Transfer Learning for Unsupervised Neural Machine Translation

Alexandra Chronopoulou

achron@cis.lmu.de

CIS, LMU Munich

11/07/2023

Presentation Outline

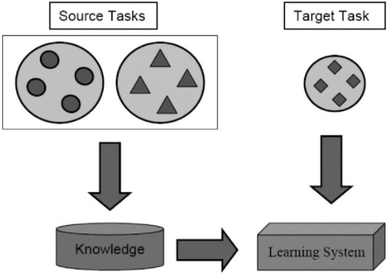
1. Motivation for Transfer Learning
2. Recap: Unsupervised Neural Machine Translation
3. Transfer Learning for NMT
4. Transfer Learning for Unsupervised NMT
   - Language Model Pretraining
   - Bilingual Language Model Pretraining
   - Continual Pretraining
   - Parallel Data from Similar Language Pairs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

Motivation for Transfer Learning

Machine learning

Problems (especially in deep learning):

- Scarcity of labeled data
- Models trained on small datasets often fail to generalize in test data $\rightarrow$ overfit



Transfer learning:

- Uses knowledge from a *learned* task to improve the performance on a *related* task
- Scarcity of labeled data $\rightarrow$ implicit data augmentation
- Helps a model generalize $\rightarrow$ avoid overfitting

Motivation for Transfer Learning

Natural language processing & Machine Translation

In Natural Language Processing tasks:

- **Out-of-context** pretrained word representations were used (*word2vec, fasttext) to initialize the* **embedding layer**

Motivation for Transfer Learning

Natural language processing & Machine Translation

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**

Presentation Outline

Recap: Unsupervised Neural Machine Translation

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)

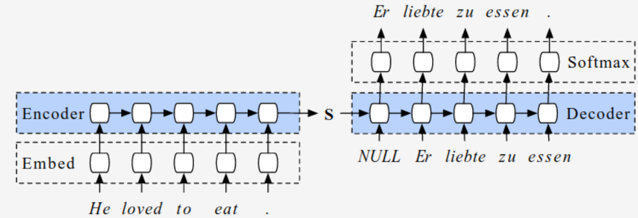


Figure: Seq2seq architecture for En-De NMT. Figure from https://smerity.com/articles/2016/google_nmt_arch.html

Recap: Unsupervised Neural Machine Translation

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.)

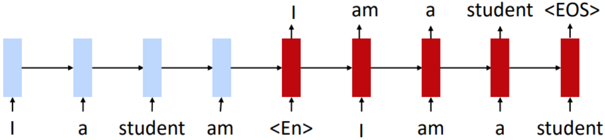
- **Goal**: uncover latent structure in unlabeled data
- **Unsupervised NMT** is not 100% realistic but...
- it serves as a very good baseline for extensions with parallel data (Semi-supervised Learning)

Recap: Unsupervised Neural Machine Translation

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)



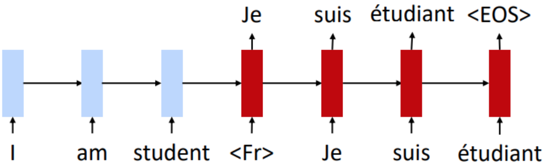
Recap: Unsupervised Neural Machine Translation

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



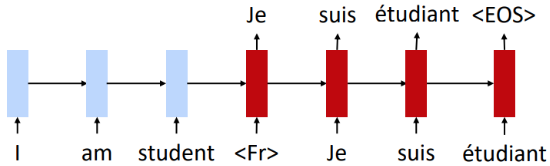
Recap: Unsupervised Neural Machine Translation

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

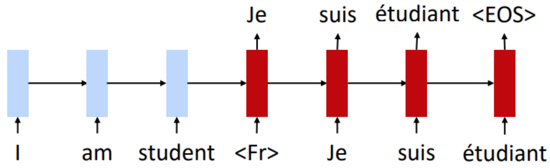
Recap: Unsupervised Neural Machine Translation

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 → en
- Then use as a **pseudo-supervised** example to train en → fr

