言語処理学会 年次大会 B2:機械翻訳(2)

ニューラルネットを用いた多方言の翻訳と類型分析

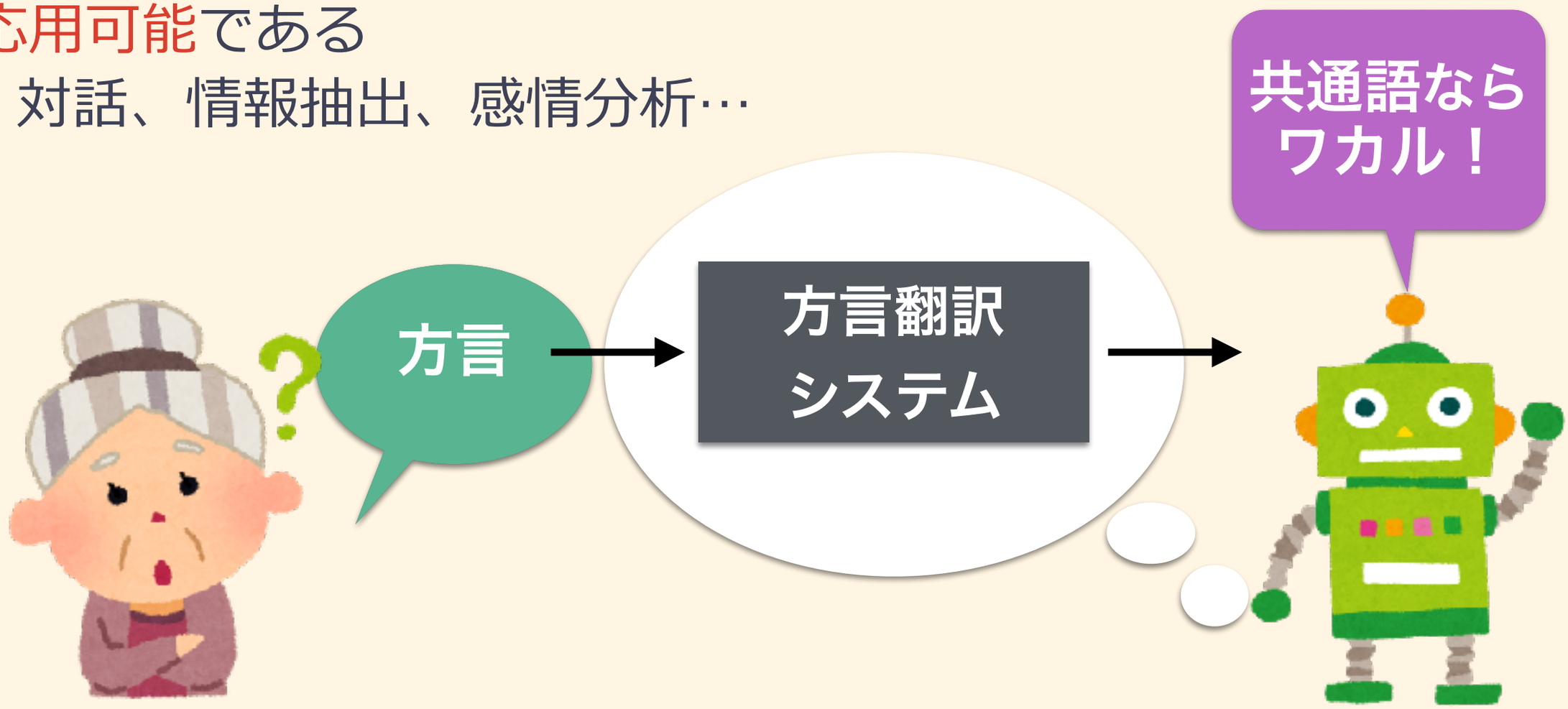
阿部 香央莉、松林 優一郎、岡崎 直観、乾 健太郎
東北大学工学部 乾研究室

本研究（方言研究）の背景

- 近年、音声あるいはテキストで人間と対話するシステムがより身近な存在になってきている
- しかし、それらの多くは**共通語**を前提として作られている
  ＝**方言で話しかけた場合の精度が低い**
- 介護ロボットなどの需要が高まる中で、
  方言を解析できないことは大きな問題となる

本研究の目的：方言→共通語の翻訳

- **方言→共通語の翻訳**ができれば、
  共通語を前提として作成された様々なシステムに
  応用可能である
  - 対話、情報抽出、感情分析…



- 人間でも訛りが強い地域の方言は理解するのが難しい
  →ロボットが理解できれば、コミュニケーションの手助けが可能

方言処理の先行研究

- **山形方言 ⇄ 共通語の統計的機械翻訳**[柴田, 2013]
  独自に作成した対訳コーパス＋ルールベースで作成した
  擬似対訳コーパスで、統計的機械翻訳(SMT)を行う
  <問題点>
  他の方言に対応するには、別の方言翻訳器を作成することになる
  → 各方言に対し十分なコーパスを用意するコストが高い

- **事前学習による標準語から関西弁へのニューラル翻訳**[長谷川, 2017]
  Twitterの関西弁を収集して得た単一言語コーパスで事前学習→
  クラウドソーシングで作成した対訳コーパスで再学習
  <問題点>
  小規模なコーパスでは、SMTのほうが一般的に精度が良いとされているが、現時点でSMTとの比較検証は行われていない

本研究のモチベーション

多言語翻訳の発想を方言に適用する

複数の言語のコーパスを**まとめて**ニューラルネットで学習

→「言語」に共通する普遍的な特徴を学び、

小規模なコーパスの言語に対しても翻訳が行えるようになる

これを**方言**に↓適用すれば…

- **小規模でも複数の方言の対訳コーパス**が存在すれば、多言語翻訳と同様に翻訳を行うことができるのでは？
- **ニューラルネットの機構を活かした多言語NMT**で、SMTの翻訳精度を超えられるのでは？

方言翻訳のための言語資源

- 国立国語研究所発行「全国方言談話データベース」
  - **30分の対話**をテキストに書き起こしたデータ**×48地点**分(※)
  （※47都道府県、沖縄は本土・離島の2地点）

