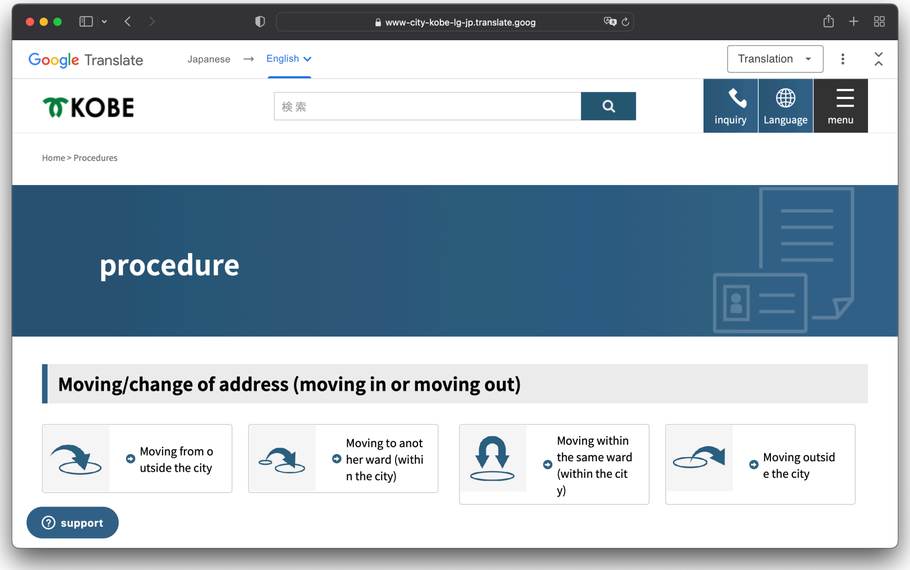
AAMT 2024, Tokyo
2024年12月3日

自治体情報の多言語化を支える翻訳資源とテクノロジー

宮田玲（東京大学）

現状と課題

- 自治体現場での翻訳のニーズ
  - 多様な言語的背景を持つ住民への情報提供
  - 対象言語と対象文書の拡大が課題
- 機械翻訳（自動翻訳）の導入
  - 多言語化のために自治体ウェブサイトで機械翻訳の導入
  - 2019年時点で、基礎自治体の2/3（1,157団体）は機械翻訳のみの利用 (宮田, 2020)
  - 誤訳が引き起こしうる問題も指摘されている



神戸市ウェブサイト（Google翻訳による英語版）より
https://www.city.kobe.lg.jp

一部の情報のみを人手で翻訳　or　機械翻訳をそのまま使う

さまざまな解決策

- 多言語化にかかる予算を増やす
- 自治体に翻訳専門職を配置する
- コミュニティ翻訳の人材を増やす
- 多言語化・多文化共生に関する法令を作る
- 翻訳プロセスを技術的に支援する **今日のテーマ**
- ・・・

ここからのお話

I. 研究プロジェクト概要

- 枠組み、理念

II. 個別プロジェクトの紹介

- 基礎調査
- 翻訳資源構築
- 翻訳資源利用

III. まとめと今後の展望

- 産官学連携

I. 研究プロジェクト概要

テーマとメンバー

- 「円滑な多言語情報発信を可能にする自治体横断型翻訳資源の構築」
  - 科研費・基盤研究(B)・23K28378・2023〜2026年度
  - https://tr4lg.p.u-tokyo.ac.jp/
- メンバー／研究協力者

宮田玲	（東京大学）	朴惠	（東京大学）
阪本章子	（関西大学）	夏目和子	（東京大学）
藤田篤	（情報通信研究機構）	山浦育子	（愛知県立大学）
香川璃奈	（産業技術総合研究所）	島津美和子	（立教大学）

ここから先の内容は、宮田+ (2024)の発表資料に一部基づきます。
資料作成にあたっては、上記のメンバーから示唆・コメントをいただきました。

翻訳資源？

- 自治体文書の特徴
  - 過去の文書と似ている
  - 他の自治体の文書と似ている
- 共通資産＝**翻訳資源**
  - 対訳文書
  - 対訳文（翻訳メモリ）
  - 対訳用語

転入時の手続き

浜松市

Entering Japan

When

Within 14 days of entering Japan and deciding on an address

Who

The applicant in person, a proxy or a member of the same household.

Forms to be Submitted

Moving-In Notification (*tennyu todoke*)

Items to Bring

The Residence Card (*zairyu kado*) or Special Permanent Resident Certificate (*tokubetsu eijusha shomeisho*) and passport of the person who has entered Japan, and documents certifying their family relations (originals and Japanese translations).
In the case of a proxy, the proxy must also bring their own identification and a letter of attorney.

Cost

Free

Where

Ward Municipal Services Division

https://www.city.yokohama.lg.jp/lang/residents/en/notifications/default.html より

横浜市

Procedures for moving to Yokohama City from overseas (notification of moving in)

Notification period

Within 14 days of the day of moving in

Applicant

Self, head of household, a member of the same household, proxy (A letter of proxy is required in the case of a proxy.)

Counter

Family registry division of the ward office of the ward you have moved to

Things you need to submit with your notification

- Resident card or Special Permanent Resident Card (In the event you are not issued with a resident card at the airport, please bring a passport stating that a resident card will be issued at a later date.)
- Passport
- If you are submitting notification of resident registration as a family of non-Japanese residents, a document confirming your family relationship (If this document is written in a foreign language, please include a translation of this document with the translator clearly identified.)
- Letter of proxy (in the event a proxy is submitting notification on your behalf)

https://www.city.yokohama.lg.jp/lang/residents/en/notifications/default.html より

タイミング
申請者
Residence Card
持ち物
場所
Resident card

うまく共有・活用できないか？

長期的目標

翻訳資源の構築

翻訳資源の利用

このサイクルを支援するエコシステムの提案・構築

↓

日本の自治体における多言語情報発信の支援

↓

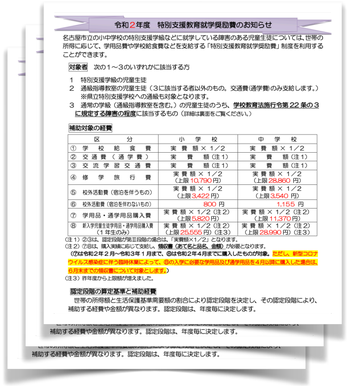
言語格差の解消

3種類の翻訳資源

- 対訳文書アーカイブ（文書レベル）
  - 各種の情報を付与し、整理した対訳文書の集合
- 拡張翻訳メモリ（文レベル）
  - 内容・スタイル・用語を統制し、テンプレート化した対訳文の集合
- 統制対訳用語集（用語レベル）
  - 標準的な表記を定めた対訳用語集

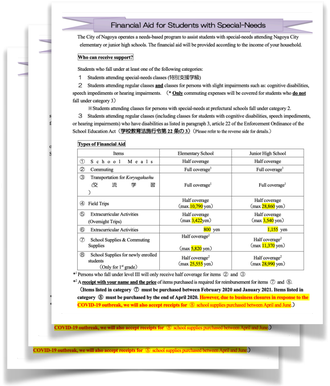
翻訳資源構築の流れ

(1) 翻訳文書アーカイブ



起点文書

対応付け



目標文書

名古屋市翻訳資源より
https://github.com/tr4lg/nagoya-dataset

抽出
組織化

(2) 拡張翻訳メモリ

出典	カテゴリ	文書構造	日本語	英語	…
…		…	…		…
○○市	印鑑登録	見出し	印鑑登録証明書	Personal Seals Registration Certificate	…
○○市	印鑑登録	本文	区役所の戸籍住民課または特別出張所の窓口に申請してください。	Please complete application procedures at the Family and Resident Registration Division of the City Office or any Branch Office.	…
…		…	…	…	…

自治体固有情報の変数化
用法付与

用語抽出
承認語の定義

(3) 統制対訳用語集

日本語	承認語	英語	根拠	用法
印鑑登録証明書		personal seal registration certificate	印鑑条例	印鑑登録証との混同に注意
印鑑証明書	→印鑑登録証明書			
…	…	…		

