単語アラインメントの最前線

- WSPAlign: スパン予測に基づく単語対応の事前訓練 -

永田昌明 (NTTコミュニケーション科学基礎研究所)

東大のQiyu Wuさん(D2)と鶴岡教授との共同研究

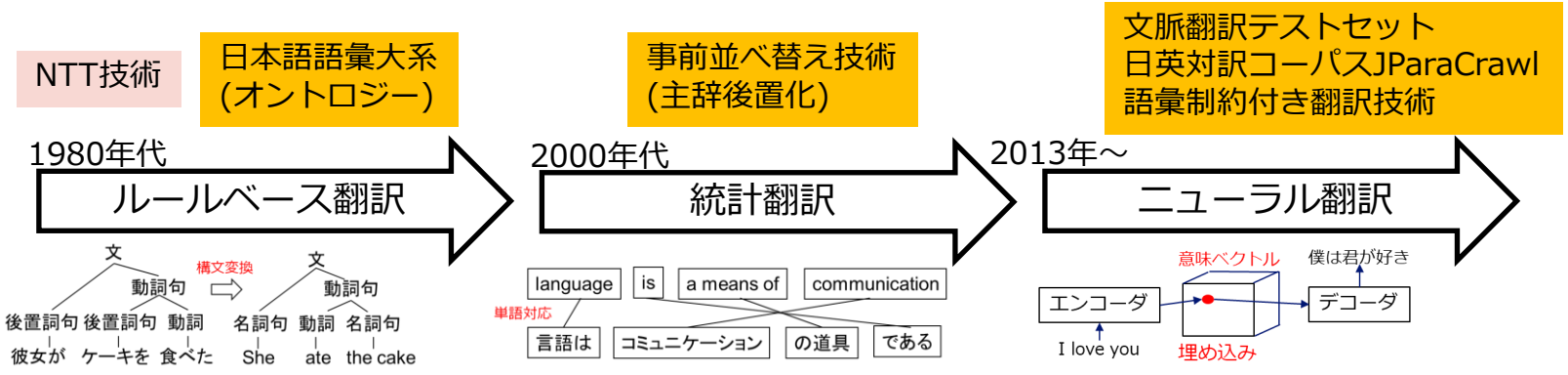
ACL2023論文の筆頭著者

はじめに

統計翻訳の時代(2010-2015年頃)
事前並べ替えの正解を作るために
単語対応データをたくさん作った

単語対応 (word alignment)

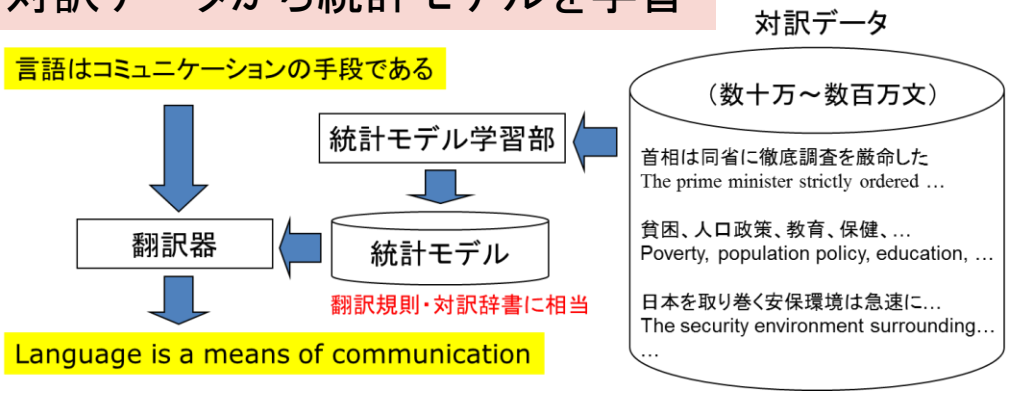
- 互いに翻訳になっている単語を結び付けること
- 一般的には多対多対応であり、英語の冠詞や日本語の助詞のように相手言語に対応する単語がない場合もある
- 2000年代に統計翻訳モデルに基づく手法(GIZA++, fast_align, etc.)が確立
- ニューラル翻訳モデルでは単語対応問題を解決できず約20年間進歩がなかった