- 発話の内容（方言）：表音的カタカナ表記
  その共通語訳：漢字かな混じり表記　で記録

例）青森県の対訳文

方言	オラホノ　マズサ　キスタガステー
	↓
共通語訳	私の　町に　来ましたかね

注：コーパスの空白区切りは便宜的なもの。これ以降「文節」と呼ぶ

方言翻訳のための言語資源

- このままだと、**方言→共通語**というタスクの他に
  カタカナ→漢字というタスクも増えてしまい、難しい

（※47都道府県、沖縄は本土と離島の2地点）

- 発話の内容（方言）：表音的カタカナ表記
  その共通語訳：漢字かな混じり表記　で記

そのため…

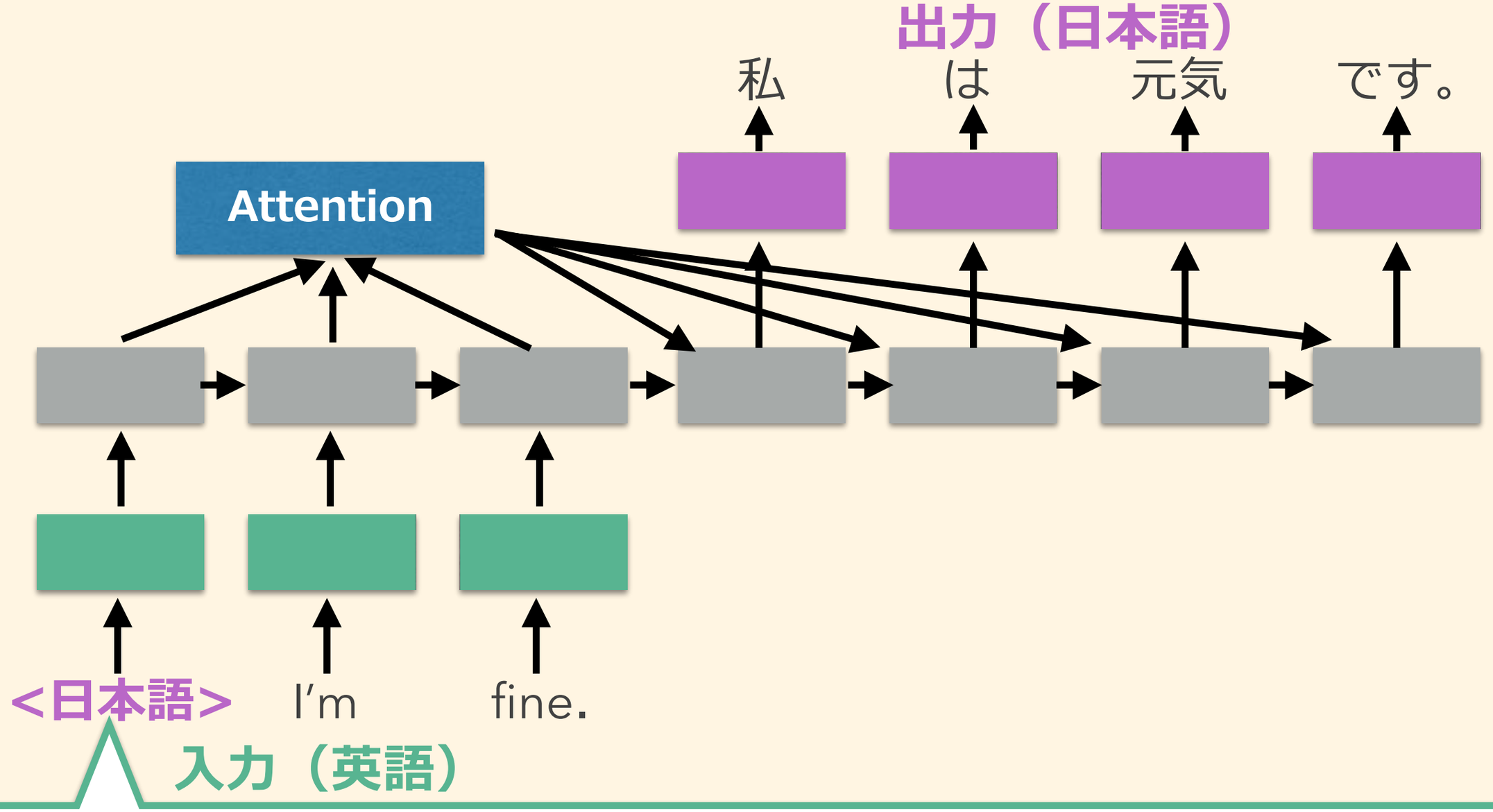
例）青森県の対訳文

方言	おらほの　まずさ　きすたがすてー
	↑どちらもひらがなにすることで対応↓
共通語訳	わたしの　まちに　きましたかね

注：コーパスの空白区切りは便宜的なもの。これ以降「文節」と呼ぶ

既存手法：多言語ニューラルネット翻訳器

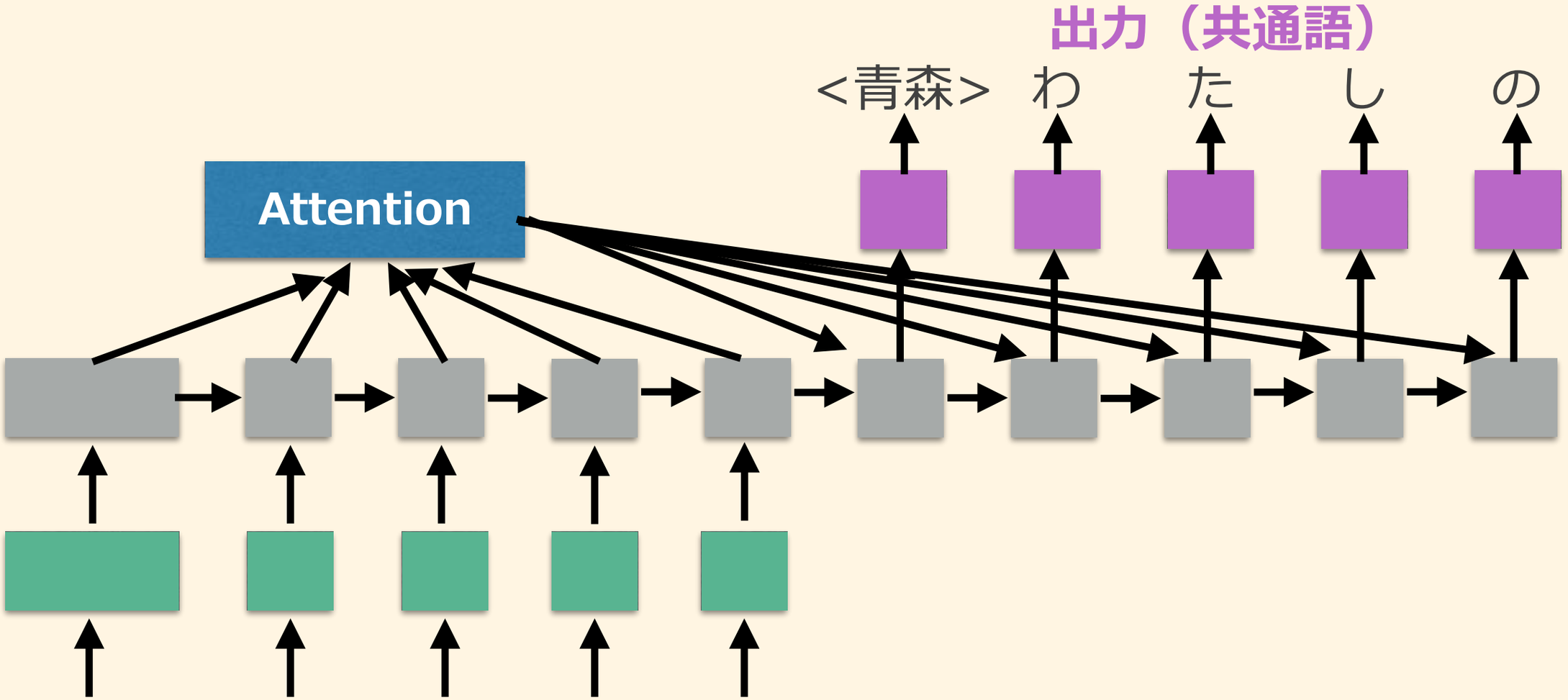
- Google’s Multilingual Translation System[Johnson+, 2016]



