産業翻訳に役立つ自然言語処理技術についての議論の足場 (テーマセッション開催趣旨説明)

藤田 篤
情報通信研究機構 (NICT)

山田 優
関西大学

影浦 峡
東京大学

翻訳に関わるコミュニティ間の橋渡し

- これまでの企画
  - 『文理・産学を越えた翻訳関連研究』@NLP2016
  - 『翻訳の質と効率:
    実社会におけるニーズと工学的実現可能性』@NLP2017
  - 『コミュニティの現場を支援する
    翻訳通訳テクノロジー』@JAITS2017
  - 『翻訳における人間と機械の協働:
    for what, how, where, when and why?』@NLP2018
  - 『翻訳におけるテクノロジーを考える』@JAITS2018
  - 『産業翻訳に役立つ自然言語処理技術』@NLP2019

産業翻訳 (実務翻訳とも)

- 産業活動に関わる文書の翻訳
  - 市場の80%以上 [日本翻訳連盟, 17]
  - 専門分野
    - 製造，医療，法務，知財，金融，IR情報，広報など
  - テキストタイプ
    - 公開書類: 特許出願書類，IR関係報告書，マニュアルなど
    - 内部書類: 契約書等，メールなど

- 産業翻訳以外の翻訳
  - 出版翻訳: 各種図書，ニュース記事など
  - 映像翻訳: 字幕，吹替など

本テーマセッションで議論したいこと

- 産業翻訳にかかわる人が恩恵にあずかれる技術とは?
  - ゼロサムではなくwin-win
    - 産業翻訳の担い手: 翻訳企業, 翻訳者
    - 産業翻訳の利用者: 翻訳の依頼者(情報発信者), 受信者
  - 翻訳のワークフローのデザイン
    - 人間中心の既存のワークフローの効率化
      - 技術の適用場所 (定式化)
      - 「役立つ」ということの条件
    - まったく新しいワークフローの創出

- 技術の「出口」として応用先を考えるのではなく, 「目的」からスタートして技術を考えよう [金出, 16]

本発表の内容

- 議論の足場がため
  - 用語の整理
  - 産業翻訳のワークフローの共有
  - 技術の有用さの基準
  - 論点の列挙

産業翻訳における前提の確認

用語の整理

- 異業種間のコミュニケーションの前提
  - 翻訳 [岩波国語辞典第7版, 09]

    ある言語で表現された文章の内容を
    原文に即して他の言語に移しかえること.
  - 機械翻訳 (MT)

    ある言語表現を，計算機を用いて別の
    言語に変換すること．自動翻訳とも.
  - スコポス [Vermeer, 89]