従来研究 (1/2):
統計翻訳モデルに基づく単語対応

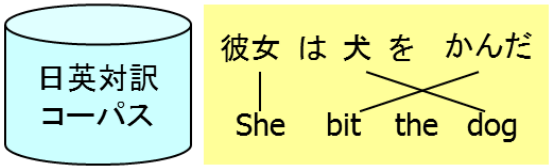
統計翻訳: 大量の対訳データから統計モデルを学習

[Brown+, CL-1993]



翻訳規則(語順変換)や対訳辞書(二言語対応)に関する確率をデータから明示的に学習

対訳コーパスの単語対応付け



（文対応付き）

自動学習

期待値最大化アルゴリズム
(EMアルゴリズム)

単語翻訳確率 P(f|e)

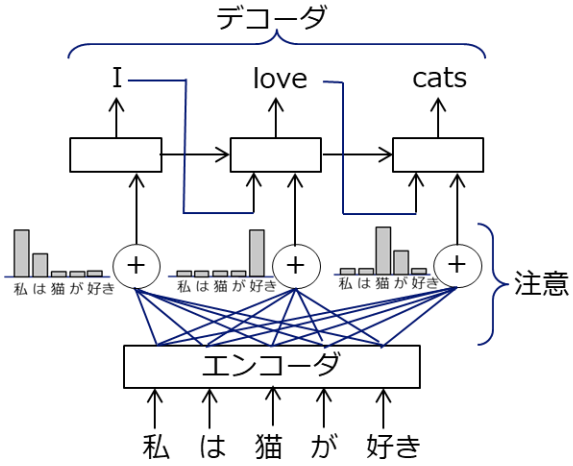
彼女		犬		かんだ	
she	0.7	dog	0.8	bit	0.5
her	0.2	dogs	0.1	chewed	0.3
girlfriend	0.1	puppy	0.1	blew	0.2

従来研究 (2/2):
ニューラル翻訳の注意機構に基づく単語対応

[Bahdanau, ICLR-2015]

注意機構 (attention mechanism)

- 出力文の各単語を生成する際に、注目すべき入力文の単語を動的に選択する仕組み
- 入力文の各単語に対するエンコーダの内部状態を重み付きで加算してデコーダへ入力
- 注意の大きさはエンコーダの各単語の内部状態とデコーダの各単語の内部状態の類似度で決まる



注意≠単語対応

- 注意から求めた単語対応は、GIZA++ (統計翻訳モデル)より精度が低い

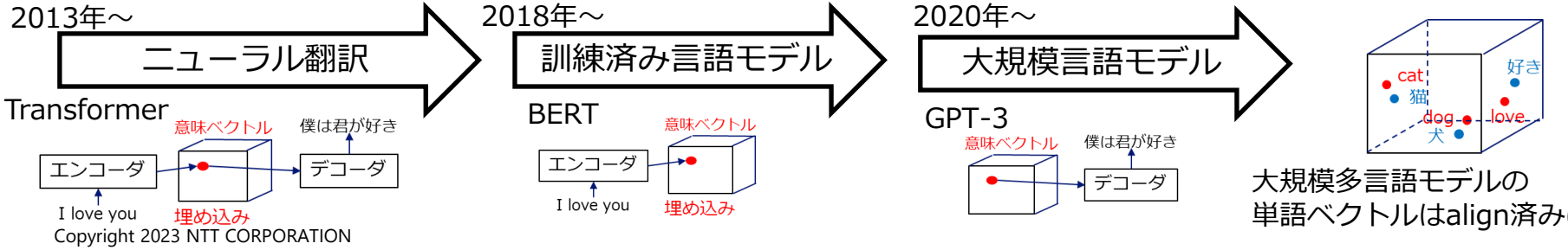
[Zenkel+, ACL-2020] End-to-End Neural Word Alignment Outperforms GIZA++

提案法の背景(1/2):
言語モデルの多言語性 (multilinguality)

