自己紹介と研究紹介

東北大学　篠原・吉仲研究室
修士1年 田崎 浩史

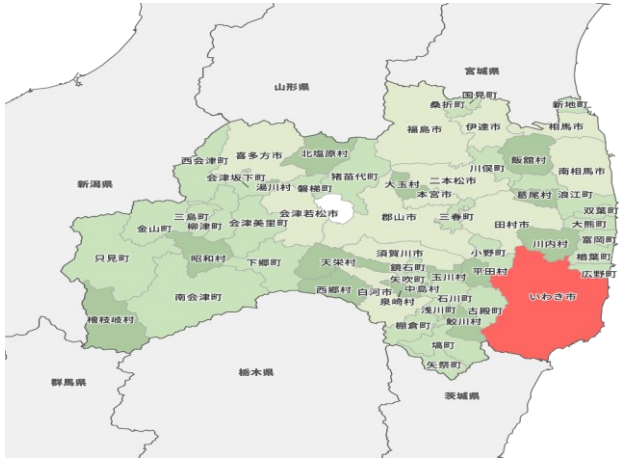
自己紹介

名前

- 田崎 浩史

出身

- 福島県いわき市



趣味

- 漫画・ライトノベル・ゲーム
- カラオケ
- ランニング
- 飼い犬の世話

研究概要


入力
出力例
翻訳モデル（AI）
（ 私は眠いです。 ， I’m sleepy. ）
入力と出力例から
日->英の翻訳方法を学習する
二言語データセット
(私は眠いです。, I’m sleepy.)
(彼は犬が好きです。, He loves dog.)
etc.

研究概要

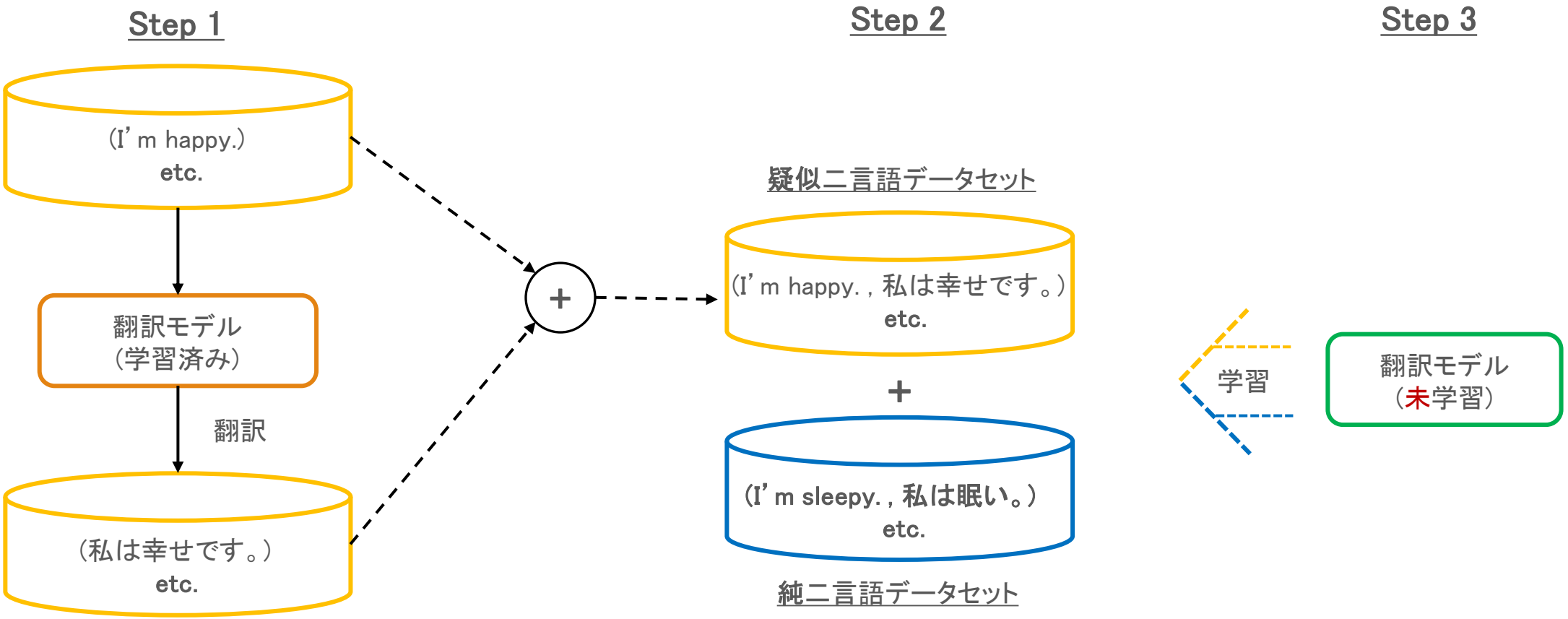
出力例の作成方法

方法A　　入力 -> 翻訳家（人間） -> 出力例　　（純二言語データ）

方法B　　入力 -> 翻訳モデル（AI） -> 出力例　　（疑似二言語データ）

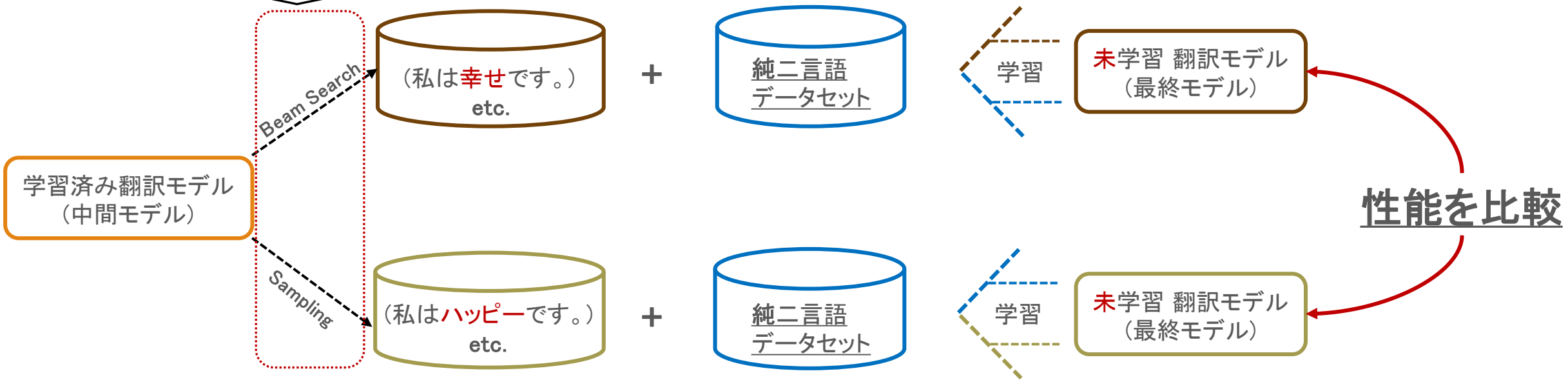
この研究では、方法Bで二言語データを作成する。　　**（逆翻訳）**

研究概要 （逆翻訳）


Step 1
Step 2
Step 3
(I’m happy.)
etc.
翻訳モデル
(学習済み)
翻訳
(私は幸せです。)
etc.
+
疑似二言語データセット
(I’m happy.，私は幸せです。)
etc.
+
(I’m sleepy.，私は眠い。)
etc.
純二言語データセット
学習
翻訳モデル
(未学習)