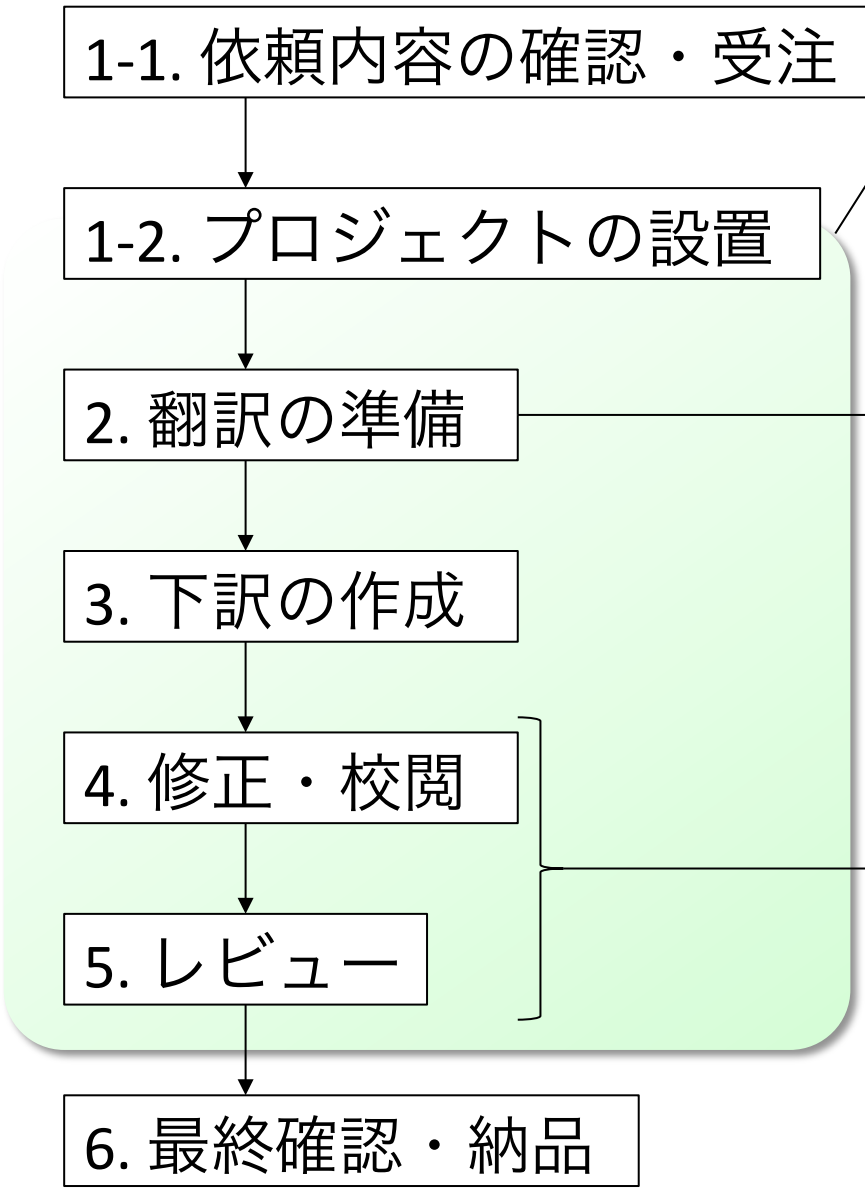
    翻訳の目的．これにより翻訳成果物，人間の行為と
    その下位範疇としての翻訳が決定される.
    - 目標文化社会の環境における目標テキストの目的
    - 目標テキストの受け手の目的

翻訳とMTの違い

	翻訳	MT
処理対象	言語表現された内容	言語表現
処理 レイヤー	文書や表現の社会的位置付け, 表現の実在性, 概念, 意味, 表現	言語表現と「意味」
表現上の 処理単位	固有表現や専門用語など 社会的な属性を担う表現要素	文や句など, まれに起点テキスト全体
スコポス	考慮される前提	明示的には扱われない
品質	保証される前提	保証できない

- 現在のMTの処理はX文Y訳 [影浦, 17]
  - cf. 精度100%の代替技術
    - そろばん → 電卓
    - 電話交換手 → 電話交換機
  - X文Y訳が100%できる = 翻訳に携わる大前提

