DSTA Executive Education Course

Unsupervised Machine Translation

Graham Neubig

Carnegie Mellon University
Language Technologies Institute

Many slides by Sachin Kumar

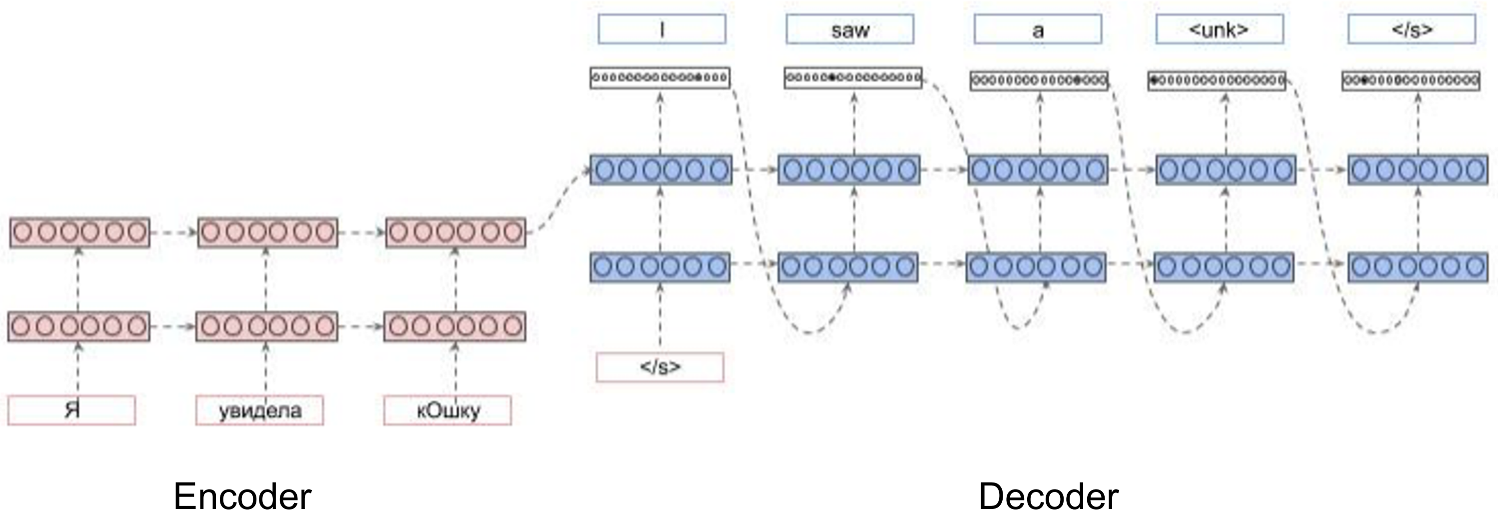
Conditional Text Generation

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

Input X	Output Y (Text)	Task
English	Hindi	Machine Translation
Image	Text	Image Captioning
Document	Short Description	Summarization
Speech	Transcript	Speech Recognition

[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?

Input X	**Output Y**	**Task**
Image (Photo)	Image (Painting)	Style Transfer
Image (Male)	Image (Female)	Gender Transfer
Text (Impolite)	Text (Polite)	Formality Transfer
Positive Review	Negative Review	Sentiment Transfer
English	Sinhalese	Machine Translation

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 Joe Biden”
- Asking annotators to generate text doesn’t usually lead to good quality datasets