日本語	英語
[X]の窓口に申請してください。	Please complete application procedures at [X].
区役所の電話番号は[X]です。	The telephone number for the City Office is [X].
[X]内で予防接種が受けられます。	Vaccinations are available within [X].
…	…

- 文書自体の集積と組織化
- 文書を前提としたテキストの処理

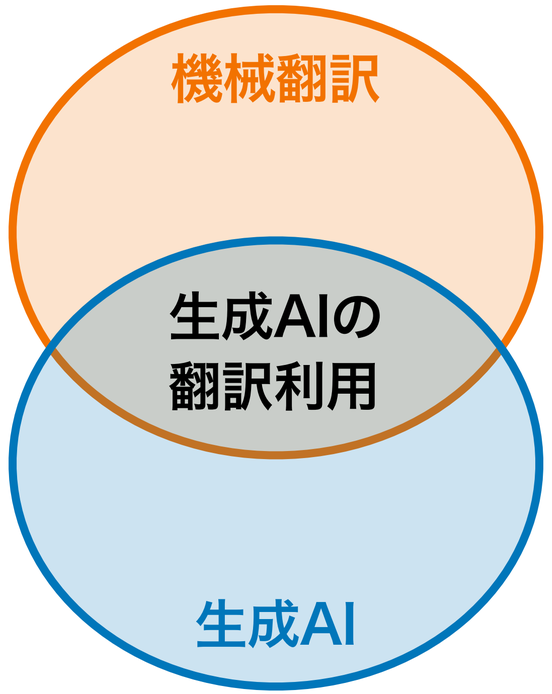
翻訳資源を使ってできそうなこと

- 過去の資源の参照
  - 辞書引き
  - 文書の再利用
- 機械翻訳や生成AI用のデータ
  - 追加訓練
  - 動的なバイアス付与
- 翻訳に留まらない様々な応用可能性
  - 文書生成
  - 多言語質問応答

テクノロジーの活用

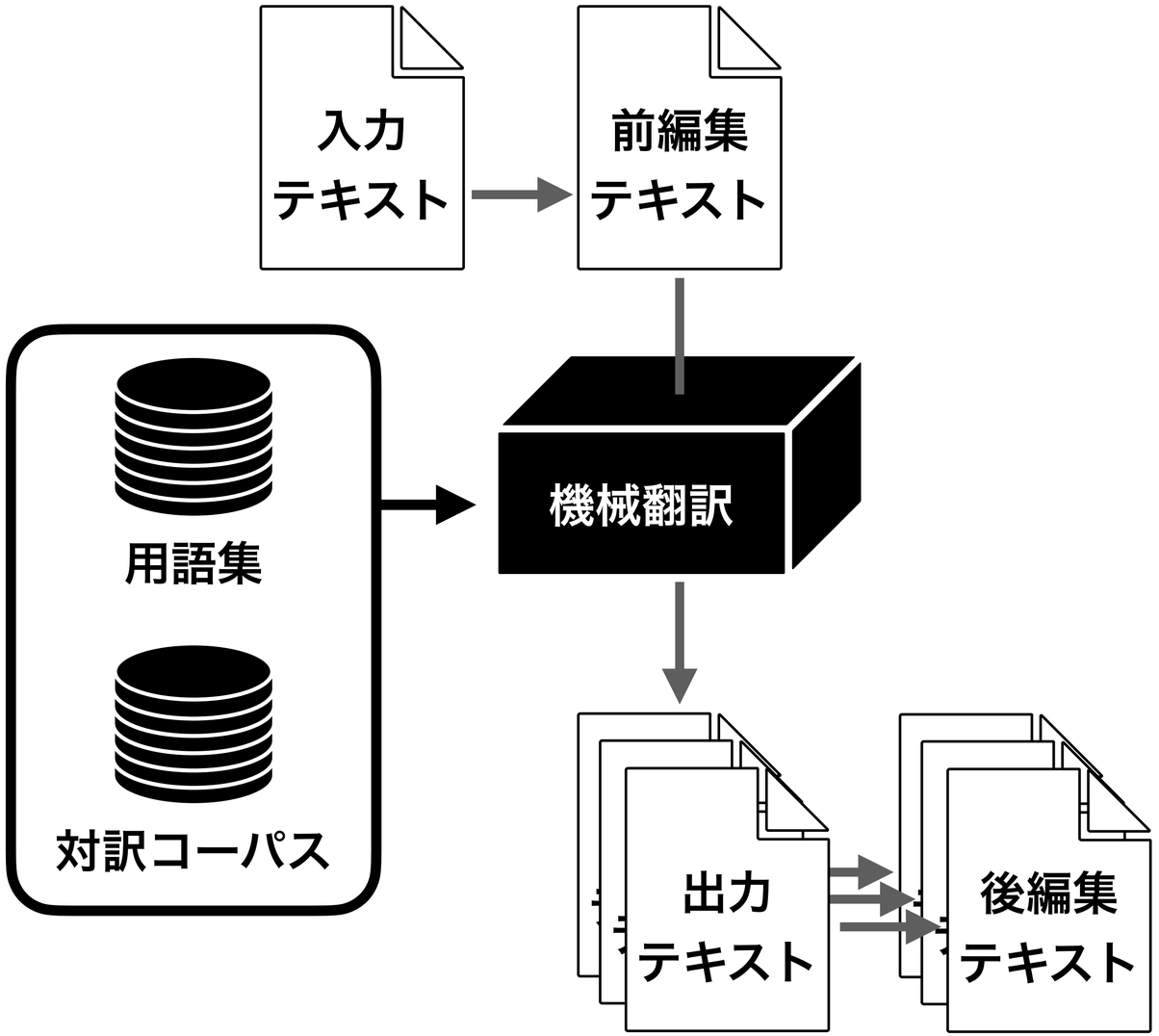
機械翻訳と生成AI

- 機械翻訳の変遷
  - ルールベース機械翻訳：自治体では長らくこの方式の機械翻訳が利用
  - 統計的機械翻訳　　：2016年までのGoogle翻訳など
  - ニューラル機械翻訳　：深層学習ベースのいわゆる「AI翻訳」
- 生成AI（大規模言語モデル；LLM）の翻訳利用
  - ChatGPTなどの汎用サービスを翻訳器として利用
  - プロンプトを工夫することで様々な調整が可能
    - 文脈内学習 (In-Context Learning; ICL)
    - 検索拡張生成 (Retrieval-Augmented Generation; RAG)
  - 自治体における利用事例はこれから
    - cf. 神戸市 (2024)