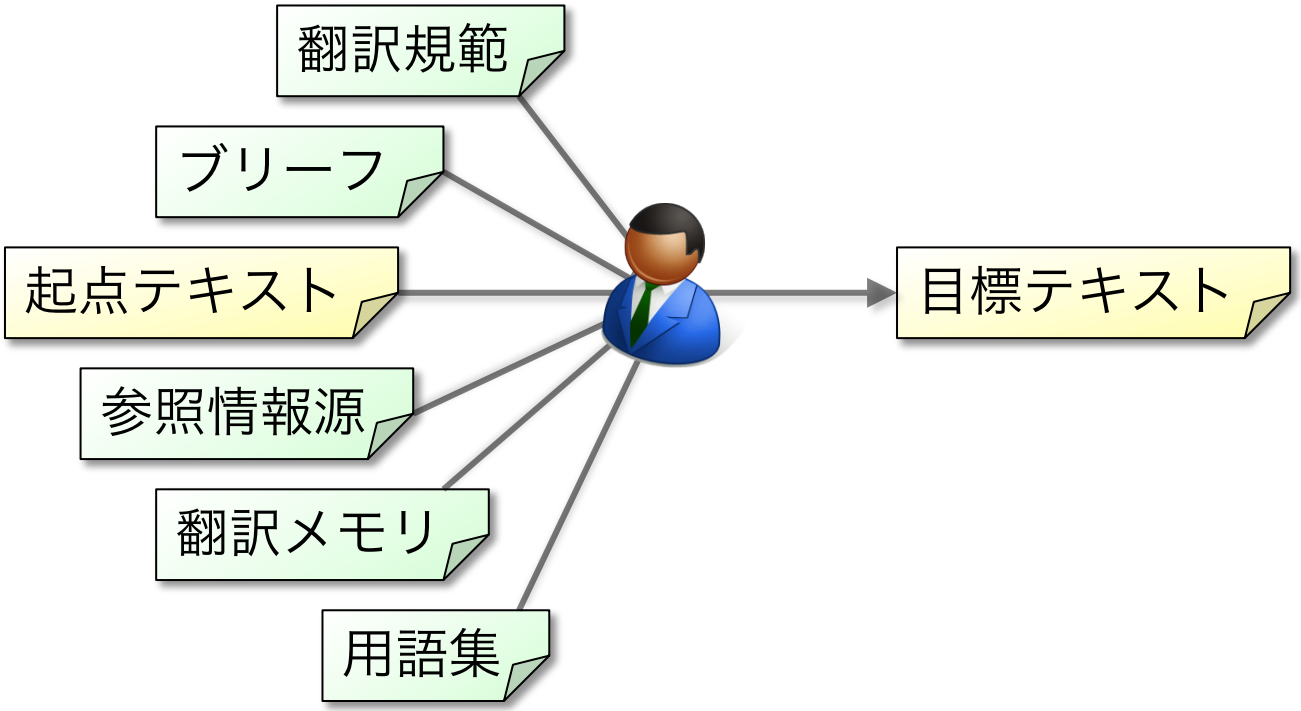
産業翻訳のワークフロー



- プロジェクト [ISO, 15]
  - 短い納期
  - 複数人による役割分担
- 準備が大事
  - 2a. 調査一般
  - 2b. 用語集の作成
  - 2c. 翻訳メモリの設定
- 品質保証が大事
  - 4. 起点テキストも参照
  - 5. 目標テキストのみ参照

翻訳とMTの参照情報の違い

- 少なくとも現在のワークフローが考慮する範囲で



産業翻訳にかかわる人が
恩恵にあずかれる技術とは?

何を扱う技術か?

目的/用途　　内容

社会的位置づけ　　知識

規範/慣例　　概念　　意味

表現の実在性　　言語構造

言語表現

技術の有用さの基準

- 翻訳の品質を人間が担保する前提で
  種々のコストに見合う効率化を技術的に実現する
  - **翻訳者の知識不足を補う**
    - ステップ2「翻訳の準備」の不可欠さ
      - 知識源: 一般の辞書，用語集，翻訳メモリなど
      - 参照情報源: 起点テキストの関連情報，情報間の関係など
  - **翻訳者の意思決定を支援する**
    - クライアントに対する説明責任
      - 技術そのものは品質の瑕疵に対する責任を取れない
      - 翻訳者力 [Kiraly, 00][影浦+, 16]
        - ✓ e.g., 特定の原文要素をなぜこのように訳したのか? の説明
        - ✓ e.g., プロジェクトのメンバ間の協調

2つのアプローチ

- 人間中心のワークフローにおける支援技術
  - 手続きの自動化/効率化，気づきの示唆
  - 言語処理的なアプローチの例

ワークフロー	自動化が期待される手続き
1-2. プロジェクトの設置	メンバの割り当て
2b. 用語集の作成	テキストデータからの用語抽出 対訳データからの既訳発見
2c. 翻訳メモリの設定	既存の対訳の再利用 翻訳成果物の動的登録 [Denkowski+, 14]
3. 下訳の作成	訳文候補の生成
4. 修正・校閲	スペルチェックや文法誤り検出 対応関係等のの可視化

- まったく新しいワークフローの創出
  - e.g., インタラクティブMT [Green, 14]

これらは本当に有用か?

- MTによる下訳の生成
  - 平均的な精度は，個々の翻訳の是非を保証しない
    - e.g., BLEUスコア70，TOEIC 900点相当
    - それゆえにポストエディットが重要
      - cf. 街中の音声翻訳の場合，訳の是非をユーザが判断
  - パラドクス: MTが助けになる ⇔ MTを超えられない
    - MT出力文を読み続けると言語感覚が狂う [井口, 19]ほか