Unsupervised 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

Statistical MT

Neural MT

1. 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

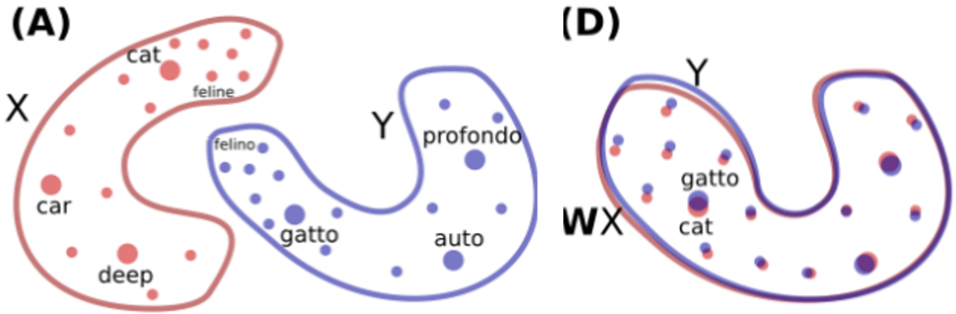
Step 1: Initialization

- 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

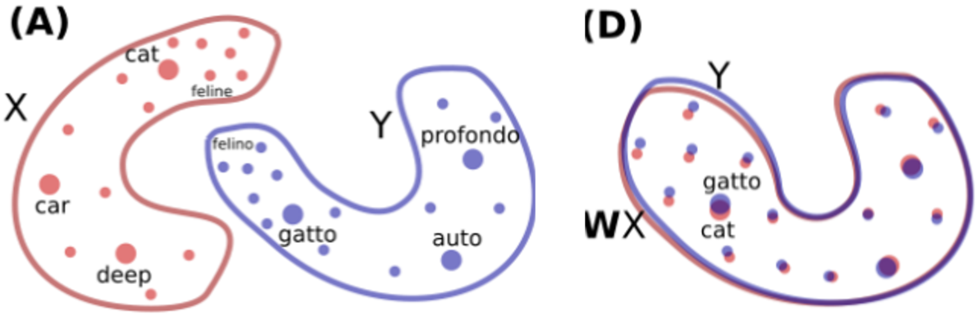
- Hypothesis: Word embedding spaces in two languages are isomorphic
  - One embedding space can be linearly transformed into another
  - Give monolingual embeddings X and Y, learn a (orthogonal) matrix, such that, WX = Y



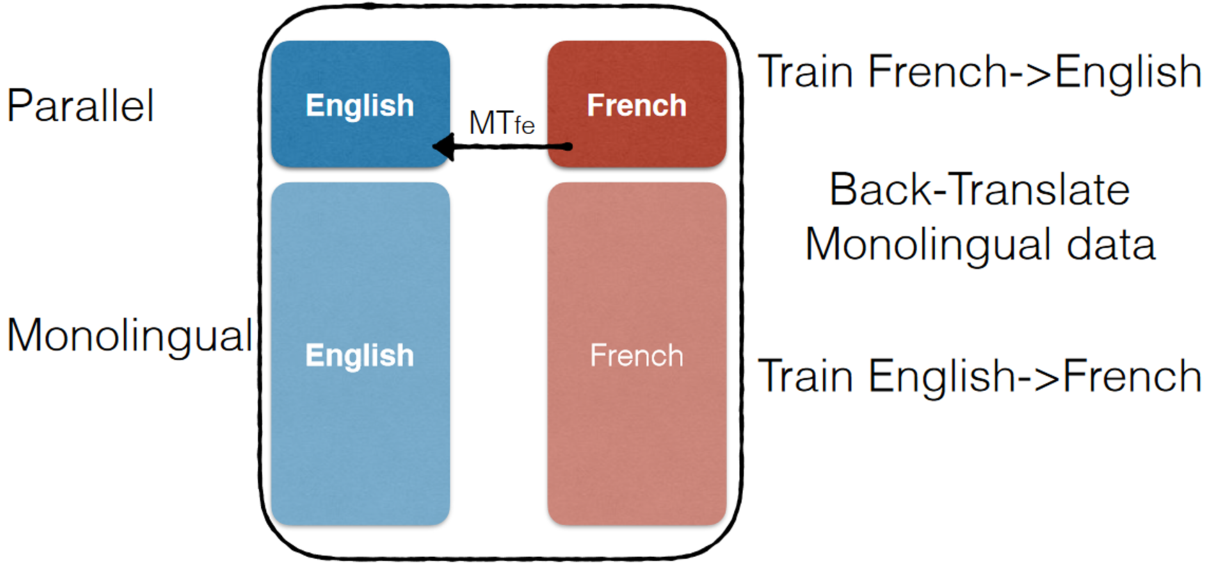
Word Translation Without Parallel Data. Conneau and Lample. ICLR 2018
A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings. Artetxe et al. ACL 2018

