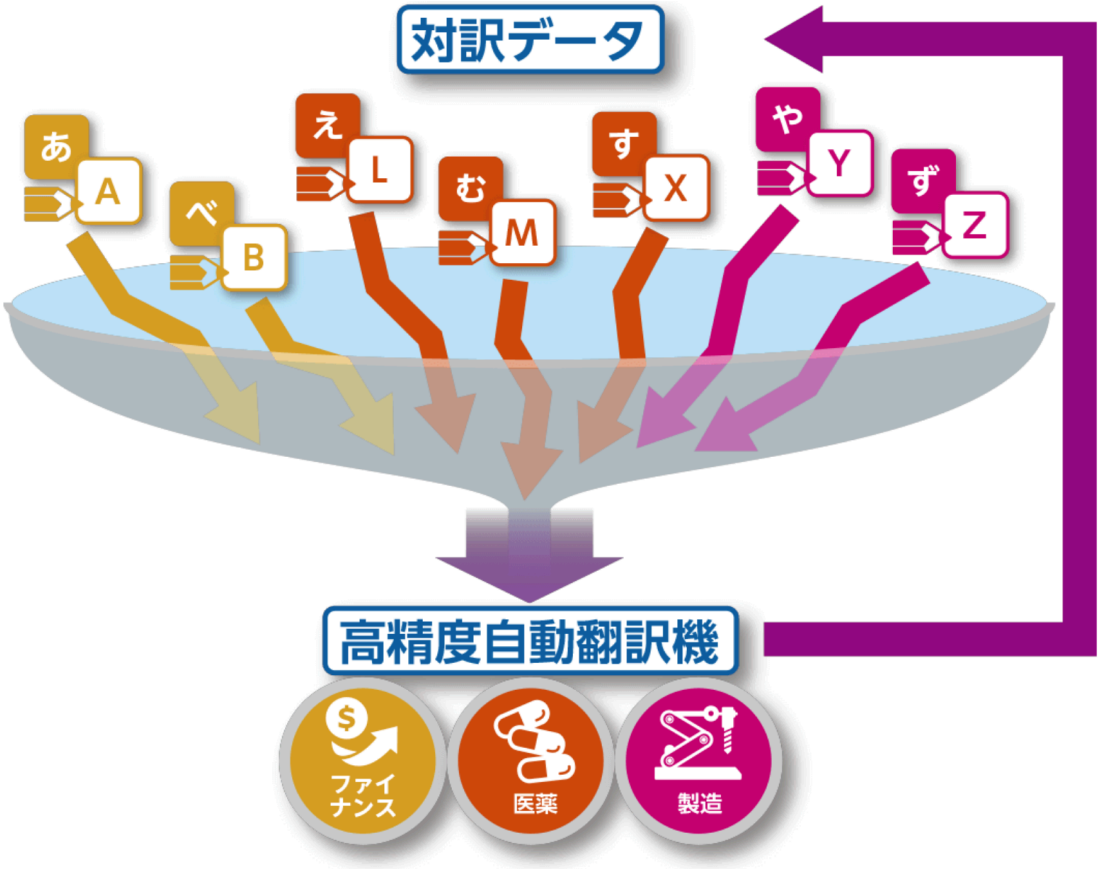
数億語(≒数百万文)で
下訳になりうるレベル
(一部のリッチな分野)

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
- 既訳の収集
  - 欧州議会文書: 数100万文 [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]
  - 企業等の資産の集積: e.g., SAP
  - 翻訳バンク [NICT, 17b]
- 単言語コーパスの活用
  - 逆翻訳で擬似的な対訳文対の生成 [Sennrich+, 16a]
  - 対訳っぽいフレーズ対，文対の抽出 [Marie+, 17a; 17b; 18a]

翻訳バンク@NICT

- 2017年9月8日に運用開始
  - 当面の目標は1億文対の集積



MTができそうなこと = X文Y訳 + α

- X文Y訳: Text-to-text の変換の範疇
  - e.g., 英文和訳
  - NMTは一旦テキストから離れるけれども…
    - テキストの「意味」を「理解」しているわけではない
    - 人間の翻訳の際の“deverbalization”でもない
  - テキストの社会的・歴史的位置付けは理解できない
- +α: 翻訳対象を取り巻く他モーダルの考慮
  - テキスト内の情報
  - テキストの外界の情報

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
  - 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]

