Transfer Learning for Unsupervised Neural Machine Translation

Alexandra Chronopoulou
achron@cis.lmu.de
CIS, LMU Munich
11/07/2023
1 Motivation for Transfer Learning
2 Recap: Unsupervised Neural Machine Translation
3 Transfer Learning for NMT
4 Transfer Learning for Unsupervised NMT
   • Language Model Pretraining
   • Bilingual Language Model Pretraining
   • Continual Pretraining
   • Parallel Data from Similar Language Pairs
   • Limitations
5 Conclusion
Motivation for Transfer Learning

Problems (especially in deep learning):
- Scarcity of labeled data
- Models trained on small datasets often fail to generalize in test data → overfit

Transfer learning:
- Uses knowledge from a *learned* task to improve the performance on a *related* task
- Scarcity of labeled data → implicit data augmentation
- Helps a model generalize → avoid overfitting
In Natural Language Processing tasks:

- **Out-of-context** pretrained word representations were used (word2vec, fasttext) to initialize the embedding layer
In Natural Language Processing tasks:

- **Out-of-context** pretrained word representations were used (word2vec, fasttext) to initialize the embedding layer
- Recently: **contextual** representations from language models (ChatGPT, GPT3, RoBERTa) are used to initialize the full model
Recap: Unsupervised Neural Machine Translation

Presentation Outline

1. Motivation for Transfer Learning
2. Recap: Unsupervised Neural Machine Translation
3. Transfer Learning for NMT
4. Transfer Learning for Unsupervised NMT
   - Language Model Pretraining
   - Bilingual Language Model Pretraining
   - Continual Pretraining
   - Parallel Data from Similar Language Pairs
   - Limitations
5. Conclusion
Supervised Learning methods in NMT work really well ... if a lot of parallel data available!

- We are provided the ground truth
- We use encoder-decoder models to
  - encode a sentence written in language x (hidden representation s)
  - provide s to decoder, it generates the sentence in language y → y'
  - compute training loss (by comparing translation y’ to ground truth y)

Figure: Seq2seq architecture for En-De NMT. Figure from https://smerity.com/articles/2016/google_nmt_arch.html
Why do we care about **Unsupervised Learning**?

- NMT models work very well, provided a lot of parallel data
- The size and domain of parallel data is limited

**Monolingual** data is easier to acquire and abundant (for most lang.)

**Goal**: uncover latent structure in unlabeled data

**Unsupervised NMT** is not 100% realistic but...

it serves as a very good baseline for extensions with parallel data (Semi-supervised Learning)
How does Unsupervised NMT work?

We use **two** new objectives:

1. Learn the structure of each language... How?

   **Denoising auto-encoding**

   (Language Model (LM) + noise + swap words)
How does Unsupervised NMT work?

We use **two** new objectives:

1. **Force the representation to be good at translating too...without parallel data.**

   **Iterative backtranslation**

   - First translate fr $\rightarrow$ en
   - Then use as a pseudo-supervised example to train en $\rightarrow$ fr

   Why does this work? We initialize the model with word translations from a dictionary created with bilingual word embeddings.
How does Unsupervised NMT work?

We use **two** new objectives:

2. Force the representation to be good at translating too...without parallel data. How?

**Iterative backtranslation**

- First translate fr $\rightarrow$ en
How does Unsupervised NMT work?

We use two new objectives:

2. Force the representation to be good at translating too... without parallel data. How?

**Iterative backtranslation**

- First translate \( \text{fr} \rightarrow \text{en} \)
- Then use as a **pseudo-supervised** example to train \( \text{en} \rightarrow \text{fr} \)
How does Unsupervised NMT work?

We use two new objectives:

1. Force the representation to be good at translating too...without parallel data. How?

   **Iterative backtranslation**

   - First translate fr → en
   - Then use as a **pseudo-supervised** example to train en → fr
   - Why does this work? We initialize the model with **word translations** from a dictionary created with bilingual word embeddings - guides first iteration
Presentation Outline

1. Motivation for Transfer Learning
2. Recap: Unsupervised Neural Machine Translation
3. Transfer Learning for NMT
4. Transfer Learning for Unsupervised NMT
   - Language Model Pretraining
   - Bilingual Language Model Pretraining
   - Continual Pretraining
   - Parallel Data from Similar Language Pairs
   - Limitations
5. Conclusion
What happens when we don’t have enough parallel data to train an NMT model?
Transfer Learning for NMT

What happens when we **don’t** have enough parallel data to train an NMT model?