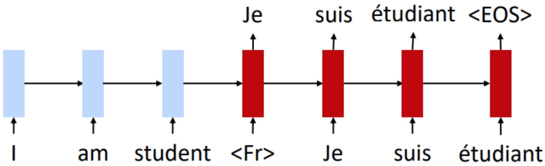
Recap: Unsupervised Neural Machine Translation

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
- 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` - guides first iteration

Presentation Outline

Transfer Learning for NMT

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

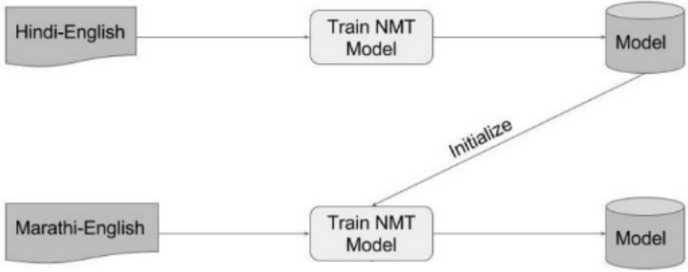
Transfer Learning for NMT

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`)

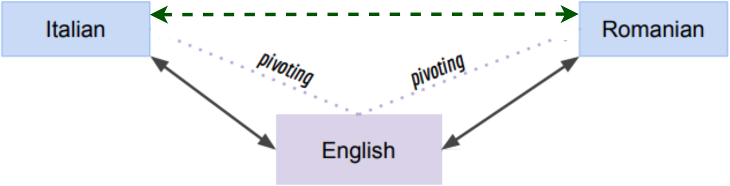


Transfer Learning for NMT

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?

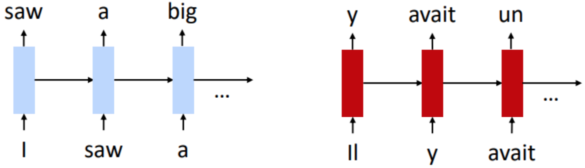
$\longrightarrow$ Unsupervised pretraining using monolingual data!

Transfer Learning for NMT

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**

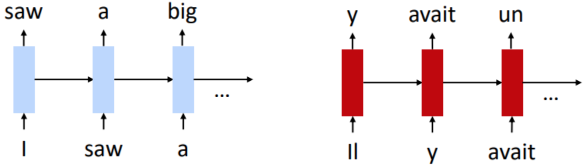


Transfer Learning for NMT

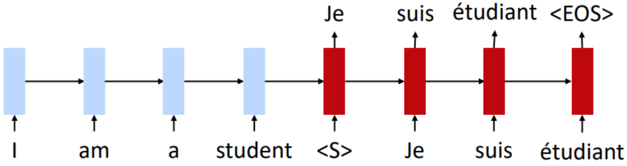
Can we use transfer learning (and specifically unsupervised pretraining) to initialize an NMT model in a better way?

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

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

1. The encoder LM learns how to produce proper `En` sentences

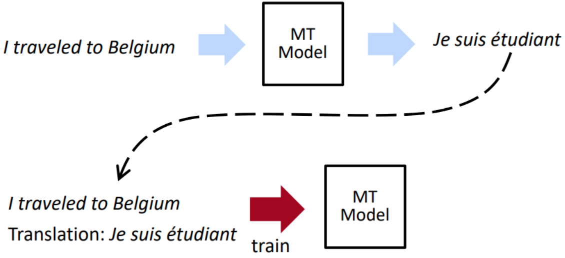
Language Model Pretraining

- 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

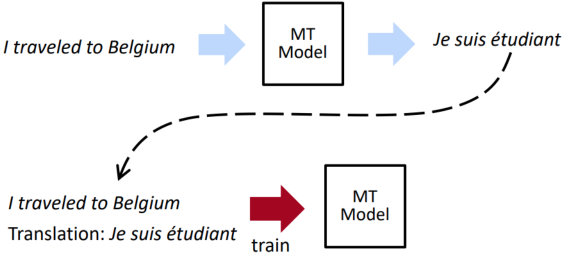
Language Model Pretraining

- 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...



