第７回
AAMT 長尾賞学生奨励賞

大語彙フレーズ翻訳機能付きSMT・NMT ハイブリッド翻訳方式

龍 梓
深圳技術大学
2020年12月2日

受賞論文

- フレーズ・トークン込みNMTモデル及びSMTによる大語彙フレーズ翻訳によるハイブリッド翻訳方式
  - 龍梓・木村龍一郎・飯田頌平・宇津呂武仁・三橋朋晴・山本幹雄,
  - 電子情報通信学会論文誌, Vol.J102-D, No.3 (2019年3月), pp.104-117
  - 学生論文特集秀逸論文

ニューラルネットワーク機械翻訳（NMT)と大規模語彙のはなし

- Bidirectional LSTM時代： 2014年 ~ 2016年
  - Bahdanau et al., Neural Machine Translation by Jointly Learning to Align and Translate. 2015
  - NMTが大語彙に対応できない問題が露見：何万程度の語彙でしか扱えない
  - 大語彙に対応する手法：Luong et al. Addressing the Rare Word Problem in Neural Machine Translation. 2015
    - 未知単語をトークンに変更してから学習を行う
    - 出力文にある未知語トークンはNMT以外の手段によって翻訳
    - 単語単位でフレーズを対応できない

ニューラルネットワーク機械翻訳（NMT)と大規模語彙のはなし

- 深層LSTM時代： 2016年 ~ 2017年
  - Wu et al., Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation. 2016
  - NMTモデルの学習力向上によって、未知単語を細かく分割することによる語彙拡大の手法が可能
    - word piece / sentence piece
    - 高頻度の文字列・機能語列は1トークンに分割し、低頻度語を細かく分割することによって未知語をなくする手法
    - 非構成的なフレーズの翻訳に弱い

大規模フレーズに対応するNMTモデル：
統計的機械翻訳による大語彙フレーズ翻訳との併用によるニューラルネットワーク機械翻訳システム

ニューラルネットワーク機械翻訳（NMT)と大規模語彙のはなし

- Transformer時代： 2017年 ~ 現在
  - Vaswani et al., Attention is all you need. 2017
  - 学習力は強くて、sentence pieceと併用すれば、未知語問題はほぼ解決
  - 非構成的な複合語フレーズを正確に翻訳するNMTモデル
    - J. Feng and et.al. Neural Phrase-to-Phrase Machine Translation. https://arxiv.org/abs/1811.02172. 2018 (rejected by ICLR 2020)
    - フレーズ単位でエンコードとデコード、および、Transformerによるsegment attention機構
    - Transformerモデルを勝っているという報告
  - 2018年以後：無敵なPretrainedモデルによってフレーズの翻訳問題を解決する
    - Song et al., MASS: Masked Sequence to Sequence Pre-training for Language Generation. 2019

ニューラルネットワーク機械翻訳（NMT)と大規模語彙のはなし

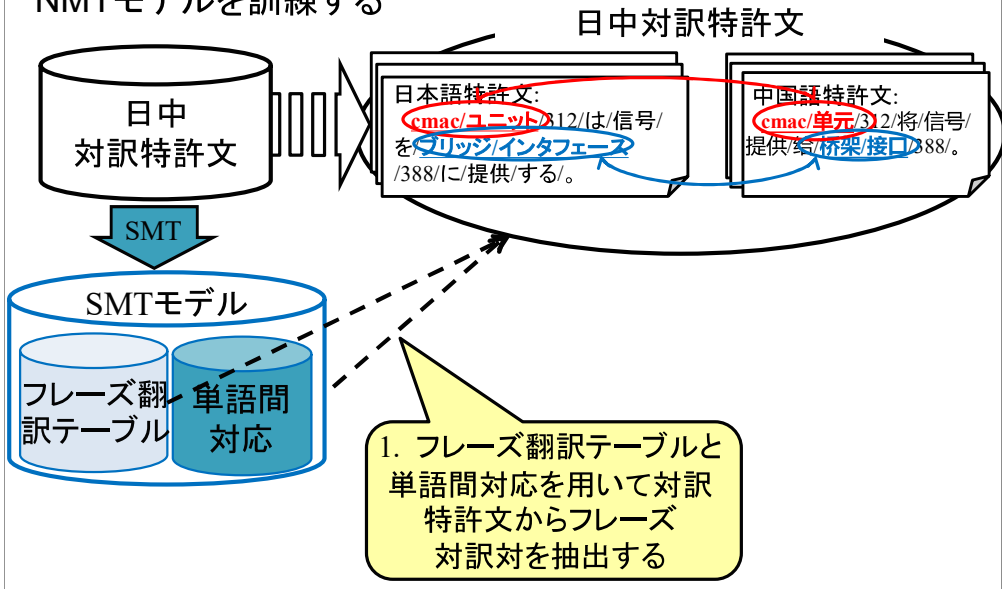
- 深層LSTM時代： 2016年 ~ 2017年
  - Wu et al., Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation. 2016
  - NMTモデルの学習力向上によって、未知単語を細かく分割することによって語彙拡大の手法が可能
    - word piece / sentence piece
    - 高頻度の文字列・機能語列は1トークンに分割し、低頻度語を細かく分割することによって未知語をなくする手法
    - 非構成的なフレーズの翻訳に弱い