翻訳先の言語を表すトークンを入力文の先頭に付加するだけ

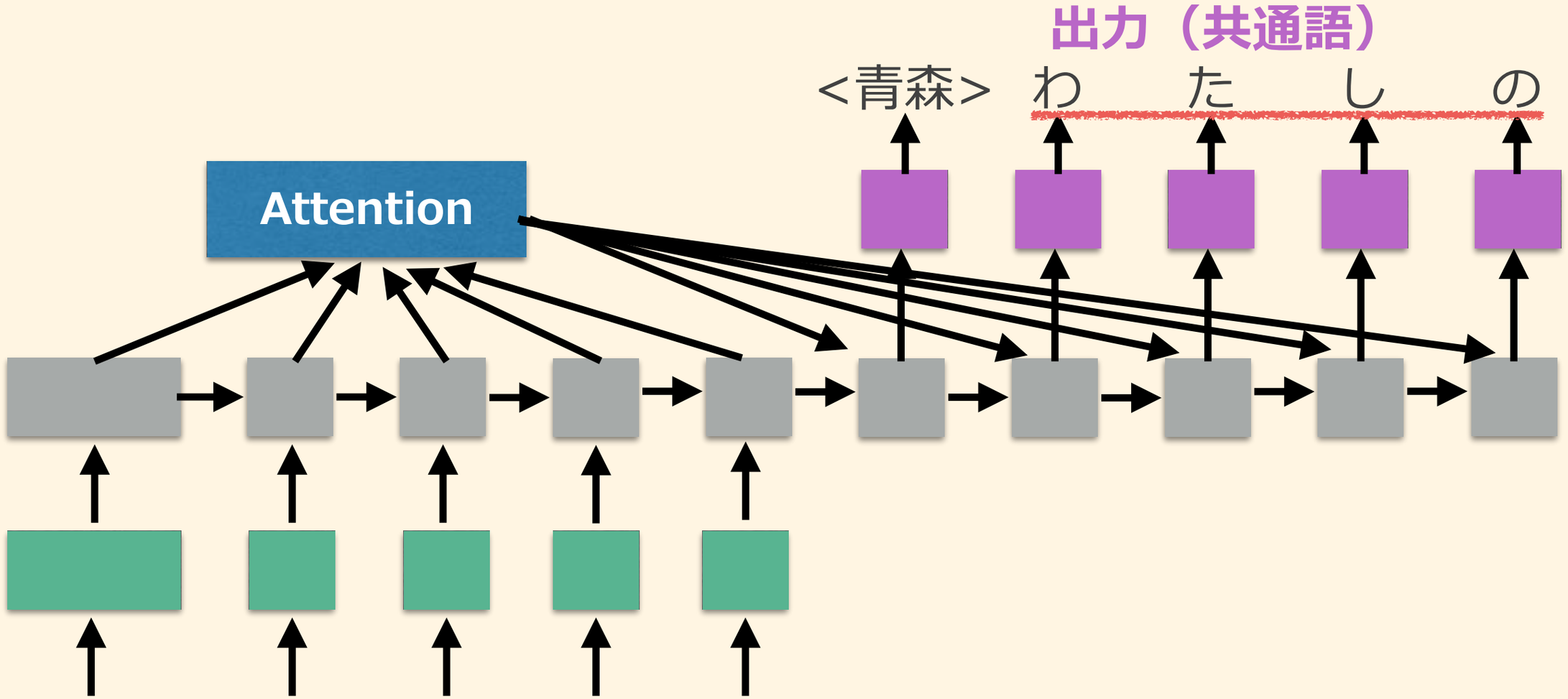
- 言語モデルを共有しながら、翻訳先の言語の特徴を学習でき、多言語NMTを実現する

提案手法：多方言ニューラルネット翻訳器



① **翻訳元の方言を表すトークン（地域トークン）**を付加する
- 一般的な多言語翻訳より、翻訳元の方言同士が似ているため同じような語彙・文法を共有して学習することができる

提案手法：多方言ニューラルネット翻訳器



② 本研究では翻訳元言語（方言）と翻訳先言語（共通語）の間に
語順変化がほぼないため、**単語→文字、文→文節**として
文節ごとに逐次的に翻訳を行う

実験における仮説

仮説❶ 複数の方言コーパスを一緒に学習することで、
方言間で共通する語彙や文法を手に入れることができ、
多方言NMT > 単一方言NMT, SMTとなる

仮説❷ 多言語翻訳の手法に基づいて、
翻訳する方言の地域を表すトークンを付与することで、
方言ごとの特徴を学ぶことができる

仮説❸ 翻訳元（方言）と翻訳先（共通語）の間で
語順変化がないことから、**文節ごとに逐次的に翻訳**する
ようなアーキテクチャでも十分翻訳が可能である

実験設定

- **コーパス**
  全国方言談話データベース「日本のふるさとことば集成」（国研）
  全48地域の対訳コーパス、34,117文（116,928文節）
  - 8:1:1の割合で分割し、それぞれ訓練・開発・評価セットとした