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



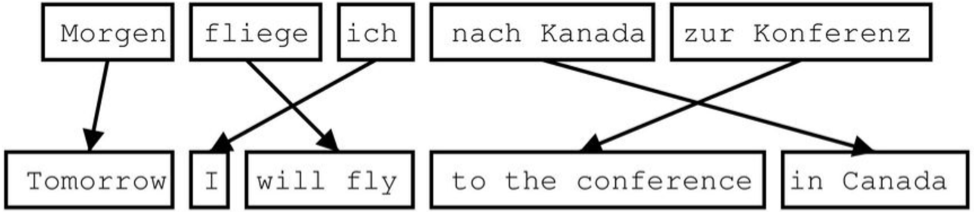
Step 2: Back-translation



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

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

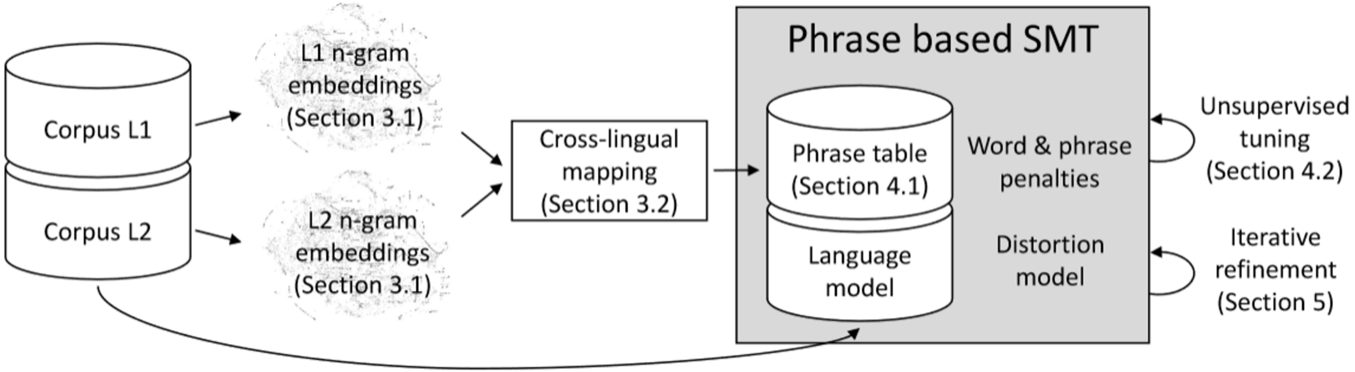
Needs parallel data :(

Only monolingual data needed :)

[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]

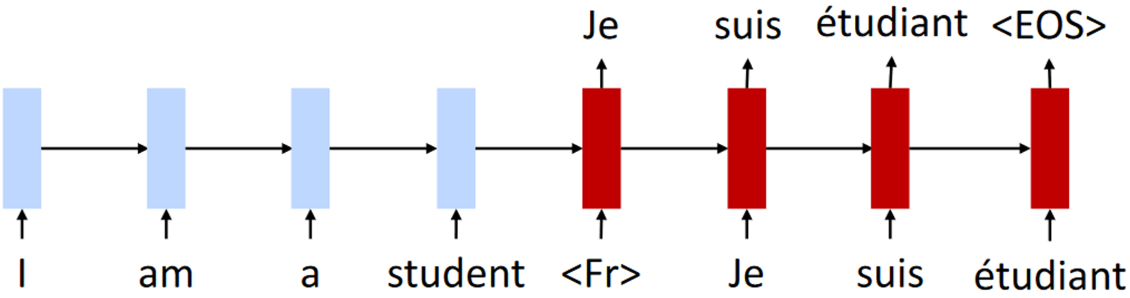
Unsupervised Statistical MT

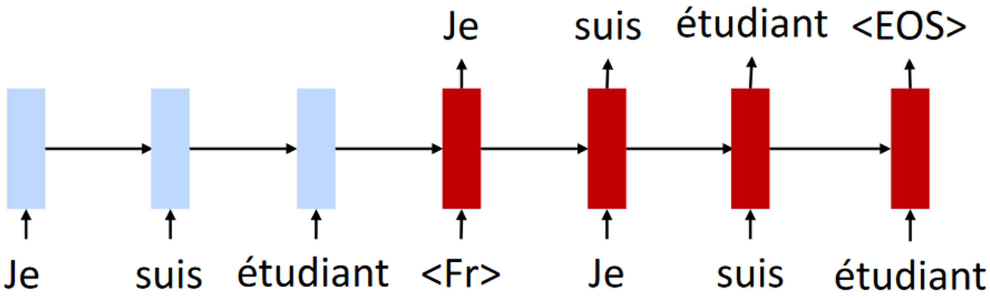
	en → fr	fr → en	en → de	de → en	en → ro	ro → en	en → ru	ru → en
Unsupervised PBSMT								
Unsupervised phrase table	-	17.50	-	15.63	-	14.10	-	8.08
Back-translation - Iter. 1	24.79	26.16	15.92	22.43	18.21	21.49	11.04	15.16
Back-translation - Iter. 2	27.32	26.80	17.65	22.85	20.61	22.52	12.87	16.42
Back-translation - Iter. 3	27.77	26.93	17.94	22.87	21.18	22.99	13.13	16.52
Back-translation - Iter. 4	27.84	27.20	17.77	22.68	21.33	23.01	13.37	**16.62**
Back-translation - Iter. 5	**28.11**	27.16	-	-	-	-	-	-