Transformer [Vaswani+, NeurIPS-2017]
- 注意機構を用いてRNN相当の計算を並列化 ⇒ 大規模データで訓練可能
- 多言語(複数の言語対)の対訳データで一つの翻訳モデルを訓練 ⇒ 訓練データにない言語対も翻訳可能

BERT [Devlin+, NAACL-2019]
- Transformerの符号化器(encoder)を単語穴埋めタスクで訓練した言語表現モデル
- 多言語の単言語データで一つのモデル訓練すると、言語横断の転移学習が可能

GPT-3 [Brown+, NeurIPS-2020]
- Transformerの複合器(decoder)を次単語予測タスクで訓練した言語生成モデル
- ほとんどの訓練データは英語なのに、なぜか翻訳もできる ⇒ ChatGPTは日本語で違和感なく利用可能



提案法の背景(2/2): スパン予測による質問応答

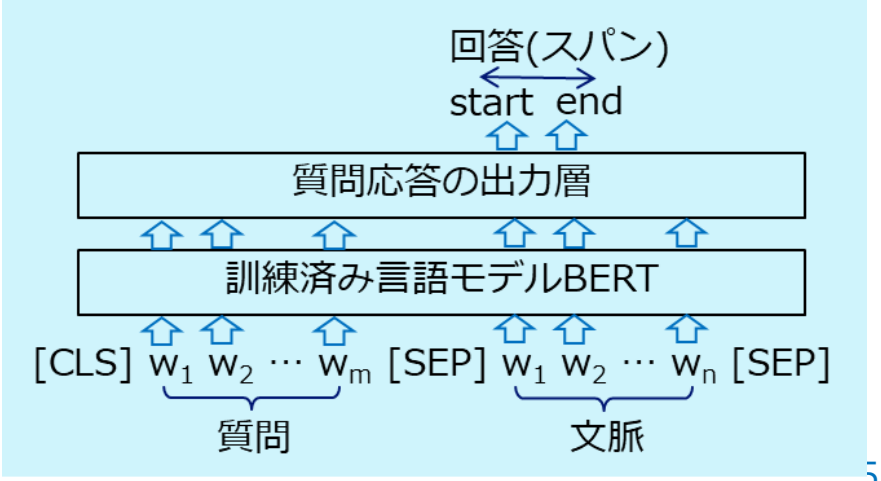
SQuAD (Stanford Quation Answering Dataset) [Rajpurkar+, EMNLP-2016]

- Wikipedia記事に対して回答が記事の部分文字列 (=スパン) であるような質問を人手で作成した質問応答データ

質問応答データSQuADを使ってBERTをファインチューン [Devlin+, NAACL-2019]

- 人間並みの質問応答(=スパン予測)の精度を実現

文脈 (context):
In meteorology, precipitation is any product of the condensation of atmospheric water vapour that falls under gravity. The main forms of precipitation include drizzle, rain, sleet, snow, graupel and hail ...
質問 (question):
What causes precipitation to fall?
回答 (answer):
gravity

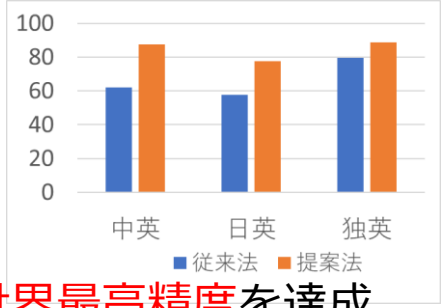


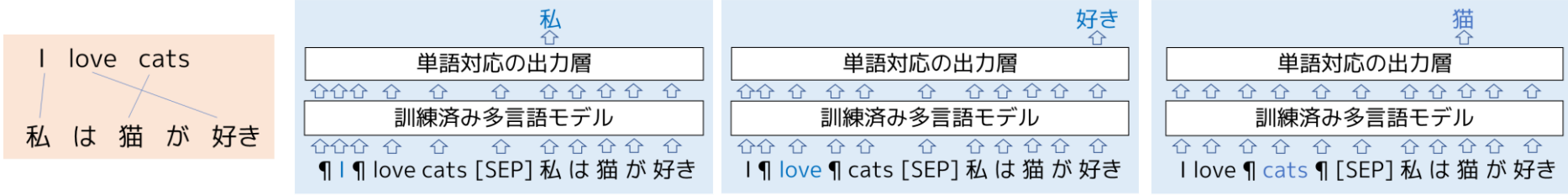
提案法1a:
言語横断スパン予測に基づく単語対応

[Nagata+, EMNLP-2020]

