サブセット探索を用いた
高速なkNNニューラル機械翻訳

NAIST渡辺研D3 / NICT翻訳研

出口 祥之

2024/03/22: AAMT 若手翻訳研究会

deguchi.hiroyuki.db0@naist.ac.jp

背景: ニューラル機械翻訳のドメイン適応

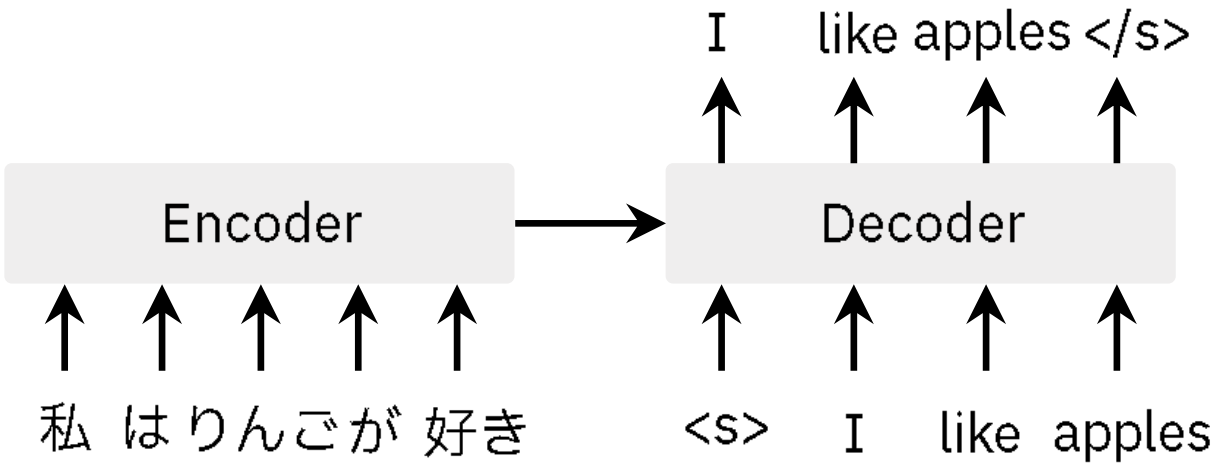
- ニューラル機械翻訳 (NMT) の課題
  - ドメイン内の翻訳精度は高い一方で，訓練データと異なるドメインの翻訳精度は低い.
  - ドメインごとに特化したエンジンを訓練するのはコストがかかる.
- 先行研究
  - 翻訳用例を活用 (Zhang+, NAACL2018; Gu+, AAAI2018; Khandelwal+, ICLR2021)
    - 用例ベース翻訳 (Nagao, 1984) のアイデアを基本とする.
    - 翻訳用例を検索・参照して翻訳することで，ドメイン翻訳の精度を改善.
  - 中でも，kNN-MT (Khandelwal+, ICLR2021) は，
    - 汎用エンジンをそのまま利用でき,
    - 追加訓練することもなく,
    - ドメイン適応翻訳において最高翻訳性能を達成.

Guiding Neural Machine Translation with Retrieved Translation Pieces (Zhang+, NAACL2018)
Search Engine Guided Neural Machine Translation (Gu+, AAAI2018)
Nearest Neighbor Machine Translation (Khandelwal+, ICLR2021)
A framework for a mechanical translation between Japanese and English by analogy principle (Nagao, 1984)