大規模フレーズに対応するNMTモデル：
統計的機械翻訳による大語彙フレーズ翻訳との併用によるニューラルネットワーク機械翻訳システム

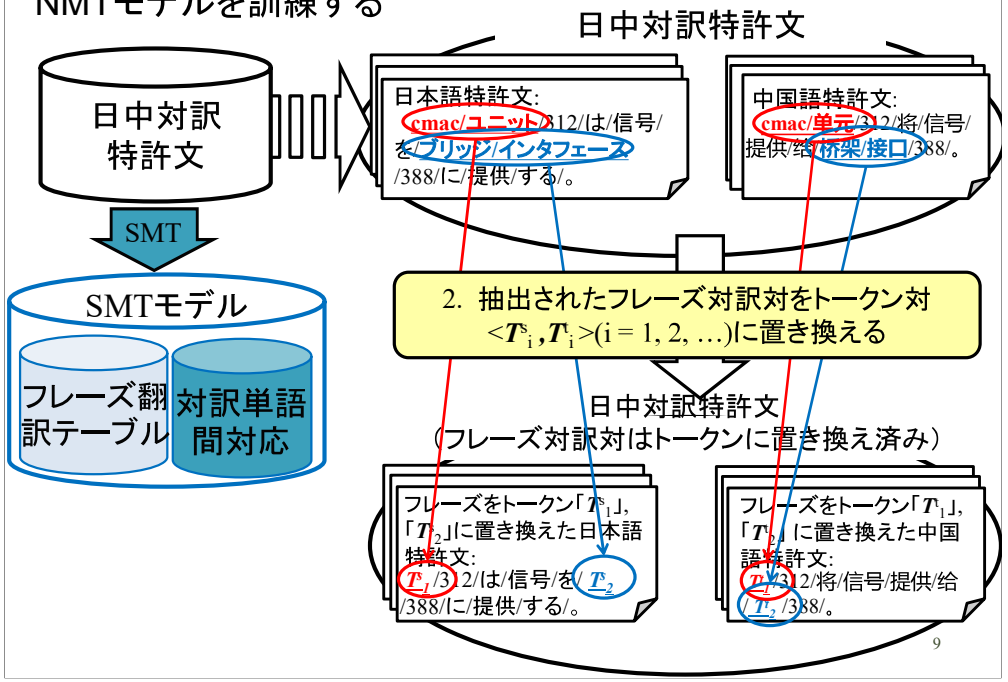
統計的機械翻訳による大語彙フレーズ翻訳との併用によるニューラルネットワーク機械翻訳システム

- ステップ 1：大規模フレーズ語彙に対応したNMTモデルを訓練
- ステップ2：生成されたNMTモデルを用いて訳文生成
  - NMTモデルによる訳文生成およびSMTモデルによるフレーズ翻訳の併用