教師あり単語対応 (supervised word alignment)

- 区切り記号¶で囲まれた文中の単語に対応する翻訳先言語の単語列(スパン)を予測
- スパン予測は、SQuAD形式の質問応答と同じ実装
- 多言語(multilingual)BERTを使って言語横断を実現
- 二方向のスパン予測を対称化してノイズを除去
- 約300文対の正解データがあれば従来法の精度を大きく上回り世界最高精度を達成





論文に書いていない知見: 中英で訓練したモデルで日英単語対応を求められる

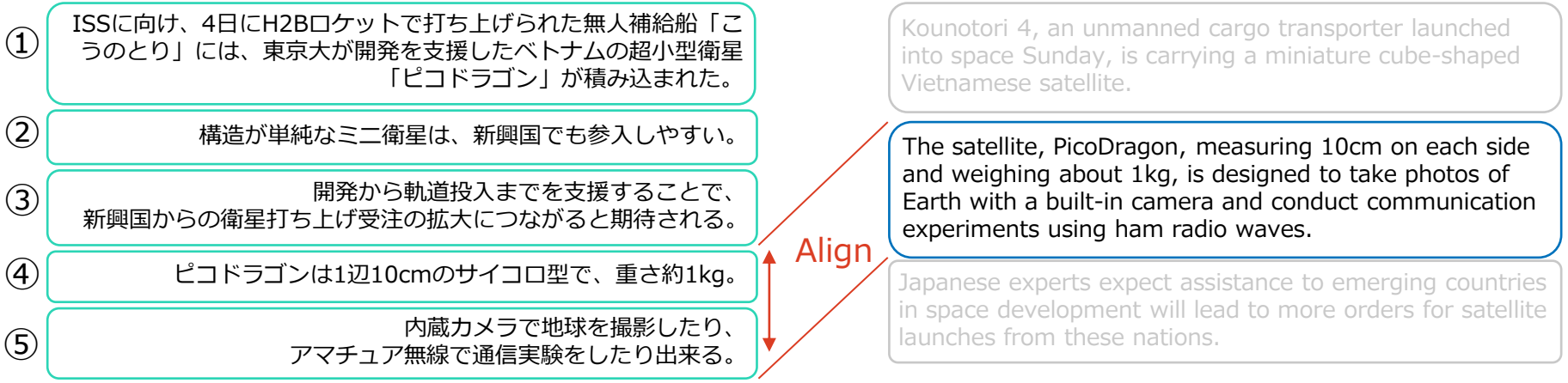
提案法1b:
SpanAlign: スパン予測に基づく文対応

[Chousa+, COLING-2020]

ほぼ同じやり方で、文対応もできる

- そもそもJParaCrawl等の対訳コーパス作成技術として文対応を先に着手
- 以後は、単語対応もSpanAlignと呼ぶ。

2018年夏: BERT以前、QAの流行り始め頃



論文に書いていない知見: 疑似データ(自動文対応で作成)を増やせば、どんどん精度が上がる

関連研究:
教師なし単語対応・半教師あり単語対応

訓練済み多言語モデルを用いた単語対応が盛り上がる (2020～)

- 教師なし(unsupervised): SimAlign [Sabet+, EMNLP findings-2020]
  - 訓練済み多言語モデル(mBERTとXLM-R)の文脈化単語埋め込みの類似度で対応関係を決定
- 半教師あり(semi-supervised): AWESOME [Dou and Neubig, EACL-2021]
  - 単語対応の精度を向上するような目標関数で、対訳データで訓練済み多言語モデルをファインチューン
  - マスク言語モデル化(MLM)、翻訳言語モデル化(TLM)、自己訓練目標関数(SO)など