前提: 典型的なニューラル機械翻訳（エンコーダ・デコーダ型）

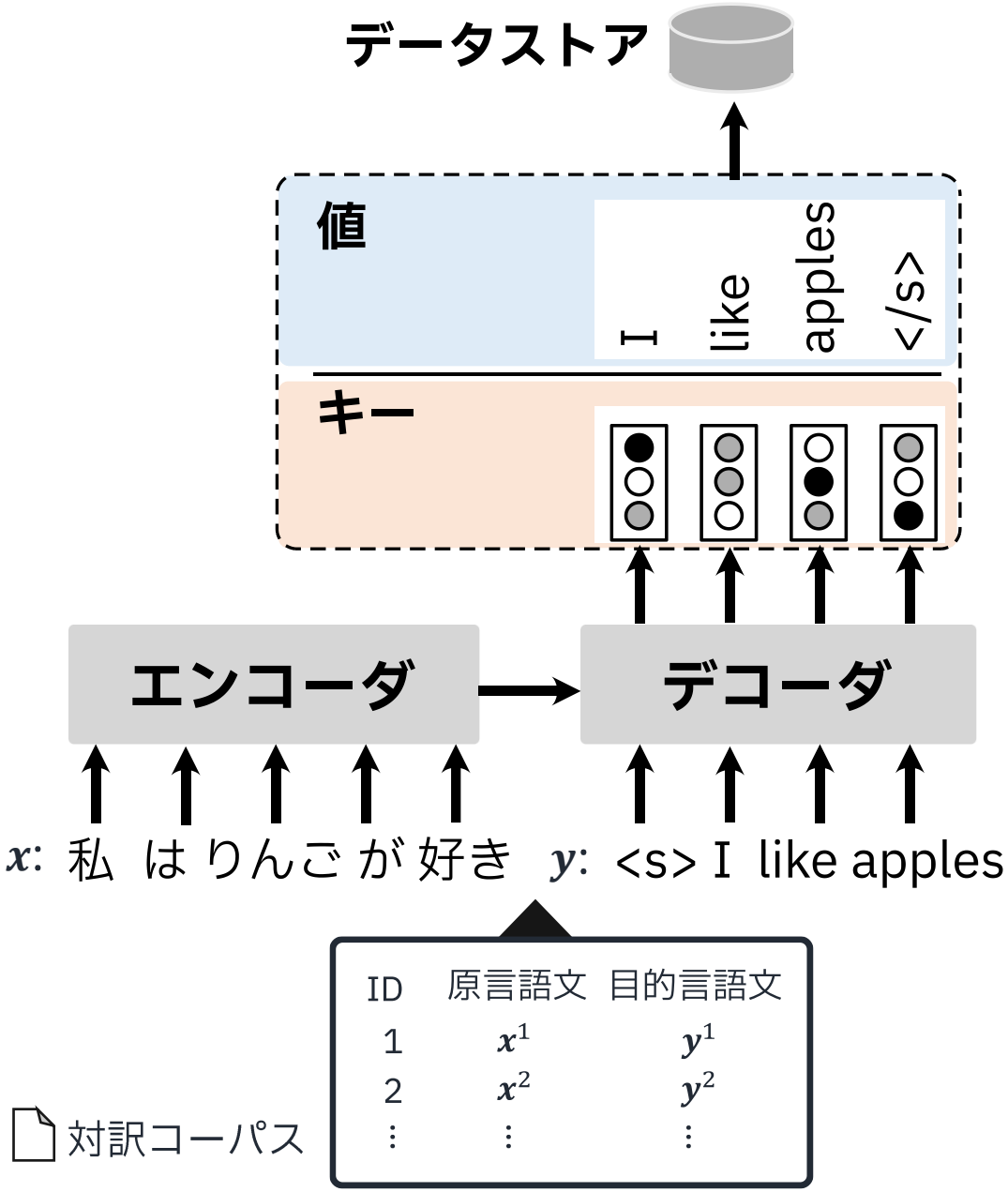
- エンコーダは入力文を受け取る
- デコーダは
  - エンコードされた入力文
  - それまでのデコーダの単語出力履歴

  をもとに，一単語ずつ出力文を生成



kNN-MT (Khandelwal+, ICLR2021)：データストア構築

- データストア構築
  - 訓練済みNMTに対訳文対を入力して計算
  - **値**: 正解の目的言語単語
  - **キー**: 各目的言語単語の中間表現ベクトル
- データストアの大きさ
  - 対訳コーパスの目的言語側の全単語数
    - ▶ 同じ語彙でも文脈毎に異なるデータとして扱われる
  - 例: WMT'19独英対訳: 29.5M文対，862.6M単語
    - ▶ 32-bit x 1024-次元 x 1B-単語 ≈ 3.7 TiB



Nearest Neighbor Machine Translation (Khandelwal+, ICLR2021)