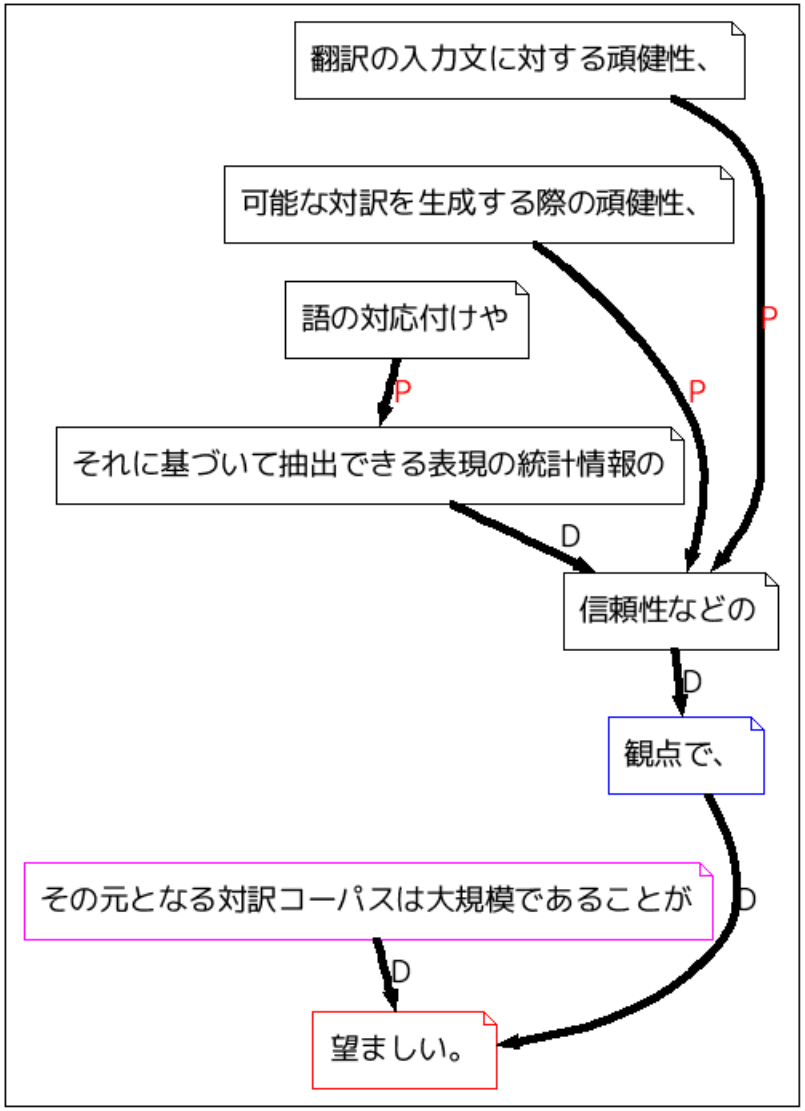
- Webでの情報検索
  - 文法性: ヒット数は近似になるか?
  - 固有表現の既訳: 検索結果の上位に表示される保証は?
  - 参照情報の検索: キーサードサーチのみでは不十分

