

機械翻訳における訳語一貫性 評価用データセットの構築

阿部 香央莉¹, 鈴木 潤^{1, 2, 3}, 永田 昌明³, 乾 健太郎^{1, 2}

乾・鈴木研究室
東北大学大学院 情報科学研究科

1. 東北大学
2. 理化学研究所 AIP センター
3. NTT コミュニケーション科学基礎研究所

背景

- 一文単位の機械翻訳精度はコミュニケーション補助ができる程度まで到達した
- 次の目標：ビジネスに使える翻訳
 - 文脈や状況を考慮した忠実な翻訳

コミュニケーションの補助



日常会話



賢い電子辞書
・フレーズ集

ビジネスに使える翻訳



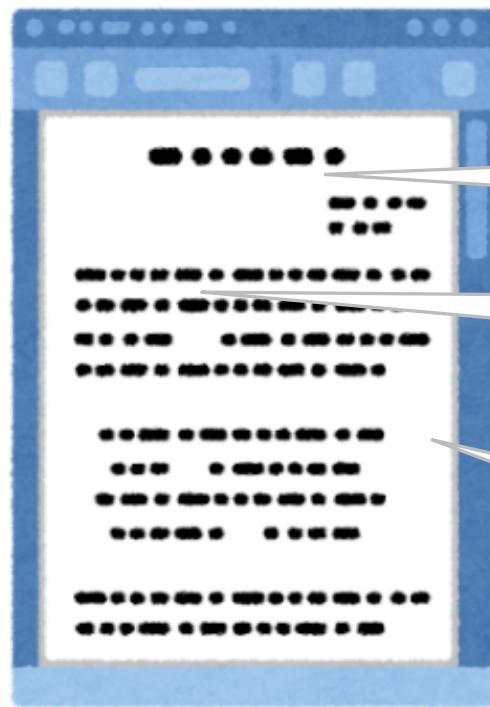
会議通訳



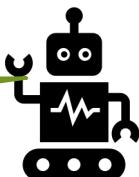
文書翻訳

課題：「訳語一貫性」を担保する翻訳

- ビジネスで使える文書翻訳を考えた際、訳語に一貫性を持たせたい対象が存在



翻訳します



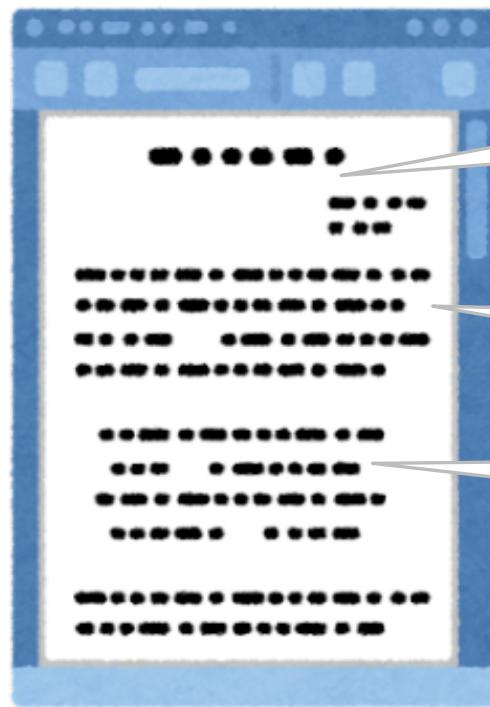
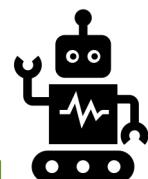
Title: “Our Company’s New Product:
○○”

In our company, we announced the
new product ○○ ...

