Translating Pro-Drop Languages with Reconstruction Models

Longyue Wang • Zhaopeng Tu • Shuming Shi • Tong Zhang • Yvette Graham • Qun Liu

ADAPT Centre, Dublin City University • Tencent AI Lab

Problem Statement
- Pronouns are frequently omitted in pro-drop languages (e.g., Chinese, Japanese) especially in informal genres.

Novelty of Work
- We show that although NMT models advance SMT models on translating pro-drop languages, there is still large room for improvement;
- Little attention has been paid to the problem within NMT. We introduce a reconstruction-based approach (+3.28 BLEU);
- We release a large-scale bilingual dialogue corpus (2.2M Chinese–English sentence pairs).

DP Annotation and Generation
Annotation: as large parallel corpora are usually available, we automatically annotate DP using alignment information.

Reconstructor
The reconstructor reads a sequence of hidden states and the labelled source sentence, and outputs a reconstruction score.

Experiments
- Data
  - Model | #Params | Speed | Bilingual BLEU
  - Baseline (±DPs) | 86.7M | 1.08K | 2.61 | 31.80 | -
  - Baseline (±DPs) | 86.7M | 1.05K | 2.63 | 32.67 | +0.87 |
  - Baseline (±DPs) | 86.7M | 0.73K | 2.63 | 33.67 | +1.87 | +1.00
  - Baseline (±DPs) | 86.7M | 0.45K | 2.18 | 33.49 | +1.87 | +1.00
  - Baseline (±DPs) | 86.7M | 0.75K | 2.16 | 35.08 | +3.28 | +2.41
  - Multi-source (Zoph and Knight 2016) | 20.7M | 1.17K | 1.27 | 32.81 | +1.01 | +0.14
  - Multi-layer (Wu et al. 2016) | 22.0M | 0.53K | 1.32 | 33.40 | +1.82 | +0.79
  - Multi-Dec (Lee et al. 2017) | 22.0M | 0.87K | 2.26 | 33.08 | +2.26 | +0.41

- Main Results

- Contribution Analysis

- Effect of Reconstruction

- Length Analysis

- Error Analysis

The ADAPT Centre for Digital Content Technology is funded under the SFI Research Centres Programme (Grant 13/RC/2106) and is co-funded under the European Regional Development Fund.