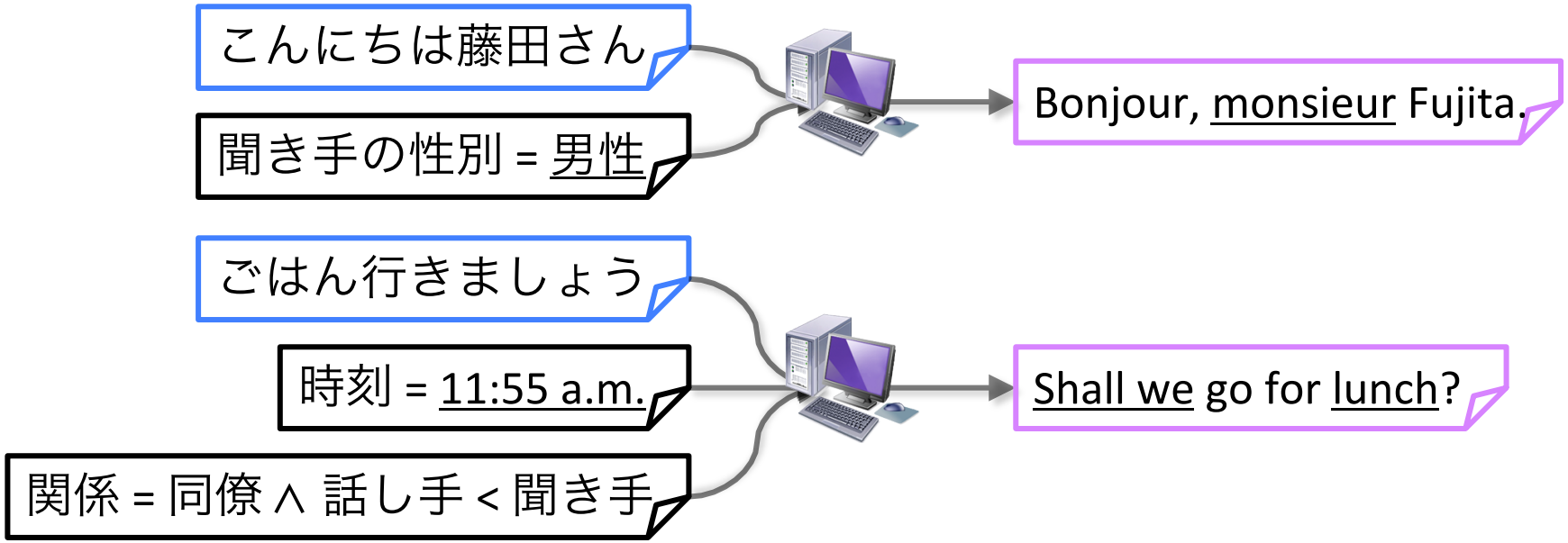
    - 文脈の情報に基づく曖昧性解消 [Tiedemann+, 17][Voita+, 18]

    - 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]

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
  - 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



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?

(現状の延長の) MTにはできないこと (1/2)

- テキストの社会的・歴史的位置付けの理解
  - i.e., X文Y訳を始めて良いかどうかの判断
  - e.g., 国際条約の文書 [影浦, 17]
    1. 社会的・歴史的に固有の役割を担う文書
    2. 目標言語の言語圏の国々における批准の有無
    3. 批准しているならば公式の訳が存在を確認
    4. 公式の訳を用いるという判断
       - スコポスに基づく例外もある: e.g., 子供向け
- 必要に応じた動的な既訳の検索
  - e.g., 翻訳メモリのカバレージの限界

Genius is the man of average ability who makes an effort.

天才とは努力する凡才のことである。

Albert Einstein

(現状の延長の) MTにはできないこと (2/2)

- スコポスに応じた情報の取捨選択・補完・意訳
  - e.g., 前提となる知識の違いに応じた情報の補完

柿くへば鐘が鳴るなり法隆寺　正岡子規

When I ate a persimmon, the bell tolled at Horyuji.

“Kaki (persimmon) kue (eat) ba (when), kane-ga naru-nari (the bell toll) Horyuji (a name of temple).” is a haiku composed by Shiki Masaoka at latest in 1895. This is the most famous haiku among his over 200k haikus. This haiku illustrates that he was notified that it was already fall when the bell of Horyuji tolled beside he was eating a persimmon after visiting the temple.