Strategy of Our company:

課題：「訳語一貫性」を担保する翻訳

- ビジネスで使える文書翻訳を考えた際、訳語に一貫性を持たせたい対象が存在
- しかし、現状の翻訳は一貫性を担保しない



タイトル：「**弊社**の新商品
〇〇について」

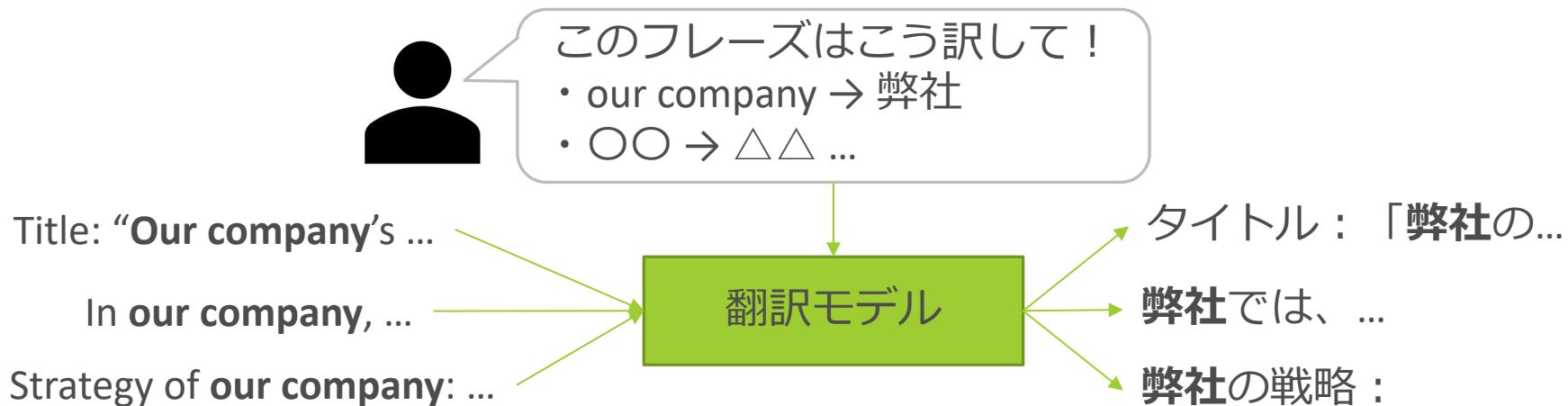
我が社 **弊社**では、この度新
商品の〇〇を発表した

当社 **弊社**の戦略：～



本研究のアプローチ

- 大目的：「訳語一貫性」を実現したニューラルネット機械翻訳



- これを実現するために、まずは以下を整備
 1. 訳語一貫性のある評価用データセット
 2. 訳語一貫性を適切に評価する自動評価指標

既存のデータセットの問題点

- 既存のデータセットにおける参照訳は、
訳語一貫性を評価するのに不適切
- データセットは、一般的に様々なドメイン
 - ・用途で記された文章を混合して作られる
→ 同じ概念を指すものでも、訳語に揺れが存在
- 例：KFTT¹（Wikipediaから京都に関する記事を抽出）
 - “幕府”→ (“Bakufu”, “Shogunate”, “the samurai government” ...)

Tokugawa shogunate ([Edo bakufu](#)からのリダイレクト)

The Tokugawa Shogunate, also known as the Tokugawa **Bakufu** (徳川幕府) and the **Edo Bakufu** (江戸幕府), was the last feudal Japanese military **Edo** period

wars", and popular enjoyment of arts and culture. The **shogunate** was officially established in **Edo** on March 24, 1603, by Tokugawa Ieyasu. The

¹ 京都フリー翻訳タスク : <http://www.phontron.com/kftt/index-ja.html>

1. 訳語一貫性のあるデータセット

「訳語一貫性」評価用データセットの要件

ユーザ



太字斜体の候補に
訳語を統一したい！

一貫性訳語候補リスト

翻訳元	翻訳先
源氏	<i>Genji</i> , Minamoto clan
宇多 源氏	<i>Uda-Genji</i>
源	<i>Minamoto no</i> , Minamoto
宇多	<i>Uda</i> , UDA

元のデータセット

源氏 は ...
宇多 源氏 の
...