How can we build systems that provide accurate translations between **low-resource** languages?
Transfer Learning for NMT

Transferring a model trained on a **lot** of parallel data to a model that has only **small** amounts of parallel data gives a large performance boost! (e.g. Hindi-English $\rightarrow$ Marathi-English)
Transfer Learning for NMT

We can also use **pivot translation**!

We want to build an Italian-Romanian translation system
(low-resource - we don’t have a lot of parallel corpora available)

We have **En-It** and **En-Ro** parallel corpora!

We can pretrain two NMT systems, that are then **transferred** to the final NMT system
Transfer Learning for NMT

- Transfer learning from an NMT system pretrained on large parallel corpora to an NMT system with small parallel corpora has limitations
- Parallel corpora are hard to find
- For some languages, there are no closely related high-resource languages
- How can we overcome this problem?
Transfer Learning for NMT

- Transfer learning from an NMT system pretrained on large parallel corpora to an NMT system with small parallel corpora has limitations
- Parallel corpora are hard to find
- For some languages, there are no closely related high-resource languages
- How can we overcome this problem?

→ Unsupervised pretraining using monolingual data!
Can we use transfer learning (and specifically unsupervised pretraining) to initialize an NMT model in a better way?

**Idea:**

1. Separately **Pretrain** Encoder and Decoder as **Language Models**
Can we use transfer learning (and specifically unsupervised pretraining) to initialize an NMT model in a better way?

Idea:

1. Separately **Pretrain** Encoder and Decoder as **Language Models**

2. Then **Train Jointly** on Bilingual Data (NMT)

(Figures from Kevin Clark’s talk)
Presentation Outline

1. Motivation for Transfer Learning
2. Recap: Unsupervised Neural Machine Translation
3. Transfer Learning for NMT
4. Transfer Learning for Unsupervised NMT
   - Language Model Pretraining
   - Bilingual Language Model Pretraining
   - Continual Pretraining
   - Parallel Data from Similar Language Pairs
   - Limitations
5. Conclusion
Transfer Learning for Unsupervised NMT

1. Motivation for Transfer Learning

2. Recap: Unsupervised Neural Machine Translation

3. Transfer Learning for NMT

4. Transfer Learning for Unsupervised NMT
   - Language Model Pretraining
   - Bilingual Language Model Pretraining
   - Continual Pretraining
   - Parallel Data from Similar Language Pairs
   - Limitations

5. Conclusion
Language Model Pretraining

Remember that we use **word translations** obtained by bilingual word embeddings to initialize the unsupervised NMT model.

How can we improve this?
Language Model Pretraining

- Pretraining the encoder and decoder using two separate language models is not *directly* applicable to unsupervised NMT.
Language Model Pretraining

- Pretraining the encoder and decoder using two separate language models is not *directly* applicable to unsupervised NMT.
- There is no “interaction” between the two languages during pretraining.
Language Model Pretraining

- Pretraining the encoder and decoder using two separate language models is not **directly** applicable to unsupervised NMT.
- There is no “interaction” between the two languages during pretraining.
  - The encoder LM learns how to produce proper En sentences.
Pretraining the encoder and decoder using two separate language models is not **directly** applicable to unsupervised NMT.

There is no “interaction” between the two languages during pretraining:

1. The encoder LM learns how to produce proper En sentences
2. The decoder LM learns how to produce proper Fr sentences
Pretraining the encoder and decoder using two separate language models is not **directly** applicable to unsupervised NMT.

There is no “interaction” between the two languages during pretraining:

1. The encoder LM learns how to produce proper En sentences
2. The decoder LM learns how to produce proper Fr sentences

If we directly applied it to unsupervised NMT...

---

**Diagram:**

- **Input:** *I traveled to Belgium*
- **Output:** *Je suis étudiant*
- **Process:**
  - **Encoder:** MT Model
  - **Decoder:** MT Model (Training)

---

Translation: *Je suis étudiant*
Language Model Pretraining

- Pretraining the encoder and decoder using two separate language models is not **directly** applicable to unsupervised NMT.
- There is no “interaction” between the two languages during pretraining.
  1. The encoder LM learns how to produce proper En sentences.
  2. The decoder LM learns how to produce proper Fr sentences.