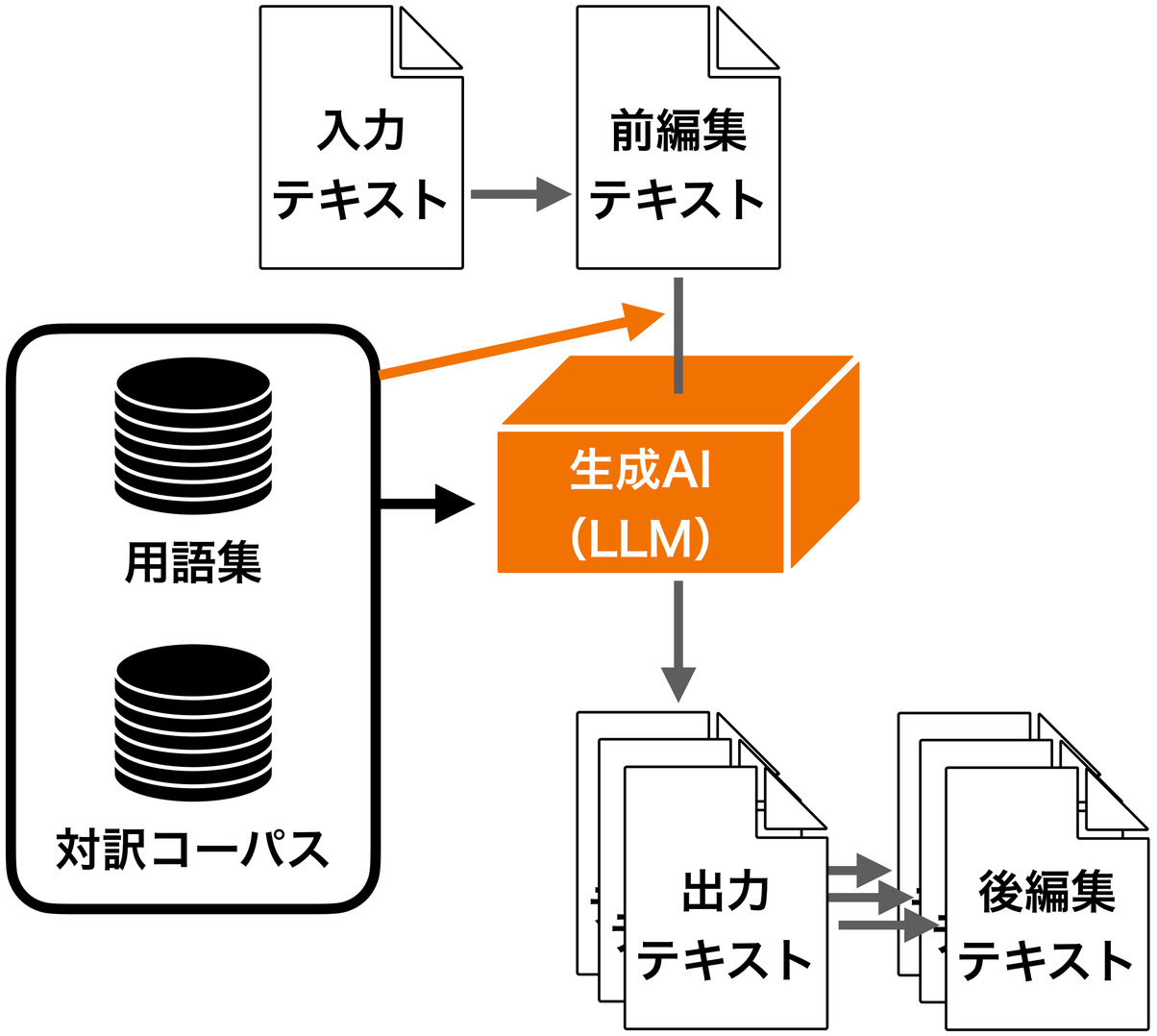
機械翻訳活用への介入点 (Miyata, 2020)

- 機械翻訳への入力
  - 前編集（プリエディット）
- 機械翻訳そのもの
  - 用語集登録
  - 追加訓練（ドメイン適応）
- 機械翻訳の出力
  - 後編集（ポストエディット）



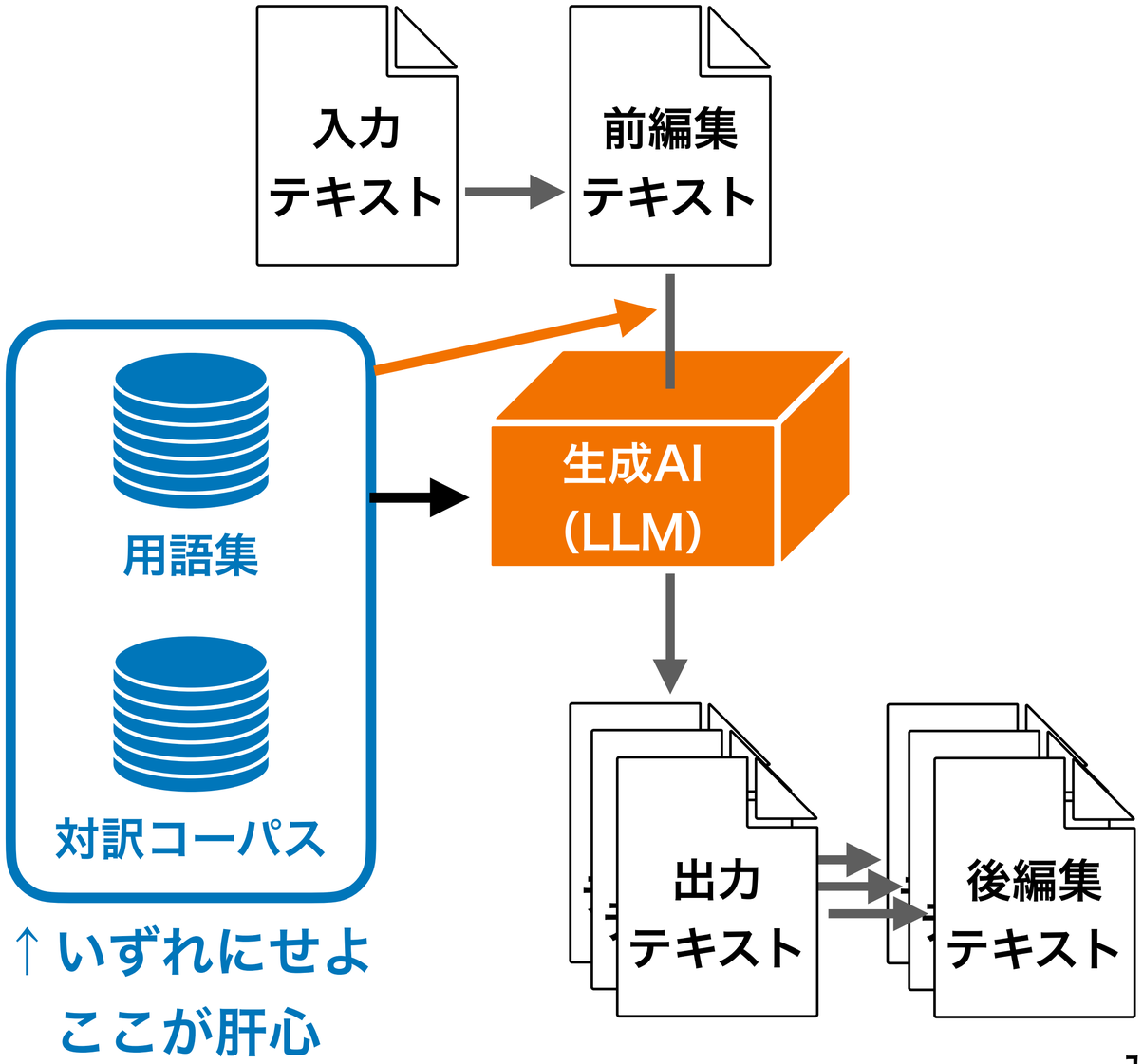
生成AI活用への介入点

- **LLM（大規模言語モデル）**への入力
  - 前編集（プリエディット）
  - **プロンプト（用語や類似訳文）**
- **LLM**そのもの
  - 追加訓練（ドメイン適応）
- 機械翻訳の出力
  - 後編集（ポストエディット）



生成AI活用への介入点

- **LLM（大規模言語モデル）**への入力
  - 前編集（プリエディット）
  - **プロンプト（用語や類似訳文）**
- **LLM**そのもの
  - 追加訓練（ドメイン適応）
- 機械翻訳の出力
  - 後編集（ポストエディット）