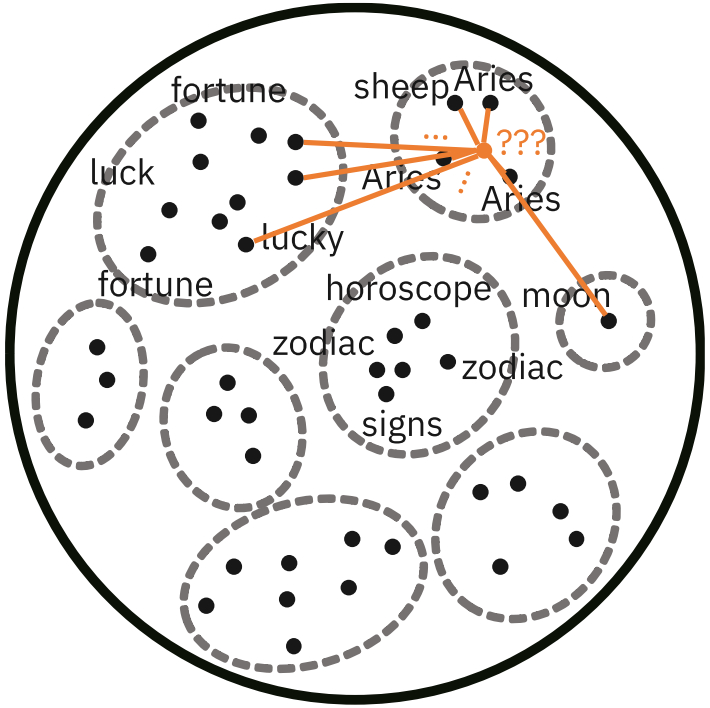
kNN-MT (Khandelwal+, ICLR2021)：翻訳文生成

- キーと同じ層の中間表現ベクトルを検索クエリとし，各時刻でデータストアからk近傍事例を探索
- 検索結果から単語出力スコアに変換
  - クエリとの距離が近い単語ほどスコアを高くする
  - クエリ近傍に同じ単語が多く存在するほどスコアを高くする

$$p_{k\mathrm{NN}}(y_t|\boldsymbol{x}, \boldsymbol{y}_{<t}) \propto \sum_{i=1}^{k} \mathbb{1}_{y_t=v_i} \exp \frac{-\|\boldsymbol{k}_i - f(\boldsymbol{x}, \boldsymbol{y}_{<t})\|_2^2}{\tau}$$

近傍に同じ単語が多くあるほど高く

キー; $\boldsymbol{k}_i \in \mathbb{R}^D$ とクエリ; $f(\boldsymbol{x}, \boldsymbol{y}_{<t}) \in \mathbb{R}^D$ との間の距離が近いほど高く



My zodiac sign is ???

kNN-MT

私の星座は牡羊座です。

- ユーザー指定の重みを付けてkNNスコアと元のNMTの予測スコアを線形補間し，翻訳文を生成する

Nearest Neighbor Machine Translation (Khandelwal+, ICLR2021)

事前実験: 医療ドメインにおける kNN-MT の翻訳性能（独英翻訳）

モデル	↑BLEU	↑単語/秒
Base MT	42.1	4392.1
kNN-MT	48.2 (+6.1)	19.8 (×1/222)

☺ 追加訓練なしで +6.1 BLEU 改善

☹ 翻訳速度は222倍低下

先行研究

- 数単語分まとめて検索 (4x高速化)
  (Martins+, EMNLP2022)
- 入出力言語間の単語対応をもとに検索する単語を制限 (10x高速化)
  (Meng+, ACLFindings2022)
  - Base MTの5%の翻訳速度では実用的といえない

項目		設定
データ	データストア	医療ドメインを含む複数ドメインの対訳コーパスから構築 (31M文対, 896M単語)
モデル	Base MT	Transformer big WMT'19独英対訳コーパスで訓練済み
	kNNスコア重み	$\lambda = 0.5$
	検索近傍数	$k = 16$
評価	翻訳精度	↑sacreBLEU (%)
	翻訳速度	↑生成単語数/秒