Unsupervised Neural MT

Step 3: Bidirectional Modeling

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

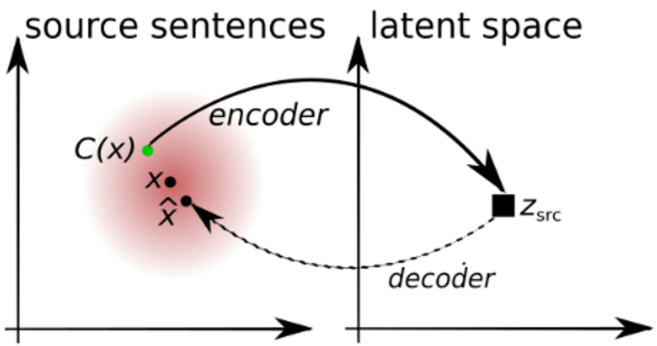




[Slide credits: Kevin Clark]

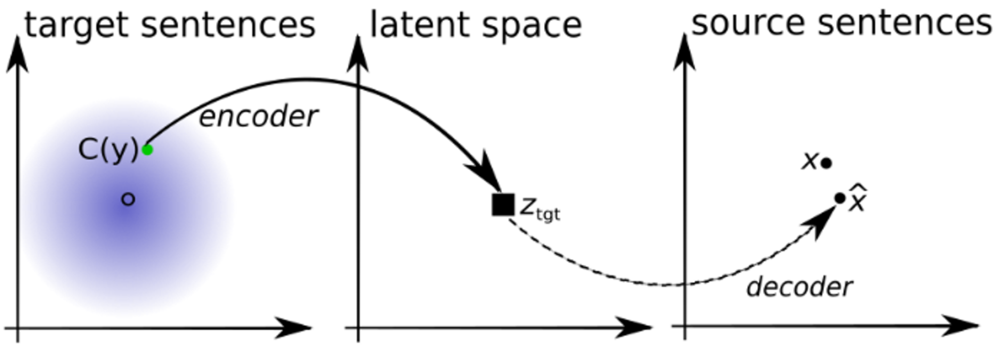
Unsupervised MT: Training Objective 1

Denoising autoencoder



Unsupervised NMT: Training Objective 2

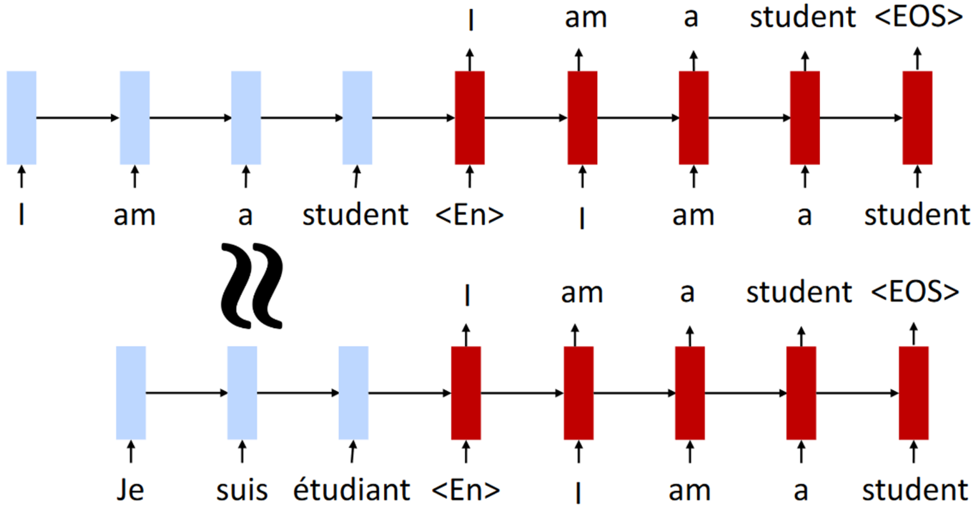
- Back-translation
    - Translate target to source
    - Use as a “supervised” example to translate source to target