理念

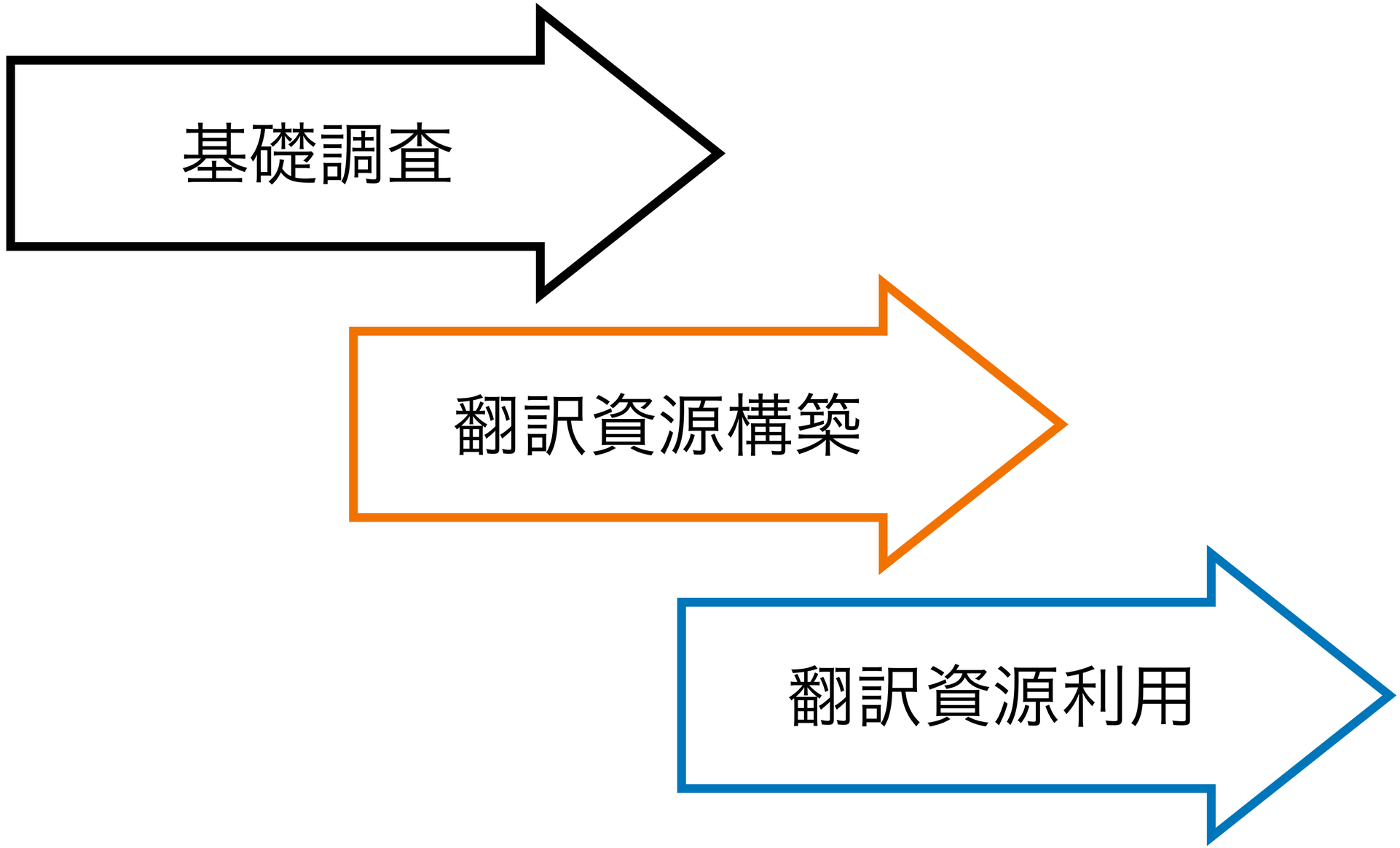
- **公共の利益**を追求する
  - 産業界と連携することも非常に重要
- **サービスの改善・拡充**＞コストの削減
  - 機械翻訳はオプションの一つ
  - 翻訳に必要なお金をかけるのは当たり前、にしていく
- **持続可能なエコシステム**をつくる
  - できるところから、できる範囲で、徐々に
  - 「データを提供しないと成果を利用できない」ということはしたくない

学術的な意義

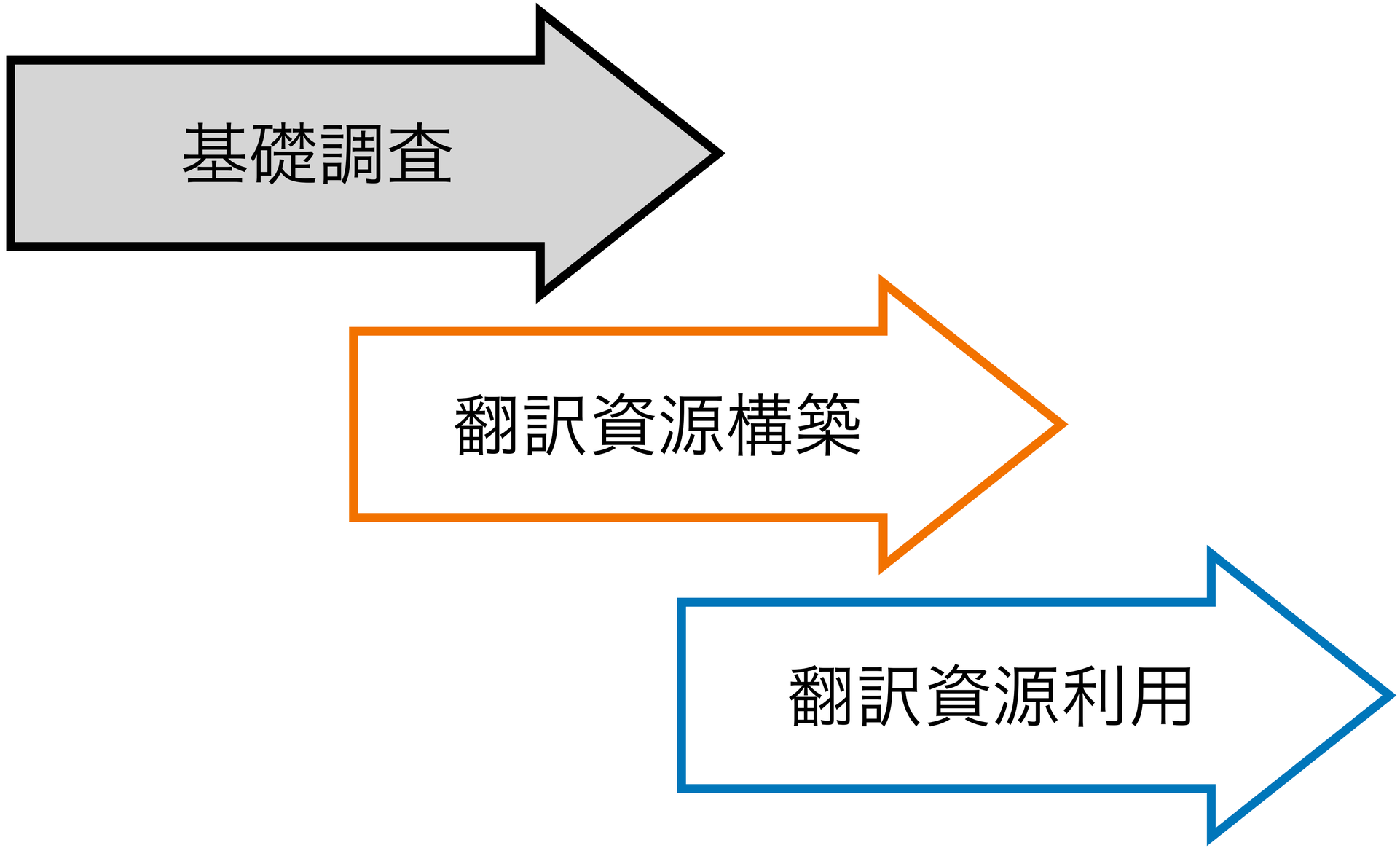
- 資源を作って公開するだけ？
- 組織化され公開された資源は、**外在化された社会的知識**である
  - cf. 個人の用語集と辞書の違い
  - cf. 図書館情報学→ドキュメントを対象とした社会的知識の生産・蓄積・流通の様態
  - cf. 翻訳メタ言語 (Miyata+, 2022)→翻訳知の外在化
- 文から**文書**へ
  - cf. センテンスサラダ現象 (Bedard, 2000)、コラージュ翻訳 (Mossop, 2006)
  - cf. 文書レベルMT (Maruf+, 2022)