Minamoto clan is ...
... of Uda-Genji ...
...

- 日英（KFTT）・英日翻訳（ASPEC）で、既存のデータセットを改良することを考える（例はKFTT）

1. 訳語一貫性のあるデータセット

「訳語一貫性」評価用データセットの要件

ユーザ



太字斜体の候補に
訳語を統一したい！

一貫性訳語候補リスト

翻訳元	翻訳先
源氏	<i>Genji</i> , Minamoto clan
宇多 源氏	<i>Uda-Genji</i>
源	<i>Minamoto no</i> , Minamoto
宇多	<i>Uda</i> , UDA

改良したデータセット

源氏 は … 。
宇多 源氏 の
… 。

{Genji/Minamoto clan} is
… .
… of {Uda-Genji} ...

注：“Uda-Minamoto clan”
とはならない

改良したデータセット：

各文で、対象の訳語部分に対し、文脈上適切な
訳語候補を最終的に人手で判断し付与

1. 訳語一貫性のあるデータセット

「訳語一貫性」評価用データセットの要件

ユーザ



太字斜体の候補に
訳語を統一したい！

一貫性訳語候補リスト

翻訳元	翻訳先
源氏	<i>Genji</i> , Minamoto clan
宇多 源氏	<i>Uda-Genji</i>
源	<i>Minamoto no</i> , Minamoto
宇多	<i>Uda</i> , UDA

改良したデータセット

源氏は…。
宇多 源氏の
…。

{Genji/Minamoto clan} is
... .
... of {Uda-Genji} ...

指定した候補に従い
「正解」の文を作成

実験で用いるデータセット

源氏は…。
宇多 源氏の ...。

Genji is
... of **Uda-Genji**
family

入力文

参照訳

実験で用いるデータセット：

ユーザが指定した候補を、改良したデータセットに付与された訳語候補から選び、参照訳とする

既存の評価指標の問題点

- 機械翻訳で一般的に用いられる評価指標
 - BLEU : 参照訳とのN-gram一致を測る
 - METEOR : 参照訳と近い意味を持つ単語を許容
- これらの評価指標は、
「特定のフレーズを常に同じ訳語に翻訳できて
いるか」ということを厳密に評価できない
→ 訳語一貫性に厳密に着目する評価指標を
考える

評価指標: TERを用いたF値

- TER (Translation error rate)を測る際の
参照訳と出力文のアライメントを利用して、
一貫性訳語に対しF値を測定

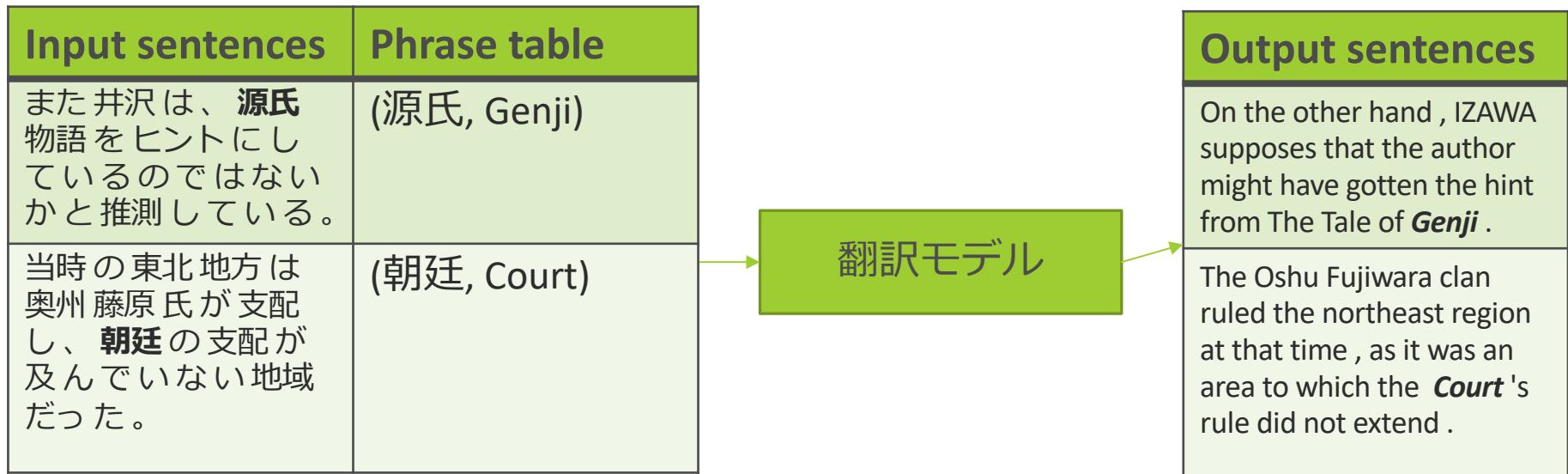
Reference	it fell precisely on the 100th day after the death of takauji ashikaga
M1 output	it was the 100th day after takauji ashikaga 's death .
Reference	it fell precisely on the 100th day after the death of takauji ashikaga
M2 output	ashikaga was the 100th day after takauji ashikaga 's death

