One-Shot Lexicon Learning for Low-Resource Machine Translation

Anjali Kantharuban[1] and Jacob Andreas[2]

[1]Department of Electrical Engineering & Computer Science, University of California, Berkeley
[2]Department of Electrical Engineering & Computer Science, Massachusetts Institute of Technology

Background

Machine translation models struggle to translate tokens that appear rarely in the training data. In languages with lots of data, this problem is avoided because even rare words appear many times. However, in low-resource languages, or languages without lots of data, rare words may be seen only once or twice.

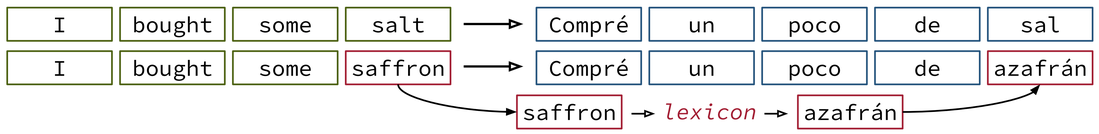


Fig. 1: An example of a word translated using the lexical translation mechanism. The context is familiar.

This has been addressed in the past using a **lexical translation mechanism**, which copies tokens from the source sequence into the right spot in the target sequence using a word-level translator and contextual information (Akyürek et al. 2021). This improves performance when the context is common, but the specific word is not. These word-level translators are called **lexicons** and are traditionally build manually, which requires human labor, or with statistical methods, which require many examples to be reasonably accurate.

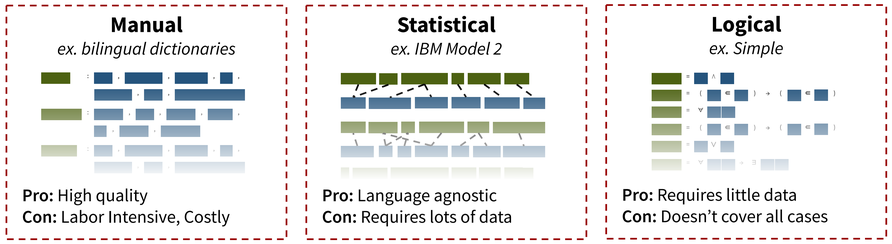
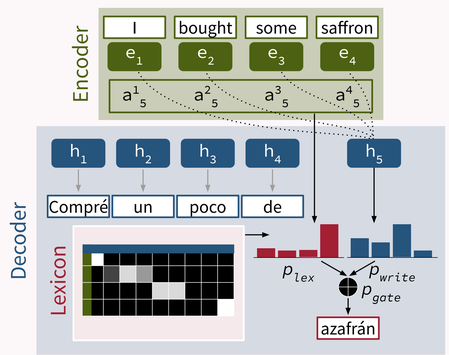


Fig. 2: Popular lexicons. These are generally pre-generated and stay static during the training of the translation network.

Lexical Alignment Model (LAM)



***Our Contribution:* Using statistical aligners to bootstrap a neural lexicon learning procedure.**

- When training the LAM, tokens are randomly masked using the *<unknown>* token to prepare the model to align words that are outside of the vocabulary
- Representations are based on both the token and the context around the token
- The model is trained on high-confidence statistical alignments from IBM Model 2's fast-align

Fig. 3: Lexical translation mechanism.

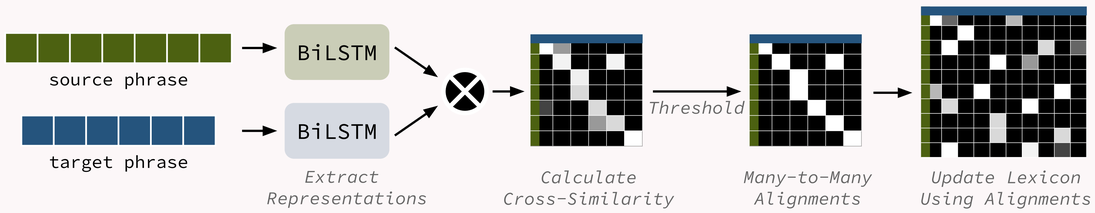


Fig. 4: Lexicon is generated from training samples by aligning tokens in each sentence and adjusting scores using alignments.

Experiments

English-Spanish
484,080 sentence pairs

- Relatively similar sentence structure, making it easier to align on a sentence level
- High-quality data from a variety of sources
- Dataset size reduced to simulate a low-resource domain

English-Tamil
684,029 sentence pairs

- Different word ordering and morphology, making it difficult to align on a sentence level
- Lower-quality data, primarily from government and religious documents

All data is from the Tatoeba dataset (Tiedemann 2020)

Results

	English-Spanish		English-Tamil	
	Full	*One-Shot*	*Full*	*One-Shot*
No Lexicon	21.48	18.82	6.12	4.79
IBM Model 2 (Brown et al., 1993)	26.32	24.58	7.68	6.17
Pre-Trained LAM	26.42	24.78	8.40	6.72

Fig. 5: BLEU scores for each dataset. BLEU scores are a way of determining the quality of translated text, where higher scores are better. It works by counting the number of segments in the output sentence that perfectly match the reference sentence.

Currently, we are comparing our alignment model against two baselines.

1. A base encoder-decoder translation model with no lexical translation mechanism.
2. The same base model with a lexicon generated using IBM Model 2's fast align, adapted to generate vocabulary level alignments.

We compare across the full test dataset and a dataset which contains only sentences which have words that appear only once in the training data (a.k.a. our one-shot dataset).

Discussion

At the moment our model shows modest performance improvements on English→Spanish and greater performance improvements on English→Tamil. We plan to do further experiments across both dataset size and language typology to see what features are contributing to this difference. High performance on this task provides the ability for translation models to be more accessible for speakers of low-resource languages.

Future Work

There are a few ablations of interest I would like to explore:

1. Testing performance on words seen once, twice, thrice, etc. in order to understand whether the improvements caused by a lexicon are limited to rare words
2. Jointly training the lexical alignment model alongside the translation network in order to explore a case where other alignment methods are too unreliable to offer training data
3. Evaluation on more languages with features differing from both Spanish and Tamil in order to see what typological features lend themselves to word-level translation being useful

Acknowledgements

MSRP MIT Summer Research Program
MIT
Berkeley UNIVERSITY OF CALIFORNIA