If we directly applied it to unsupervised NMT...

The first sentence is in En, the second sentence is in Fr, **but** the Fr sentence is not a translation of the En sentence!
Extension of idea, specifically for unsupervised NMT:

- Training two language models (LMs) separately does not permit “interaction” between the two languages
Extension of idea, specifically for unsupervised NMT:

- Training two language models (LMs) separately does not permit “interaction” between the two languages
- Training one LM on two languages could be more helpful
Extension of idea, specifically for unsupervised NMT:

- Training two language models (LMs) separately does not permit “interaction” between the two languages
- Training one LM on two languages could be more helpful

→ We want to pretrain a bilingual LM on monolingual data of 2 languages simultaneously to obtain better initial translations
Language Model Pretraining

Extension of idea, specifically for unsupervised NMT:

- Training two language models (LMs) separately does not permit “interaction” between the two languages
- Training **one** LM on **two** languages could be more helpful

→ We want to pretrain a bilingual LM on **monolingual** data of 2 languages simultaneously to obtain better initial translations

And not just a “regular” LM...but the “parent” of most LLMs nowadays: **BERT**
Extension of idea, specifically for unsupervised NMT:

- Training two language models (LMs) separately does not permit "interaction" between the two languages
- Training one LM on two languages could be more helpful

→ We want to pretrain a bilingual LM on monolingual data of 2 languages simultaneously to obtain better initial translations

And not just a "regular" LM...but the "parent" of most LLMs nowadays: BERT
1 Motivation for Transfer Learning

2 Recap: Unsupervised Neural Machine Translation

3 Transfer Learning for NMT

4 Transfer Learning for Unsupervised NMT
   - Language Model Pretraining
   - Bilingual Language Model Pretraining
   - Continual Pretraining
   - Parallel Data from Similar Language Pairs
   - Limitations

5 Conclusion
Bilingual Language Model Pretraining

- We want to use transfer learning for unsupervised NMT.
- A LM that provides \textit{contextual} word representations in both languages we care about gives far better initial translations than a simple dictionary obtained from bilingual word embeddings.
- Then, we can initialize an \textbf{unsupervised} encoder-decoder NMT model with the pretrained bilingual LM!
Pretrain BERT simultaneously on 2 languages (without the next sentence prediction task)

Large amounts of training data:

Not Parallel!!!
Bilingual Language Model Pretraining

- We have a shared encoder and decoder (for both En→Fr and Fr→En)
**Bilingual Language Model Pretraining**

- We have a shared encoder and decoder (for both $\text{En} \rightarrow \text{Fr}$ and $\text{Fr} \rightarrow \text{En}$).
- We initialize the encoder and the decoder with a **bilingual masked language model** (pretrained on a lot of monolingual data)!

![Diagram of bilingual language model pretraining](image)

- Denoising autoencoder loss
- Machine translation loss

We train the NMT model using as training objectives (losses) denoising auto-encoding and iterative backtranslation.
We have a shared encoder and decoder (for both En→Fr and Fr→En).

We train the NMT model using as training objectives (losses) denoising auto-encoding and iterative backtranslation.
Bilingual Language Model Pretraining

- We have a shared encoder and decoder (for both En→Fr and Fr→En)
- We train the NMT model using as training objectives (losses) **denoising auto-encoding** and **iterative backtranslation**
We have a shared encoder and decoder (for both En→Fr and Fr→En)

We train the NMT model using as training objectives (losses) **denoising auto-encoding** and **iterative backtranslation**
Bilingual Language Model Pretraining

- We have a shared encoder and decoder (for both En→Fr and Fr→En)
- We train the NMT model using as training objectives (losses) **denoising auto-encoding** and **iterative backtranslation**
### Unsupervised NMT Results

<table>
<thead>
<tr>
<th>Model</th>
<th>En-Fr</th>
<th>En-De</th>
<th>En-Ro</th>
</tr>
</thead>
<tbody>
<tr>
<td>UNMT</td>
<td>25.1</td>
<td>17.2</td>
<td>21.2</td>
</tr>
<tr>
<td>UNMT + Pre-Training</td>
<td>33.4</td>
<td>26.4</td>
<td>33.3</td>
</tr>
<tr>
<td>Current supervised State-of-the-art</td>
<td>45.6</td>
<td>34.2</td>
<td>29.9</td>
</tr>
</tbody>
</table>

Table from Kevin Clark’s talk.
Bilingual Language Model Pretraining

Why does training a LM jointly on 2 languages help?