教師信号の質や量によって単語対応の精度は変わる

統計翻訳モデルの
期待値最大化法(EM)と類似

- 教師あり(SpanAlign) > 半教師あり(AWESOME) > 教師なし(SimAlign)
  - 精度: 人手で作成した単語対応の正解 > 対訳データ > 訓練済み多言語モデル
  - データ入手の容易さ: 人手で作成した単語対応の正解 < 対訳データ < 訓練済み多言語モデル

多くの場合、AWESOMEやSimAlignの方が使い勝手がよい⇒SpanAlignを教師なし(正解データ不要)にしたい

提案法2 (1/2): WSPAlign

WhisperやCLIPの事前訓練を参考にした

スパン予測に基づく単語対応の事前訓練

[Wu+, ACL-2023]

弱教師あり学習: 正解データの制約を緩める

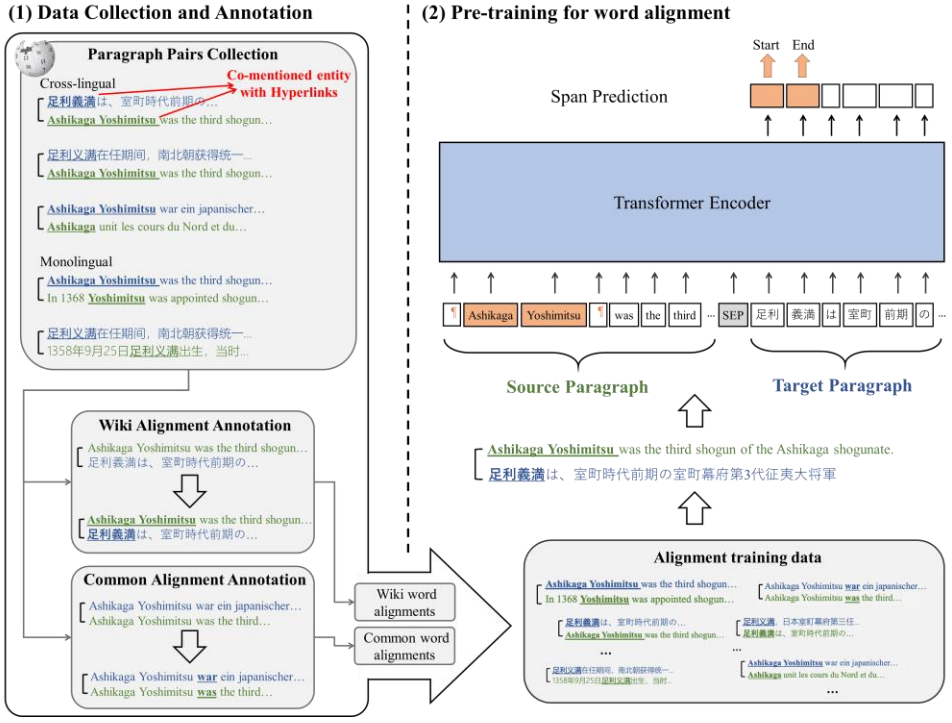
- 誤りを含んでもよい (correct → noisy)
- 部分的な対応でよい (fully-aligned → partial)
- 対訳文でなくてよい (parallel → non-parallel)

Wikipediaの言語間リンク

- Wikidataから求めた異なる言語のCo-mentionを文対応および単語対応とみなす
- ほとんどが固有表現

多言語単語埋め込みに基づく自動単語対応

- SimAlignと同じ方法で、Comentionを持つ文対に単語対応を付与
- 英語のPOS taggerを使って、品詞に基づき、一般語の単語対応のみを事前訓練に使用



WSPAlign: Weakly Supervised span Prediction pretraining for word Alignment

提案法2 (2/2): WSPAlign
スパン予測に基づく単語対応の事前訓練 (2/2)