[Lample et al 2018, Artetxe et al 2018]

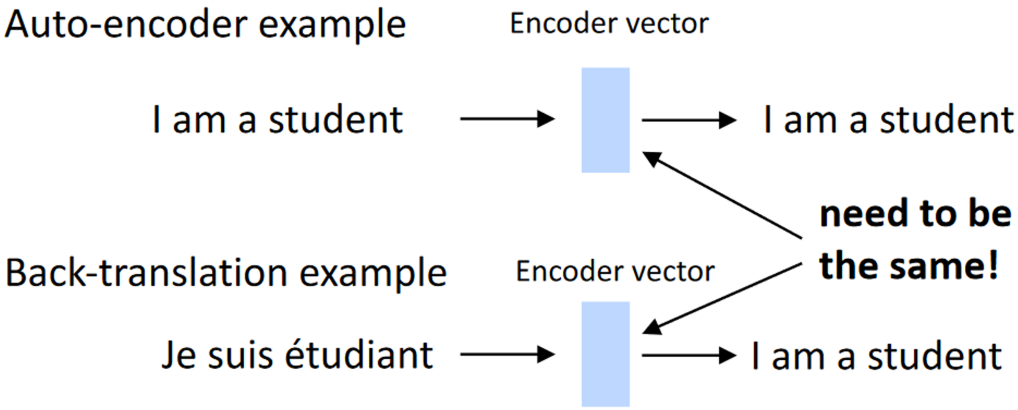
How does it work?

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



Unsupervised NMT: Training Objective 3

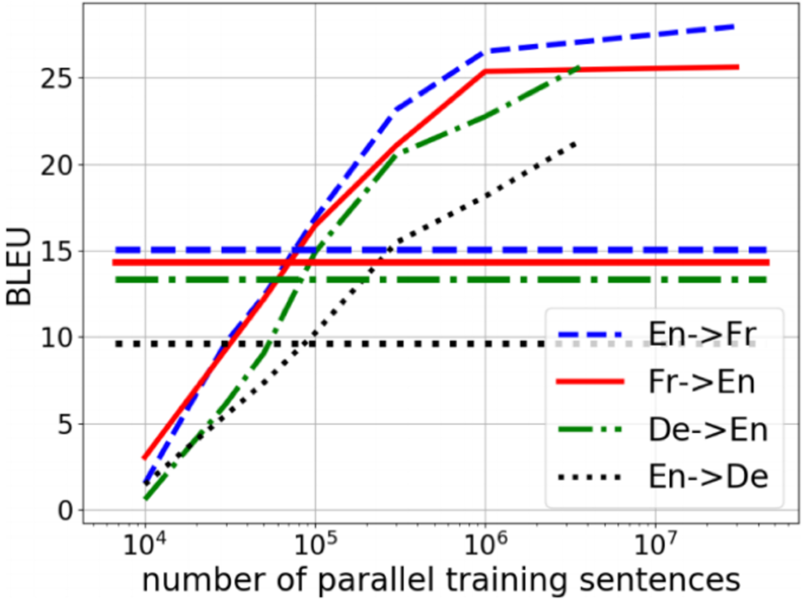
- Training Objective 3: Adversarial
  - Constraining the encoder to map the two languages in the same feature space



[Lample et al 2018]

Performance

- Horizontal lines are purely unsupervised, rest are purely 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.