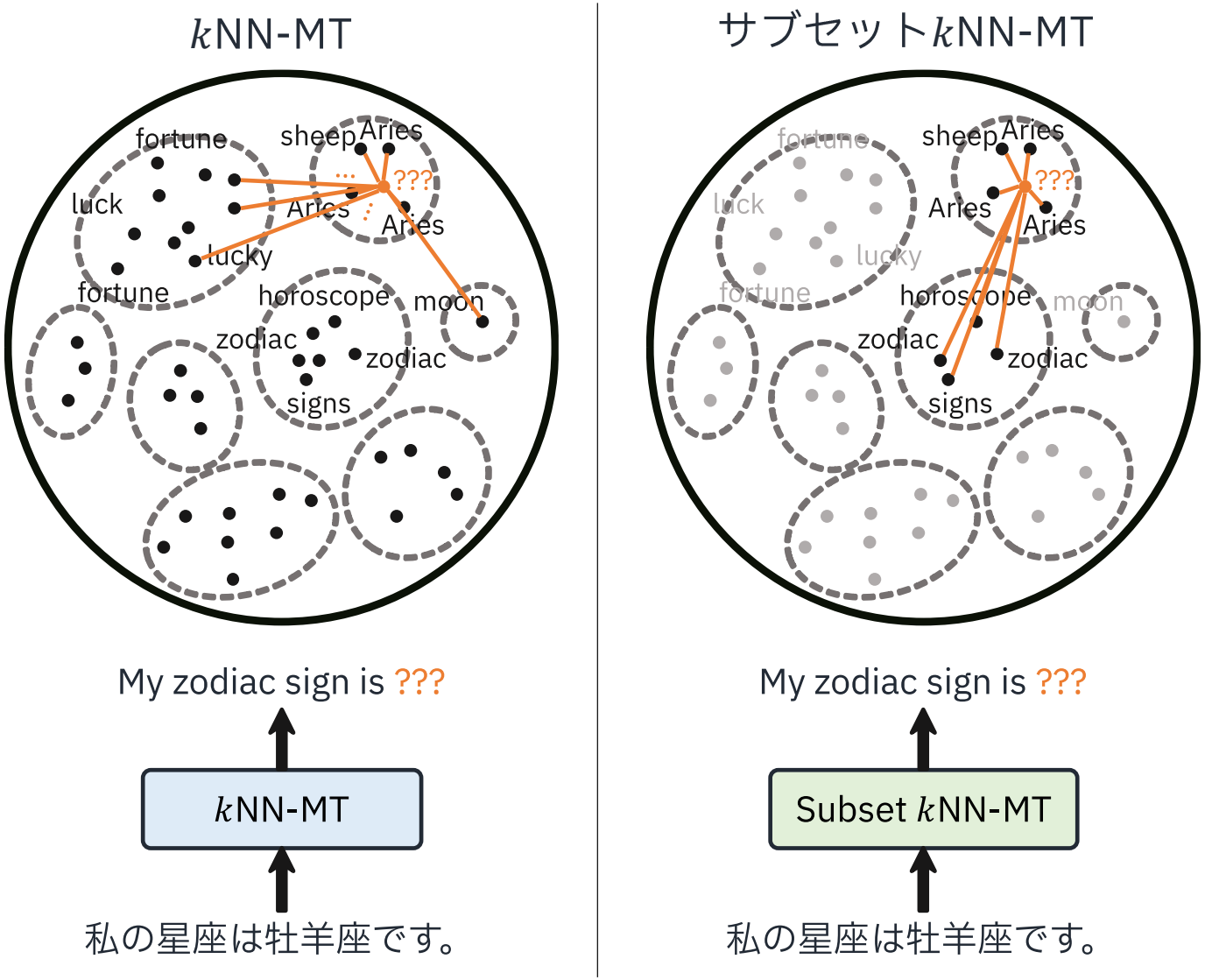
Chunk-based Nearest Neighbor Machine Translation (Martins+, EMNLP2022)
Fast Nearest Neighbor Machine Translation (Meng+, ACL Findings2022)

提案: サブセットkNN-MT