II. 個別プロジェクト紹介

研究の流れ



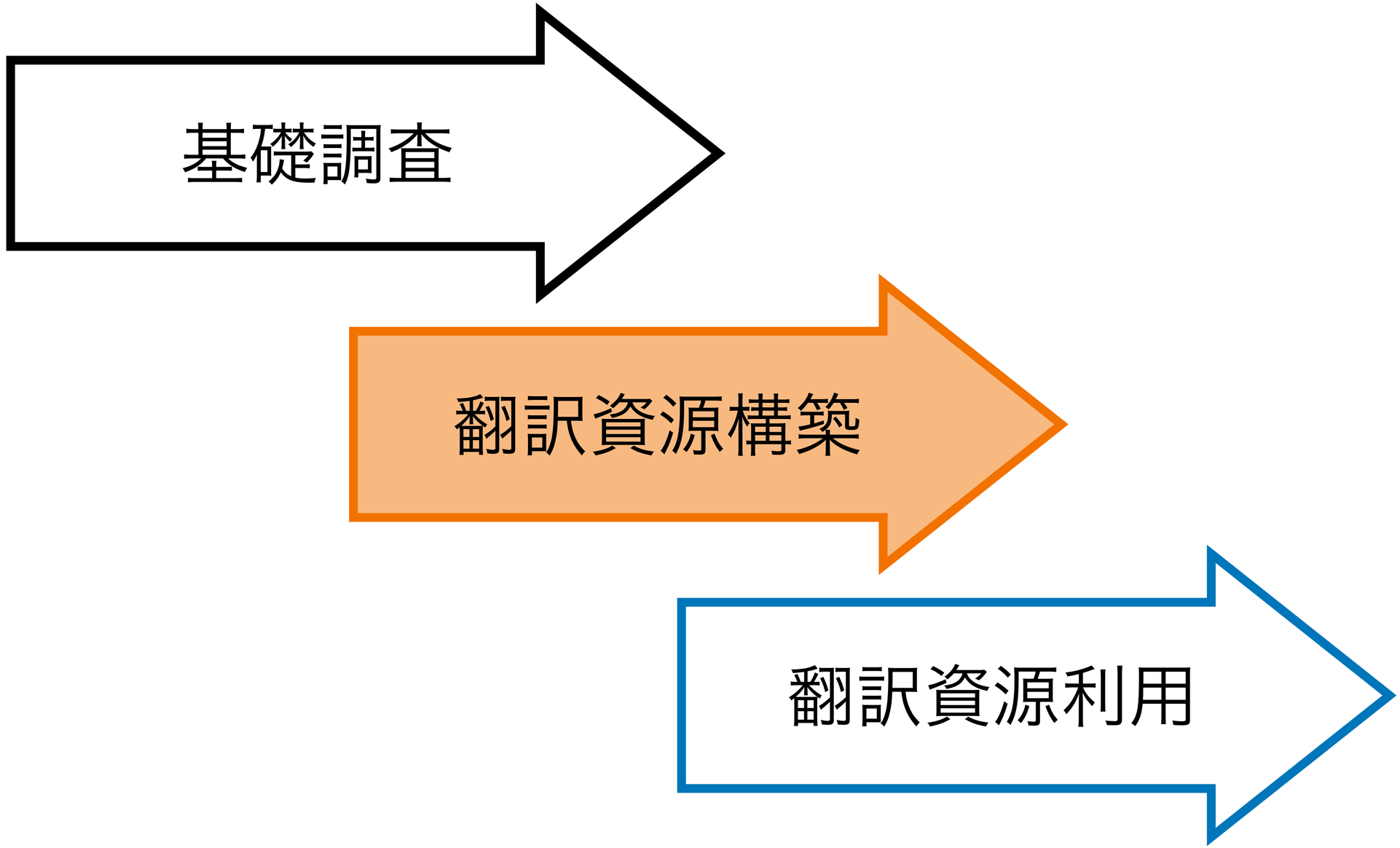
研究の流れ



自治体アンケート調査

- 自治体における翻訳業務の課題とニーズの洗い出し
  - どのような翻訳業務に関する課題があるか
  - どのようなツールを使っているか／があると便利か
- 対象
  - 部署向け
  - 個人向け
- 協力自治体
  - 中規模自治体：1団体
  - 小規模自治体：1団体
- 結果は集計中

研究の流れ



自治体からのデータ提供

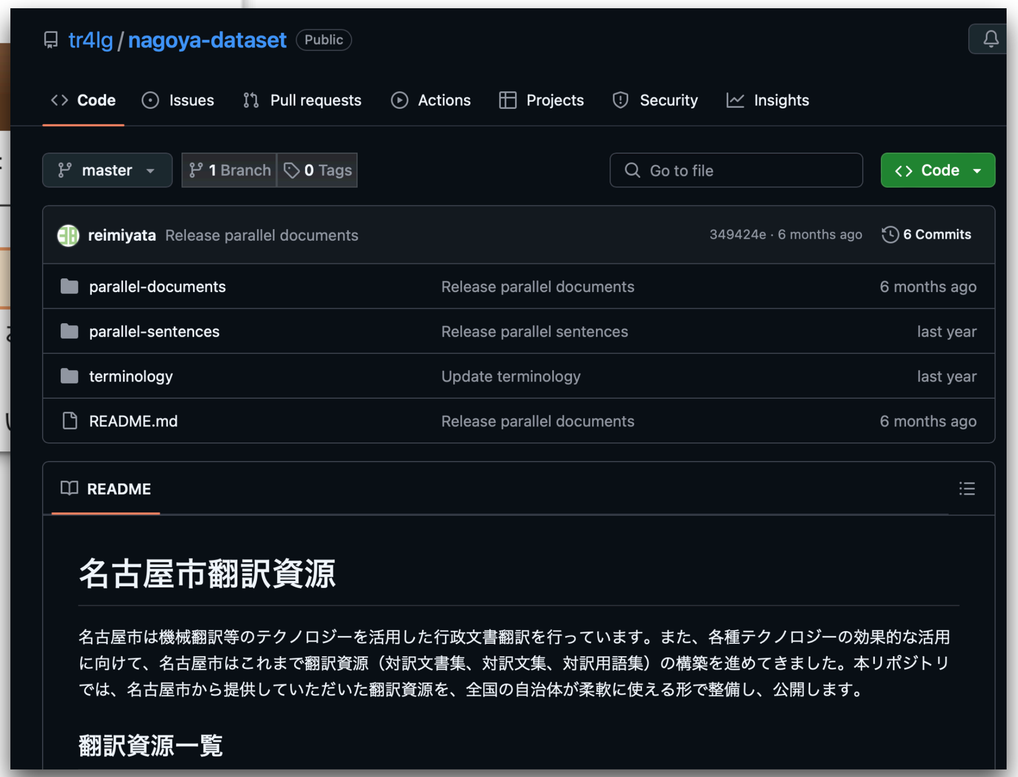
- 名古屋市、S市、Y市からの翻訳文書等の資源の提供
  - 著作権処理済
  - データの整備は現在進行形（まずは現状ベースでCC BYで公開）
- 名古屋市の場合
  - 対訳文書：約90文書対
  - 対訳文　：約800文対
  - 対訳用語：約3000用語対

日本語	名古屋城本丸御殿
英語	Nagoya Castle Hommaru Palace
中国語	名古屋城本丸御殿
ベトナム語	Cung điện Hommaru Lâu đài Nagoya
ネパール語	नागोया महल होम्मारु दरबार
ポルトガル語	Palácio de Hommaru do Castelo de Nagoya
ハングル	나고야성 혼마루 고텐
スペイン語	Palacio de Honmaru en el Castillo de Nagoya
フィリピン語	Kastilyo ng Nagoya ng Palasyo ng Hommaru