- **NMT, SMTのアーキテクチャ**
  どちらもオープンソースのものを使用
  - NMT：OpenNMT-py
    (https://github.com/OpenNMT/OpenNMT-py)
  - SMT：Moses(http://www.statmt.org/moses/)

- **パラメータ**
  - デフォルト値を使用

- **評価指標**
  BLEUで評価（文節単位で翻訳した結果を文に統合し、
  文字n-gramのBLEUを測定）

実験結果：定量評価(BLEU)

- 各種モデルの翻訳精度を比較

学習コーパス	MTの種類	(山形弁) BLEU	(全地域) BLEU
全地域	**地域トークンありNMT**	**75.79**	**77.10**
	地域トークンなしNMT	68.70	72.66
山形県のみ	NMT	19.75	
全地域	SMT	60.32	74.74
山形県のみ		60.47	

- 提案手法が最も翻訳精度が高いという結果になった

実験における仮説

仮説❶ 複数の方言コーパスを一緒に学習することで、
方言間で共通する語彙や文法を手に入れることができ、
多方言NMT > 単一方言NMT, SMTとなる

仮説❷ 多言語翻訳の手法に基づいて、
翻訳する言語の地域を表すトークンを付与することで、
方言ごとの特徴を学ぶことができる

S

仮説❸ 翻訳元（方言）と翻訳先（共通語）の間で
語順変化がないことから、**文節ごとに逐次的に翻訳**する
ようなアーキテクチャでも十分翻訳が可能である

実験結果：定量評価(BLEU)

- 各手法における、**山形弁のみの**翻訳精度を比較

学習コーパス	MTの種類	(山形弁) BLEU	(全地域) BLEU
全地域	地域トークンあり**NMT**	**75.79**	77.10
	地域トークンなし**NMT**	**68.70**	72.66
山形県のみ	NMT	19.75	
全地域	**SMT**	60.32	74.74
山形県のみ		**60.47**	

- **複数地域の対訳コーパスで学習したNMT**のほうが、
  単一地域の対訳コーパスのみで学習したSMTよりも、
  翻訳精度が高い

実験結果：定量評価(BLEU)

- 各手法における、**山形弁のみの**翻訳精度を比較

学習コーパス	MTの種類	（山形弁）BLEU	（全地域）BLEU
全地域	地域トークンあり**NMT**	**75.79**	77.10
	地域トークンなし**NMT**	**68.70**	72.66
山形県のみ	**NMT**	**19.75**	
全地域	SMT	60.32	74.74
山形県のみ		60.47	

- **複数地域の対訳コーパスで学習したNMT**のほうが、
  単一地域の対訳コーパスのみで学習したNMTよりも、
  圧倒的に翻訳精度が高い

実験における仮説

仮説❶ 複数の方言コーパスを一緒に学習することで、
方言間で共通する語彙や文法を手に入れることができ、
多方言NMT > 単一方言NMT, SMTとなる

仮説❷ 多言語翻訳の手法に基づいて、
翻訳する言語の地域を表すトークンを付与することで、
方言ごとの特徴を学ぶことができる

仮説❸ 翻訳元（方言）と翻訳先（共通語）の間で
語順変化がないことから、**文節ごとに逐次的に翻訳**する
ようなアーキテクチャでも十分翻訳が可能である

実験結果：定量評価(BLEU)

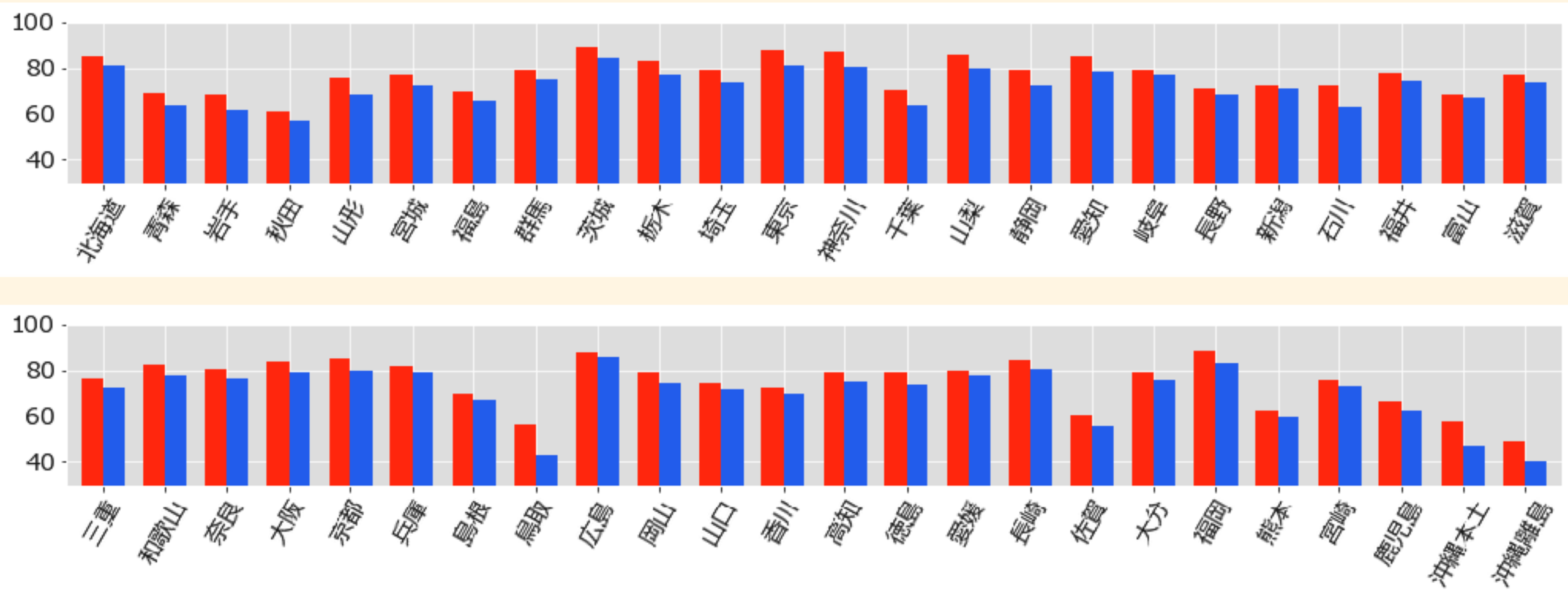
- **地域トークンの有無**に着目して比較

学習コーパス	MTの種類	(山形弁) BLEU	(全地域) BLEU
全地域	**地域トークンあり**NMT	**75.79**	**77.10**
	地域トークンなしNMT	68.70	72.66
山形県のみ	NMT	19.75	
全地域	SMT	60.32	74.74
山形県のみ		60.47	