	En-Turkish	Ne-En	Si-En
Supervise d	20	7.6	7.2
UNMT	4.5	0.2	0.4

[When Does Unsupervised Machine Translation Work? Marchisio et al 2020,
Rapid Adaptation of Neural Machine Translation to New Languages. Neubig and Hu. EMNLP 2018]

Reasons for this poor performance

1. Small monolingual data for low-resource languages -> bad embeddings
2. Different word frequencies/morphology hurt bilingual lexicon induction
3. Different content makes sentence-level distribution matching difficult

Open Problems

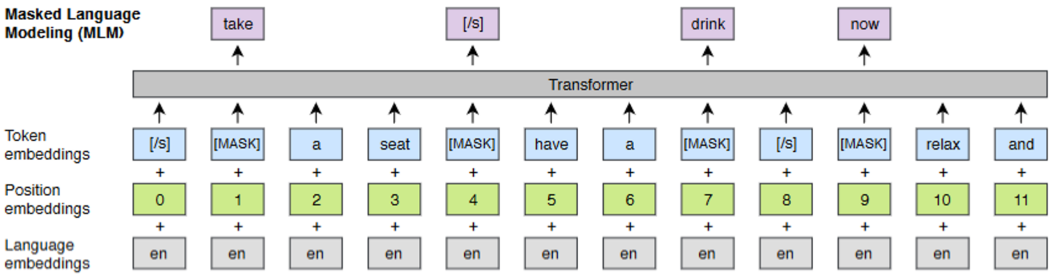
- Diverse languages and domains.
  - Better cross-lingual initialization: better data selection/regularization in pretraining language models

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

[When and Why is Unsupervised Neural Machine Translation Useless? Kim et al. 2020]

Better Initialization: Cross Lingual Language Models

- Cross Lingual Masked Language Modelling

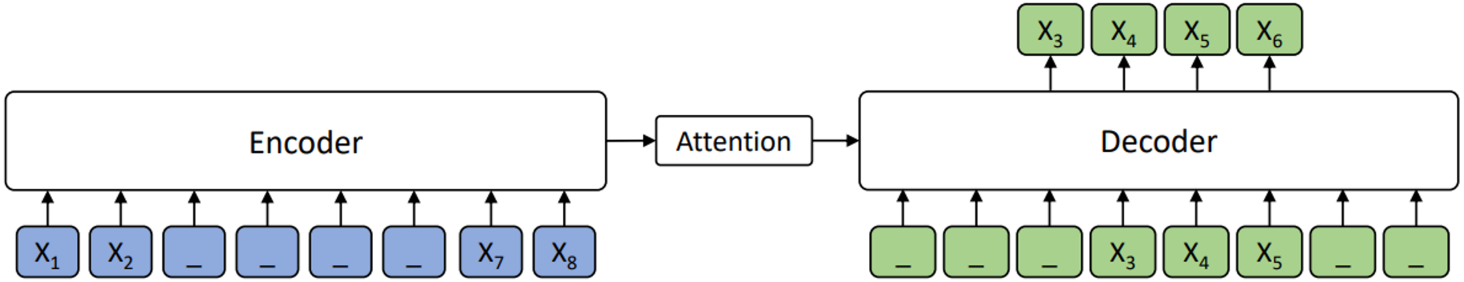


- Initialize the entire encoder and decoder instead of lookup tables
- Alignment comes from shared sub-word vocabulary

[Cross-lingual Language Model Pretraining. Lample and Conneau. 2019]

Better Initialization: Masked Sequence to Sequence Model (MASS)

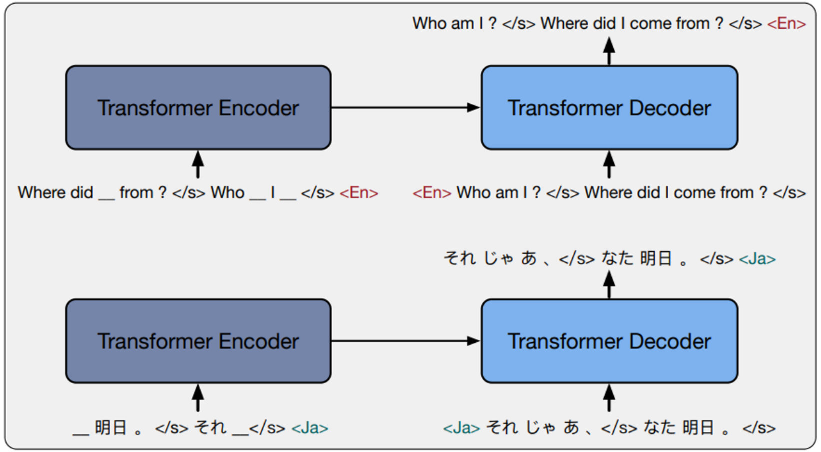
- Encoder-decoder formulation of masked language modelling