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!

Language Model Pretraining

Extension of idea, specifically for unsupervised NMT:

- Training two language models (LMs) separately does not permit "interaction" between the two languages

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

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

$\longrightarrow$ 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

$\longrightarrow$ 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**

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

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

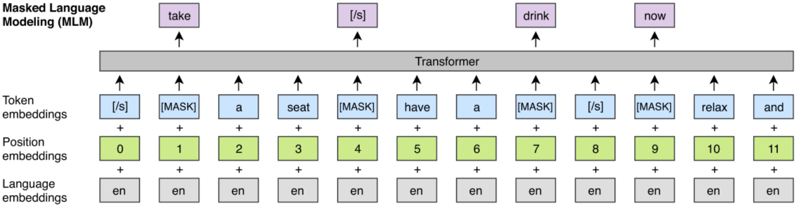
And not just a "regular" LM...but the "parent" of most LLMs nowadays: **BERT**

Bilingual Language Model Pretraining

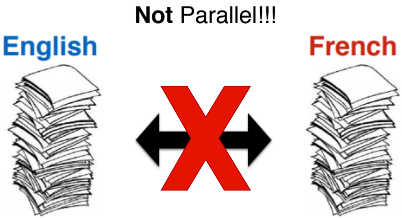
- We want to use transfer learning for unsupervised NMT
- A LM that provides *contextual* word representations in both languages we care about gives far better initial translations than a simple dictionary obtained from bilingual word embeddings
- Then, we can initialize an **unsupervised** encoder-decoder NMT model with the pretrained bilingual LM!

Bilingual Language Model Pretraining

- **Pretrain BERT simultaneously on 2 languages** (without the next sentence prediction task)



Large amounts of training data:

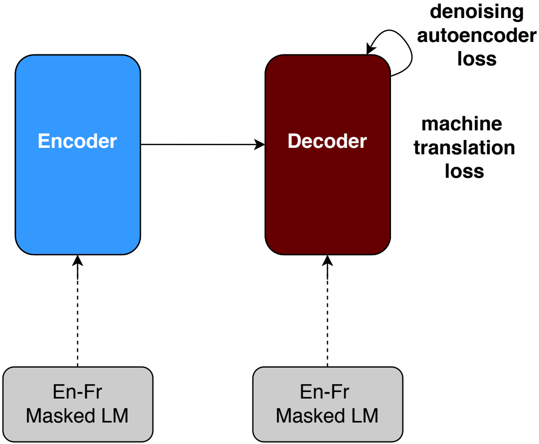


Bilingual Language Model Pretraining

- We have a shared encoder and decoder (for both En→Fr and Fr→En)

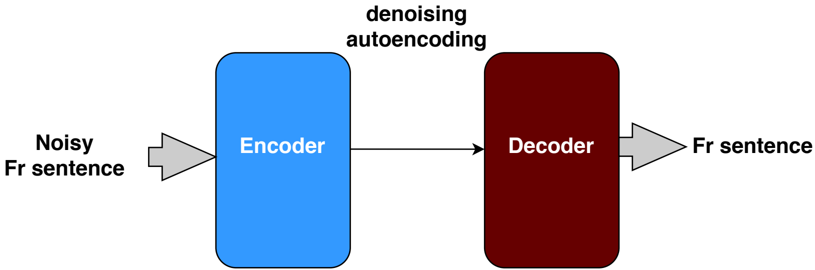
Bilingual Language Model Pretraining

- We have a shared encoder and decoder (for both En→Fr and Fr→En)
- We initialize the encoder **and** the decoder with a **bilingual masked language model** (pretrained on a lot of monolingual data)!