- **地域トークンを付加して**学習した多方言NMTが**一番精度が良い**

実験結果：定量評価

- BLEUスコアは、全48地域において
  地域トークンあり > **地域トークンなし**
  →地域トークンにより、各方言の特徴を学んだと考えられる



実験における仮説

仮説❶ 複数の方言コーパスを一緒に学習することで、
方言間で共通する語彙や文法を手に入れることができ、
多方言NMT > 単一方言NMT, SMTとなる

仮説❷ 多言語翻訳の手法に基づいて、
翻訳する言語の地域を表すトークンを付与することで、
方言ごとの特徴を学ぶことができる

仮説❸ 翻訳元（方言）と翻訳先（共通語）の間で
語順変化がないことから、**文節ごとに逐次的に翻訳**する
ようなアーキテクチャでも十分翻訳が可能である

実験結果：定量評価(BLEU)

- **入力系列を文あるいは文節**にした場合の翻訳精度比較

学習コーパス	入力系列	MTの種類	BLEU
全地域	**文**	**地域トークンありNMT**	71.64
	文節	**地域トークンありNMT**	**77.10**
		地域トークンなしNMT	72.66

- 地域トークンを付与した状態では、
  入力系列が文節単位である方が精度が良い
  →語順が変化しないため、文節ごとに逐次的に翻訳しても良い

実験結果：定性評価（翻訳例）

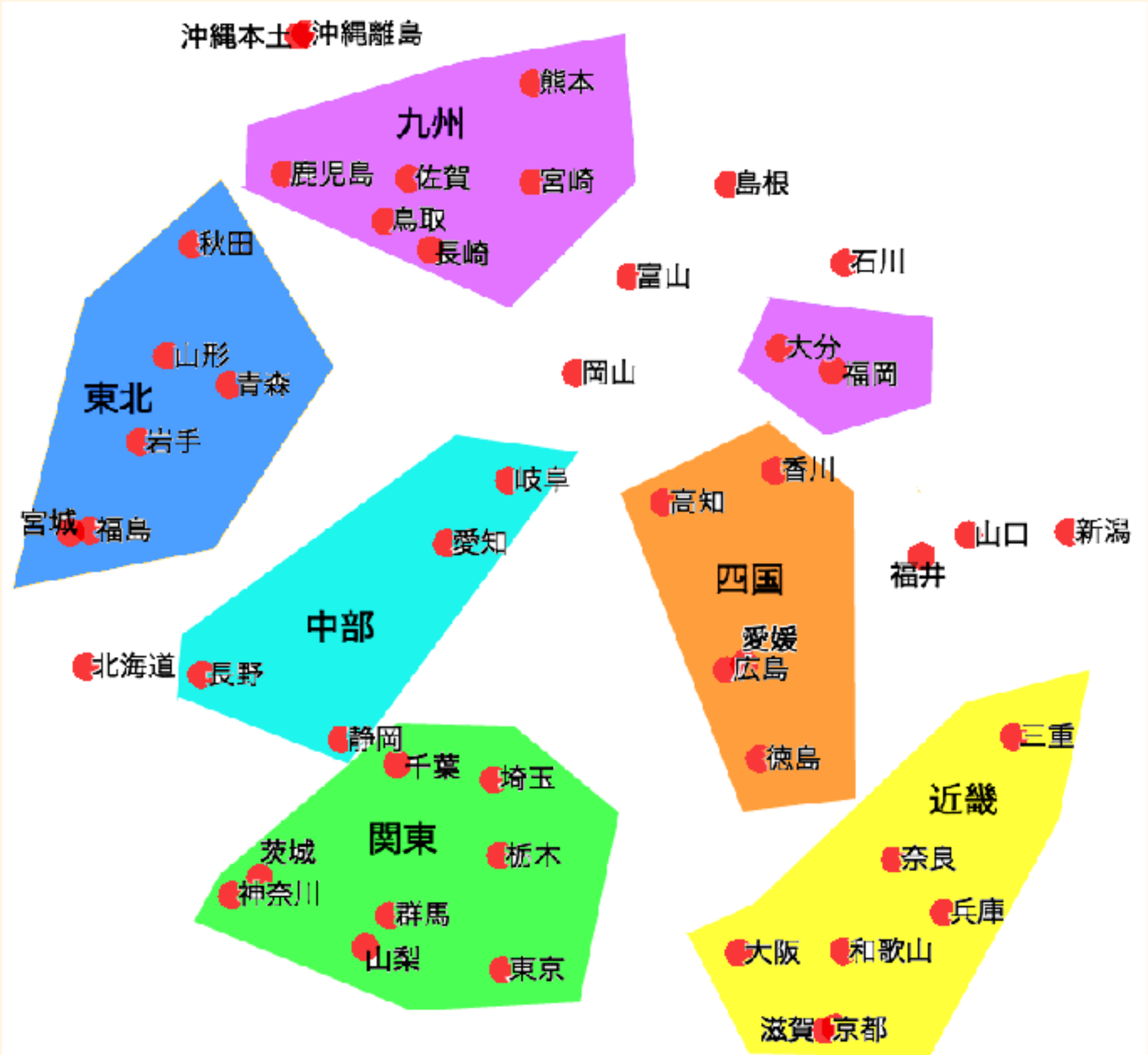
- 翻訳できなかった例

	岩手方言（遠野）
入力 方言文	**ががが**/とった/ときぁ/いま/えってる
参照文	**つま**/もらった/ときは/いま/はなしている
モデル 出力文	**かか**/とった/ときは/いま/いっている

- 「ががが（かか）」→「妻」は学習コーパス中でも現れるが、文節ごとに逐次的に翻訳するため、他の文脈で現れる「ががが」と区別ができず、翻訳できなかったと考えられる
  → **文脈を考慮した訳し分けはできていない**