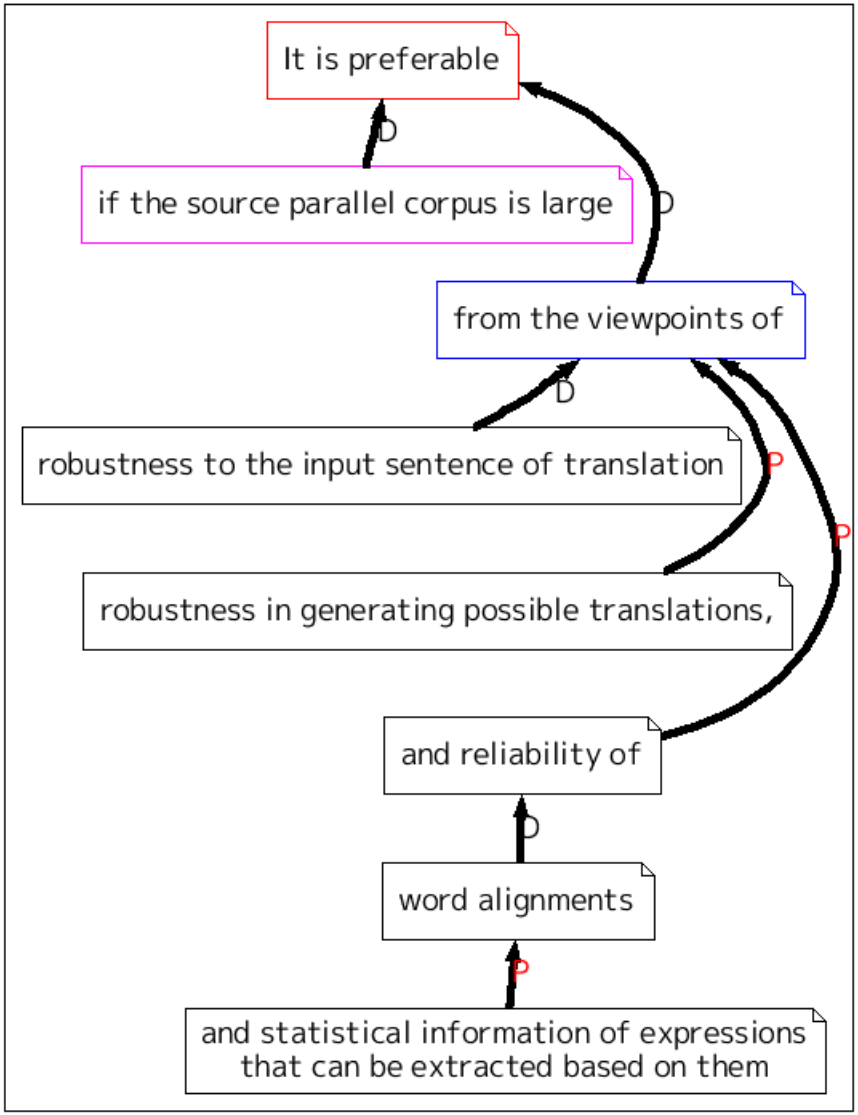
知識不足と意思決定の補助

- 人間が陽に行ってきたことの(全/半)自動化
  - e.g., そろばん → 電卓
  - 2b. 用語集の作成: 起点テキストからの用語抽出 [新田, 19]
  - 3. 下訳～5. レビュー: 用語の訳出の一貫性の検査
- 人間が暗に行ってきたことの明示化
  - 3. 下訳～5. レビュー: 言語表現の構造の可視化
    - 起点テキストの構造
    - 起点テキストと目標テキストの対応
- 人間にはできない，お節介未満の情報提供
  - e.g., Google Web検索の「もしかして」

可視化の例: 原文要素と訳文要素の対応

- e.g., 依存構造, 要素間の対応





可視化の例: エンティティ・リンキング

- 起点テキストにおける同一エンティティの色塗り
  - (ゼロ)代名詞，共参照など

名古屋大学は、1939年に創設された名古屋帝国大学を直接の母体とする国立大学である。(φの)前身の名古屋帝国大学は9番目（内地では7番目）に設立され、(φは)内地・外地を通じて「最後の帝国大学」であった。名古屋帝国大学創設当初は医学部と理工学部の2学部を設置し、(φは)1942年には理工学部を理学部と工学部に分離した。第二次世界大戦後の旧制学制残滓期間内に、(φは)法経学部と文学部の2学部を設置した。新制名古屋大学となった後も(φは)教育学部、農学部、情報文化学部等の学部や大学院研究科および附属研究教育施設を順次設置し続け、(φは)2018年時点、9学部・13研究科・3附置研究所を擁している。

本テーマセッションで議論したいこと

- 技術の有用性の条件
  - 必要な「知識」とは何か?
  - 「意思決定」に必要なものは何か?
- 技術に対する信頼の条件
  - 理想的には精度100%
  - 不具合の検出の容易さ: e.g., QE技術とポストエディット
- 技術の実現方法
  - 様々な翻訳作業環境，APIとしての共有化
- 外部サーバ利用時のデータ保護
  - MTサービスやクラウド上の翻訳作業環境
- 業界の持続可能性
  - 市場拡大，人材育成，業種の多様化(生存戦略)

本テーマセッションの発表一覧

- 前半: 14:50-16:30
  - F4-1 藤田+: 開催趣旨説明
  - F4-2 宮田+: 制限言語，執筆と翻訳の同時最適化
  - F4-3 新田: NMTの誤りを修正する全自動/半自動処理技術
  - F4-4 土井+: NMTの分野適応と評価
  - F4-5 橋本+: XMLによる文書構造情報，NMTの適用と評価
- 後半: 16:50-17:30
  - F4-6 平岡+: YouTube字幕翻訳，プリエディット
  - F4-7 渡部+: NMTの誤り分析，プリポストエディット
- 総合討論: 17:30-18:30 (最長19:30)