Data Selection with Fewer Words

Amitai Axelrod
Philip Resnik
Xiaodong He
Mari Ostendorf

University of Maryland & Johns Hopkins
University of Maryland
Microsoft Research
University of Washington
Domain* Adaptation

* Defined by construction.

Ideally based on some notion of textual similarity:

- Lexical choice
- Grammar
- Topic
- Style
- Genre
- Register
- Intent

Domain = particular contextual setting.
Here we use “domain” to mean “corpus”.

Amittai Axelrod  Data Selection with Fewer Words  WMT 2015 2
Domain Adaptation

- Training data doesn’t always match desired tasks.
- Have bilingual:
  - Parliament proceedings
  - Newspaper articles
  - Web scrapings
- Want to translate:
  - Travel scenarios
  - Facebook updates
  - Realtime conversations
- Sometimes want a specific kind of language, not just breadth!
Data Selection

- "filter Big Data down to Relevant Data"

- Use your regular pipeline, but improve the input!

- Not all sentences are equally valuable...
Data Selection

• For a particular translation task:
  • Identify the most relevant training data.
  • Build a model on only this subset.

• Goal:
  • Better task-specific performance
  • Cheaper (computation, size, time)
Data Selection Algorithm

- Quantify the domain
- Compute similarity of sentences in pool to the in-domain corpus
- Sort pool sentences by score
- Select top n%
Data Selection Algorithm

- Quantify the domain
- Compute similarity of sentences in pool to the in-domain corpus
- Sort pool sentences by score
- Select top n%
- Use n% to build task-specific MT system
- Combine with system trained on in-domain data (optional)
- Apply task-specific system to task.
Perplexity-Based Filtering

- A language model $LM_Q$ measures the likelihood of some text by its perplexity:

\[
ppl_{LM_Q}(s) = 2^{-\frac{1}{N} \sum_{i=1}^{N} \log LM_Q(w_i|h_i)} = 2^{H_{LM_Q}(s)}
\]

- Intuition: Average branching factor of LM
- Cross-entropy $H$ (of a text w.r.t. an LM) is $\log(\text{ppl})$. 
Cross-Entropy Difference

- Perplexity-based filtering:
  - Score and sort sentences in pool by perplexity with in-domain LM.
  - Then rank, select, etc.
- However! By construction, the data pool does not match the target task.
Cross-Entropy Difference

- Score and rank by cross-entropy difference:

$$\text{argmin}_{s \in \text{POOL}} H_{LM_{\text{IN}}}(s) - H_{LM_{\text{POOL}}}(s)$$

(Also called "XEDiff" or "Moore-Lewis")

- Prefer sentences that both:
  - Are **like** the target task
  - Are **unlike** the pool average.
Bilingual Cross-Entropy Diff.

- Extend the Moore-Lewis similarity score for use with bilingual data, and apply to SMT:

\[
(H_{L1}(s_1, LM_{IN}) - H_{L1}(s_1, LM_{POOL})) + (H_{L2}(s_2, LM_{IN}) - H_{L2}(s_2, LM_{POOL}))
\]

- Training on only the most relevant subset of training data (1%-20%) yields translation systems that are smaller, cheaper, faster, and (often) better.
Using Fewer Words

- How much can we trust rare words?
- If a word is seen 2 times in the general corpus and 3 in the in-domain one, is it really 50% more likely?
- Low-frequency words often ignored (Good-Turing smoothing, singleton pruning...)
Hybrid word/POS Corpora

- In stylometry, syntactic structure = proxy for style.

- POS-tag n-grams used as features to determine authorship, genre, etc.

- Incorporate this idea as a pre-processing step to data selection:
Hybrid word/POS Corpora

- In stylometry, syntactic structure = proxy for style.
- POS-tag n-grams used as features to determine authorship, genre, etc.
- Incorporate this idea as a pre-processing step to data selection:

  Replace rare words with POS tags
Hybrid word/POS Corpora

- Replace rare words with POS tags:
  - an earthquake in **Port-au-Prince**
  - an earthquake in **NNP**
Hybrid word/POS Corpora

- Replace rare words with POS tags:
  - an earthquake in Port-au-Prince
  - an NN in NNP
Hybrid word/POS Corpora

- Replace rare(?) words with POS tags:
  - an earthquake in Port-au-Prince

  - DT  NN  IN  NNP
Hybrid word/POS Corpora

- Replace rare words with POS tags:
  - an earthquake in Port-au-Prince
  - an earthquake in NNP
  - an earthquake in Kodari
Hybrid word/POS Corpora

- Replace rare words with POS tags:
  - an earthquake in Port-au-Prince
  - an earthquake in NNP
  - an earthquake in Kodari
- Threshold: (if $Count < 10$) in either corpus
Using Fewer Words

- Use the hybrid word/POS texts instead of the original corpora.

- Train LMs on the corpora, compute sentence scores, and re-rank the original general corpus.

- Standard Moore-Lewis / Cross-entropy diff, but with different corpus representation.
TED Zh-En Translation

- Task: Translate TED talks, Chinese-to-English, using LDC data (6m sentence pairs).

- Vocabulary reduction from TED+LDC: Eliminate 97% of the vocabulary

- What happens to SMT performance?

<table>
<thead>
<tr>
<th>Lang</th>
<th>Vocab</th>
<th>Kept</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>En</td>
<td>470,154</td>
<td>10,036</td>
<td>2.1%</td>
</tr>
<tr>
<td>Zh</td>
<td>729,283</td>
<td>11,440</td>
<td>1.5%</td>
</tr>
</tbody>
</table>
Slightly better scores, despite (much) smaller selection vocab!
In-Domain Lexical Coverage

- Up to 10% more in-domain coverage
General-Domain Coverage

- Hybrid-selected data covers 10-15% more of the general lexicon.

Amitai Axelrod

Data Selection with Fewer Words

WMT 2015
Hybrid Word/POS Selection

- Must re-compute for every task/pool, but vocabulary statistics are easy.

- Aggregating the statistics for rare terms allows generalizing to other unseen words.

- Perhaps preserving sentence structure, picking up words that fill similar roles/patterns in the sentence?
Hybrid Word/POS Selection

- Replace all rare words with POS tags, then run regular data selection.

- Reduces active lexicon by 97%, to ~10k words with robust statistics

- Potentially helpful for algorithms bound by vocabulary size "V"

- Selection LM is 25% smaller
Questions?
[ this slide intentionally left blank ]