研究概要 （実験）

探索アルゴリズムによって、生成される文章が異なる。
=> 疑似二言語データセットの品質が変わる。
=> 学習結果が変わる。



研究概要 （目標）

- 探索アルゴリズムの違いによる最終モデルの学習への影響を調べる。
- 逆翻訳において、最適な探索アルゴリズムを求める。

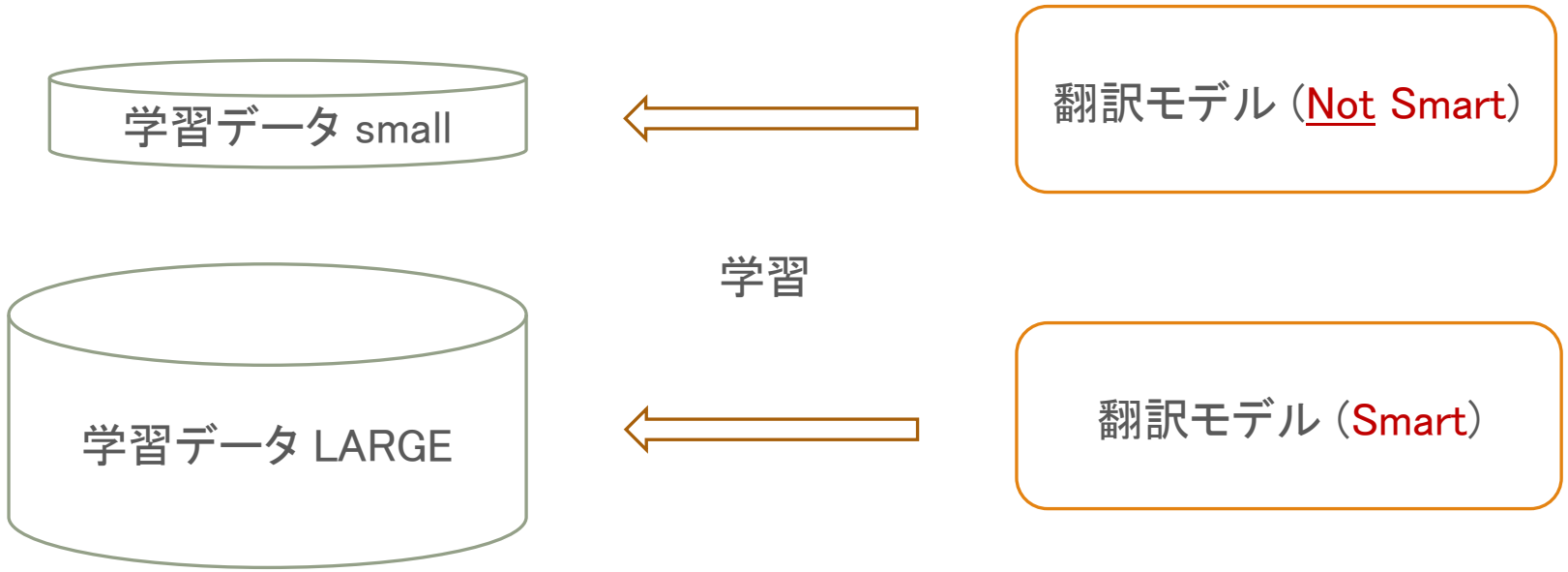
- 先行研究で取り扱われていなかった探索アルゴリズムの有効性を調べる。

Note: 探索アルゴリズムを変化させるのは、中間モデルだけです。
最終モデルの探索アルゴリズムは、Beam Searchに固定します。
スペース上の制約から、スライド6で英文と訳文(日本語)の結合操作の図を省略しました。

なぜ逆翻訳なのか？

なぜデータを増やす必要があるのか？

=> 翻訳モデルの性能向上に直接つながるから



純二言語データを増やすことに伴う問題点

純二言語データを作成するのに必要なコスト

- プロの翻訳家を雇うための人件費が掛かる。
- 大量の二言語データを人の手で作成することで膨大な時間が掛かる。

=> 新たに純二言語データを作成することは、あまり現実的ではない。

（一部の場合を除く）

=> 逆翻訳を使うと、この問題を解決できる！

逆翻訳のメリット

メリット

- 翻訳家の代わりに翻訳モデルを使うことで、高速かつ低コストで訳文を作成できる。
- 翻訳モデルを使ってデータを増やす他の手法に比べて、シンプルで実装しやすい。
- 中間モデルとして既製品(ChatGPT等)を採用すれば、中間モデルの学習コストが掛からない。
- 過去の研究結果等から、一定程度の性能向上が保証されている。

