Unsupervised Machine Translation

Sachin Kumar

Carnegie Mellon University
Language Technologies Institute
Conditional Text Generation

- Generate text according to a specification: $P(Y|X)$

<table>
<thead>
<tr>
<th>Input X</th>
<th>Output Y (Text)</th>
<th>Task</th>
</tr>
</thead>
<tbody>
<tr>
<td>English</td>
<td>Hindi</td>
<td>Machine Translation</td>
</tr>
<tr>
<td>Image</td>
<td>Text</td>
<td>Image Captioning</td>
</tr>
<tr>
<td>Document</td>
<td>Short Description</td>
<td>Summarization</td>
</tr>
<tr>
<td>Speech</td>
<td>Transcript</td>
<td>Speech Recognition</td>
</tr>
</tbody>
</table>

[Slide Credits: Graham Neubig]
Modeling: Conditional Language Models

How to estimate model parameters?

- Maximum Likelihood Estimation
- Needs supervision -> parallel data! Usually millions of parallel sentences
What if we don’t have parallel data?

<table>
<thead>
<tr>
<th>Input X</th>
<th>Output Y</th>
<th>Task</th>
</tr>
</thead>
<tbody>
<tr>
<td>Image (Photo)</td>
<td>Image (Painting)</td>
<td>Style Transfer</td>
</tr>
<tr>
<td>Image (Male)</td>
<td>Image (Female)</td>
<td>Gender Transfer</td>
</tr>
<tr>
<td>Text (Impolite)</td>
<td>Text (Polite)</td>
<td>Formality Transfer</td>
</tr>
<tr>
<td>English</td>
<td>Sinhalese</td>
<td>Machine Translation</td>
</tr>
<tr>
<td>Positive Review</td>
<td>Negative Review</td>
<td>Sentiment Transfer</td>
</tr>
</tbody>
</table>
Can’t we just collect/generate the data?

- Too time consuming/expensive 🤑

- Difficult to specify what to generate (or evaluate the quality of generations)
  - Generate text like Donald Trump

- Asking annotators to generate text doesn’t usually lead to good quality datasets
Unsupervised Machine Translation

Previous Lectures:

1. How can we use monolingual data to improve an MT system
2. How can we reduce the amount of supervision (or make things work when supervision is scarce)

This Lecture:

Can we learn WITHOUT ANY supervision
Outline

1. Core concepts in Unsupervised MT
   a. Initialization
   b. Iterative Back Translation
   c. Bidirectional model sharing
   d. Denoising auto-encoding

2. Open Problems/Advances in Unsupervised MT

Unsupervised machine translation using monolingual corpora only. Lample et al. ICLR 2018
Phrase-Based & Neural Unsupervised Machine Translation. Lample et al. EMNLP 2018
Unsupervised Neural Machine Translation. Artetxe et al ICLR 2018
Step 1: Initialization

- Prerequisite for unsupervised MT:
  - To add a good prior to the state of solutions we want to reach
  - Kickstarting the solution - use approximate translations of sub-words/words/phrases

- the context of a word, is often similar across languages since each language refers to the same underlying physical world.
Initialization: Unsupervised **Word** Translation

- **Hypothesis:** Word embedding spaces in two languages are isomorphic
  - One embedding space can be linearly transformed into another
  - Give monolingual embeddings $X$ and $Y$, learn a (orthogonal) matrix, such that, $WX = Y$
Unsupervised **Word** Translation: Adversarial Training

- Use adversarial learning to learn W:
  - If WX and Y are perfectly aligned, a discriminator shouldn’t be able to tell
  - Discriminator: Predict whether an embedding is from Y or the transformed space WX.
  - Train W to confuse the discriminator
Step 2: Back-translation

- Models never see bad translations only bad inputs
- Generate back-translated data, train model in both directions, repeat: iterative back-translation

[Slide credits: Graham Neubig]
Applying these steps to non-neural MT
One slide primer on phrase-based statistical MT

- Foreign input is segmented in phrases
  - any sequence of words, not necessarily linguistically motivated
- Each phrase is translated into English
- Phrases are reordered

[Statistical Phrase-Based Translation. Koehn, Och and Marcu. NAACL 2003]
Unsupervised Statistical MT

- Learn monolingual embeddings for unigram, bigram and trigrams
- Initialize phrase-tables from cross-lingual mappings
- Supervised training based on back-translation
- Iterate

[Artetxe et al 2018, Lample et al 2018]
### Unsupervised Statistical MT

<table>
<thead>
<tr>
<th></th>
<th>en → fr</th>
<th>fr → en</th>
<th>en → de</th>
<th>de → en</th>
<th>en → ro</th>
<th>ro → en</th>
<th>en → ru</th>
<th>ru → en</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unsupervised phrase table</td>
<td>-</td>
<td>17.50</td>
<td>-</td>
<td>15.63</td>
<td>-</td>
<td>14.10</td>
<td>-</td>
<td>8.08</td>
</tr>
<tr>
<td>Back-translation - Iter. 1</td>
<td>24.79</td>
<td>26.16</td>
<td>15.92</td>
<td>22.43</td>
<td>18.21</td>
<td>21.49</td>
<td>11.04</td>
<td>15.16</td>
</tr>
<tr>
<td>Back-translation - Iter. 2</td>
<td>27.32</td>
<td>26.80</td>
<td>17.65</td>
<td>22.85</td>
<td>20.61</td>
<td>22.52</td>
<td>12.87</td>
<td>16.42</td>
</tr>
<tr>
<td>Back-translation - Iter. 4</td>
<td>27.84</td>
<td>27.20</td>
<td>17.77</td>
<td>22.68</td>
<td>21.33</td>
<td>23.01</td>
<td>13.37</td>
<td>16.62</td>
</tr>
<tr>
<td>Back-translation - Iter. 5</td>
<td><strong>28.11</strong></td>
<td>27.16</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>
Unsupervised Neural MT
Step 3: Bidirectional Modeling

- Model: **same** encoder-decoder used for both languages
- Initialize with cross-lingual word embeddings

[Diagram showing bidirectional modeling with I am a student and Je suis étudiant]<br>

[Slide credits: Kevin Clark]
Unsupervised MT: Training Objective 1

Denoising autoencoder
Unsupervised NMT: Training Objective 2

- Back-translation
  - Translate target to source
  - Use as a “supervised” example to translate source to target

[Lample et al 2018, Artetxe et al 2018]
How does it work?