- 「**望んだ位置に正しく訳語が生成されたか**」を
より厳密に評価

モデル: 制約付きデコーディング

[Hokamp and Liu (2017)]

- Grid-beam searchを用いたデコーディング手法
- 各文のデコーディング時に、1フレーズ制約を指定可能



[Hokamp and Liu (2017)] : Lexically con- strained decoding for sequence generation using grid beam search. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, pages 1535–1546.

実験

- 翻訳モデル（2種類）

- ベースライン: RNN encoder-decoder

- 制約付き: ベースライン + 制約付きデコーディング

- 一貫性訳語リストに基づき、リスト内のフレーズを含む文には必ず制約を付加

- 使用したデータセット（2種類）

- KFTT（日英）, ASPEC（英日）

- 自動評価手法（3種類）

- BLEU

- 一貫性訳語の出現率（省略）

- TERのアライメントを用いた一貫性訳語のF値

実験結果：定量評価

	BLEU		一貫性訳語のF値	
	ベースライン	制約付き	ベースライン	制約付き
KFTT	12.88	12.91	61.17	86.27
ASPEC	35.71	35.36	71.31	77.64

○KFTT (日→英)

○BLEU, F値 : ベースライン < 制約付き

○ASPEC (英→日)

○BLEU : 制約付き < ベースライン

○F値 : ベースライン < 制約付き

一貫性訳語は出力したもの
の、ベースラインが持つ
翻訳精度を損なう結果に

実験結果：定性評価

一貫性訳語：(approach, 考え方)	
入力文	The details in the establishment of "Individual horticultural products section" such as progress and approach are explained.
参照文	また、設置に当たっての進捗状況及び 考え方 など「個性園芸室」の設置を解説した。
ベースライン	「個別園芸製品部門」の確立の経緯を説明した。
制約付き	その経緯と 考え方 を解説した。

○制約を与えた訳語（“考え方”）の出力はできたが
その部分以外のフレーズが欠けている

結論

- 機械翻訳における、訳語一貫性を評価するためのデータセットを構築した
- 構築したデータセットによって既存の翻訳手法の訳語一貫性を評価し、現状の翻訳器の問題点を明示的に表した
- 今後の目標：訳語の一貫性を実現するための自動翻訳の方法論を考案する

Appendix

訳語の一貫性問題が生じやすい 翻訳方向、ドメイン

○ KFTT (日→英)

- Wikipediaの「京都」に関する記事の収集
- 日本独自の文化や概念を表す語が多く出現
→他の言語に訳す際、様々な翻訳の仕方がある
(e.g., 日本語読みに忠実なローマ字表記、
その概念の「意味」を英語で表した表記...)

