CS11-737 Multilingual NLP

Data-based Strategies to Low-resource MT

Graham Neubig

Carnegie Mellon University
Language Technologies Institute

Site
http://demo.clab.cs.cmu.edu/11737fa20/

Many slides from:
Data Challenges in Low-resource MT

• MT of high-resource languages (HRLs) with large parallel corpora → relatively good translations

• MT of low-resource languages (LRLs) with small parallel corpora → nonsense!
A Concrete Example

A system that is trained with 5000 sentence pairs on Azerbaijani and English?

**source** - Atam balaca boz radiosunda BBC Xəbərlərində qulaq aşdı.

**translation** - So I’m going to became a lot of people.

**reference** - My father was listening to BBC News on his small, gray radio.

Does not convey the correct meaning at all.
Multilingual Training Approaches

- **Transfer HRL to LRL** (Zoph et al., 2016; Nguyen and Chiang, 2017)

- **Joint training with LRL and HRL parallel data** (Johnson et al., 2017; Neubig and Hu, 2018)

  ![Diagram]

  **Problem:** Suboptimal lexical/syntactic sharing.
  **Problem:** Can't leverage monolingual data.
Data Augmentation

Available Resources

TRG-M
TRG-H
TRG-L
HRL
LRL

Augmented Data

TRG
LRL
Data Augmentation 101: Back Translation

TRG -> LRL

TRG-M

TRG-H

TRG-L

HRL

LRL

LRL-M

TRG-M
Back Translation Idea

1. Train TRG → LRL System

2. Back-translate TRG → LRL

3. Train LRL → TRG

- Some degree of error in source data is permissible!

How to Generate Translations

- How to generate translations?
- **Beam search** (Sennrich et al. 2016)
  - Select the highest scoring output
  - Higher quality, but lower diversity, potential for data bias
- **Sampling** (Edunov et al. 2018)
  - Randomly sample from back-translation model
  - Lower overall quality, but higher diversity
- Sampling has shown to be more effective overall

Iterative Back-translation

1. Train LRL → TRG System
2. Forward-translate LRL-TRG
3. Train TRG → LRL System
4. Back-translate TRG-LRL
5. Final LRL → TRG System

Back Translation Issues

• Back-translation fails in low-resource languages or domains

  • Use other high-resource languages

  • Combine with monolingual data (maybe with denoising objectives, covered in following class)

  • Perform other varieties of rule-based augmentation
Using HRLs in Augmentation

English -> HRL Augmentation

- **Problem:** TRG-LRL back-translation might be low quality

- **Idea:** also back-translate into HRL
  - more sentence pairs
  - vocabulary sharing of source-side
  - syntactic similarity of source-side
  - improves target-side LM

TRG: Thank you very much.

TUR: Çok teşekkür ederim.

AZE: Ho Ho Ho.
Available Resources + TRG-LRL and TRG-HRL Back-translation

TRG -> LRL

TRG-M

TRG-H

TRG-L

LRL-M

HRL-M

TRG-M
Augmentation via Pivoting

- **Problem:** HRL-TRG data might suffer from lack of lexical/syntactic overlap

- **Idea:** Translate existing HRL-TRG data
  - Translate from HRL to LRL

**TUR:** Çok teşekkür ederim.  
**TRG:** Thank you so much.  
**AZE:** Çok sağ olun.  
**TRG:** Thank you so much.
Available Resources + TRG-LRL and TRG-HRL Back-translation + Pivoting

TRG -> LRL

TRG -> HRL

HRL -> LRL

LRL-M → TRG-M

LRL-H → TRG-H
Back-Translation by Pivoting

• **Problem:** TRG-HRL back-translated data also suffers from lexical or syntactic mismatch

• **Idea:** TRG-HRL-LRL
  ○ Large amount of English monolingual data can be utilized

TRG: Thank you so much.

TUR: Çok teşekkür ederim.

TRG: Thank you so much.

AZE: Çox sağ olun.

TRG: Thank you so much.
Data w/ Various Types of Pivoting

TRG-M

TRG-H

TRG-L

HRL

LRL

HRL-M

LRL-M

LRL-MH

LRL-H

TRG-M

TRG-H

TRG-M

TRG-H
Monolingual Data Copying
**Monolingual Data Copying**

- **Problem:** Back-translation may help with structure, but fail at terminology
- **Idea:** Use monolingual data *as-is*
  - Helps encourage the model to not drop words
  - Helps translation of terms that are identical across languages

---

Heuristic Augmentation Strategies
Dictionary-based Augmentation

1. Find rare words in the source sentences
2. Use a language model to predict another word that could appear in that context

<table>
<thead>
<tr>
<th>Sentence [original / substituted]</th>
<th>Plausible</th>
</tr>
</thead>
<tbody>
<tr>
<td>My sister drives a [car / motorbike]</td>
<td>yes</td>
</tr>
<tr>
<td>My uncle sold his [house / motorbike]</td>
<td>yes</td>
</tr>
<tr>
<td>Alice waters the [plant / motorbike]</td>
<td>no (semantics)</td>
</tr>
<tr>
<td>John bought two [shirts / motorbike]</td>
<td>no (syntax)</td>
</tr>
</tbody>
</table>

3. Replace, and aligned word with translation from dictionary

An Aside: Word Alignment

- Automatically find alignments between source and target words for dictionary learning, analysis, supervised attention etc.

- **Traditional symbolic methods:** word-based translation models trained using EM algorithm
  - GIZA++: https://github.com/moses-smt/giza-pp
  - FastAlign: https://github.com/clab/fast_align

- **Neural methods:** use model like multilingual BERT or translation and find words with similar embeddings
  - SimAlign: https://github.com/cisnlp/simalign
Word-by-word Data Augmentation

- Even simpler, translate sentences word-by-word into target sentence using dictionary
  
  J'ai acheté une nouvelle voiture
  ↓ ↓ ↓ ↓ ↓
  I bought a new car

- **Problem:** what about word ordering, syntactic divergence?
  
  私 は 新しい 車 を 買った
  ↓ ↓ ↓ ↓ ↓
  I the new car a bought

Word-by-word Augmentation w/ Reordering

- **Problem:** Source-target word order can differ significantly in methods that use monolingual pre-training.

- **Solution:** Do re-ordering according to grammatical rules, followed by word-by-word translation to create pseudo-parallel data.

In-class Assignment
In-class Assignment

• Read one of the cited papers on heuristic data augmentation


• Try to think of how it would work for one of the languages you're familiar with

• Are there any potential hurdles to applying such a method? Are there any improvements you can think of?