目的: kNN-MTの翻訳速度の改善を狙う

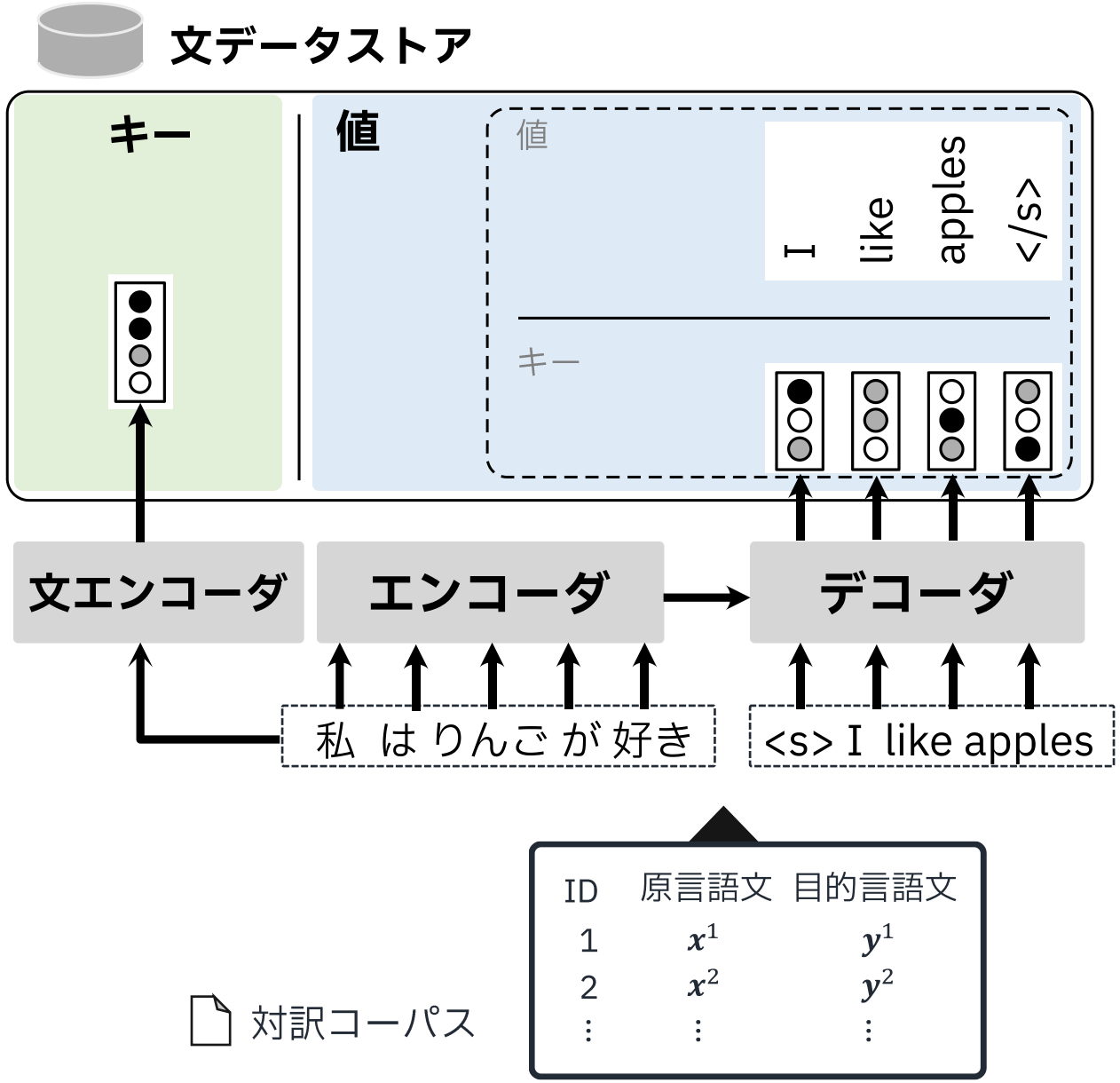
- 入力文に関連する事例のみにkNN探索空間を絞り込む
- ルックアップテーブルを用いて，ベクトル間距離を効率的に計算
  - 既存の大規模kNN探索アルゴリズムは全探索に特化している (Matsui+, ACMMM2018)
  - 入力文ごとに検索対象が変わるサブセットkNN-MTに適した距離計算法を採用