○ ASPEC (英→日)

- 科学論文のデータセット
- 日常的意味と異なる意味を持つ専門用語が多く出現
- また、翻訳者や時代により、同じ概念を表す
専門用語でも表現が異なる場合もある

データセット作成方法(1)

- 1. 既存のデータセットのうち、訳語揺れがある（ぶれている）対象を見つける
- 具体的には、単語アライメント (KFTT: 人手アナリシジョン, ASPEC: GIZA++で自動) を利用し、翻訳元のある単語に対し翻訳先で複数の訳語候補があるものを取得
- 例:
 - ("examination", {"検討", "実験", "検査"})
 - ("源氏", {"Genji", "Minamoto", "clan"})

データセット作成方法(2)

- 2. 訳語揺れがある単語に対し、フレーズレベルに直せるものは直す
 - 例: (“源氏”, {"Genji", "Minamoto", "clan"})
→ (“源氏”, {"Genji, “Minamoto clan”})
- 3. 2で得られた単語/フレーズペアが含まれる文それぞれについて、正解の訳語を付加
 - 「正解かどうか」は最終的に人手で判断

実際のデータセット

OKFTT

(もっと見やすくする予定)

d	ja_line	en_line	source_word1	target_word1	source_word2	target_word2	source_word3	target_word3
1	浄土真宗（じょうどしんしゅう、shin - buddhism , pure land buddhism）は、日本の仏教の宗派のひとつで、鎌倉時代初期、法然の弟子・親鸞が、法然の教え（浄土宗）を継承発展させた教団である。	Jodo Shinshu (Shin-Buddhism / True Pure Land Sect) is one of the sects of Japanese Buddhism , and a religious community that Shinran , an apprentice of Honen , succeeded and which developed Honen 's doctrine (Jodo Shu / Pure Land Buddhism) in the early Kamakura period .	弟子	apprentice, disciple	教え	doctrine, teachings	時代	era,period
2	宗派名の成り立ちの歴史的経緯から、現在、同宗に属する宗派の多くが宗旨名としては真宗を名乗る。	Due to the historical background of the origin of the sect name , most sects belonging to this religion call themselves Shinshu as a sect name .	真宗	Shinshu				
3	過去には一向宗、門徒宗とも通称された。	It used to also be called Ikko Shu and Monto Shu .	門徒	Monto	一向宗	Ikko Shu		

OASPEC

	en_line	ja_line	source_word1	target_word1	source_word2	target_word2
1	Details of dose rate of "Fugen Power Plant" can be calculated by using DERS software .	D E R S ソフトウェアを用いて「ふげん発電所」の線量率を詳細に計算できる。	software	ソフトウェア,ソフト	dose	線量
2	The changes of conditions for computation of the dose rate , namely , object changes , are inputted to DERS from VRdose software by using a special marker .	線量率を計算する際の状況の変化、すなわちオブジェクトの変化を VR d o s e ソフトウェアが D E R S に特別なマークにより知らせる。	software	ソフトウェア,ソフト	dose	線量
3	Responding to these changes DERS can compute new dose rate .	D E R S はこれらの変化に対応して新たな線量率を計算できる。	dose	線量		

一貫性訳語のF値

参照文の各単語	出力文の各単語	一貫性訳語に対する判定
Ashikaga	it	アライメント不一致
...	...	
Takauji	Takauji	
Ashikaga	Ashikaga	アライメント一致

○Precision

○(アライメント一致数)/(Reference中の一貫性訳語出現数)

○Recall

○(アライメント一致数)/(Output中の一貫性訳語出現数)