逆翻訳の使用において重要なこと

- 訳文の品質（翻訳モデル）　く　訳文の品質（翻訳家）

- 各探索アルゴリズムの特性を知り、状況に応じて使い分ける必要がある。

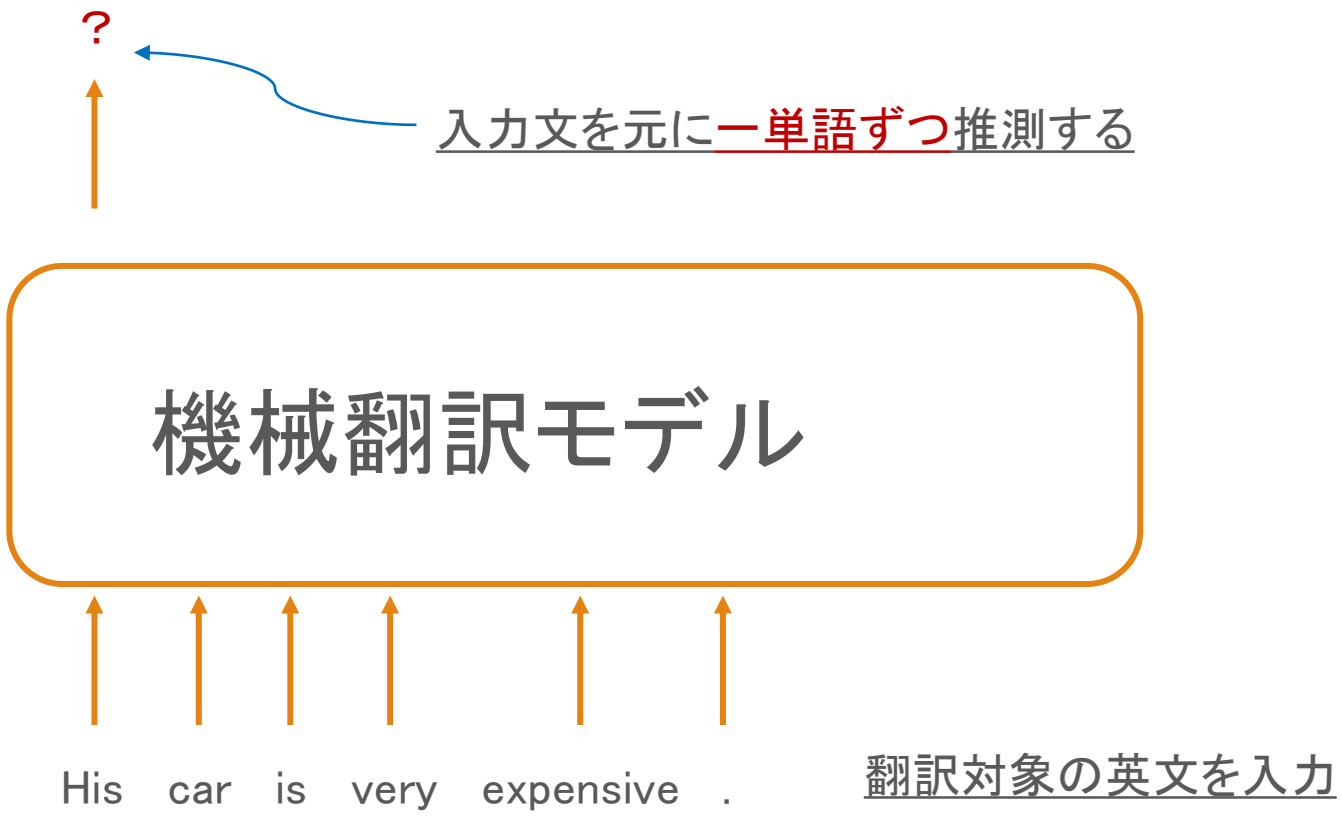
探索アルゴリズム

- 推論プロセス

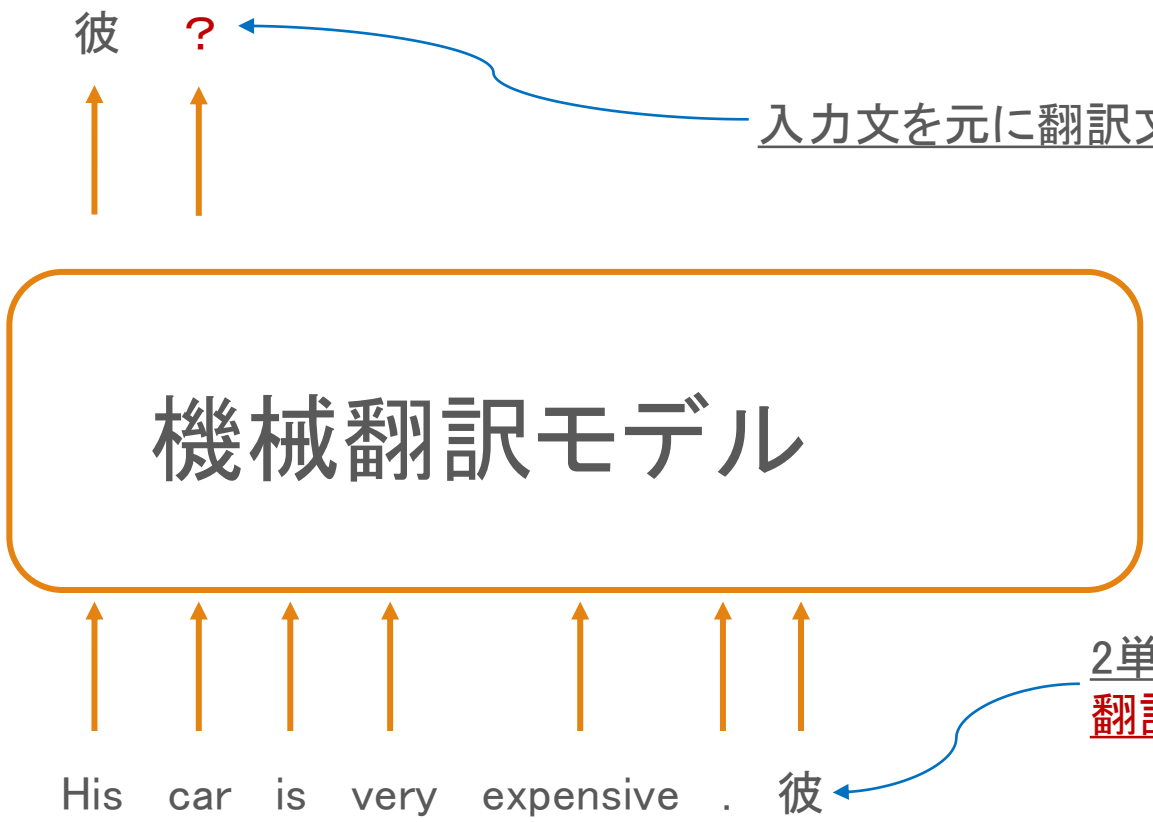
- Sampling

- Beam Search

機械翻訳モデルの推論プロセス



機械翻訳モデルの推論プロセス

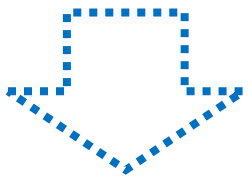


入力文を元に翻訳文の2単語目を推測する

2単語目以降の出力時には、