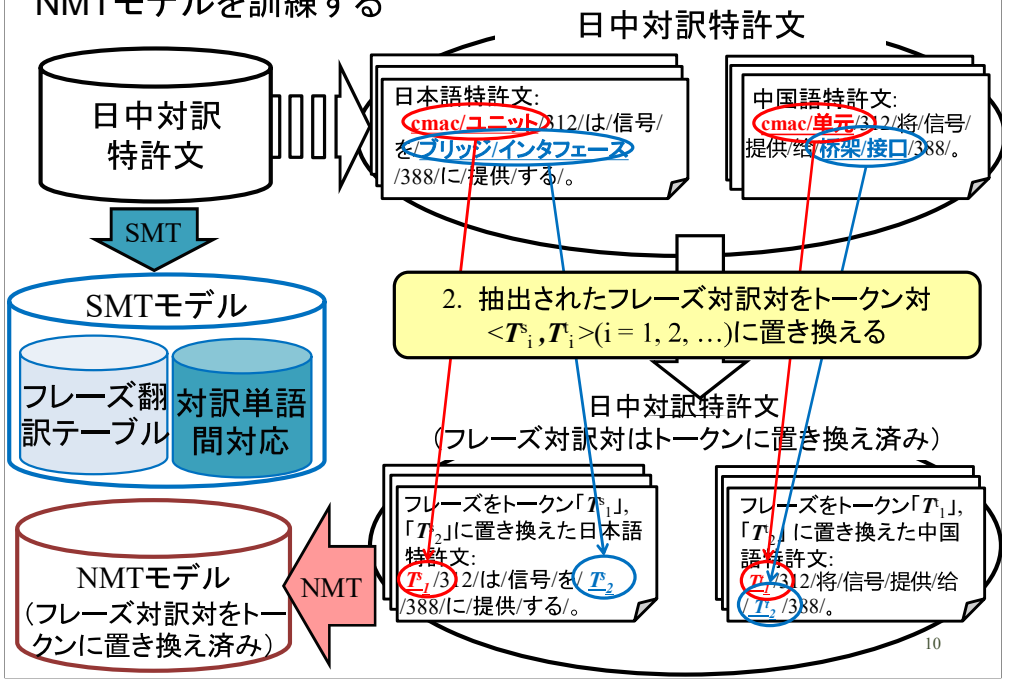
フレーズ対訳対をトークンに置き換えてからNMTモデルを訓練する



フレーズ対訳対をトークンに置き換えてから
NMTモデルを訓練する



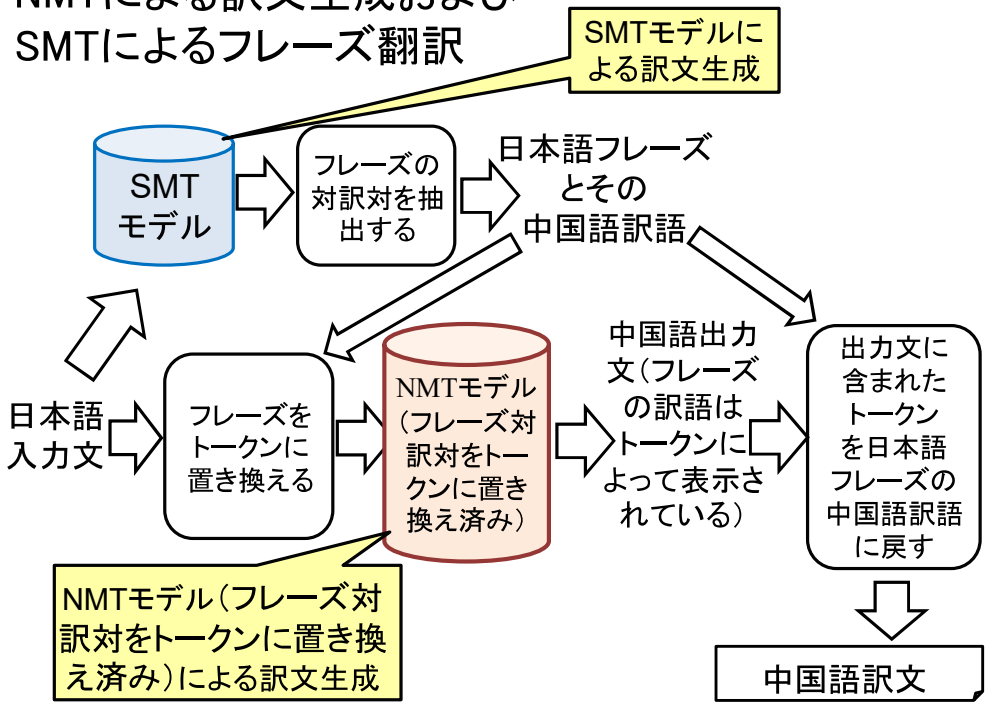
フレーズ対訳対をトークンに置き換えてから
NMTモデルを訓練する



統計的機械翻訳による大語彙フレーズ翻訳との併用によるニューラルネットワーク機械翻訳システム

- ステップ 1：大規模フレーズ語彙に対応したNMTモデルを訓練
- ステップ2：生成されたNMTモデルを用いて訳文生成
  - NMTモデルによる訳文生成およびSMTモデルによるフレーズ翻訳の併用

