Dual, Monolingual, Cross-Entropy-Delta Filtering of Noisy Parallel Data

Amittai Axelrod
DiDi Labs, Los Angeles
amittai@didiglobal.com

Anish Kumar Steve Sloto

- **One laptop, 3 hours.**
- **Nothing pretrained**
- **No parallel data**
- **No multilingual tools**
- Uses relative informativeness of data
- Planet-friendly, and surprisingly decent!

(…considering you have no parallel data)

Inspiration:

Filtering with Dual Conditional Cross-Entropy

[Junczys-Dowmunt, 2018]

Dual MT systems:
parallel inputs → identical likelihoods

Works because systems trained on the same parallel corpus. Both halves contain the same information = both systems contain same info.

What if no parallel data?

Dual monolingual corpora:
Same amount of information in each language,
…even if it is different information.

Adding half of a parallel sentence should also add same amount of information to both sides.

Score sentences based on monolingual entropy deltas: **are both halves equally informative** with respect to their monolingual set?

For Sinhala-English shared task: Randomly pick English data to match size of Sinhala, but same domain ratio. Use SentencePiece to force identical vocab size. Train LM, verify perplexities match on held-out (provided) parallel test set.

Corpus	Lines	Tokens
Sinhala Wikipedia	156k	4.7M
Sinhala Common Crawl	5.2M	110M
EN Random Wikipedia	150k	5.5M
EN Random Common Crawl	6M	123M

Corpus	valid	test
Sinhala, untok	1,759.8	1,565.9
English, untok	1,069.2	985.3
Sinhala, tok=SP	**320.5**	**299.2**
English, tok=SP	**302.5**	**292.7**

Cross-Entropy Delta

Measures incremental change in information of a single model. Not the same as cross-entropy difference, which is between two static models. Used as the selection criterion inside Cynical data selection (Axelrod, 2017).

How much would the perplexity of a test set change if a new sentence *s* was added to the training set of a language model?

$$\underset{n \to n+1}{\Delta H} = H_{n+1} - H_n$$

$$\underset{n \to n+1}{\Delta H} = \underbrace{\log \frac{W_n + w_{n+1}}{W_n}}_{Penalty} + \underbrace{\sum_{v \in V_{\text{REPR}}} \frac{C_{\text{REPR}}(v)}{W_{\text{REPR}}} \log \frac{C_n(v)}{C_n(v) + c_{n+1}(v)}}_{Gain}$$

Entropic penalty for increasing size of training set, and potential entropy improvement for adding new information to corpus.

Use monolingual estimates of information gain instead of translation model likelihoods!

$$|\Delta H_{En}(s_{En}|\text{MONO}_{En}) - \Delta H_{Si}(s_{Si}|\text{MONO}_{Si})|$$
$$+\frac{1}{2}\left(\Delta H_{En}(s_{En}|\text{MONO}_{En}) + \Delta H_{Si}(s_{Si}|\text{MONO}_{Si})\right)$$

Results:

Successfully filtered parallel data with no parallel data! Not the best system, but it works!
Bonus: Everything runs quickly on a single laptop!

System	1M SMT	1M NMT	5M SMT	5M NMT
Rank 1	4.27	6.39	4.94	4.44
DiDi	2.53	0.19	3.70	0.20
Rank 10	0.92	0.03	2.73	0.10

What happened with NMT?

User error! Length ratio feature too fancy, did not work.
System cutoffs only based on English side! Our submissions:
740k Si -> 1M En
3.6M Si -> 5M En
NMT hallucinates to produce enough output words;
BLEU score is zero. Important to get the heuristic scores right!