マルチソースニューラル機械翻訳における翻訳時の原言語欠落補完

西村 優汰[1], 須藤 克仁[1], Graham Neubig[2,1], 中村 哲[1]
[1]奈良先端科学技術大学院大学
[2]Carnegie Mellon University

概要

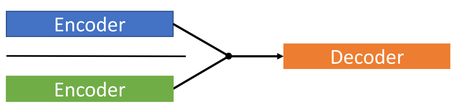
- マルチソースニューラル機械翻訳 (multi-source NMT) の**翻訳時**に，原言語側に欠落が存在する場合の問題に着手
- One-to-one NMTを用いて擬似対訳の複数候補をビーム探索によって生成し，欠落を補完する目的として最適な擬似対訳を選択する手法を提案
- 提案手法は擬似対訳を選択する手法として有効であるということが実験によって示された

1. 従来法と提案法について

従来法

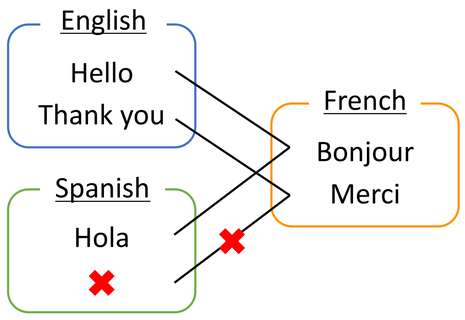
Multi-source NMT (Zoph and Knight, 2016)

Multi-source NMTは，2つ以上の原言語を用いる手法



Multi-source NMTは，全ての言語の対訳が揃っていることを前提にしている

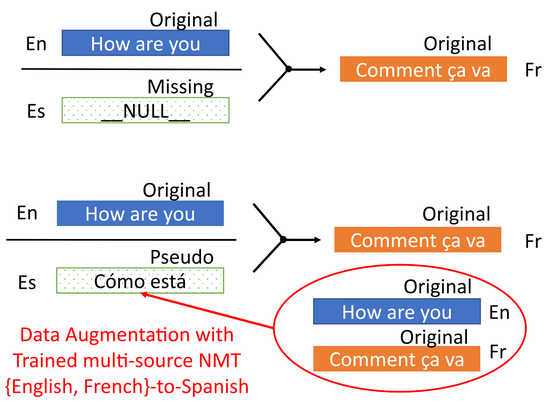
→欠落が存在する対訳は使用することができない



Multi-source NMT with missing data (Nishimura et.al., 2018)

→ 特殊記号による置換

欠落部分に特殊記号を置換し，欠落が存在する対訳も使用



→ 擬似対訳による補完

学習済みmulti-source NMTモデルを使用して擬似対訳を作成し，欠落部分を補完

提案法

従来法の問題点

モデルの学習時を考慮しており，翻訳時（テスト時）を考慮していない

特殊記号による置換：
翻訳時も適用可能だが翻訳精度は良くない

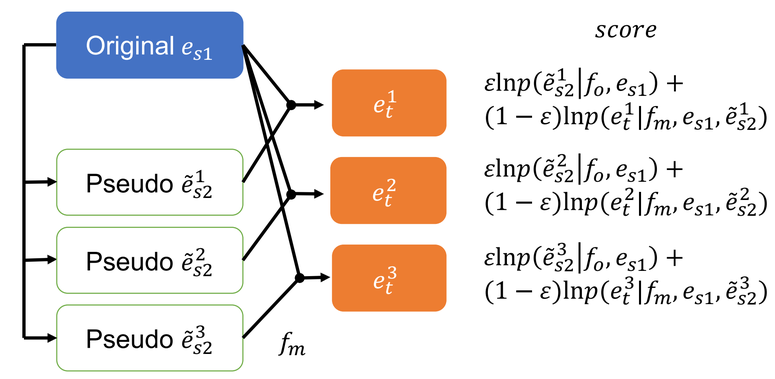
擬似対訳による補完：
擬似対訳生成時に目的言語の対訳を用いなければならない

→ One-to-one NMTによって擬似対訳の複数候補を生成し，欠落を補完する最適な擬似対訳を選択する手法を提案

1．f_oを用いて複数の擬似対訳候補をビーム探索によって生成

2．生成した擬似対訳それぞれに対して，f_mを用いて翻訳を生成

3．それぞれの翻訳結果に対し$score$を算出し最大となるものを最終的な翻訳結果とする



2. 実験

- データ

コーパス：TED Talks
言語対　：English – Croatian / Serbian
English – Slovak / Czech
English – Vietnamese / Indonesian

Pair	Trg	train	missing	test
en-hr/sr	hr	115,127	34,116 (29.6%)	1,145
	sr	129,461	48,450 (37.4%)	896
en-sk/cs	sk	58,109	16,772 (28.9%)	602
	cs	97,488	56,151 (57.6%)	1,966
en-vi/id	vi	150,829	81,945 (54.3%)	1,405
	id	77,936	9,052 (11.6%)	333

- Baseline手法
  - One-to-one NMT (English-to-X)
  - ビーム探索による1-bestの擬似対訳のみで欠落を補完したmulti-encoder NMT

- BLEUによる実験結果

Pair	Trg	Baseline		Proposed (5-best)
		One-to-one (En-to-Trg)	1-best	
en-hr/sr	hr	22.58	22.55	22.43
	sr	16.38	15.71	16.07
en-sk/cs	sk	14.16	16.57	16.59
	cs	15.13	13.63	13.85
en-vi/id	vi	22.62	22.96	23.69
	id	26.41	26.23	26.96

Proposed (5-best) vs Baseline (1-best)
ほとんどの言語対で提案手法の方が良い翻訳精度

Proposed (5-best) vs Baseline (One-to-one NMT)
言語対によって異なる→さらなる調査が必要

→ 提案手法はn-bestの擬似対訳候補からどの候補文が欠落を補完するのに適切であるかを選択する手法として有効

3. 今後の課題

- 言語の組み合わせや訓練文数の違い，欠落の度合いなどによる提案手法への影響の調査
- より良い欠落補完手法の提案

参考文献

Barret Zoph and Kevin Knight. 2016. Multi-Source Neural Translation. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 30–34, San Diego, California. Association for Computational Linguistics.
Yuta Nishimura, Katsuhito Sudoh, Graham Neubig, and Satoshi Nakamura. Multi-source neural machine translation with data augmentation. In 15th International Workshop on Spoken Language Translation (IWSLT), Bruges, Belgium, October 2018.