- Cross lingual embeddings and a shared encoder gives the model a good starting point
Unsupervised MT

- Training Objective 3: Adversarial
  - Constraining the encoder to map the two languages in the same feature space

Auto-encoder example

I am a student

Encoder vector

I am a student

Back-translation example

Je suis étudiant

Encoder vector

I am a student

need to be the same!

[Lample et al 2018]
Performance

- Horizontal Lines are unsupervised, rest are supervised
In summary

- Initialization is important
  - To introduce biases

- Need Monolingual data
  - Both of good initialization/alignments and learning a language model

- Iterative refinement
  - Noisy data-augmentation
Open Problems with Unsupervised MT
When Does Unsupervised Machine Translation Work?

- In sterile environments
  - Languages are fairly similar languages written with similar writing systems.
  - Large monolingual datasets are in the same domain and match the test domains
- On less related languages, truly low resource languages, diverse domains, or less amounts of monolingual data UMT performs less well.

<table>
<thead>
<tr>
<th></th>
<th>En-Turkish</th>
<th>Ne-En</th>
<th>Si-En</th>
</tr>
</thead>
<tbody>
<tr>
<td>Supervised</td>
<td>20</td>
<td>7.6</td>
<td>7.2</td>
</tr>
<tr>
<td>UNMT</td>
<td>4.5</td>
<td>0.2</td>
<td>0.4</td>
</tr>
</tbody>
</table>

Reasons for this poor performance

Graham Neubig
@gneubig

Reasons why real low-resource languages make unsupervised MT difficult: (1) small monolingual data -> bad embeddings, (2) different word frequencies/morphology hurt bilingual lexicon induction, (3) different content makes sentence-level distribution matching difficult.
Open Problems

● Diverse languages and domains.
  ○ Better cross-lingual initialization: better data selection/regularization in pretraining language models

● What if no (or very little) monolingual data is available.
  ○ A tiny amount of parallel data goes a long way than massive monolingual data: Semi-supervised learning
  ○ Make use related languages

[When and Why is Unsupervised Neural Machine Translation Useless? Kim et al. 2020]
Better Initialization: Cross Lingual Language Models

- Cross Lingual Masked Language Modelling

- Initialize the entire encoder and decoder instead of lookup tables
- Alignment comes from shared sub-word vocabulary

[Cross-lingual Language Model Pretraining. Lample and Conneau. 2019]
Masked Sequence to Sequence Model (MASS)

- Encoder-decoder formulation of masked language modelling

[MASS: Masked Sequence to Sequence Pre-training for Language Generation. Song et al. 2019]
Multilingual BART

- Multilingual Denoised Autoencoding
- Corrupt the input and predict the clean version. Type of noise
  - Mask or swap words/phrases
  - Shuffle the order of sentences in an instance
Multilingual Unsupervised MT

- Assume, three languages X, Y, Z:
  - Goal: Translate X to Z
  - We have parallel data in (X, Y) but only monolingual data for Z.
  - (If we have parallel data for (X, Z) or (Y, Z): zero-shot translation; covered in last lecture)

- Pretrain using MASS

- Two translation objectives:
  - Back-translation: \( P(x \mid y(x)) \) [Monolingual data]
  - Cross-translation: \( P(y \mid z(x)) \) [Parallel data (x, y)]

- Shows improvement for dissimilar languages with less monolingual data

Multilingual UNMT

- Shows improvements on low resource languages

[Harnessing Multilinguality in Unsupervised Machine Translation for Rare Languages. Garcia et al. EMNLP 2020]
If some parallel data is available?

- Semi-supervised Learning

- Train the model first with unsupervised method and fine tune using the parallel corpus OR more commonly, train the model using the parallel corpus and update with iterative back-translation
Related Area: Style Transfer

- Rewrite text in the *same* language but in a different “style”

<table>
<thead>
<tr>
<th>Relaxed ↔ Annoyed</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Relaxed</strong></td>
</tr>
<tr>
<td><strong>Annoyed</strong></td>
</tr>
<tr>
<td><strong>Annoyed</strong></td>
</tr>
<tr>
<td><strong>Relaxed</strong></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Male ↔ Female</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Male</strong></td>
</tr>
<tr>
<td><strong>Female</strong></td>
</tr>
<tr>
<td><strong>Female</strong></td>
</tr>
<tr>
<td><strong>Male</strong></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Age 18-24 ↔ 65+</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>18-24</strong></td>
</tr>
<tr>
<td><strong>65+</strong></td>
</tr>
<tr>
<td><strong>65+</strong></td>
</tr>
<tr>
<td><strong>18-24</strong></td>
</tr>
</tbody>
</table>

Discussion Question

Pick a low resource language or dialect and argue whether unsupervised MT will be a suitable for translating to it (from English).

If yes, why? If not, what could be potential solutions