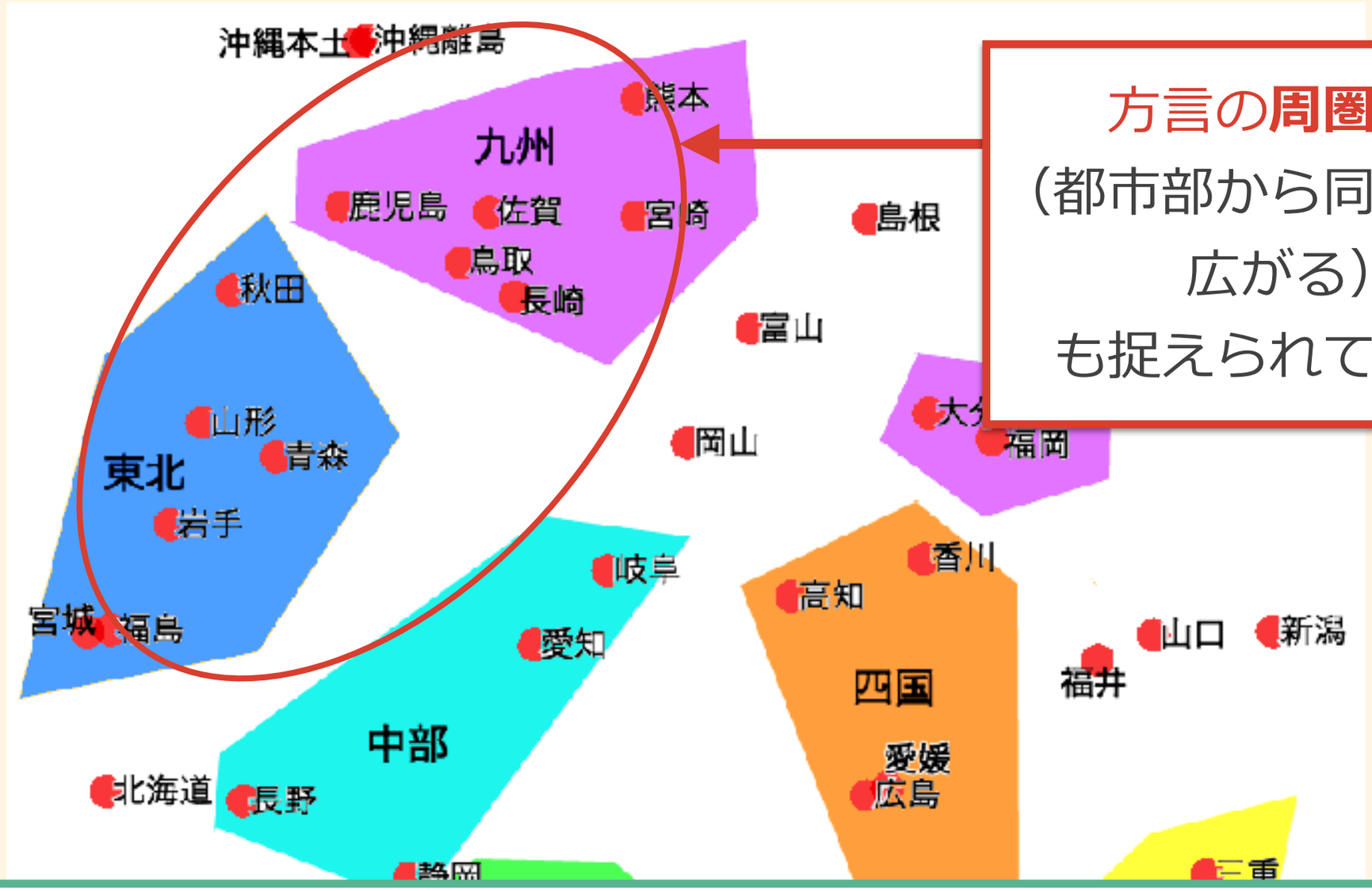
地域トークン埋め込み表現の可視化

- **地域トークン**をt-SNEで可視化すると、地方区分に沿うクラスタを得た



地域トークン埋め込み表現の可視化

- **地域トークン**をt-SNEで可視化すると、地方区分に沿うクラスタを得た



方言の**周圏分布**
（都市部から同心円状に広がる）
も捉えられている？

仮説❷ 多言語翻訳の手法に基づいて、
翻訳する言語の地域を表すトークンを付与することで、
方言ごとの特徴を学ぶことができる

本研究の貢献

- 48地域の方言対訳コーパスを用い、多言語翻訳の手法に基づき、全都道府県に対応する**多方言翻訳器**を初めて作成した
- 複数の方言が似通った特徴を持っていることから、別々の方言で現れる同じような語彙・文法を共有する多方言NMTが単一方言のみ学習させたSMT, NMTよりも精度が良いことを示した
- 入力に付与する地域トークンの埋め込み表現を可視化することで、方言学の知見に合致する**類型分析**が行える可能性を示した

引用

- 柴田直由, 横山昌一, 井上雅史. 統計的手法を用いた双方向方言機械翻訳システム. 言語処理学会第 19 回年次大会 発表論文集, pp. 126–129, 2013.
- 長谷川駿, 田中駿, 山本悠二, 高村大也, 奥村学. 事前学習と汎化タグに おける方言翻訳の性能向上. 情報処理学会研究報告, Vol. 2017-NL-23, No. 12, pp. 3–8, 2017.
- M. Johnson, M. Schuster, Quoc V. Le, M. Krikun, Y. Wu, Z. Chen, N. Thorat, F. Vi´egas, M. Wattenberg, G. Corrado, M. Hughes, and J. Dean. Google’s Multilingual Neural Machine Translation System: Enabling Zero-Shot Translation. TACL, Vol. 5, pp. 339–351, 2016.