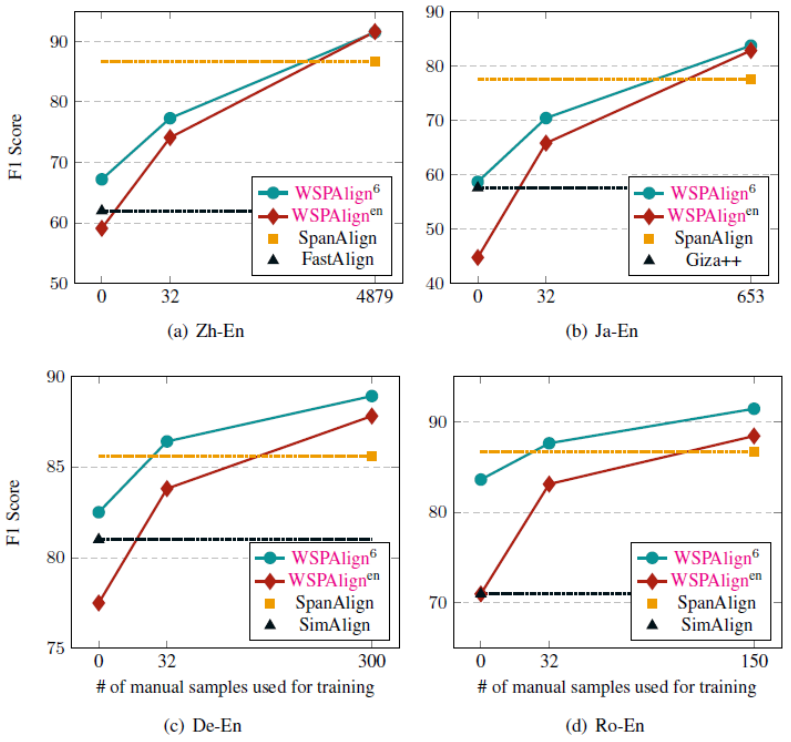
[Wu+, ACL-2023]

スパン予測に基づく事前訓練

- 中英、日英、独英、ル英、英仏: 各200万対
  - M6: 5言語対、E: 英語のみ
- 中英と日英はmBERT、その他はXLM-R

単語対応

- 人手で作成した正解でファインチューン
- ベースライン(SpanAlign)から単語対応精度(F1)が3.3-6.1 ポイント向上
- ゼロショット(ファインチューンなし)でも SimAlignより精度が高い
  - 事例が少ない段階では言語を横断する知識が重要
- 英語のみの事前訓練でもSpanAlignの精度を超える
  - 言語に関係なく、スパン予測の能力を高めることが重要



応用先:
機械翻訳の誤り検出 (単語レベルの品質推定)

機械翻訳の後編集 (Post-Editing)

- ニューラル翻訳になって精度は大きく向上したが、誤りがゼロにはならないので、医療や特許など誤りが許されない分野では、人間が誤りを修正する作業は不可欠

機械翻訳の単語レベルの品質推定 (Word-level Quality Estimation)

- 単語ごとに、原文と機械翻訳文から参照訳なしで機械翻訳の正(OK)/誤(BAD)を推定。後編集において修正すべき場所を予測 (国際会議WMTの共通タスクの一つ)
- 後編集文を介して原文と機械翻訳文の単語対応とOK/BADの正解を作成し、翻訳品質タグ推定モデルを教師あり学習
  - › 原文と後編集文: 二言語単語対応、後編集文と機械翻訳文: 単言語単語対応

まとめ

単語対応を教師あり学習する新しい方法SpanAlignを考案

- 言語横断スパン予測に基づく単語対応
- 単語対応を独立なスパン予測の集合として定式化
- 人手で作成した正解データが300文対程度あれば、非常に高い精度を達成

スパン予測に基づく単語対応の新しい事前訓練法WSPAlignを考案

- 正解データの制約を緩和: 対訳ではない文対の一部に自動的に対応を付与
- ゼロショットで教師なし単語対応SimAlignより精度が高い
- 正解データでファインチューンするとSpanAlignより精度が高い

今後の課題

- 必ずしも意味的に等価でない、同じ/異なる言語の文対の単語対応 (semantic diff)
- 高速で使い易い単語対応ツールの開発
- 機械翻訳の後編集支援 (機械翻訳の誤り検出)への応用