Multi-Task Learning for Cross-Lingual Abstractive Summarization

Sho Takase and Naoaki Okazaki
Tokyo Institute of Technology

Overview

Task: cross-lingual summarization (CLS)

Generate a summary in a target language

E.g., Japanese → English

議会共和党とホワイトハウスのトップレベルの予算交渉担当者は、日曜の午後遅くに交渉を終え…

→ Negotiators likely to extend 3rd deadline.

Proposal: use related-task data for improvement

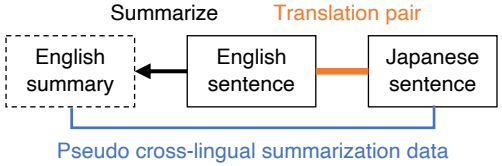
Problem and existing approaches

Problem: few training data for neural EncDec model

Existing approaches:

- Pipeline e.g., translation → summarization
- Pseudo data construction

  For example, summarize translation pairs



Proposal

Use existing data of translation and summarization since these are sub-tasks of CLS

Translation: CLS without any compression

Summarization: CLS within the same language

Use special token to distinguish tasks

<Trans>: translation

<Summary>: summarization including CLS

The proposed method can be used for translation, summarization, and CLS

<Trans> or 議会共和党と → Top-level budget …

<Summary> ホワイトハウスの… → Negotiators …

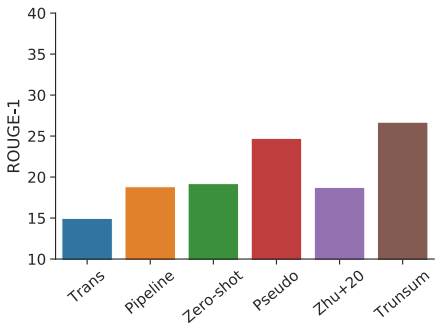
<Summary> Top-level budget … → Negotiators …

Compare methods

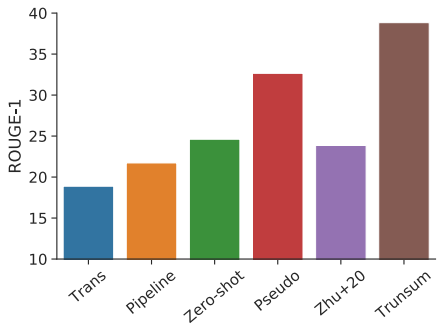
- Trans: translation of an input
- Pipeline: translation → summarization
- Zero-shot: training without pseudo CLS data
- Pseudo: training with only pseudo CLS
- Zhu+20: existing CLS method
- Trunsum: the proposed method

We used Transformer to construct each model

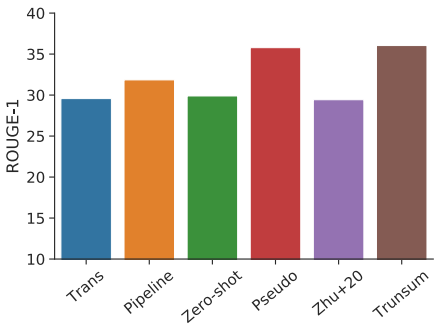
Experiments on cross-lingual summarization task in various language pairs



Chinese-English summarization



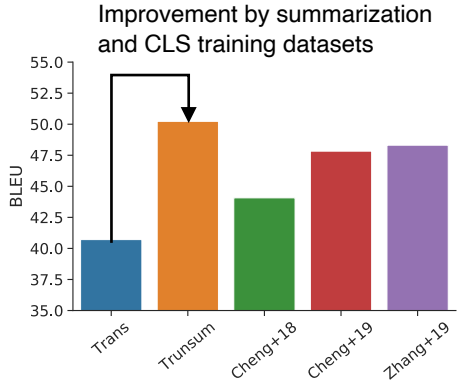
English-Japanese summarization



Arabic-English summarization

Trunsum (the proposed method) outperformed existing methods in all language pairs

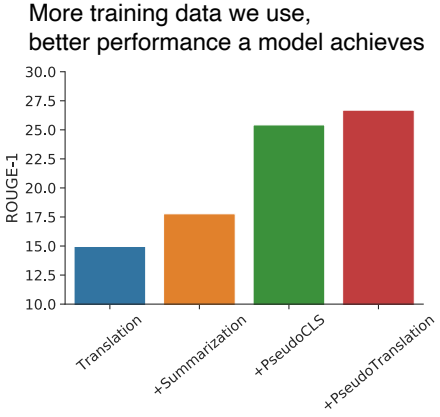
Results on machine translation



Chinese-English translation

Training data of related tasks also has a positive effect on MT

Effectiveness of each data



Chinese-English summarization

To improve the performance, the amount of data is important

Conclusion

Task: cross-lingual summarization (CLS)

Generate a summary in a target language

Problem: lack of training data for EncDec

Two Existing approaches:

- Pipeline
- Pseudo data construction

Proposal: use related-task training data

- Use translation and summarization data for the training of CLS
- Introduce special tokens to distinguish each task
- Trunsum outperformed existing methods in various language pairs
- Trunsum improved the performance of CLS and related tasks such as MT

Trunsum (the proposed method) is simple and effective