Reconfigurable Inverted Index (Matsui+, ACMMM2018)



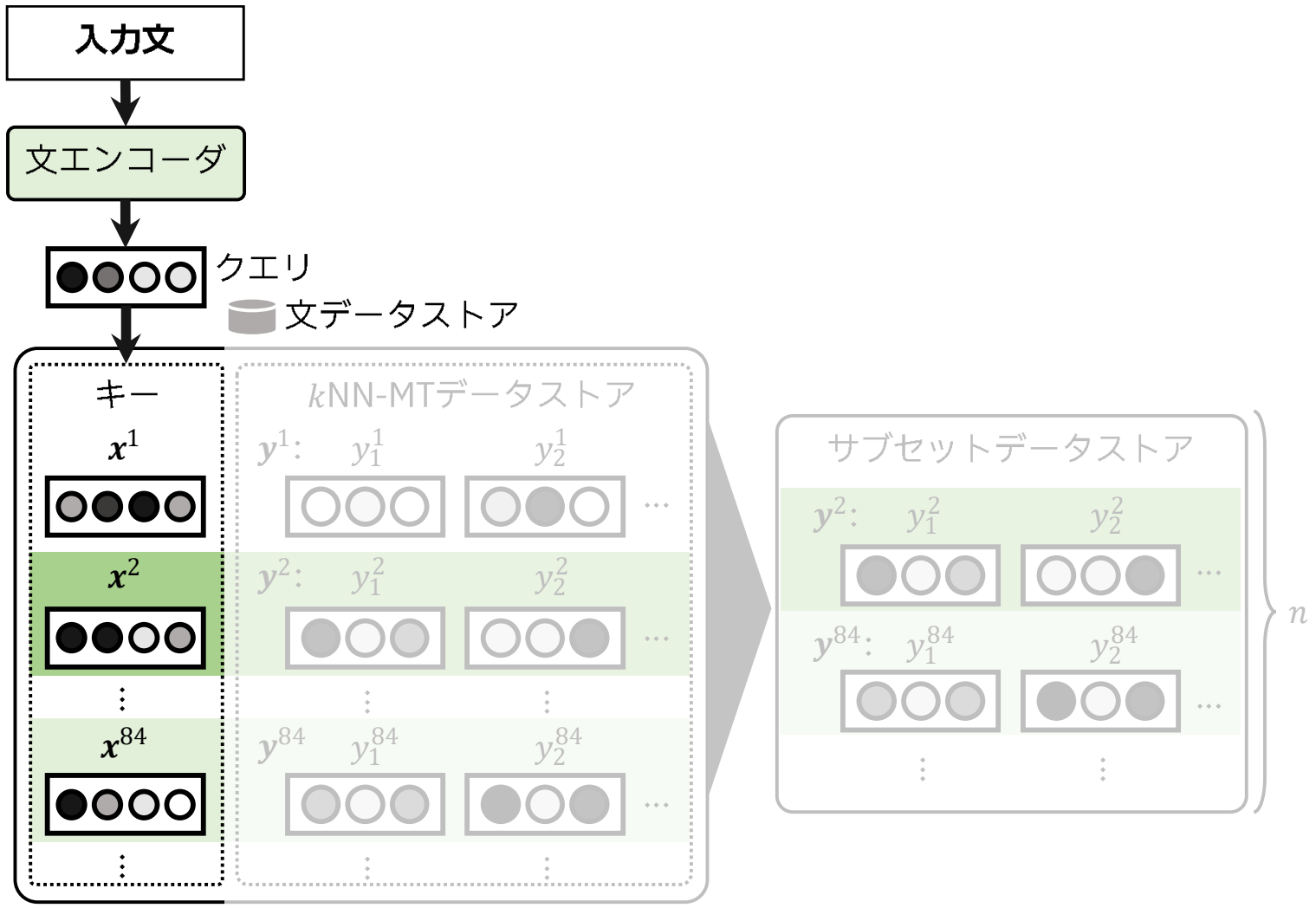
文データストア構築

- 文データストア
  - kNN-MTデータストアを文単位で切り出せるようにした構造
  - 原言語文とそれに対応する目的言語文のkNN-MTデータストアを紐づける
- キー: 原言語文の文ベクトル
- 値: キーに対応する目的言語文内の各単語のkNN-MTデータストア