Appendix

表音的カタカナ表記について

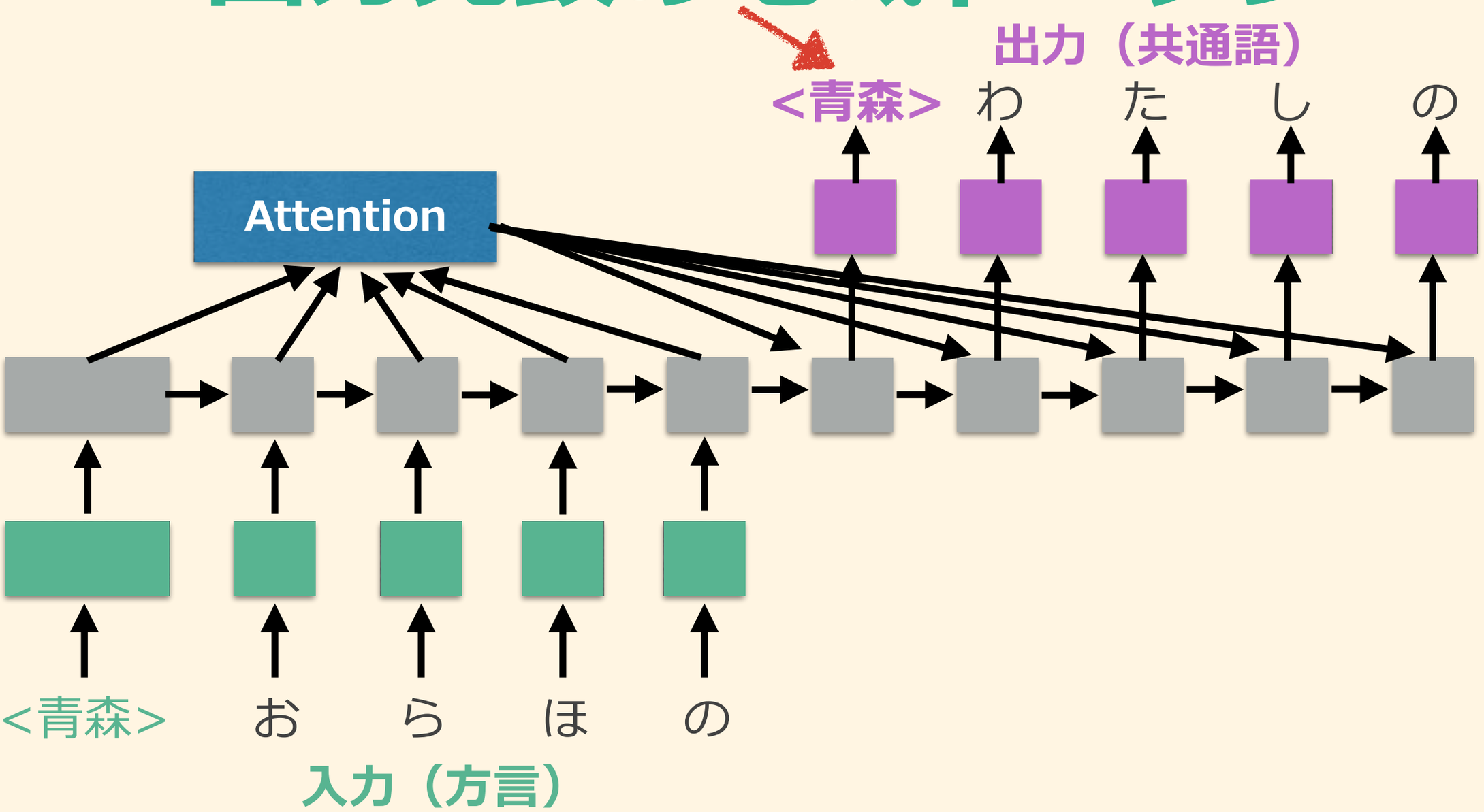
- 発話に対して忠実な音で書き起こしたカタカナ表記
- 例）
  - 伸ばす音（長音）→「ー」

  - 助詞「は」「を」「へ」→「ワ」「オ」「エ」

  - 鼻濁音（東北等で見られる、か行の特別な濁音）
    →「か゜」「き゜」「く゜」「け゜」「こ゜」

  - 「時には」が訛って「トキニァ」 等

出力先頭の地域トークン



- 出力先頭に地域トークンを付与しない場合でも同様に実験を行ったが、付与した方がBLEUスコアが高かった

方言と共通語の差分

- 共通語→方言の変化は、言語処理的視点で見ると
  文字レベル or 単語レベルの変化に分けることができる

文字レベルの変化	単語レベルの変化
・音の変化（母音変化、濁音化） - いい感じ → **ええ**感じ ・撥音, 母音の追加または削除 - めずらしい → め**ん**ずらしい - がっこ**う** → がっこ ・特徴的な助詞・助動詞 - ～っぺ, ～だべ （終助詞） - 押さ**さる**（助動詞）	・表層形だけ変化した単語 - 私 → わ, わい, おら, … - とても → なまら, たげ, … ・標準語にない概念を表す単語 - あずましい（青森） - いずい（宮城）