翻訳資源の一部公開：名古屋市

名古屋市公式ウェブサイトより（↑）

https://www.city.nagoya.jp/kankobunkakoryu/page/0000162160.html



GitHubリポジトリ「名古屋市翻訳資源」より（→）

https://github.com/tr4lg/nagoya-dataset/

「名古屋市翻訳資源」で検索

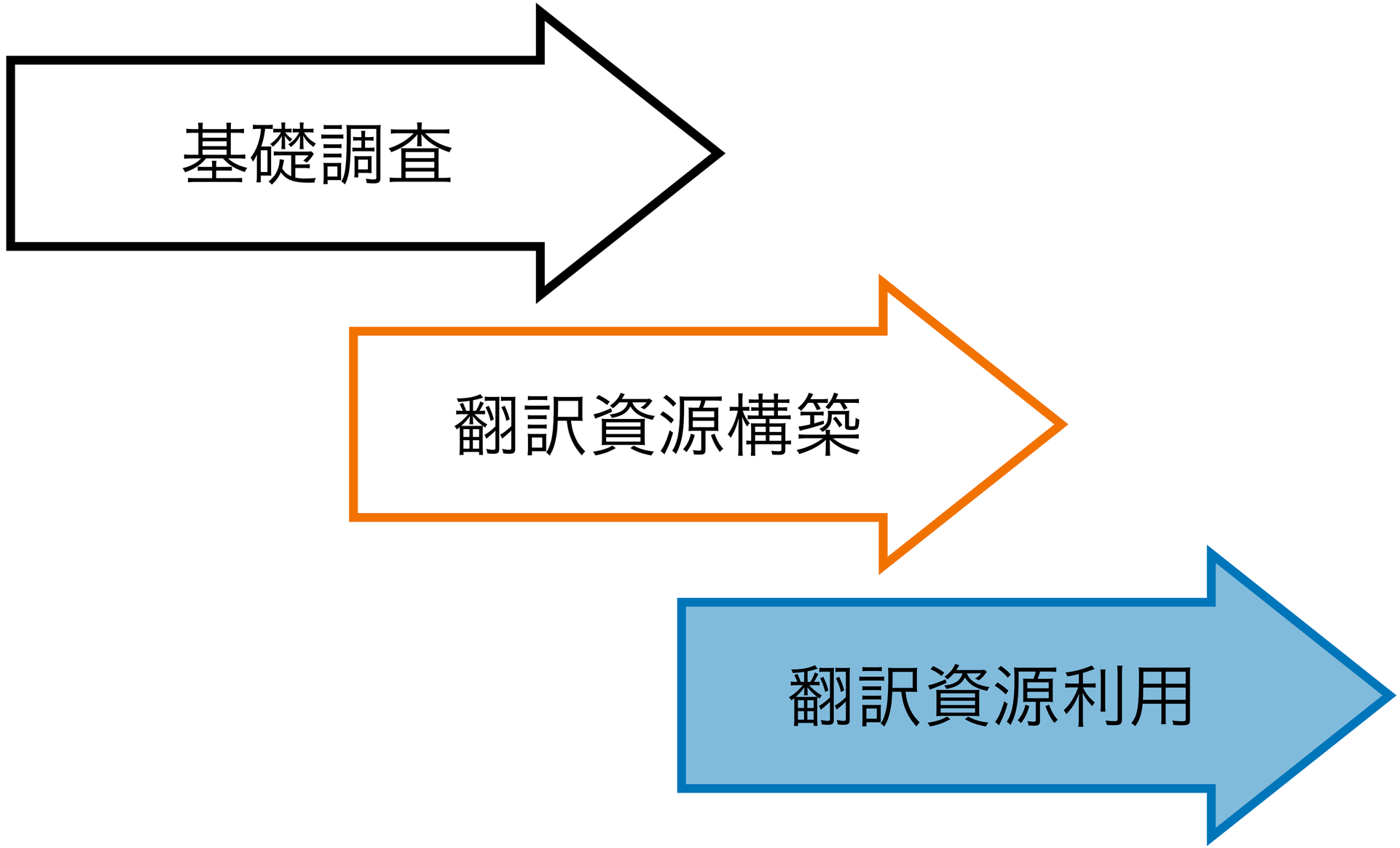
文書属性の付与 (Miyata & Miyauchi, 2022)

- 翻訳時に参照されうる情報（**メタ言語**）
- 属性の絞り込みと文書への付与

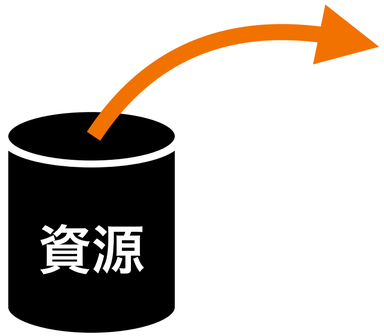
知識属性	コミュニケーション属性	形態属性	テキスト属性
(K01) 主題分野	(C01) 発信	(F01) 伝達媒体	(T01) 言語
(K02) トピック	(a) 発信者	(F02) 記号系	(T02) レジスター
(K03) ジャンル	(i) 依頼者	(F03) ファイル	(a) モード
(K04) 内容の難易度	(ii) 執筆者	(a) 分量	(b) 形式度合い
(K05) 背景知識	(b) 発信の時	(b) フォーマット	(T03) 方言
(a) 学問分野	(c) 発信の場所	(c) マークアップ	(a) 地域方言
(b) 前提	(C02) 受信	(d) 編集可能性	(b) 時代方言
(K06) 関連リソース	(a) 受信者	(F04) 構造	(c) 社会方言
(a) 出典	(i) 対象読者	(a) 文書構造	(T04)スタイル
(b) 専門用語集	(ii) 潜在読者	(b) 内容構造	(a) スタンス
	(b) 受信の時		(b) 感情強度
	(c) 受信の場所		(c) 文学性
	(C03) 発信者と受信者の関係		(d) 個人的特徴／独創性
	(C04) コミュニケーション分野		(T05) 品質
	(C05) 機能		(a) 結束性
	(C06) 目的		(b)一貫性
	(C07) 背景状況		(c) 読みやすさ
			(d) 話しやすさ
			(e) 誤り度合い
			(T06) 表現形式

出典：TML-SDProperty-v1-JA, https://github.com/tntc-project/document-properties

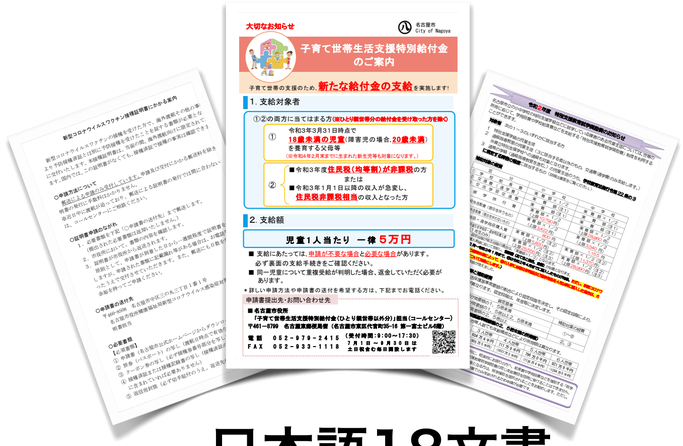
研究の流れ



機械翻訳研究へのデータの展開

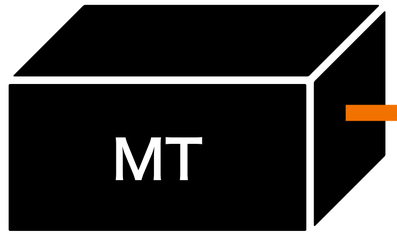


対訳文書



日本語18文書
(1045 文)

https://github.com/tr4lg/nagoya-dataset



機械翻訳システム
- TexTra
- Google翻訳

PE

(Fujita, 2024)

アノテーション

MT	How many **tickets** do you need?
MQM	**type:** Accuracy/Mistranslation
	severity: Major

MTPEdocs-MQM (翻訳イシュー付与データ)

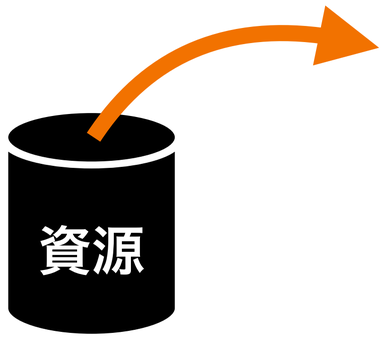
https://github.com/tr4lg/MTPEdocs-MQM

ST	何枚必要ですか？
MT	How many **tickets** do you need?
PE	How many **copies** do you need?

