MADLIBS: A Novel Multilingual Data Augmentation Algorithm for Low-Resource Neural Machine Translation

Introduction

- While **natural language processing (NLP)** algorithms have grown rapidly, these advancements have been **almost exclusively English-centric,** relying on **gigantic amounts of data** to train
- **Low-data and multilingual settings** are of the most prominent challenges of the NLP field
- Most languages are left behind, and as **language loss** occurs at an accelerated rate, technology can help endangered languages
- I aim to develop and **inclusive** method to support **underserved communities**

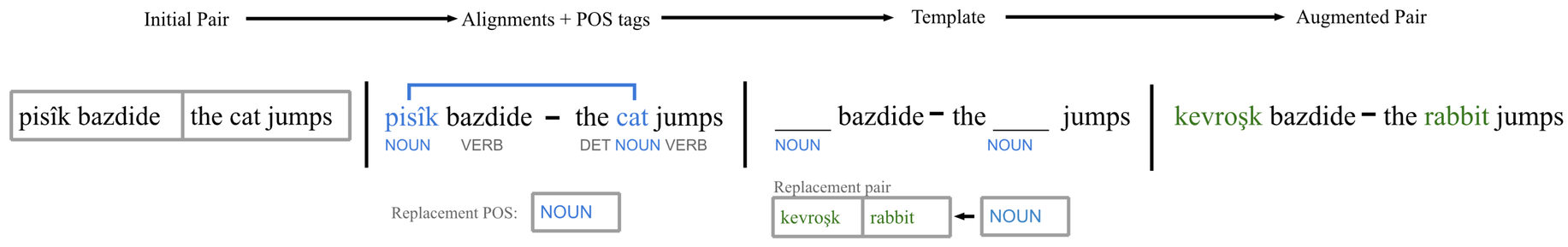
Background

- Generating synthetic multilingual text from existing sentences through **data augmentation (DA)** is valuable for low-resource neural machine translation (NMT)
- Many existing multilingual DA methods use auxiliary data and have thus been primarily applied for high-resource settings
- Effective DA methods:
  - Augment both sides of the data
  - Maintain equivalence across the languages
  - Enhance lexical and syntactic diversity

Algorithmic Design & Methodology

MADLIBS (**M**ultilingual **A**ugmentation of **D**ata with A**li**gnment-**B**ased **S**ubstitution) generates diversified and semantically consistent sentence pairs without auxiliary data.

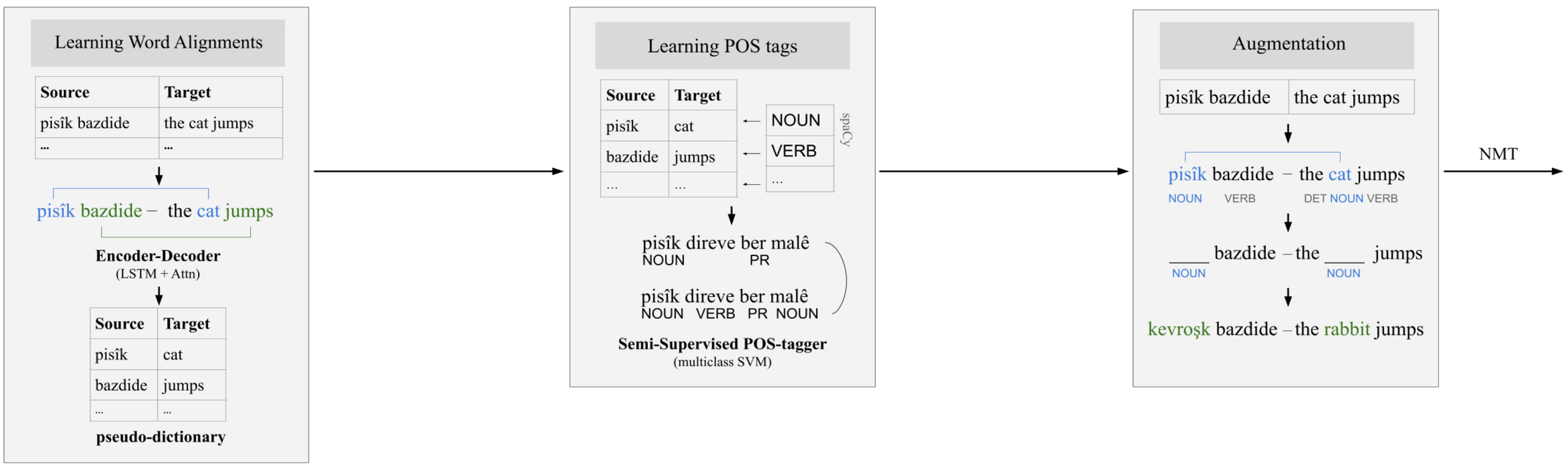
MADLIBS replaces word pairs from existing bilingual sentences simultaneously from both languages with suitable substitutions, following foundational linguistic principles. This process, inspired by the children's word game, is conducted automatically in a low-resource through multiple key NLP algorithms to learn **word alignments**, categorize **parts-of-speech**, and **construct sentences**.



Word Alignments. Word alignments across the languages are first learned from the small parallel corpora. The model follows an attention-based encoder-decoder architecture, trained jointly in a cross-lingual space for efficiency. I build a pseudo-bilingual dictionary from high-likelihood alignments.

POS-tagging. I generate a partially annotated dataset using the pseudo-dictional by exploiting the known parts-of-speech in the target high-resource side of the data with spaCy's POS-tagging, and mapping to the low-resource language. I then train a SVM-based semi-supervised POS-tagging model.

Template Pipeline. Given a source-target sentence pair (x,y), the template generator selects a POS and replaces a word alignment pair (x_i,y_j) of that POS within the sentences with a random pair of the same POS (x_i',y_j') from the pseudo-dictionary, weighted for rare words. This is repeated multiple times.



Results

Dataset Pairs	eng→uig 143K	eng→pag 146K	eng→mri 221K	eng→ita 200K
Baseline	0.5	2.6	7.8	19.5
BT	0.4 (-0.1)	5.8 (+3.2)	8.8 (+1.0)	20.9 (+1.4)
MADLIBS	**1.0** (+0.5)	**6.0** (+3.4)	**9.3** (+1.5)	**22.5** (+3.0)

BLEU performance of NMT systems with gains from baseline across languages. eng→uig, eng→pag, and eng→mri refer to translation from English to Uyghur, Pangasinan, and Māori. eng→ita is the simulated low-resource setting to Italian.

NMT performance with MADLIBS consistently shows significant improvements upon the baseline across a range of diverse language tasks. Without the use of any external data, the approach surpasses the gains of back-translation, one of the current most established and commonly employed DA method in NMT. In fact, in the eng→uig task, MADLIBS surpasses the current top results as reported on the OPUS-MT leaderboard for the test set.

Conclusion

- I have proposed an effective approach to augment the training data of multilingual NLP for low-resource languages without the use of auxiliary data
- I generate new diversified sentence pairs with aligned substitutions, enhancing data diversity
- The method surpassed the results of the most popular and established existing data augmentation method for NMT, even without the use of external data or transfer learning, making the method highly valuable
- The work is a fundamentally unique approach towards one of the greatest persisting and pervasive challenges of deep learning–low-resource learning–and in one of the most complex areas for data augmentation–multilingual textual data
- This method, expanded for accessibility for global communities, can be used to support endangered and minority languages

all figures created by student researcher