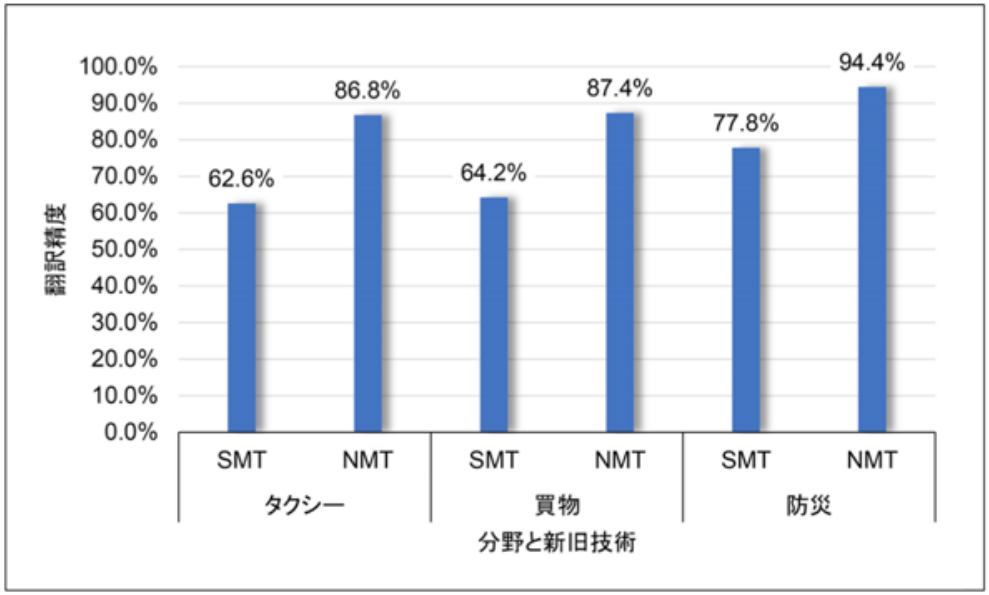
  - e.g., 人間並みの推論, プラグマティクス
    - X文Y訳に矮小化して処理している場合もある

ドアの開閉にご注意ください。

Watch your fingers.

MTの研究 = X文Y訳の範疇での精度向上

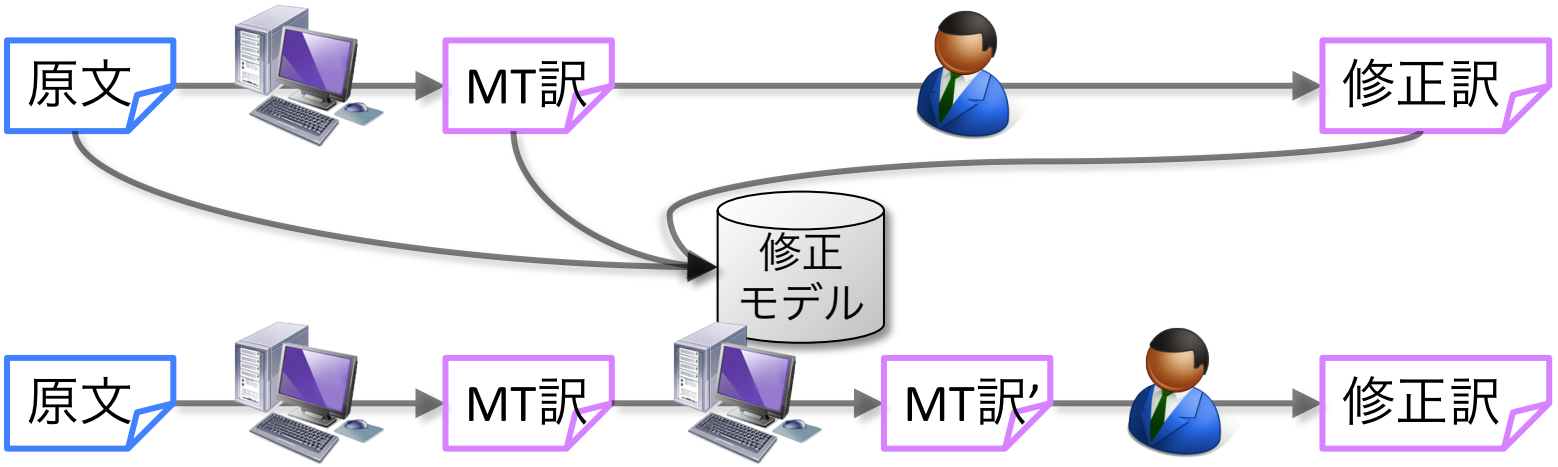
- とはいえその範囲についても「完璧」には程遠い
  - e.g., NICTの英日NMTの精度 [NICT, 17a]