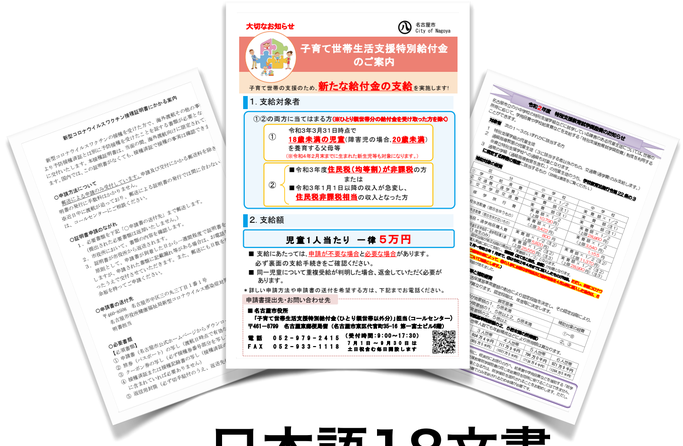
MTPEdocs (MTPEの研究データ)

https://github.com/tntc-project/MTPEdocs

機械翻訳研究へのデータの展開

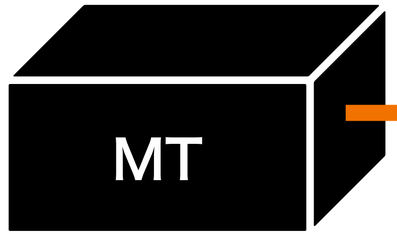


対訳文書



日本語18文書
(1045 文)

https://github.com/tr4lg/nagoya-dataset



機械翻訳システム
- TexTra
- Google翻訳

PE

ST	[GAP] 何 枚 必 要 で す か 。
	BAD OK OK OK OK OK OK OK OK
MT	How many **tickets** do you need ?
	OK OK BAD OK OK OK OK

QEdatasetJaEn (翻訳品質推定用データ)

https://github.com/tntc-project/QEdatasetJaEn

(島田+, 2024)

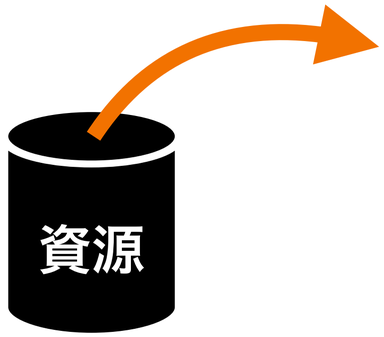
アノテーション

ST	何枚必要ですか？
MT	How many **tickets** do you need?
PE	How many **copies** do you need?

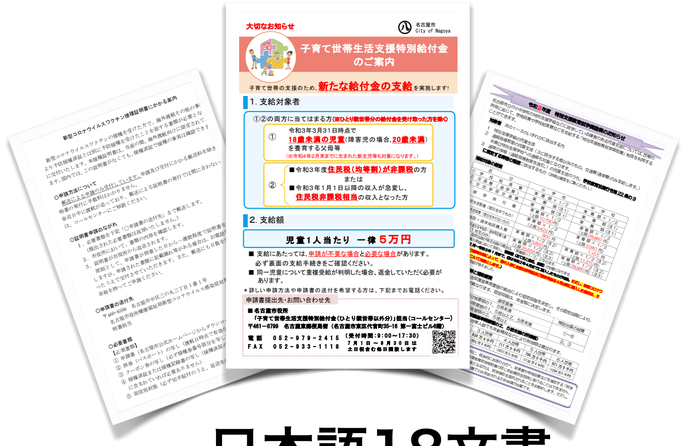
MTPEdocs (MTPEの研究データ)

https://github.com/tntc-project/MTPEdocs

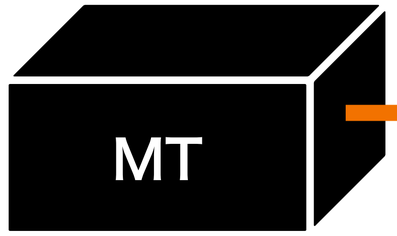
機械翻訳研究へのデータの展開



対訳文書



日本語18文書
(1045 文)



機械翻訳システム
- TexTra
- Google翻訳

PE

https://github.com/tr4lg/nagoya-dataset

MT	Please write your life **situation**.
中間文	Please write your **life conditions**.
PE	Please write your **living** conditions.

DecomposedMTPE (編集事例分解データ)

https://github.com/tntc-project/DecomposedMTPE

(Yamaguchi+, 2024)

アノテーション

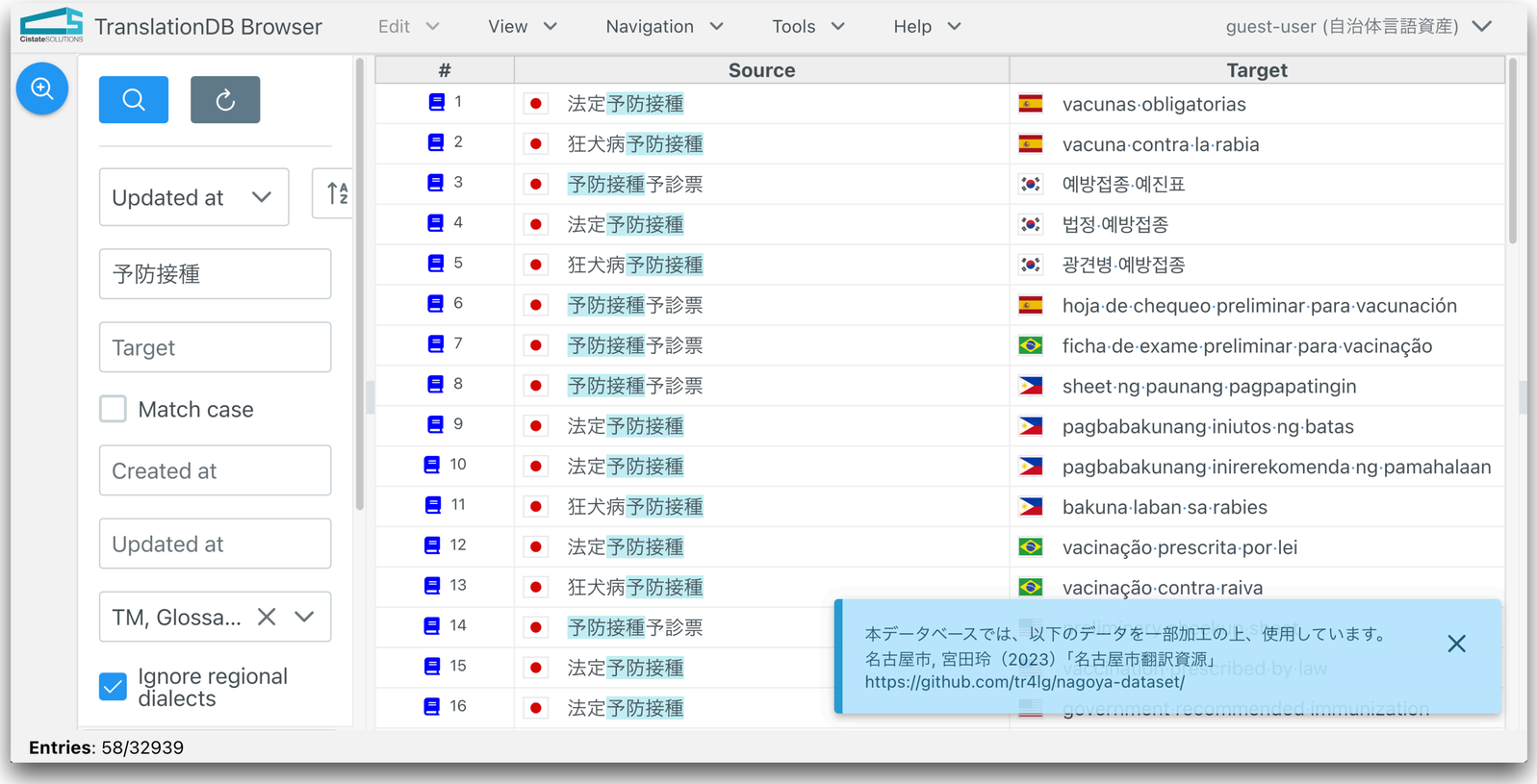
ST	生活状況を書いてください。
MT	Please write your life situation.
PE	Please write your living conditions.

