An Operation Sequence Model for Explainable Neural Machine Translation

Felix Stahlberg, Danielle Saunders, Bill Byrne

UNIVERSITY OF CAMBRIDGE
Department of Engineering

Introduction

- Achieve explainable AI by changing the output representation to explain itself
  - For NMT: Output sequence conveys both the translation and an “explanation” in form of alignments
- Potentially works on any Seq2Seq architecture

Operation Sequence NMT (OSNMT)

2000 hr の 安定 動作 を 確認 し た → Seq2Seq → SET_MARKER 2000 POP_SRC hr POP_SRC JMP_BWD SET_MARKER of POP_SRC JMP_BWD stable POP_SRC operation POP_SRC JMP_FWD JMP_FWD was POP_SRC POP_SRC confirmed POP_SRC

SET_MARKER: Insert new marker X_i into target sentence

$word$: Insert $word$ into target sentence

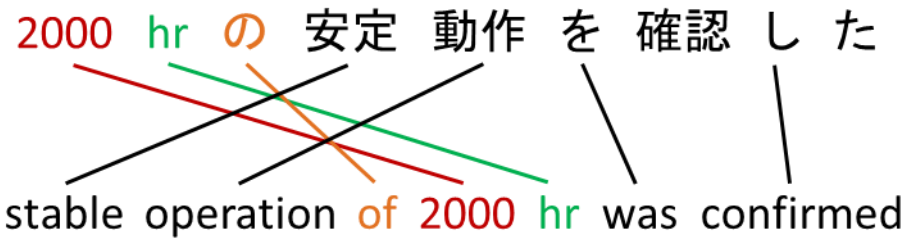
POP_SRC: Move source read head right by 1 token.

JMP_FWD|BWD: Move target write head to next marker.

Source side read head

Target side write head

	Operation	Source sentence	Target sentence (compiled)
		2000 hr の 安定 動作 を 確認 し た	X_1
1	SET_MARKER	2000 hr の 安定 動作 を 確認 し た	X_2 X_1
2	2000	2000 hr の 安定 動作 を 確認 し た	X_2 2000 X_1
3	POP_SRC	2000 hr の 安定 動作 を 確認 し た	X_2 2000 X_1
4	hr	2000 hr の 安定 動作 を 確認 し た	X_2 2000 hr X_1
5	POP_SRC	2000 hr の 安定 動作 を 確認 し た	X_2 2000 hr X_1
6	JMP_BWD	2000 hr の 安定 動作 を 確認 し た	X_2 2000 hr X_1
7	SET_MARKER	2000 hr の 安定 動作 を 確認 し た	X_3 X_2 2000 hr X_1
8	of	2000 hr の 安定 動作 を 確認 し た	X_3 of X_2 2000 hr X_1



- Generates target words in source sentence order, interspersed with reordering instructions
- Interpretable behaviour of Transformer attention heads
- OSNMT output can be seen as a parse through a formal (multitext) grammar

Representation	BLEU es-en	BLEU pt-en	BLEU ja-en dev	BLEU ja-en test
Plain	37.6	37.5	28.3	28.1
OSNMT	37.1	38.4	28.1	28.8

AER on ja-en against GIZA++ alignments

Method	BLEU	AER
RNN (max attention)	20.2	61%
Transformer (OSNMT)	**28.1**	**20%**