- We encode text in a **subword** level
- Subword token improves the alignment of embedding spaces of two languages (especially if they share the alphabet or the digits)

![Diagram showing examples of cognates, loan words, names, transliteration, and morphology]

Figure from Graham Neubig notes on MT class, Fall 2019.
1. Motivation for Transfer Learning

2. Recap: Unsupervised Neural Machine Translation

3. Transfer Learning for NMT

4. Transfer Learning for Unsupervised NMT
   - Language Model Pretraining
   - Bilingual Language Model Pretraining
   - Continual Pretraining
   - Parallel Data from Similar Language Pairs
   - Limitations

5. Conclusion
Extending a language model to more languages

Problems

- In a continual learning setting, can we add more languages to an existing LM as we get more data?
- If the new language does not have a common vocabulary with the one we trained our LM on, it will break
- How can we avoid this?
We can leverage the lexical overlap of the pretraining language and fine-tuning language to extend the vocabulary.

We add subword tokens that are randomly initialized.

**Figure 1:** RE-LM. (A) LM pretraining. (B) Fine-tuning. The embedding and the projection layer are extended using §3.2 (dark gray) and (C) Transfer to an NMT system. Dashed arrows indicate transfer of weights.
Motivation for Transfer Learning

Recap: Unsupervised Neural Machine Translation

Transfer Learning for NMT

Transfer Learning for Unsupervised NMT
- Language Model Pretraining
- Bilingual Language Model Pretraining
- Continual Pretraining
- Parallel Data from Similar Language Pairs
- Limitations

Conclusion
Leverage languages for which we have parallel data

Continuously extend the vocabulary (it converges rather fast)
Motivation for Transfer Learning

Recap: Unsupervised Neural Machine Translation

Transfer Learning for NMT

Transfer Learning for Unsupervised NMT
- Language Model Pretraining
- Bilingual Language Model Pretraining
- Continual Pretraining
- Parallel Data from Similar Language Pairs
- Limitations

Conclusion
Can UNMT replace NMT?

- Semi-supervised: continue training from the supervised baseline with BT added to the training data.
- Unsupervised NMT still lags behind fully- or semi-supervised NMT models.

(a) German→English
Can UNMT replace NMT?

- Domain matching is critical for unsupervised NMT
- If data from similar domains is not available, performance drops sharply

<table>
<thead>
<tr>
<th>Domain (en)</th>
<th>Domain (de/ru)</th>
<th>BLEU [%]</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>de-en</td>
<td>en-de</td>
</tr>
<tr>
<td>Newswire</td>
<td>23.3</td>
<td>19.9</td>
</tr>
<tr>
<td>Newswire</td>
<td>11.5</td>
<td>12.2</td>
</tr>
<tr>
<td>Politics</td>
<td>18.4</td>
<td>16.4</td>
</tr>
</tbody>
</table>

Table 3: Unsupervised NMT performance where source and target training data are from different domains. The data size on both sides is the same (20M sentences).
Can UNMT replace NMT?

- Pretraining a bilingual LM on an adequate amount of (comparable) data is very important.
- Unsupervised learning cannot build a reasonable NMT model when starting from a poor initialization.

**Figure 6**: Unsupervised NMT performance over the training data size for translation training, where the pre-training data for initialization is fixed (10k or 20M sentences).
Presentation Outline

1. Motivation for Transfer Learning
2. Recap: Unsupervised Neural Machine Translation
3. Transfer Learning for NMT
4. Transfer Learning for Unsupervised NMT
   - Language Model Pretraining
   - Bilingual Language Model Pretraining
   - Continual Pretraining
   - Parallel Data from Similar Language Pairs
   - Limitations
5. Conclusion
Unsupervised Neural Machine Translation is interesting as an extreme scenario

It cannot replace NMT (you guessed it)

In practice, we have (some) parallel data for most language pairs we want to translate to/from

We can use methods developed for UNMT to improve low-resource NMT

Pretraining multilingual unsupervised models (such as LMs) is very useful for all tasks in multilingual NLP (and not just NMT or UNMT)
Thank You for your Attention! Questions?
References I


References II


References IV