- MT界隈のアプローチ
  - MTそのものの改良: アルゴリズムの改良, データの拡充
  - MTの外側の技術: 自動ポストエディット, 品質推定

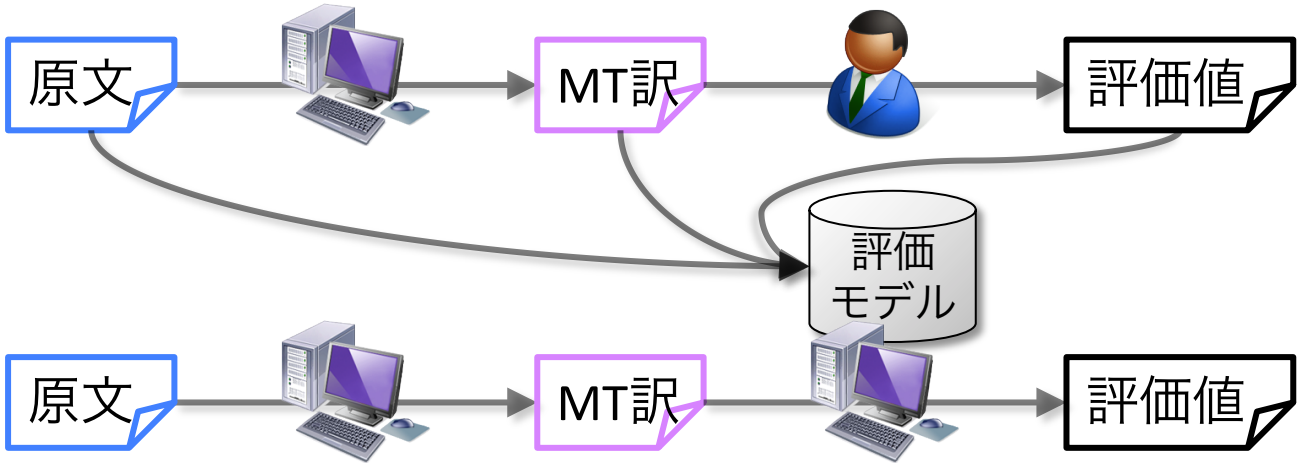
MT訳の自動ポストエディット

- APE: Automatic/Automated Post-Editing [Allen+, 00]
  - 人間のPEの前にできる範囲で修正
    - e.g., ルールベースMTが繰り返す同じ誤りへの対策
    - e.g., MT外部のターミノロジとの照合



MT訳の品質推定

- QE: Quality Estimation [Specia+, 10]
  - ユーザの立場での信頼度 (!=MTシステムの自信)

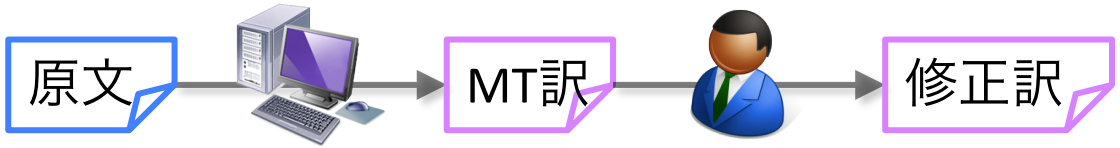


  - HTER: 人間によるPEの前の見積もり [Specia+, 11]
    - e.g., 再訳の要否やPEのコスト
  - OK/NGの2値: MT訳の直接使用の判断 [Soricut+, 10]
    - e.g., 音声翻訳器で注意を促すかどうか
      - e.g., 100%出力する場合は精度80%, 上位80%だと精度90%

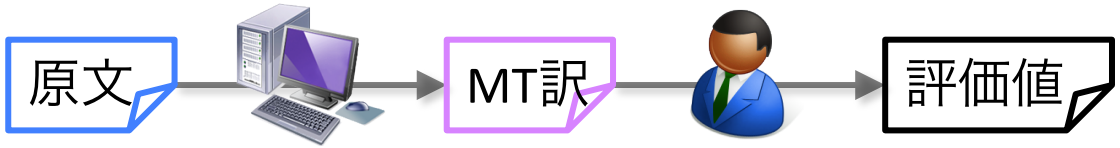
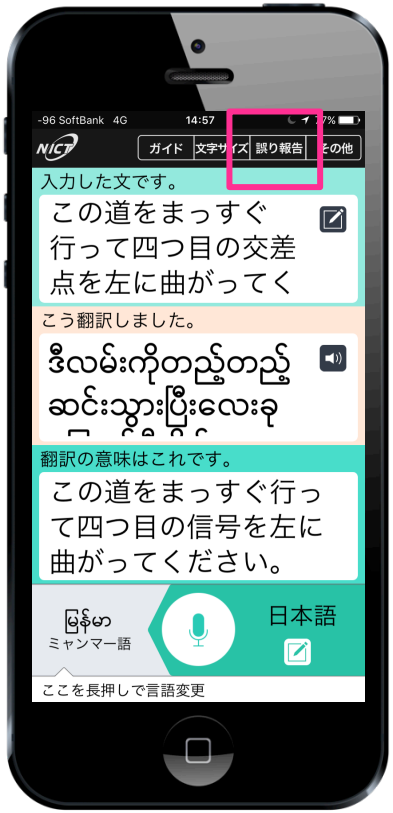
どのタスクもデータの蓄積が重要

- 種類と用途
  - 対訳 → MT

  - MT訳のPE結果 → APE, MT



  - MT訳評価結果 → QE



- NICTの取り組み
  - VoiceTraユーザからMT訳の誤り報告
  - 人手で再評価・修正