NMTによる訳文生成および
SMTによるフレーズ翻訳



評価

- WAT 2017の特許翻訳タスクにおいて配布されたデータ

	訓練用 対訳特許文	開発用 対訳特許文	評価用 対訳特許文
日中対訳文	998,054	2,000	2,000
日英対訳文	999,636	2,000	2,000

評価

- 文単位の翻訳性能の評価
  - 自動評価・4-gram BLEU (K. Papineni and et al. 2002.)
    - 全出力文と全参照文との間のn-gram類似度の調和平均
  - 人手評価・絶対評価
    - JPO評価基準に基づいて、全出力文に対して人手で評価し、スコアの平均を求める
  - 人手評価・一対評価
    - 提案手法の出力文をベースラインNMTの出力文と比較して人手で優劣を判断し、改善の文数と改悪のぶん数の差が全体における割合をスコアとする
- 名詞句翻訳性能の評価
  - 大語彙フレーズの典型例である名詞句の翻訳性能がどの程度改善したのかの評価を行う

評価結果(文単位の評価結果)

- 自動評価（BLEU）

手法	日中	中日	日英	英日
ベースラインSMT	30.0	36.2	28.0	29.4
ベースラインNMT	34.2	40.8	43.1	41.8
PosUnkによるNMT（Luong 2015)	35.0	41.0	43.3	41.8
Sentence Piece NMT	34.5	41.0	43.5	42.0
提案手法	35.6	41.6	43.9	42.5

評価結果(文単位の評価結果)

- 一対評価(ベースラインNMT との比較，スコアの範囲: −100 ～100)

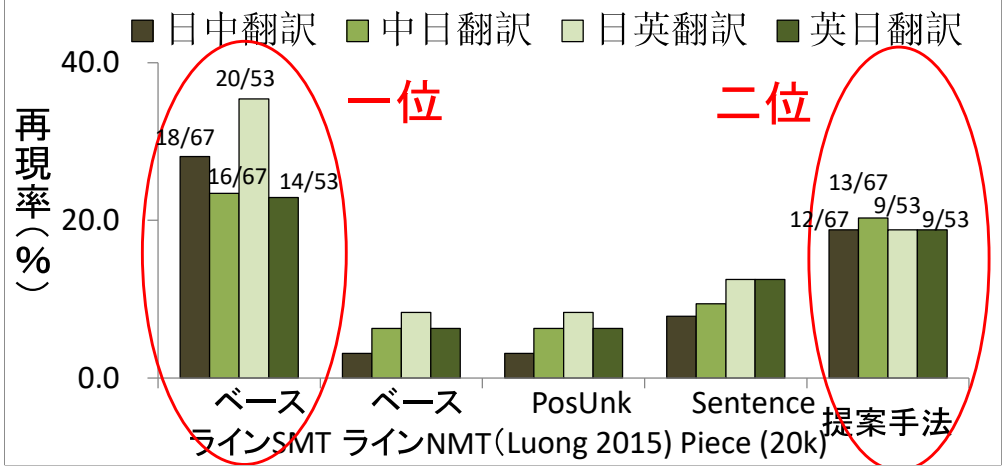
手法	日中	中日	日英	英日
PosUnkによるNMT	13	12.5	9.5	14.5
Sentence Piece NMT	19.0	18.0	11.5	16.0
提案手法	23.5	22.5	15.5	19.0

- JPO 基準に基づく絶対評価結果(スコアの範囲: 1 ～ 5)

手法	日中	中日	日英	英日
ベースラインSMT	3.1	3.2	2.9	3.0
ベースラインNMT	3.6	3.6	3.7	3.7
PosUnkによるNMT	3.8	3.9	3.9	3.9
Sentence Piece NMT	3.9	3.9	4.0	3.9
提案手法	4.1	4.1	4.2	4.1