翻訳対象の英文とそれまでの出力がセットで入力される

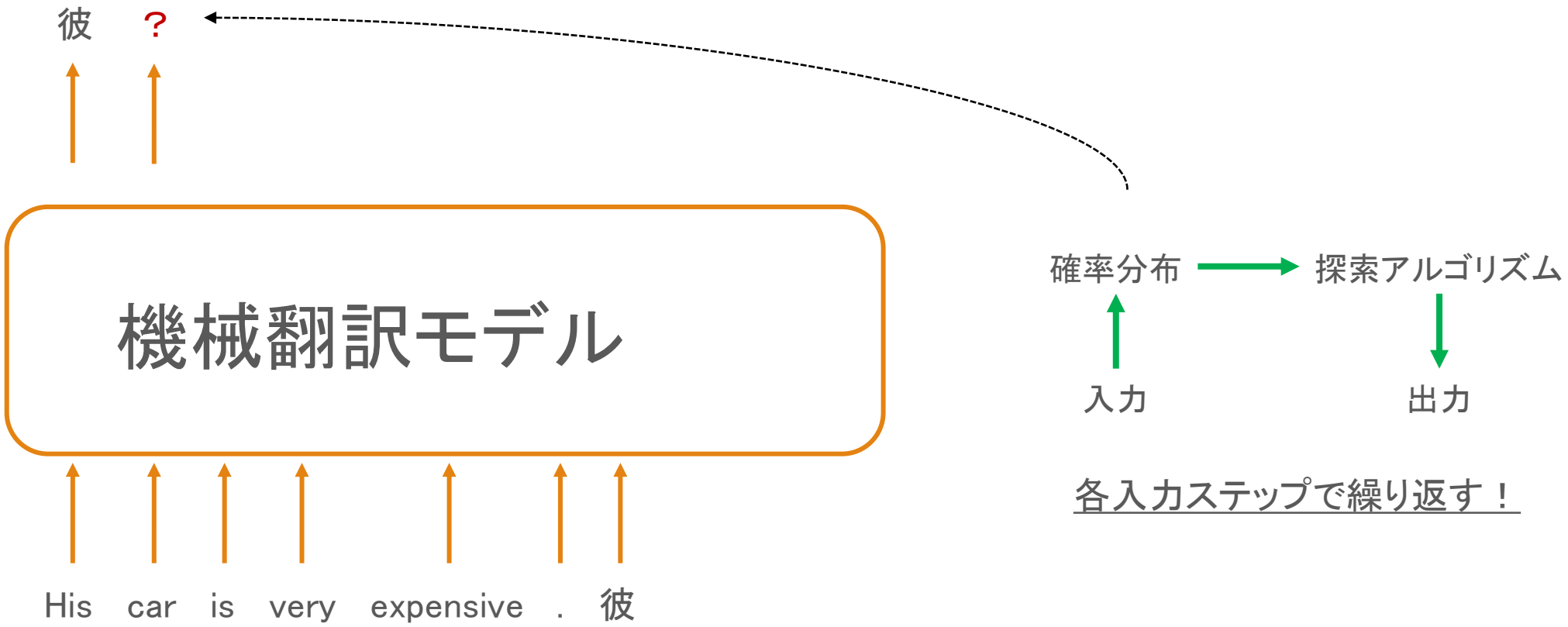
機械翻訳モデルの推論プロセス

入力文に基づいて、確率分布(モデルの語彙に含まれる各単語に対して、次の出力単語として出現する確率が与えられている)を求める。

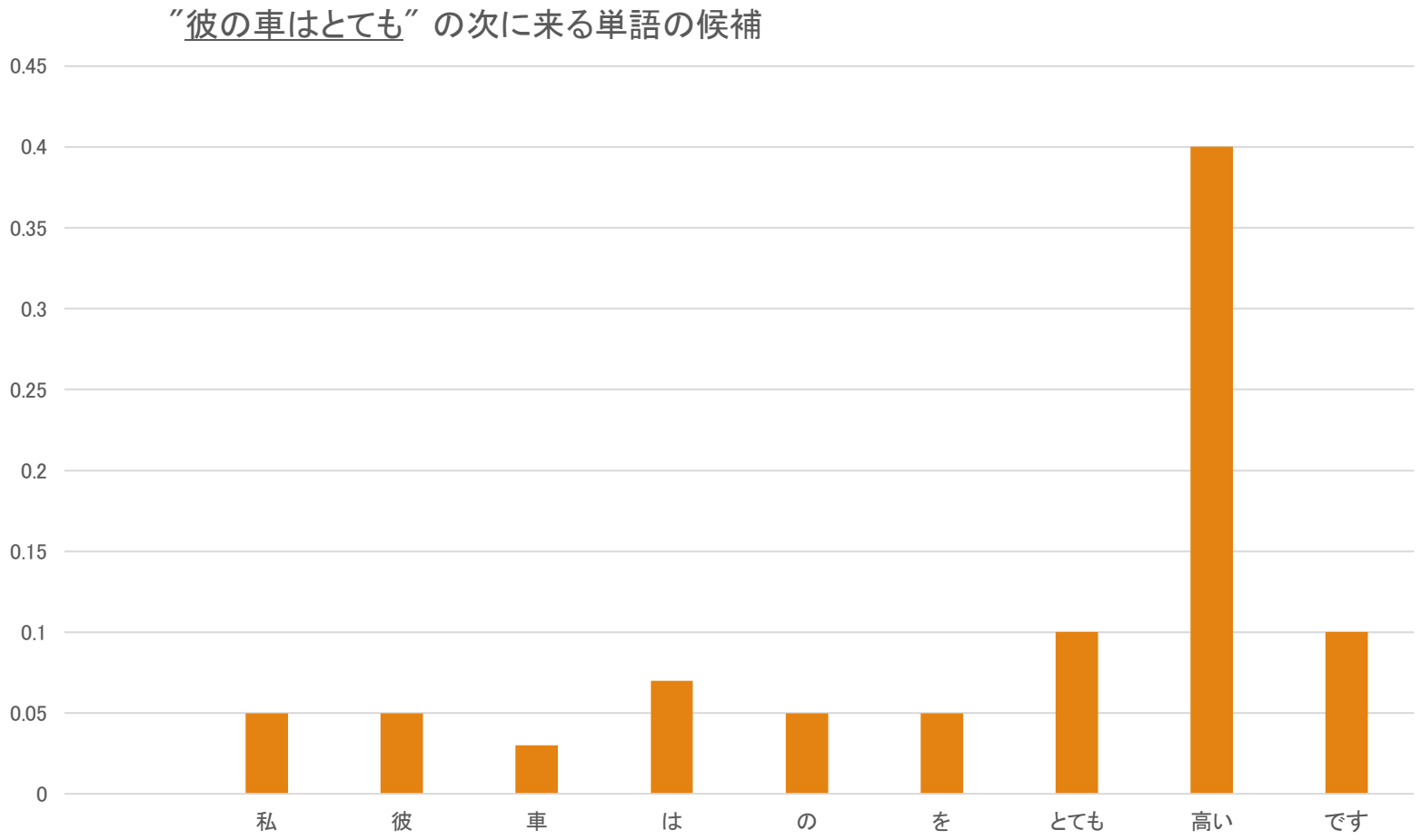


確率分布をもとにして、次の出力単語を選ぶ。

機械翻訳モデルの推論プロセス



確率分布


″彼の車はとても″ の次に来る単語の候補
0.45
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
私
彼
車
は
の
を
とても
高い
です