MTPEdocs (MTPEの研究データ)

https://github.com/tntc-project/MTPEdocs

「自治体言語資産データベース」

- 名古屋市翻訳資源を検索できるプラットフォーム
  - 公開済の資源を利用して企業が作成・公開



株式会社システートソリューションズ, https://jichitai.translation-db.com/view

令和6年能登半島地震 支援情報ナビ

- 株式会社アスコエパートナーズが運用・提供
- 能登半島地震の被災者向けの支援制度情報を集約

「石川県 令和6年能登半島地震 支援情報ナビ」より　https://aidfor.ishikawa-pref.supportnavi.jp

多言語版の提供

Cistate Translation Proxy
by 株式会社システートソリューションズ

10言語への翻訳（翻訳資源を活用した生成AIベースの機械翻訳）

「石川県 令和6年能登半島地震 支援情報ナビ」より　https://aidfor.ishikawa-pref.supportnavi.jp

教訓・課題

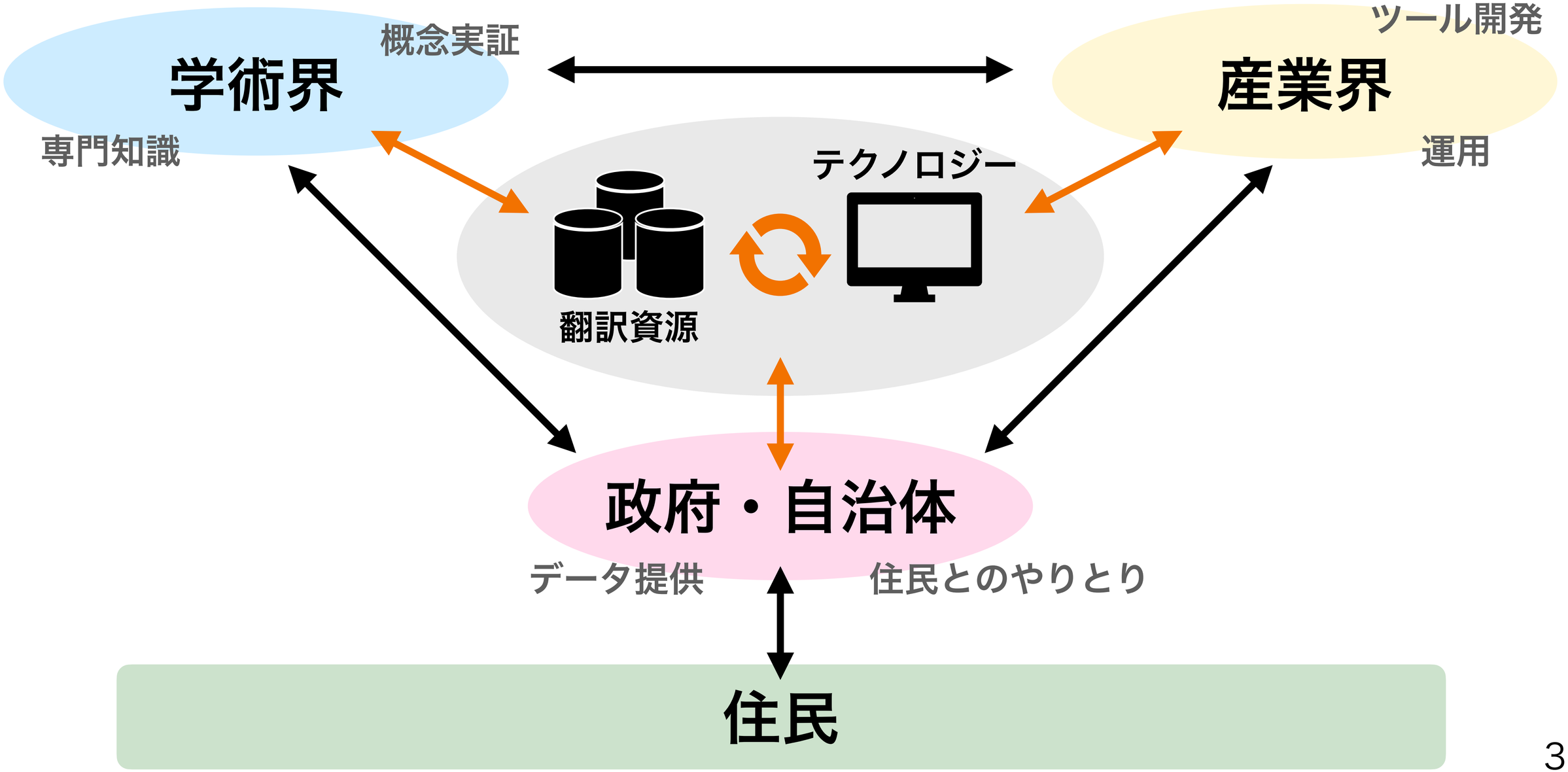
- 公開済みの翻訳資源を用いた、機械翻訳のカスタマイズの円滑化
  - 約10日ほどでウェブサイトへ多言語翻訳機能の実装
- 現在カバーできていない固有名詞、専門用語、ロケールの翻訳は難しい
  - 金沢市鞍月　→　Anazuki, Kanazawa City
  - 半壊　→　partial collapse, partial destruction, half-collapse …
  - 令和6年　→　Reiwa 6
- 用語集等の翻訳資源のさらなる拡張が必要
  - 順次公開予定！

III. まとめと今後の展望

まとめ

- 自治体横断で使える**翻訳資源**の構築
  - 文書・文・用語レベルでの翻訳知識の集積と活用
  - ただ集めて公開すればよいわけではない
    - メタ情報の付与、ガイドライン・規準の作成
- **翻訳テクノロジー**への応用
  - 微調整・RAG用データ
  - 文書レベル機械翻訳研究

翻訳資源構築・活用のエコシステムに向けて



参考文献

Bedard, C. (2000). Mémoire de traduction cherche traducteur de phrases... *Traduire*, 186, 41–49.

Fujita, A. (2024). MQM-like issue typology used for annotation. https://github.com/tr4lg/MTPEdocs-MQM/blob/main/fujita-issue-typology-en.pdf

Maruf, S., Saleh, F., & Haffari, G. (2022). A survey on document-level neural machine translation: Methods and evaluation. *ACM Computing Surveys*, 54(2), 1–36.

Miyata, R. (2020). *Controlled Document Authoring in a Machine Translation Age*. Routledge.

Miyata, R., & Miyauchi, T. (2022). Metalanguages for source document analysis: Properties and elements. *Metalanguages for Dissecting Translation Processes: Theoretical Development and Practical Applications* (pp. 63–79), Routledge.

Miyata, R., Yamada, M., & Kageura, K. (2022). *Metalanguages for Dissecting Translation Processes: Theoretical Development and Practical Applications*. Routledge.

Mossop, B. (2006). Has computerization changed translation? *Meta*, 51(4), 787–805.

Yamaguchi, D, Miyata, R, Fujita, A., Kajiwara, T., & Sato, S. (2024). Automatic decomposition of text editing examples into primitive edit operations: Toward analytic evaluation of editing systems. *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation*,1899–1914.

参考文献

神戸市 (2024).「神戸市におけるAI活用のためのルール整備」https://www8.cao.go.jp/cstp/ai/ai_kenkyu/2kai/shiryou3.pdf

島田紗裕華, 山口大地, 宮田玲, 藤田篤, 梶原智之, 佐藤理史 (2024). 機械翻訳向け原文編集の支援に向けた日英翻訳品質推定データセットの設計と構築. 言語処理学会第30回年次大会.

朴惠, 山浦育子, 宮田玲 (2024). 日中対訳用語集構築に向けた翻訳の規準と手続きの明確化：被災者支援分野を対象に. 日本通訳翻訳学会第25回年次大会.

宮田玲 (2020). 日本における自治体ウェブサイトの多言語化の現況と課題. 通訳翻訳研究, 20, 1–24.

宮田玲, 阪本章子, 藤田篤, 香川璃奈 (2024).言語情報発信を支援する自治体横断型翻訳資源の構築：プロジェクトの理念、枠組み、現状. 日本通訳翻訳学会第25回年次大会.