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

Better Initialization: Multilingual BART

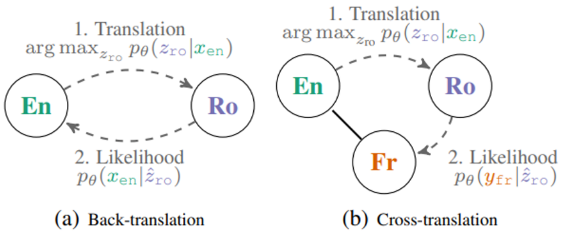
- Multilingual Denoised Autoencoding
- Corrupt the input and predict the clean version. Type of noise
  - Mask or swap words/phrases
  - Shuffle the order of sentences in an instance



[Multilingual Denoising Pre-training for Neural Machine Translation. Liu et al 2020]

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))

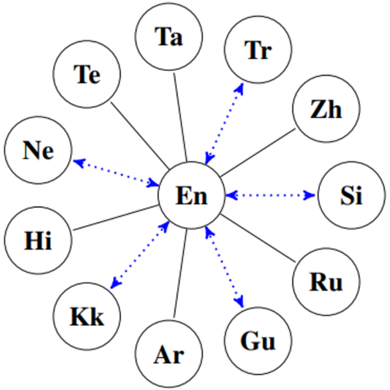


- Pretrain using seq2seq objective
- Two translation objectives:
  - Back-translation: P(x | y(x)) [Monolingual data]
  - Cross-translation: P(y | z(x)) [Parallel data (x, y)]
- Shows improvement for dissimilar languages with less monolingual data

[A Multilingual View of Unsupervised Machine Translation. Garcia et al. 2020]

Multilingual UNMT

- Improvements on low resource languages



	FLoRes devtest Ne ↔ En		*FLoRes devtest* Si ↔ En	
Unsupervised	-	-	-	-
	-	17.9*	-	8.99*
	8.3*	18.3*	0.1	0.1
Ours (Mult. Unsup.)	3.34	18.33	1.44	11.52
	8.62	20.76	7.72	15.66
	8.93	**21.68**	**7.9**	**16.23**
Supervised	-	-	-	-
	-	-	-	-
	-	-	-	-
	9.6	21.3	9.3	20.2
	8.8*	21.5*	6.5	15.1

[Harnessing Multilinguality in Unsupervised Machine Translation for Rare Languages. Garcia et al. EMNLP 2020]

How practical is the strict unsupervised scenario

- 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

[A Call for More Rigor in Unsupervised Cross-lingual Learning. Artetxe, Ruder et al 2020]

Related Area: Style Transfer

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

Relaxed ↔ Annoyed	
Relaxed	Sitting by the Christmas tree and watching Star Wars after cooking dinner. What a nice night 💗 🎄 ✨
Annoyed	Sitting by the computer and watching The Voice for the second time tonight. What a horrible way to start the weekend 😡 😡 😡
Annoyed	Getting a speeding ticket 50 feet in front of work is not how I wanted to start this month 😒
Relaxed	Getting a haircut followed by a cold foot massage in the morning is how I wanted to start this month 😊
Male ↔ Female	
Male	Gotta say that beard makes you look like a Viking...
Female	Gotta say that hair makes you look like a Mermaid...
Female	Awww he’s so gorgeous 😍 can’t wait for a cuddle. Well done 😘 xxx
Male	Bro he’s so f***ing dope can’t wait for a cuddle. Well done bro
Age 18-24 ↔ 65+	
18-24	You cheated on me but now I know nothing about loyalty 😂 ok
65+	You cheated on America but now I know nothing about patriotism. So ok.
65+	Ah! Sweet photo of the sisters. So happy to see them together today .
18-24	Ah 😂 Thankyou 💞 #sisters ❤️ happy to see them together today

[MULTIPLE-ATTRIBUTE TEXT REWRITING. Lample at al. ICLR 2019]

Discussion Question

Pick a low resource language or dialect, research **all** of the monolingual or parallel data that you can find online for it. Would unsupervised or semi-supervised MT methods be helpful? How could best use the existing resources to set up unsupervised or semi-supervised MT for success on this language or dialect?

Refer to: “When does unsupervised MT work?” (https://arxiv.org/pdf/2004.14958.pdf)