探索アルゴリズム

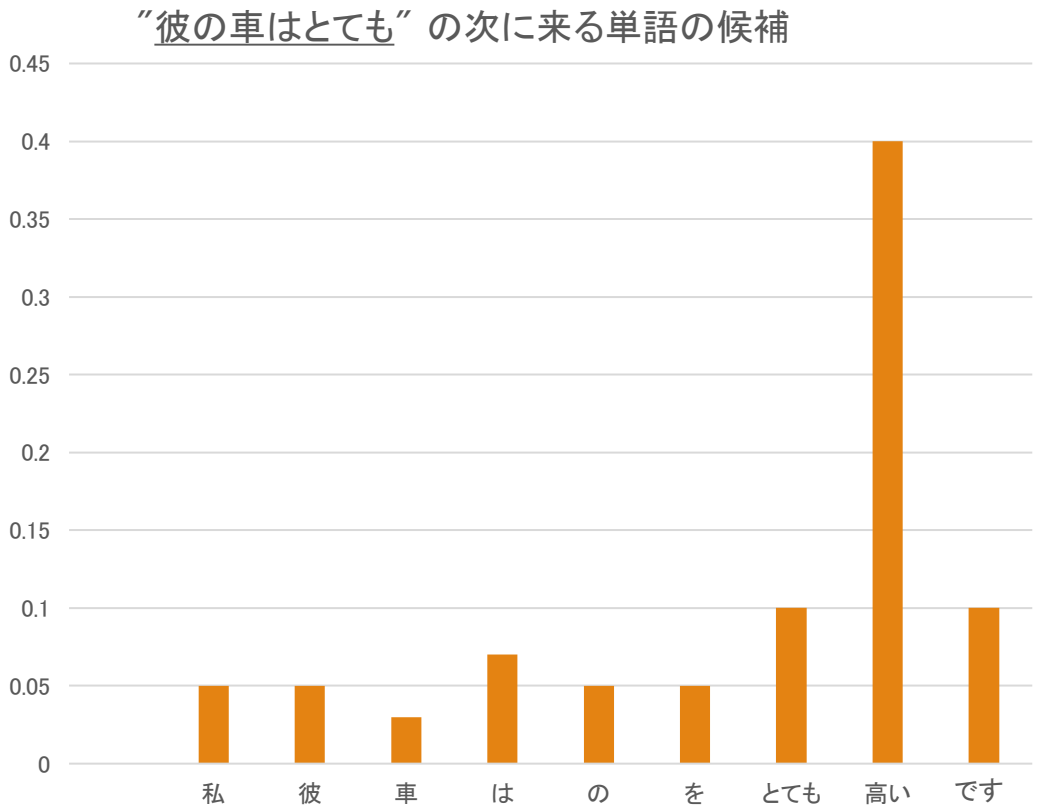
Sampling 系

- (自然度 => low, 表現力 => high)

Beam Search 系

- (自然度 => high, 表現力 => low)

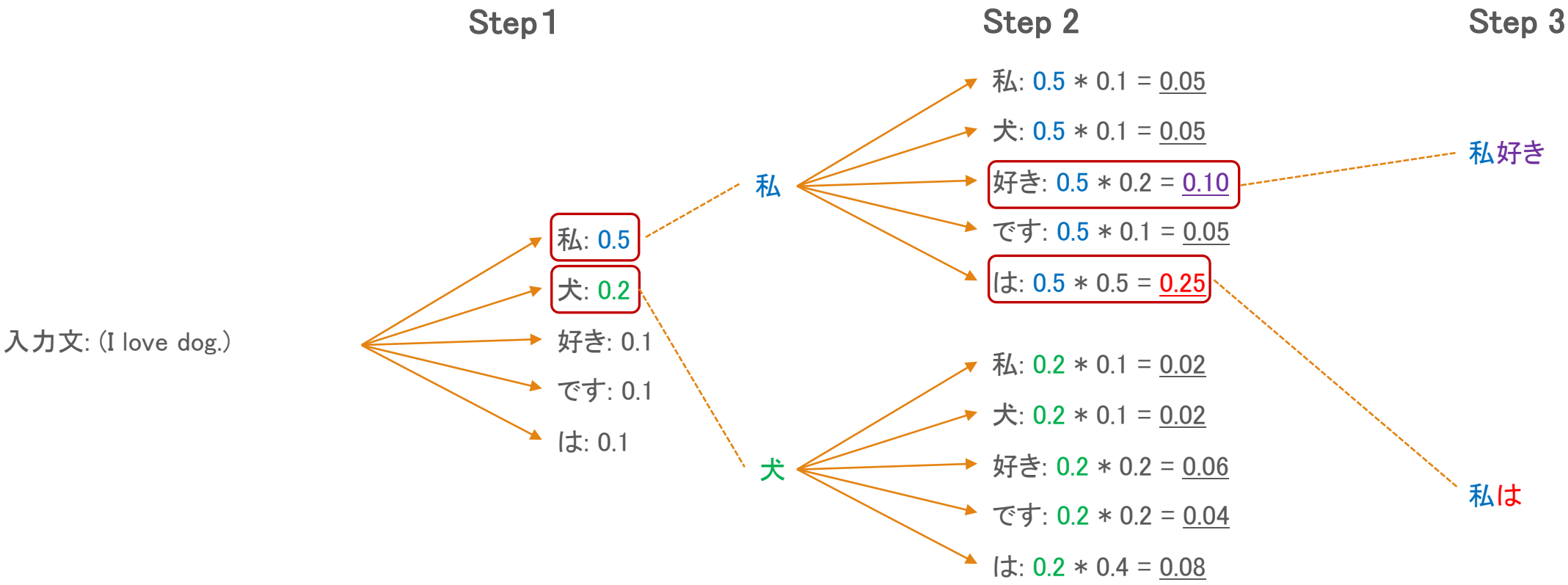
Sampling とは？



Samplingでは、各単語は割り振られた確率分だけ、次の単語として選ばれる可能性がある。

- “高い”は、0.4という確率が割り当てられているので、40%の割合で選ばれる。
- “を”は、0.05という確率が割り当てられているので、5%の割合で選ばれる。