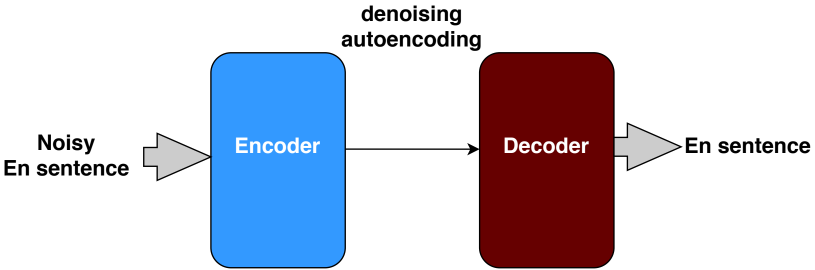
Bilingual Language Model Pretraining

- 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**



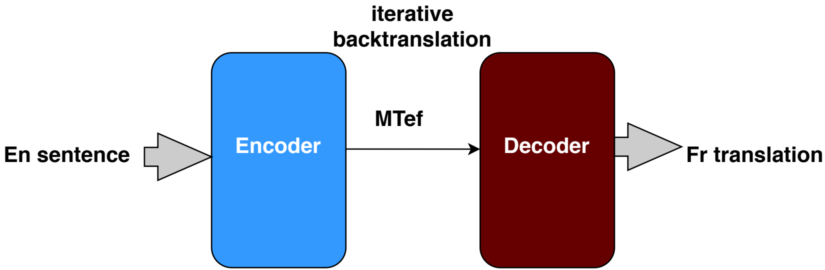
Bilingual Language Model Pretraining

- 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**



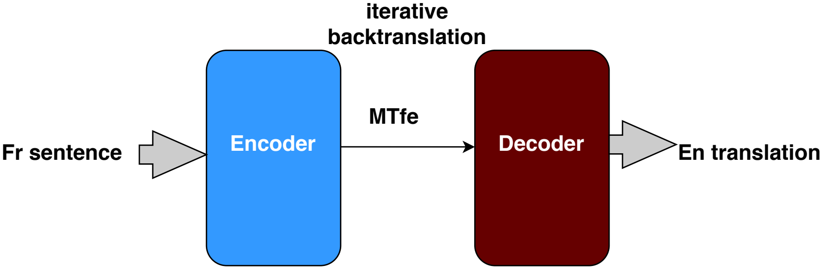
Bilingual Language Model Pretraining

- 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**



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**



Bilingual Language Model Pretraining

Unsupervised NMT Results

Model	En-Fr	En-De	En-Ro
UNMT	25.1	17.2	21.2
UNMT + Pre-Training	33.4	26.4	**33.3**
Current supervised State-of-the-art	**45.6**	**34.2**	29.9

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`)

Figure from `Graham Neubig notes` on MT class, Fall 2019.

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?

Extending a language model to more languages

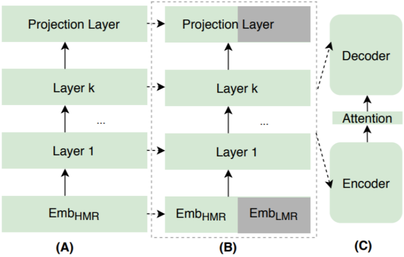


Figure 1: **RE-LM**. **(A)** LM pretraining. **(B)** Fine-tuning. The embedding and the projection layer are extended using §3.2 (dark gray) and **(C)** Transfer to an NMT system. Dashed arrows indicate transfer of weights.

- 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

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

Parallel Data from Similar Language Pairs

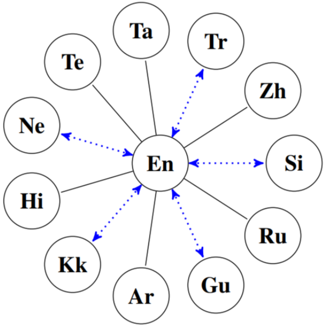
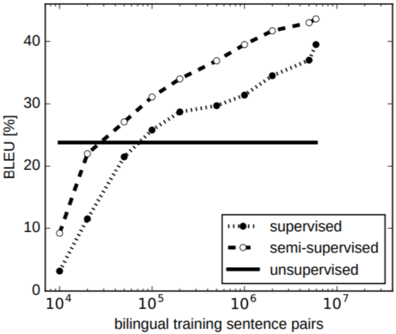


Figure 1: A pictorial depiction of our setup. The dashed edge indicates the target unsupervised language pairs that lack parallel training data. Full edges indicate the existence of parallel training data.