○F値

○ $2 * \text{Precision} * \text{Recall} / (\text{Precision} + \text{Recall})$

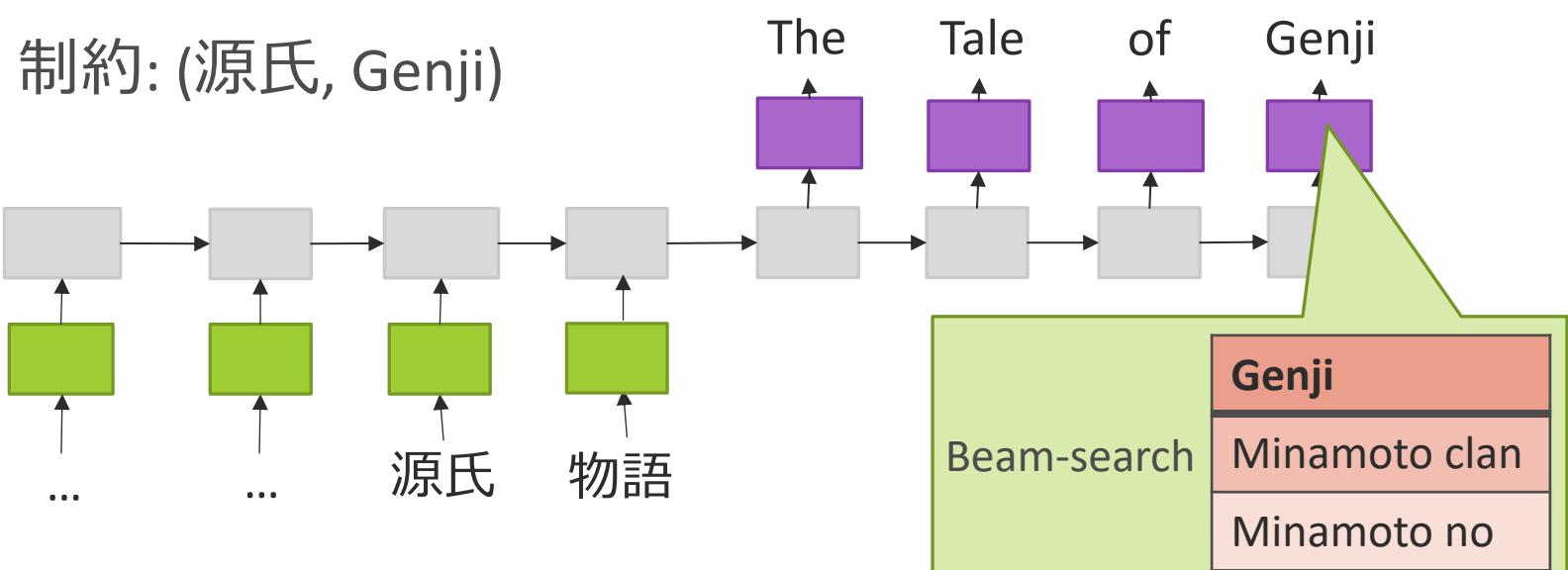
データセットの統計

	KFTT (日英)	ASPEC (英日)
一貫性訳語の種類数	84	196
A: 一貫性訳語を含む文数	457	407
B: 元データセット文数	1245	1812
A/Bの割合	36.7%	22.5%
訳語候補数の平均	2.27	2.77
訳語候補数の最大値	12	10

モデル: 制約付きデコーディング

[Hokamp and Liu (2017)]

- デコーディング時に指定した制約対象が出現した場合、beam searchの最上位にその訳語を入れる
→ 必ず "Genji" が出力される (図をより精緻なものに変える予定)



[Hokamp and Liu (2017)] : Lexically con- strained decoding for sequence generation using grid beam search. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, pages 1535–1546.

実験結果：定性評価(2)

一貫性訳語：(above, 上記)	
入力文	Two cases with the above mentioned disease in 54-year-old and 84-year-old women were reported .
参照文	54歳および84歳女性の 上記 の症例を報告した。
ベース ライン	54歳と84歳女性の 標記 疾患の症例を報告した。
制約付き	54歳と84歳女性の 上記 疾患の症例を報告した。

- うまく制約が働き、望んだ通りに「above」を「上記」と訳せている

実験結果：定性評価(3)

一貫性訳語：(本尊, Honzon)	
入力文	本尊は阿弥陀如来一仏である。
参照文	The Honzon is only Amida Nyorai .
ベース ライン	The principal image is Amida Nyorai (Amitabha Tathagata) .
制約付き	Honzon (principal image of Buddha) is Amida Nyorai (Amitabha Tathagata) .