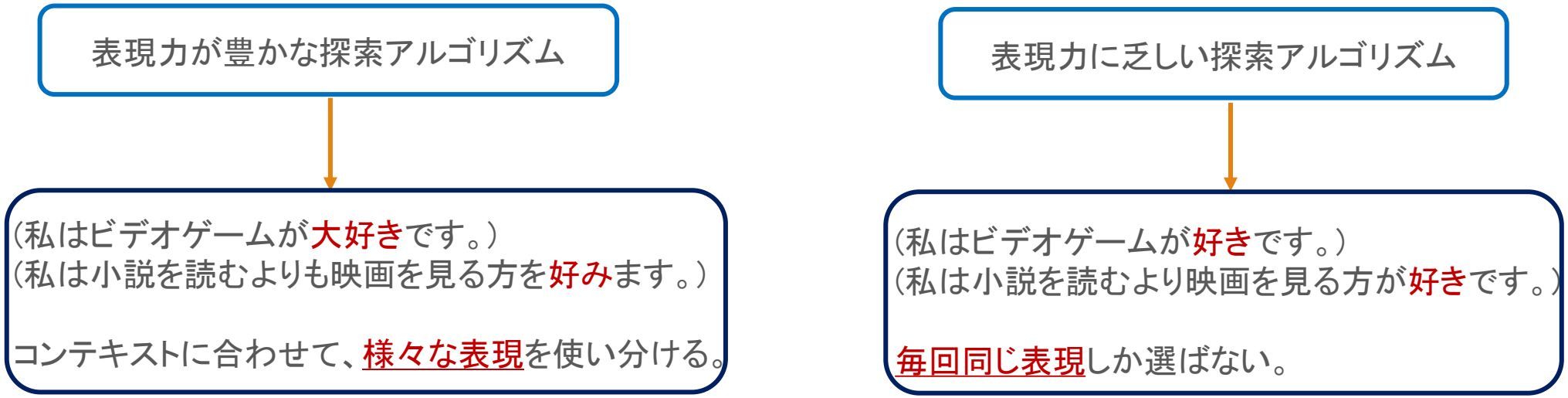
Beam Search （beam幅=2）



表現力が豊かであるとは？

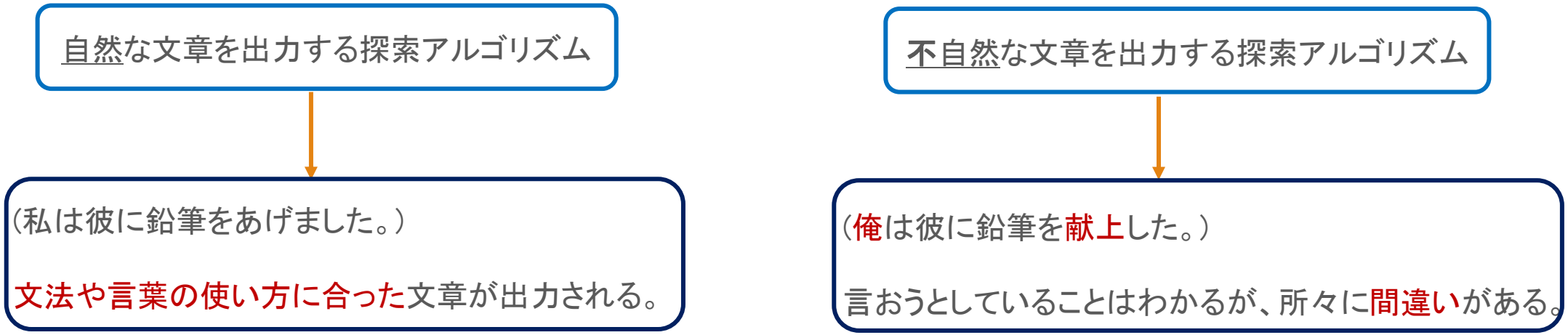
入力文:　(I love playing video games.)

(I prefer watching movie rather than reading novel.)