実験結果：定性評価（翻訳例）

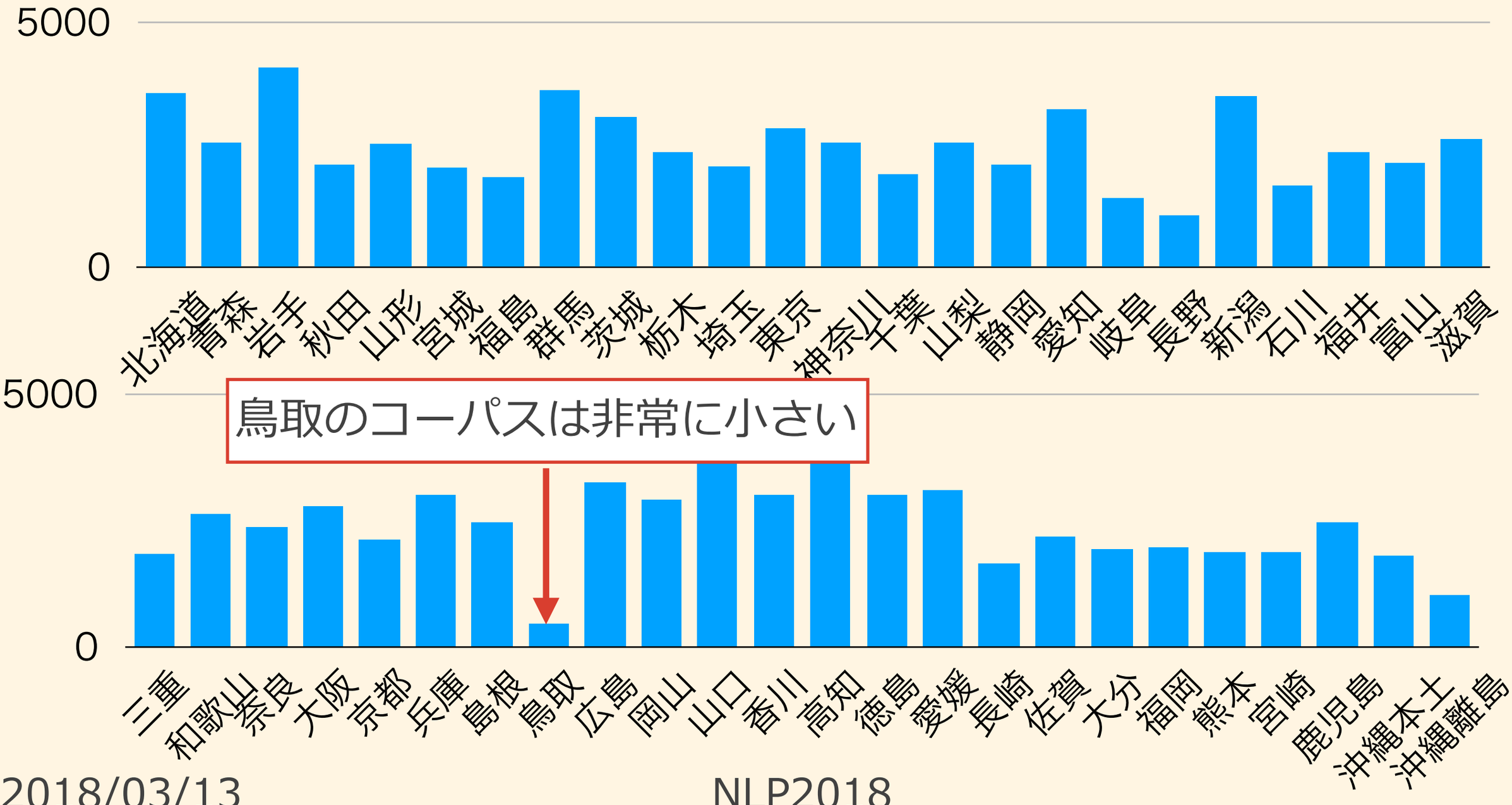
- 翻訳できた例

	青森方言（津軽）
入力方言文	あげぁ/いろこ/つだり/すて/それか ゜/ほら
参照文	あかい/いろ/ついたり/して/それが/ほら
出力文（提案手法）	あかい/いろ/ついたり/して/それが/ほら
出力文（文ごと翻訳）	あかい/いろ/**ついて**/それが/ほら
出力文（全地域SMT）	あかい/いろ**こ**/ついたり/して/それが/ほら

- **完璧な翻訳を出力したのは提案手法だけ**

コーパスにおける各地域ごとの割合

- 全地域合わせて34,117文（116,928文節）
- 地域ごとの文節数は以下の通り



多言語翻訳で付加するトークンのクラスタリング

- **多言語翻訳における言語の名前を表すトークン**も、
  t-SNEクラスタリングすると、言語類型学上知られている
  語族の区分に沿うことが判明した[Tiedemann, 2018]

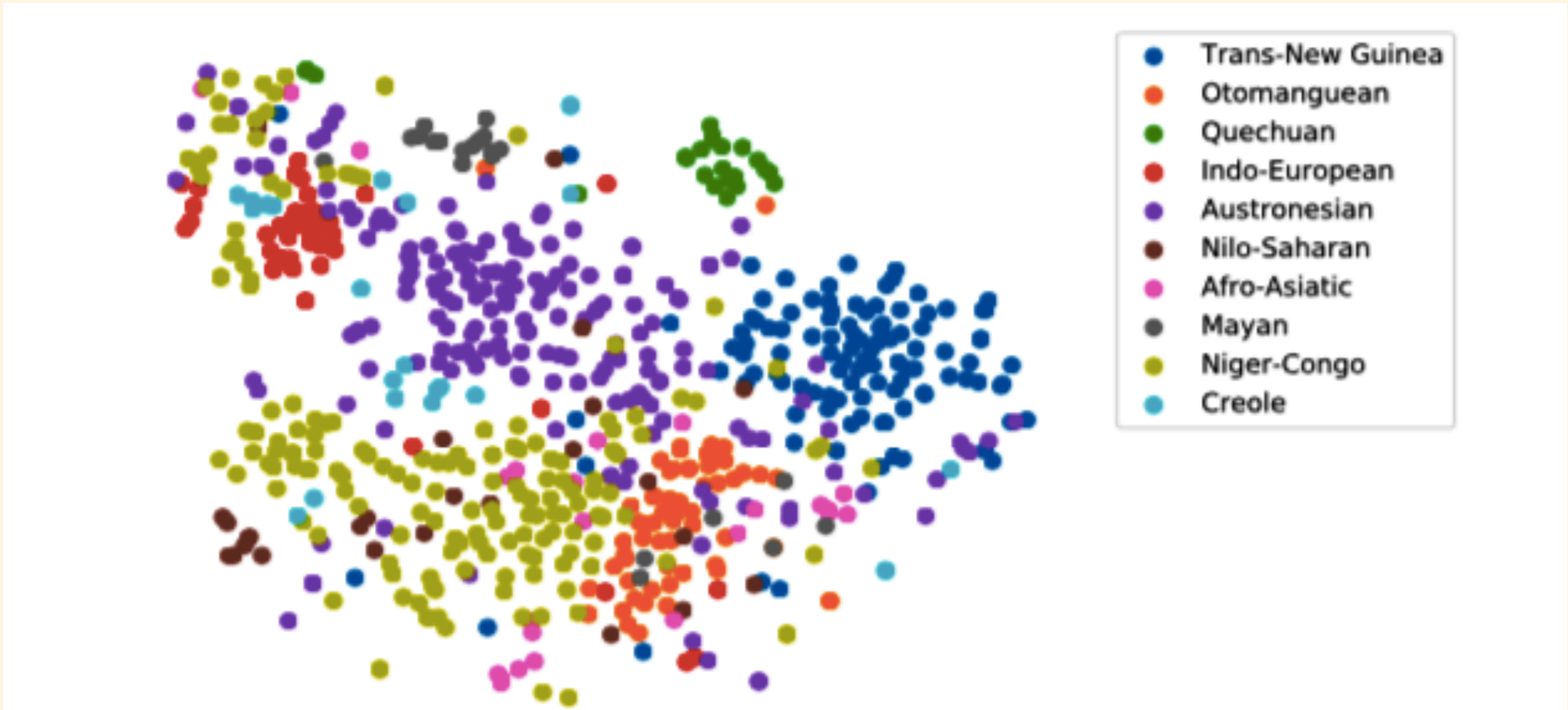


Fig. 4. Continuous language space that emerges from multilingual NMT (t-SNE plot).

J¨org Tiedemann, Emerging Language Spaces Learned From Massively Multilingual Corpora, arXiv:1802.00273