Multi-Source Neural Machine Translation with Missing Data

Yuta Nishimura[1], Katsuhito Sudoh[1], Graham Neubig[2,1], Satoshi Nakamura[1]
[1]Nara Institute of Science and Technology
[2]Carnegie Mellon University

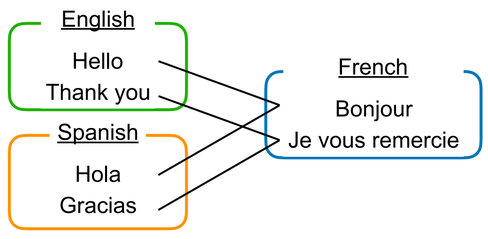
Abstract

- We approach the problem that there are some missing input sentences on multi-source translation
- We Examined a simple solution where missing inputs are replaced by a special symbol
- The experimental results with simulated (UN6WAY) and actual (TED Talks) incomplete multilingual corpora show that this method allows us to effectively use all available translations at both training and test time

1. Introduction

Conventional

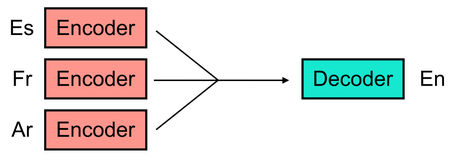
Multi-source NMT uses input in 2+ languages to improve results. Normally assumes that we have data in **all of the languages**



complete multi-source corpus

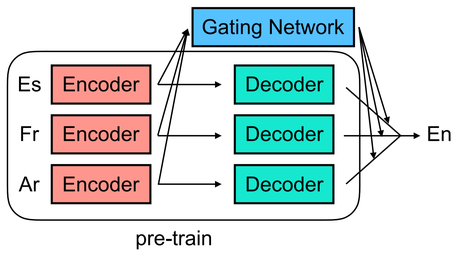
Multi-encoder NMT (Zoph and Knight, 2016)

Use **multiple encoders** corresponding to the source languages and **single decoder**



Mixture of NMT Experts (Garmash and Monz, 2016)

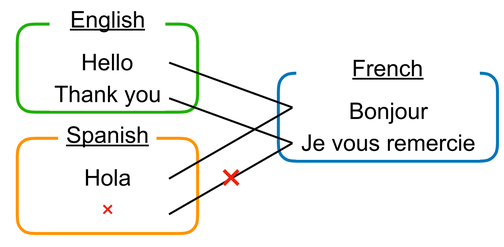
Ensemble together independently-trained encoder-decoder networks. Use sum of probabilities from one-to-one models weighted according to a **gating network**



Proposed

Problem

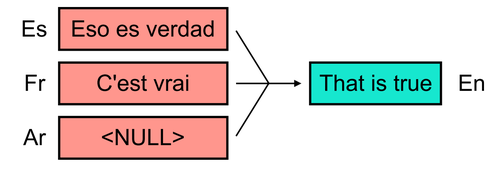
Many multilingual corpora are **not complete**
Existing studies on multi-source translation **did not explicitly handle** this situation



Incomplete multi-source corpus with missing data

Our method: Replace each missing input sentence with a **special symbol <NULL>**

We can expect the system to basically ignore the <NULL> symbol and use the other sentences



Ex) Arabic input is missing

2. Experiments

1. Pseudo-incomplete multilingual corpus (UN6WAY)

Using **pseudo incomplete corpus** created from complete corpus

Corpus : UN6WAY
Source language :
Spanish, French, Arabic
Target Language : English
Training sentences : 800K
Test set : complete

Settings of the pseudo incomplete corpus
(× means that this part was deleted)

Sentence No.	Es	Fr	Ar	En
1-200,000	×			
200,001-400,000			×	
400,001-600,000		×		
600,001-800,000				

BLEU by one-to-one and multi-source translation ({Es, Fr, Ar}-to-En})

Condition	One-to-One			Multi-encoder	Mix. of NMT Experts
	Es-En	Fr-En	Ar-En		
Complete (0.8M)	31.87	25.78	23.08	37.55 (+5.68)	33.28 (+1.41)
Complete (0.2M)	27.62	22.01	17.88	31.24 (+3.62)	32.16 (+4.54)
Pseudo-incomplete (0.8M)	30.98	25.62	22.02	**36.43** (+5.45)	32.44 (+1.47)

Pseudo-incomplete (0.8M) > complete(0.2M)

The additional use of incomplete corpora with **replacing missing sentence with <NULL>** is beneficial

2. An actual incomplete multilingual corpus (TED Talks)

Using an **actual incomplete corpus**

Corpus : Transcriptions of TED Talks
Language Pair :
{English, French, Brazilian Portuguese}-to-Spanish
{English, Spanish, Brazilian Portuguese}-to-French
{English, Spanish, French}-to-Brazilian Portuguese
Training sentences : 164K-200K (Different with languages)
Test set : incomplete

BLEU by one-to-one and multi-source NMT

Task	One-to-one (En-to-target)	Multi-encoder	Mix. NMT Experts
{En, Fr, Pt(br)}-to-Es	24.32	**26.01** (+1.69)	25.51 (+1.19)
{En, Es, Pt(br)}-to-Fr	24.54	25.62 (+1.08)	**26.23** (+1.69)
{En, Es, Fr}-to-Pt(br)	25.14	**27.36** (+2.22)	26.39 (+1.25)

Multi-source > One-to-one

The additional use of incomplete corpora is beneficial in multi-source NMTs even if test data is incomplete

3. Future Work

- The relation of the languages included in the multiple sources
- The relation of the number of missing inputs

References

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.
Ekaterina Garmash and Christof Monz. 2016. Ensemble Learning for Multi-Source Neural Machine Translation . In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1409–1418, Osaka, Japan. The COLING 2016 Organizing Committee.