- Leverage languages for which we have parallel data
- Continuously extend the vocabulary (it converges rather fast)

Can UNMT replace NMT?



(a) German→English

- 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

Can UNMT replace NMT?

Domain (en)	Domain (de/ru)	BLEU [%] de-en	en-de	ru-en	en-ru
Newswire	Newswire	23.3	19.9	11.9	9.3
	Politics	11.5	12.2	2.3	2.5
	Random	18.4	16.4	6.9	6.1

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).

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

Can UNMT replace NMT?

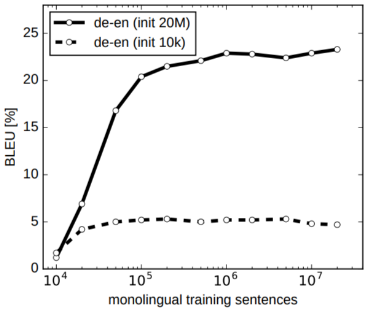


Figure 6: Unsupervised NMT performance over the training data size for translation training, where the pre-training data for initialization is fixed (10k or 20M sentences).

- 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

Presentation Outline

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

[1] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Advances in Neural Information Processing Systems*, 2013, pp. 3111–3119. (visited on 03/22/2017).

[2] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, "Enriching word vectors with subword information," *Transactions of the Association for Computational Linguistics*, vol. 5, pp. 135–146, 2017. DOI: 10.1162/tacl_a_00051. [Online]. Available: https://www.aclweb.org/anthology/Q17-1010.

References II

[3] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," in *Advances in Neural Information Processing Systems*, Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, Eds., 2014, pp. 3104–3112. [Online]. Available: http://papers.nips.cc/paper/5346-sequence-to-sequence-learning-with-neural-networks.pdf.

[4] B. Zoph, D. Yuret, J. May, and K. Knight, "Transfer learning for low-resource neural machine translation," in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, Austin, Texas: Association for Computational Linguistics, Nov. 2016, pp. 1568–1575. DOI: 10.18653/v1/D16-1163. [Online]. Available: https://www.aclweb.org/anthology/D16-1163.

References III

[5] P. Ramachandran, P. Liu, and Q. Le, "Unsupervised pretraining for sequence to sequence learning," in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, Copenhagen, Denmark: Association for Computational Linguistics, Sep. 2017, pp. 383–391. DOI: 10.18653/v1/D17-1039. [Online]. Available: https://www.aclweb.org/anthology/D17-1039.

[6] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2019, pp. 4171–4186. [Online]. Available: https://www.aclweb.org/anthology/N19-1423.

References IV

[7] G. Lample and A. Conneau, "Cross-lingual language model pretraining," in *Advances in Neural Information Processing Systems*, 2019, pp. 7057–7067. [Online]. Available: https://papers.nips.cc/paper/8928-cross-lingual-language-model-pretraining.

[8] A. Chronopoulou, D. Stojanovski, and A. Fraser, "Reusing a Pretrained Language Model on Languages with Limited Corpora for Unsupervised NMT," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Online: Association for Computational Linguistics, Nov. 2020, pp. 2703–2711. DOI: 10.18653/v1/2020.emnlp-main.214. [Online]. Available: https://aclanthology.org/2020.emnlp-main.214.

References V

[9] X. Garcia, A. Siddhant, O. Firat, and A. Parikh, "Harnessing multilinguality in unsupervised machine translation for rare languages," in *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Online: Association for Computational Linguistics, Jun. 2021, pp. 1126–1137. DOI: 10.18653/v1/2021.naacl-main.89. [Online]. Available: https://aclanthology.org/2021.naacl-main.89.

[10] Y. Kim, M. Graça, and H. Ney, "When and why is unsupervised neural machine translation useless?" In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, Lisboa, Portugal: European Association for Machine Translation, Nov. 2020, pp. 35–44. [Online]. Available: https://aclanthology.org/2020.eamt-1.5.