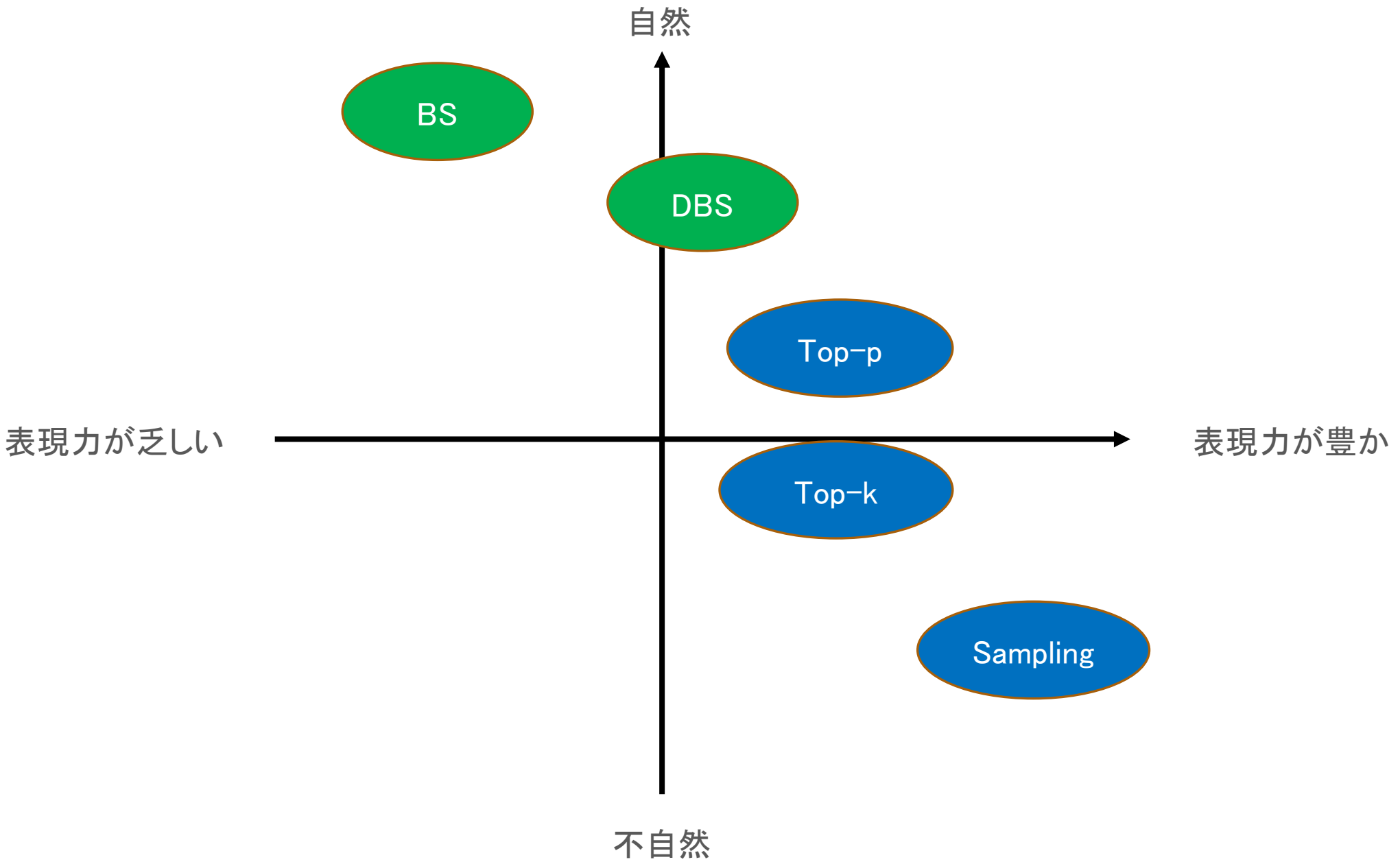
文章が自然であるとは？

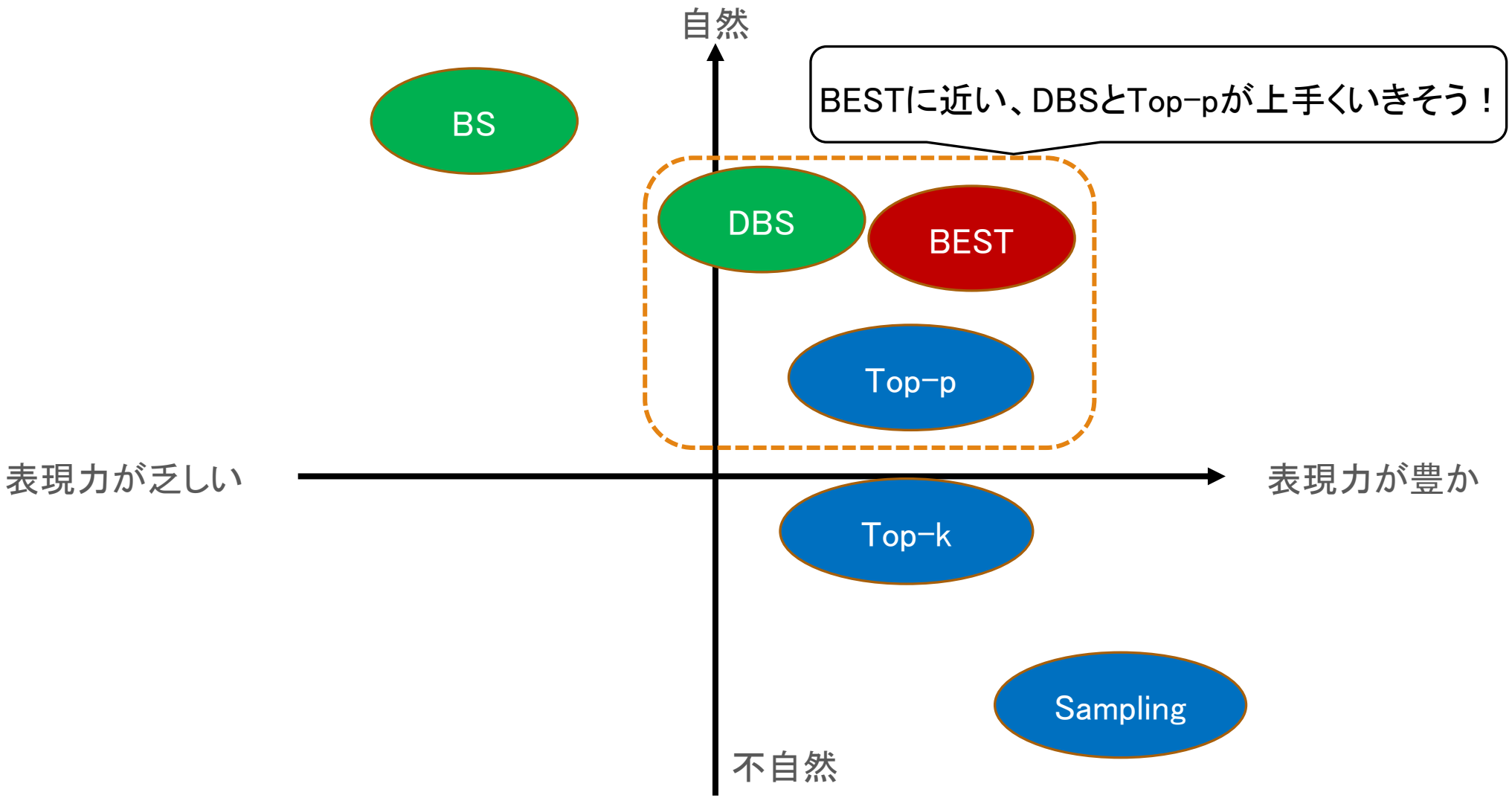
入力文:　(I gave him a pencil.)



実験で用いられた探索アルゴリズム

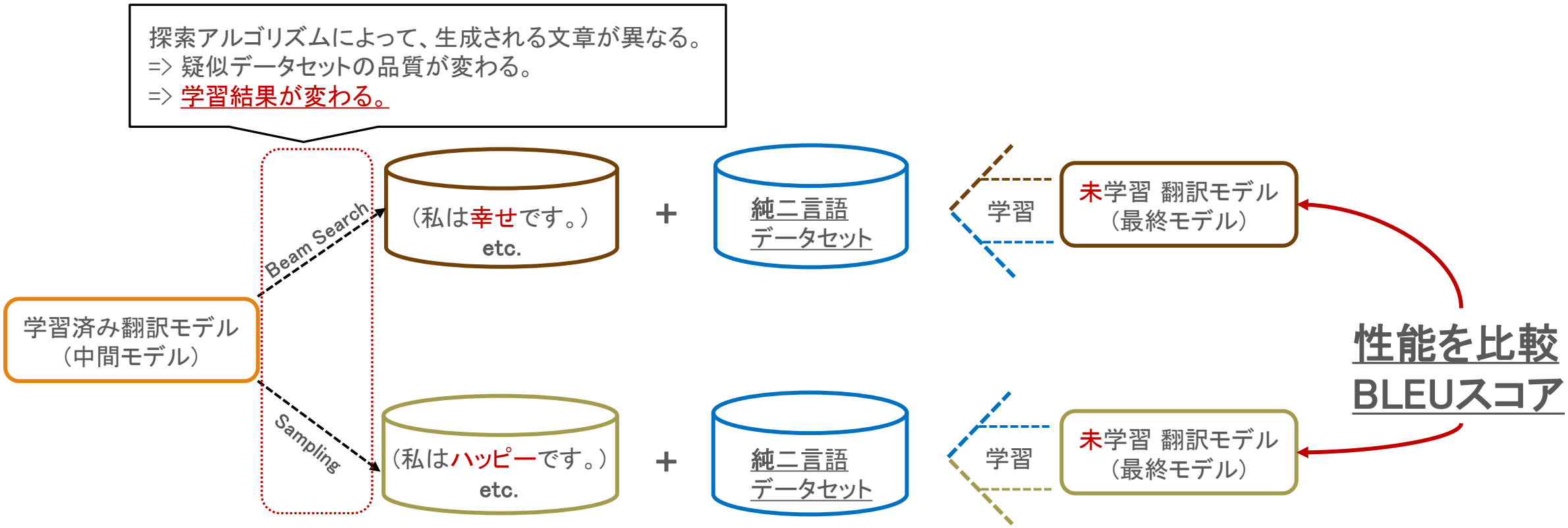
Beam Search 系	Beam Search (BS), Diverse Beam Search (DBS)
Sampling 系	Sampling Top-k Sampling (Top-k) Top-p Sampling (Top-p)