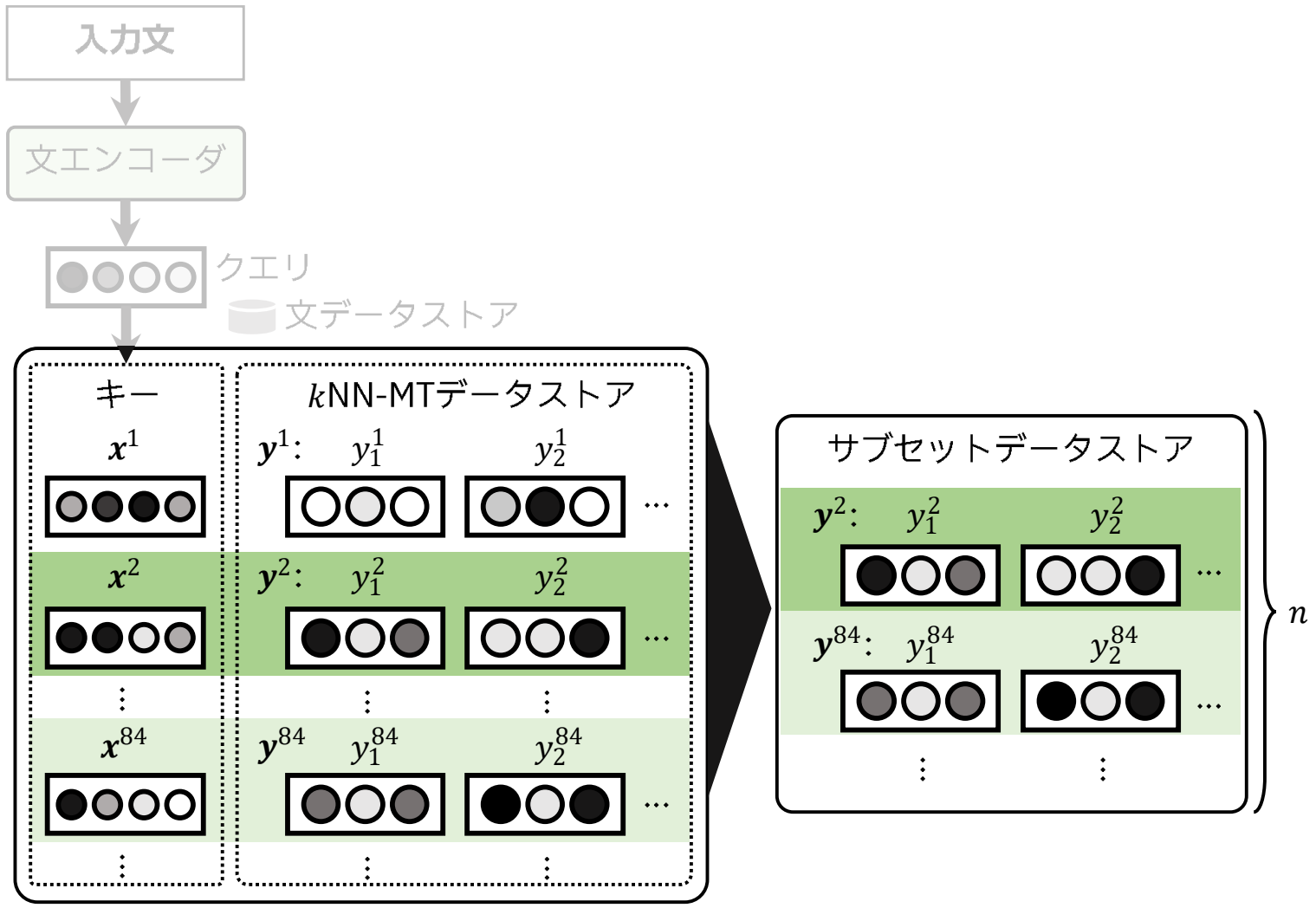
翻訳文生成

1. 入力文に類似した上位n文を文データストアから検索
2. 探索対象を類似文の事例のみに絞り込む
3. 絞り込まれたサブセットデータストアを用いて通常のkNN-MTと同じように翻訳文を生成する



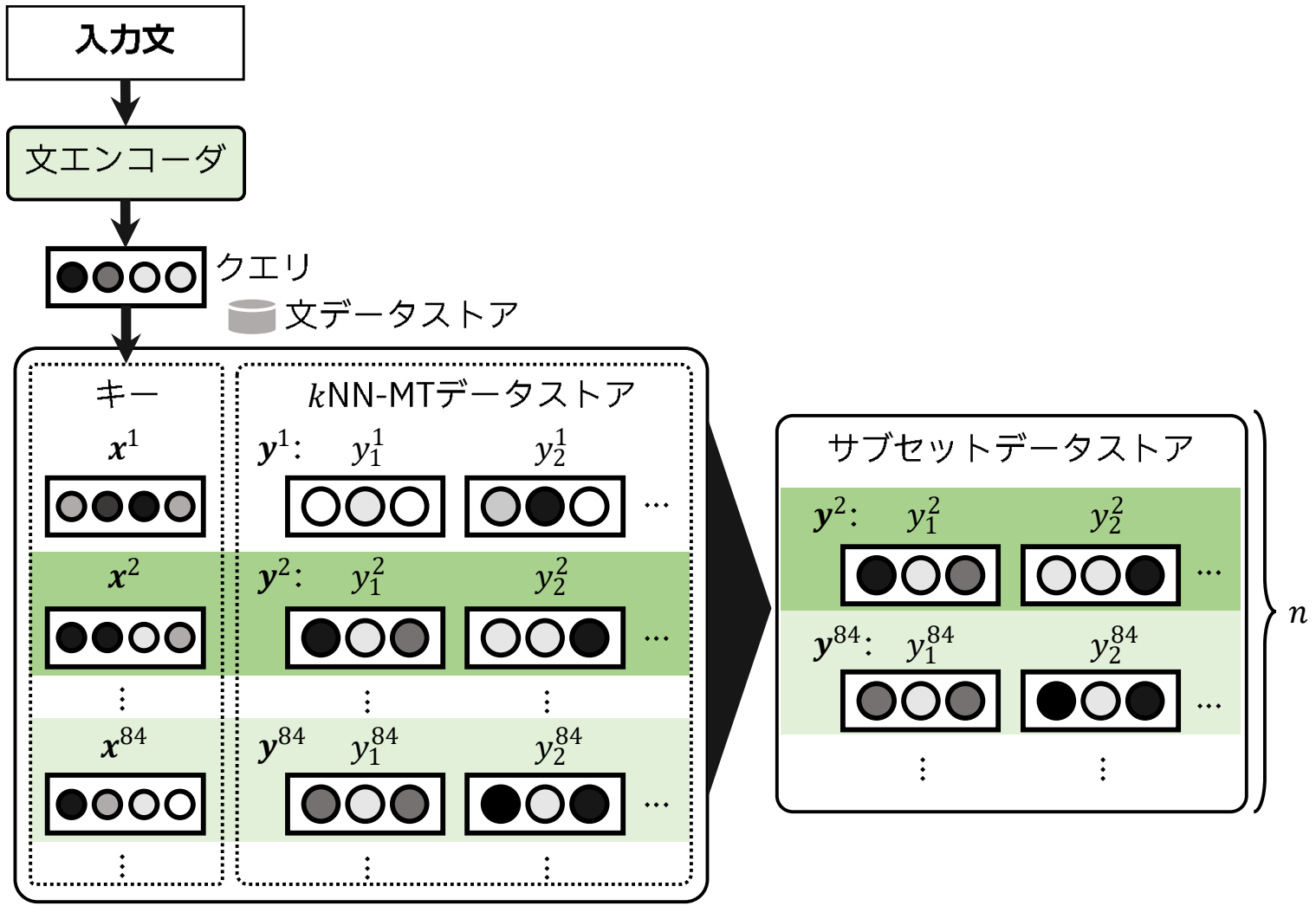
翻訳文生成

1. 入力文に類似した上位n文を文データストアから検索
2. 探索対象を類似文の事例のみに絞り込む
3. 絞り込まれたサブセットデータストアを用いて通常のkNN-MTと同じように翻訳文を生成する