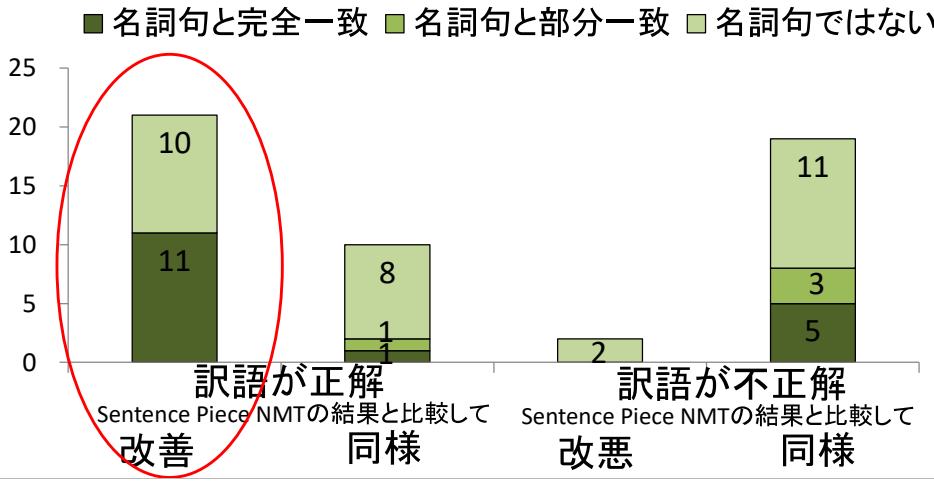
評価結果（名詞句翻訳性能の評価）

- 参照目的言語文に含まれる名詞句訳語は出力文における再現率（%）
  - 未知語を含む名詞句のみを評価対象



評価結果（名詞句翻訳性能の評価）

- トークン箇所のフレーズは正しく翻訳されたか否かの個数
  - 日中翻訳、提案手法をSentence Piece NMTと比較する



提案手法による改善例・Sentence Piece NMTと比較

日本語入力文

このとき、メインバルブ27、パイロットバルブ28及び**フェイルセーフバルブ**29を通った油液は、メインボディ35及びパイロットボディ37と締結部材82との間へ流れ、締結部材82の通路85を通って排出される。

提案手法

此时,通过主阀27、先导阀28以及**防故障阀**29的油液流向主体35及先导主体37与紧固部件82之间,通过紧固部件82的通路85排出。

正解！

Sentence Piece NMT

此时,通过主阀27、先导阀28和**失效阀**29的油液流向主体35和先导主体37和紧固构件82之间,并且排出通过紧固构件82的通道85。

正解：**防故障阀**

不正解！

Sentence Piece NMTによって「フェイルセーフバルブ」を翻訳する際に、「セーフ」の訳語が抜けて、「失效阀」（フェイルバルブ）に誤訳した

提案手法による改善例・Sentence Piece NMTと比較

日本語入力文

このとき、メインバルブ27、パイロットバルブ28及び**フェイルセーフバルブ**29を通った油液は、メインボディ35及びパイロットボディ37と締結部材82との間へ流れ、締結部材82の通路85を通って排出される。

提案手法

此时,通过主阀27、先导阀28以及**防故障阀**29的油液流向主体35及先导主体37与紧固部件82之间,通过紧固部件82的通路85排出。

正解！

Sentence Piece NMT

此时,通过主阀27、先导阀28和**失效阀**29的油液流向主体35和先导主体37和紧固构件82之间,并且排出通过紧固构件82的通道85。

正解：**防故障阀**

不正解！

提案手法において、「フェイルセーフバルブ」はトークンに置き換えて，直接にSMTによって翻訳されて、訳抜けになれなかった

Future Work

- 現状:
  - Transformer学習力は強くて、sentence pieceと併用すれば、未知語問題はほぼ解決
  - 大規模コーパスから訓練されたPretrainedモデルによってフレーズの翻訳問題を解決
    - 非構成的なフレーズの翻訳知識を大規模コーパスから学習する
- 疑問:
  - 文単位で翻訳結果を評価して、フレーズ単位で評価はしていない
    - フレーズの翻訳は本当によくなっているか
  - Pretrainedモデル訓練時、masked処理はフレーズ単位で行っているわけではない
    - 非構成的なフレーズの翻訳知識は本当に上手く学習されているか？

ご清聴ありがとうございました