自然
BS
DBS
Top-p
表現力が乏しい
表現力が豊か
Top-k
Sampling
不自然



一般的に翻訳モデルの学習が上手くいく条件
1. 二言語データの表現力が適度に豊かである。
2. 二言語データが適度に自然である。

実験方法


探索アルゴリズムによって、生成される文章が異なる。
=> 疑似データセットの品質が変わる。
=> 学習結果が変わる。
学習済み翻訳モデル
（中間モデル）
Beam Search
Sampling
（私は幸せです。）
etc.
（私はハッピーです。）
etc.
純二言語
データセット
純二言語
データセット
学習
学習
未学習 翻訳モデル
（最終モデル）
未学習 翻訳モデル
（最終モデル）
性能を比較
BLEUスコア

BLEU スコア

- 機械翻訳の評価方法

- 翻訳家の訳文と近ければ近いほどその翻訳モデルの訳文の精度は高い

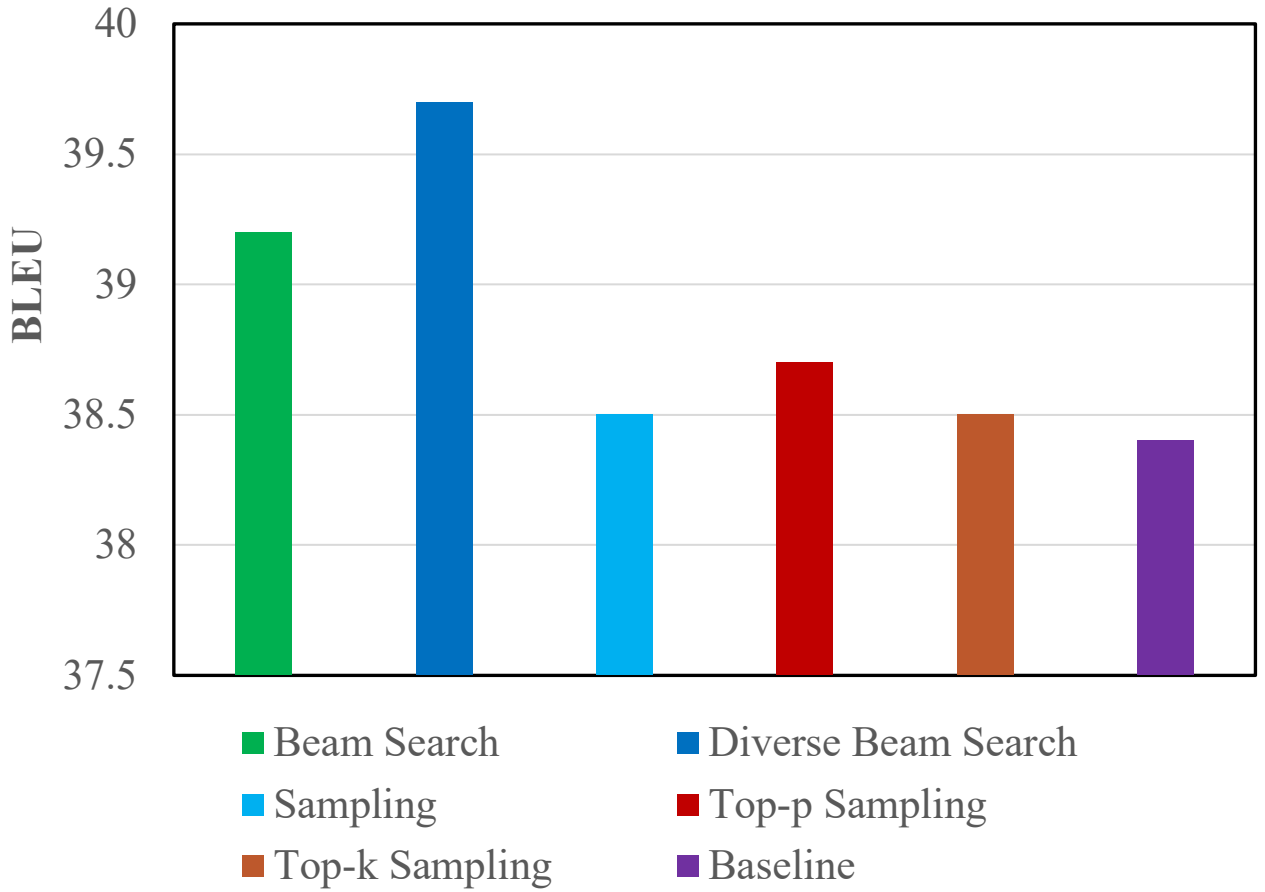
- 0 ～ 100 までの実数値をとる

BLEU スコア	翻訳精度
30 ～ 40	品質の良い訳文
40 ～ 50	高クオリティー

データセット

	データ数
純二言語データセット (英日)	20万
単一言語データセット (英)	80万
疑似二言語データセット (英日)	80万
純二言語データセット (英日) ＋ 疑似二言語データセット (英日)	100万

実験結果



1.3 BLEU スコア 性能向上 (DBS vs Baseline)

0.8 BLEU スコア 性能向上 (BS vs Baseline)

小さな性能向上 (Sampling 系 vs Baseline)

Baselineモデル
- 疑似二言語データを使わずに、
  純二言語データのみを使用

実験結果

- Sampling 系 ＜ Beam Search 系 (データセット規模 100万以下)

- データセットの規模が小さい(100万件以下)の場合には、
  DBSが最も有効である。

Reference

[1] Edunov, Sergey, Myle Ott, Michael Auli, and David Grangier. "**Understanding back-translation at scale**"

[2] Sennrich, Rico, Barry Haddow, and Alexandra Birch. "**Improving neural machine translation models with monolingual data**"

[3] Holtzman, Ari, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. "**The curious case of neural text degeneration.**"

[4] Vijayakumar, Ashwin K., Michael Cogswell, Ramprasath R. Selvaraju, Qing Sun, Stefan Lee,David Crandall, and Dhruv Batra. "**Diverse beam search: Decoding diverse solutions from neural sequence models.**"