翻訳文生成

1. 入力文に類似した上位n文を文データストアから検索
2. 探索対象を類似文の事例のみに絞り込む
3. 絞り込まれたサブセットデータストアを用いて通常のkNN-MTと同じように翻訳文を生成する



翻訳実験: ドメイン適応（独英翻訳）

項目		設定
データ	翻訳対象ドメインと各ドメインの対訳コーパスサイズ	• IT: 185K • コーラン: 15K • 法律: 451K • 医療: 210K • 字幕: 443K
	データストア	汎用+上記5ドメインの対訳コーパス 31M 文対（896M 単語）から構築
モデル	kNNスコア重み	$\lambda = 0.5$
	探索近傍数	$k = 16$
	絞り込み文数	$n = 256$
	文検索モデル	LaBSE

実験結果: ドメイン適応（独英翻訳）

モデル	IT		コーラン		法律		医療		字幕	
	BLEU	単語/秒	BLEU	単語/秒	BLEU	単語/秒	BLEU	単語/秒	BLEU	単語/秒
Base MT	38.7	4433.2	17.1	5295.0	46.1	4294.0	42.1	4392.1	29.4	6310.5
kNN-MT	41.0	22.3	19.5	19.3	52.6	18.6	48.2	19.8	29.6	30.3
サブセット kNN-MT	**41.9**	2362.2	**20.1**	2551.3	**53.6**	2258.0	**49.8**	2328.3	**29.9**	3058.4

- kNN-MTと比較して，
  - 速度: 100倍以上高速化し，Base MTの50%程度の速度で翻訳
    - ▶ **実用的に使える速度まで高速化に成功**
  - 精度: 各ドメインで1 BLEU%ほど翻訳精度が改善（最大1.6%）
    - ▶ 分析によりわかったこと: 関連する事例のみに絞り込むことで検索結果からノイズが削減され，翻訳精度が改善した

事例分析: 医療ドメイン（独英翻訳）

入力文	Eine gemeinsame Anwendung von Nifedipin und Rifampicin ist daher kontraindiziert.
参照訳	Co-administration of nifedipine with rifampicin is therefore contra-indicated.
Base MT	A joint use of nifedipine and rifampicin is therefore contraindicated.
kNN-MT	A joint use of nifedipine and rifampicin is therefore contraindicated.
サブセット kNN-MT	Co-administration of nifedipine and rifampicin is therefore contraindicated.

- サブセットkNN-MTは医療ドメインの専門用語 “Co-administration” を正確に訳出している.

事例分析: 入力類似文の検索結果

入力文	Eine gemeinsame Anwendung von Nifedipin und Rifampicin ist daher kontraindiziert.
Src-1	Die gemeinsame Anwendung von Ciprofloxacin und Tizanidin ist kontraindiziert.
Src-2	Rifampicin und Nilotinib sollten nicht gleichzeitig angewendet werden.
Src-3	Die gleichzeitige Anwendung von Ribavirin und Didanosin wird nicht empfohlen.
Tgt-1	**Co-administration** of ciprofloxacin and tizanidine is contra-indicated.
Tgt-2	Rifampicin and nilotinib should not be used concomitantly.
Tgt-3	**Co-administration** of ribavirin and didanosine is not recommended.

- 訳出してほしい “Co-administration” が検索結果上位に2件含まれている
  - 一方，通常のkNN-MTで誤って訳出した “A joint use” は類似文上位256文中に1件も含まれていなかった

→ 入力文に関連する事例のみに絞り込むことで検索結果からノイズが削減

まとめ

kNN-MTの翻訳速度を改善するサブセットkNN-MTを提案

- 提案法
  - 探索空間を入力文の関連事例のみに動的に削減
  - ルックアップテーブルを用いた効率的な距離計算法を採用
- ドメイン適応実験より，従来法と比較して，サブセットkNN-MTは，
  - 翻訳速度を最大132.2倍改善するだけでなく，
  - 翻訳精度も最大1.6BLEU%改善することを確認した．
- 今後の課題
  - 音声翻訳 (Speech-to-Text) や画像情報付き翻訳のようなマルチモーダル化を検討

追記

「詳細を知りたい」という方は,

- ACL2023 “Subset Retrieval Nearest Neighbor Machine Translation” https://aclanthology.org/2023.acl-long.10/
- ジャーナル論文: 「自然言語処理」（2024年6月号掲載予定）
  - ▶ 上記ACL論文より，実験・分析，実装詳細についての解説が充実する予定です.
- 実装: https://github.com/naist-nlp/knn-seq
  - ▶ fairseq plug-inとして実装

をご覧いただけると幸いです.