Neural Machine Translation Training in a Multidomain Scenario

Hassan Sajjad, Nadir Durrani, Fahim Dalvi, Yonatan Belinkov[1], Stephan Vogel

QCRI, Doha, Qatar [1]MIT CSAIL, Cambridge, MA, USA

Introduction

In this work, we study domain adaptation for Neural Machine Translation (NMT) and target the following **questions**:

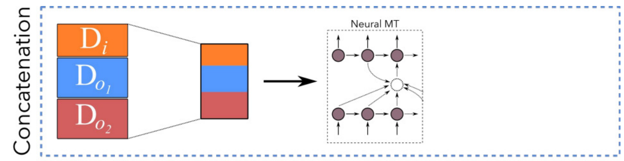
- What are different ways to combine multiple domains during a training process?
- How to build an optimal in-domain system?
- How to obtain a robust system that works best for several domains?
- What is the best strategy under time constraints?

Data and Experimental Setup

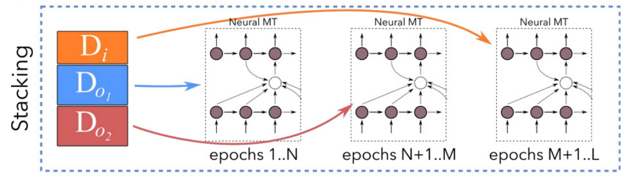
- **Arabic-English corpora**
  - TED (in-domain)
  - UN
  - OPUS
- **German-English corpora**
  - TED (in-domain)
  - EP
  - CC
- **NMT settings**
  - Nematus toolkit
  - 2-layered bidirectional LSTM with attention
  - Embedding size 512
  - Hidden layer size 1000
  - BPE 50,000
  - Vocabulary of TED talks only

Methodology

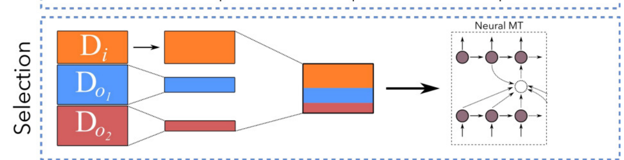
Train a system by concatenating all the available in-domain and out-of-domain data



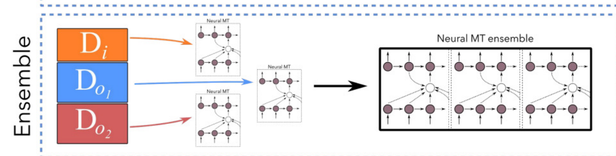
Build NMT in an online fashion starting from the most distant domain, fine-tune on the closer domain and finish by fine-tuning the model on the in-domain data



Select a certain percentage of the available out-of-domain corpora that is most closest to the in-domain data and use it for training the system



Separately train models for each available domain and combine them during decoding using balanced or weighted averaging



Results

Our Findings

- A concatenated system fine-tuned on the in-domain data achieves the most optimal in-domain system
- Model stacking works best when starting from the furthest domain, fine-tuning on closer domains and then finally fine-tuning on the in-domain data

	ALL	OD→TED	UN→OPUS→TED
Arabic-English			
tst13	36.1	37.9	36.8
tst14	30.2	32.1	31.2
avg.	33.2	35.0	34.0
German-English	ALL	OD→TED	EP→CC→TED
tst13	35.7	38.1	36.8
tst14	30.8	32.8	31.7
avg.	33.3	35.4	34.3

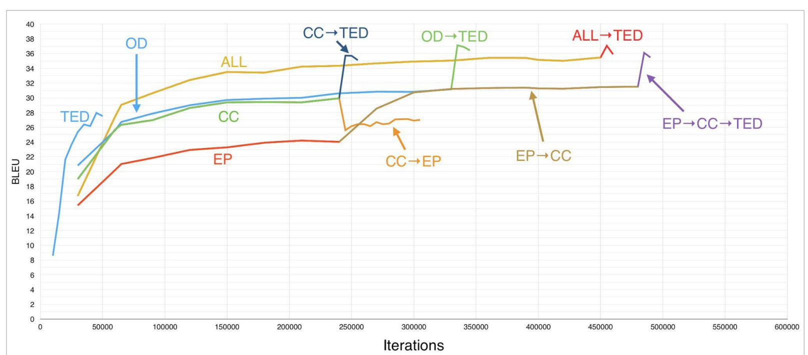
- A concatenated system on all available data results in the most robust system
- Data selection gives a decent trade-off between translation quality and training time

	Arabic-English ALL	Arabic-English Selected	German-English ALL	German-English Selected
tst13	36.1	32.7	35.7	34.1
tst14	30.2	27.8	30.8	29.9
avg.	33.2	30.3	33.3	32.0

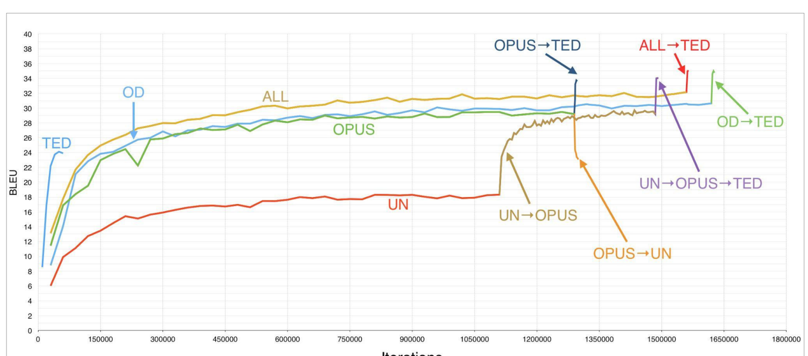
- Weighted ensemble is helpful when several individual models have been already trained and there is no time for retraining/fine-tuning

Arabic-English	OPUS	ALL	ENS_b	ENS_w
tst13	32.2	36.1	31.9	34.3
tst14	27.3	30.2	25.8	28.6
avg.	29.7	33.2	28.9	31.5

German-English



Arabic-English



Summary

- We explored several approaches to train NMT systems under multi-domain scenario: Best system is obtained by training system on the entire data and fine-tuning with the in-domain model
- Data selection is helpful under time constraint scenarios

Future Work

- We would like to explore domain adaptation under various vocabulary settings; in-domain vocabulary, out-of-domain vocabulary, large general vocabulary
- Another interesting direction to look at is to explore ways to dynamically adapt the vocabulary of an already trained model in favor of the in-domain data