- うまく制約が働き、指定した通りに”本尊”に
対応する”Honzon”という単語を出力できている

Trash

背景：文単位のNMTの精度

- ニューラルネットを用いた文単位の機械翻訳は、実用レベルのものになってきた

Google翻訳、最近精度良いよねみたいな記事のスクショ
or 文単位翻訳制度に歓喜してる人の図



課題：自然な文章翻訳

- 文章単位の翻訳を考えた時、現状のNMTは「文章全体として自然な翻訳」の要件を満たしていない
- 文章全体として自然な翻訳とは？
 - 文脈に依存した訳語選択
 - 代名詞/ゼロ照応補完
 - 訳語の一貫性

課題：自然な文章翻訳

- 文章単位の翻訳を考えた時、現状のNMTは「文章全体として自然な翻訳」の要件を満たしていない
- 文章全体として自然な翻訳とは？
 - 文脈に依存した訳語選択
 - 代名詞/ゼロ照応補完
 - 訳語の一貫性**
- 本研究では、**訳語の一貫性**に焦点を当てる

本研究のアプローチ

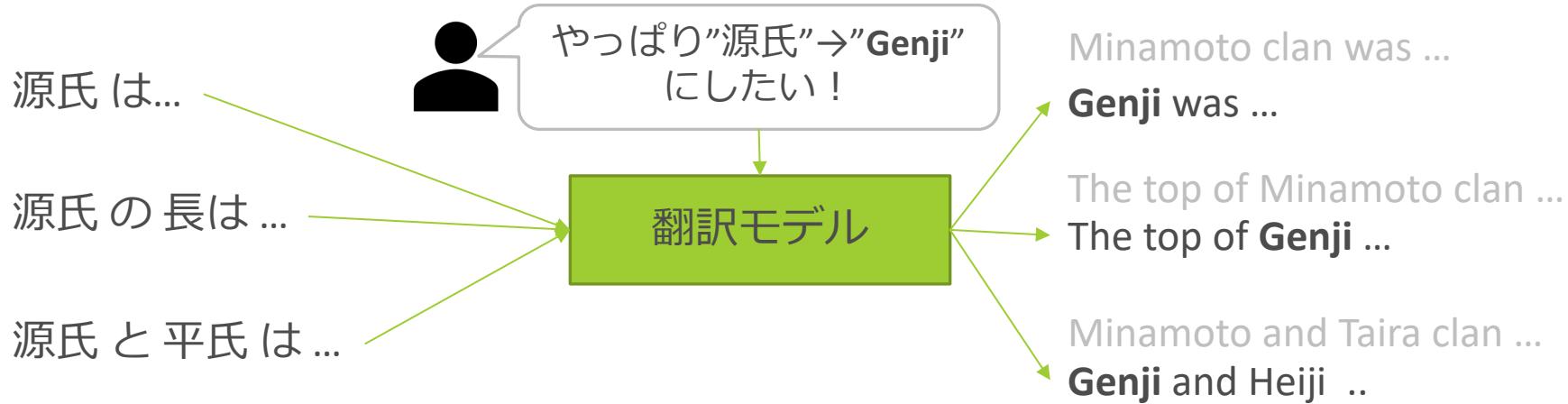
- 大目的：「訳語一貫性」を実現したニューラルネット機械翻訳



- これを実現するために、まずは以下を整備
 - 「訳語一貫性」評価用データセット
 - 「訳語一貫性」を適切に評価する自動評価指標

本研究のアプローチ

- 大目的：「訳語一貫性」を実現したニューラルネット機械翻訳



- これを実現するために、まずは以下を整備
 - 「訳語一貫性」評価用データセット
 - 「訳語一貫性」を適切に評価する自動評価指標