Curriculum Learning for Domain Adaptation in Neural Machine Translation

Xuan Zhang[1], Pamela Shapiro[1], Gaurav Kumar[1], Paul McNamee[1], Marine Carpuat[2], Kevin Duh[1]

[1]Johns Hopkins University [2]University of Maryland

Overview

Two significant **challenges** for neural machine translation (NMT):
Domain mismatch & scarce in-domain data

Existing **data selection** methods for domain adaptation:
- Assume large unlabeled-domain corpus, select subset that is similar to in-domain text
- Problem: No clear-cut way to define whether a sample is sufficiently similar to in-domain data to be included in training; need to try different thresholds

Our approach: Feed all unlabeled-domain corpus to training, with a curriculum schedule based on similarity scores. "Good" samples are fed earlier and more often.

Domain Similarity Scoring

I: in-domain corpus N: unlabeled-domain corpus s: sentence in N

Moore-Lewis Method (Moore and Lewis, 2010)

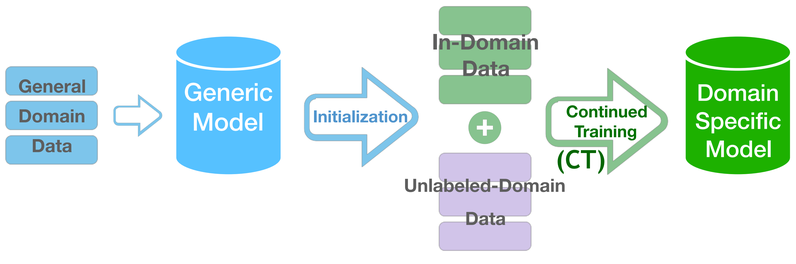
$$H_I(s) - H_N(s)$$

$H_I(s)$: *cross-entropy of s according to a language model trained on corpus I*

Cynical Data Selection (Axelrod, 2017)

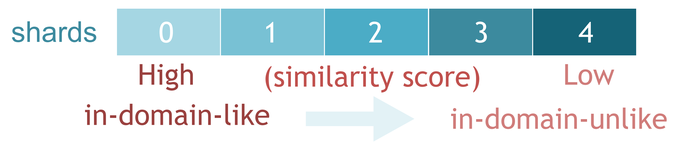
Iteratively select s which most decreases the *cross-entropy* between previously selected sentences and I.

Work Flow of Domain Apdaptation System

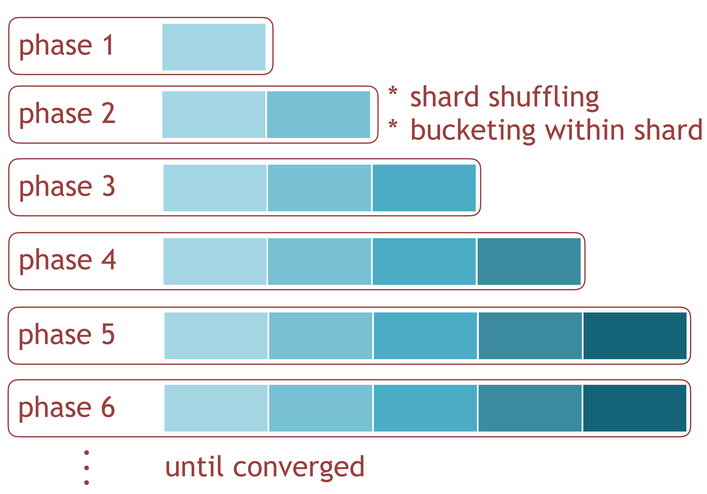


Curriculum Learning Training Strategy

- Sentences are ranked by similarity scores and distributed evenly into *shards*.



- The training process is segmented into consecutive phases, where only a subset of shards are available for training.



- The training set is increased gradually by adding shards with samples of lower similarity scores into it.

* *The presentation order of samples is not deterministic:*

(1) Shards within one curriculum phase are shuffled.

(2) Samples within one shard are bucketed by length and batches are drawn randomly from buckets.

code: https://github.com/kevinduh/sockeye-recipes/tree/master/egs/curriculum

Experiments and Results

Data

Language pairs: de-en, ru-en
- General domain data (51 million): OpenSubtitles2018, WMT2017 (Europarl, UN Parallel Copus, news commentary, Rapid corpus, Common Crawl, Yandex, Wikipedia titles)
- In-domain data (15k): TED talks, Patents
- unlabeled-domain data (13.6 million de-en, 3.7 million ru-en): web-crawled bitext from Paracrawl project

Performance of models trained with in-domain & 4096k(de) / 2048k(ru) Paracrawl samples

	TED(de)	TED(ru)	patent(de)	patent(ru)
IN: CT on in-domain only	36.16	25.04	54.70	35.61
std_rand: standard CT on a random subset of training data	35.32	24.33	50.00	34.70
std_ML: std CT with ML scores	36.02	24.73	50.40	30.96
CL_ML: CL CT with ML scores	38.78	26.45	52.91	34.18
△_ML	2.76	1.72	2.51	3.22
std_CDS: std CT with CDS scores	35.83	24.60	52.58	34.54
CL_CDS: CL CT with CDS scores	**38.88**	**26.49**	**55.51**	**36.59**
△_CDS	3.05	1.89	2.93	2.05

BLEU of models using a concatenation of in-domain and varying amounts of Paracrawl data