Neural Machine Translation
(Unsupervised, Supervised, Multilingual and Self-Supervised)

Cristina España-Bonet
DFKI GmbH

Low-Resource NLP:
Multilinguality and Machine Translation
Webinar Series — Session III
13th July 2021

Outline

1. Unsupervised MT
   - Recap on Basics & Cross-Lingual Embeddings
   - The Low-Resource Setting
2. Supervised NMT
   - Basics
   - The Low-Resource Setting
   - Multilingual Neural Machine Translation
3. Self-Supervised NMT
   - Basics
   - The Low Resource Setting (Session IV)

Recap, Unsupervised MT

Main Ingredients

1. Data

- Monolingual corpora

2. Initialisation

- **Cross-lingual embeddings**
- Deep MLM pretraining

3. Training

SMT and/or **NMT**

- **Denoising autoencoder**
- **Backtranslation**

Recap, Unsupervised MT

From Supervised Mapping to Unsupervised Self-Learning

1. **Supervised**
   - Joint learning
     - Regularization term in the loss function
     - Creating pseudo-bilingual corpora
   - **Mapping** (post-hoc alignment)

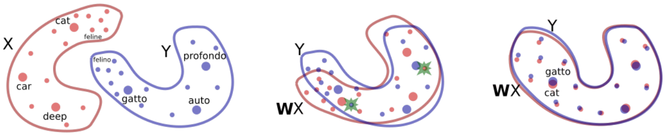
2. **Unsupervised**
   - **Mapping with self-learning**
   - Mapping with adversarial training

Recap, Unsupervised MT

Mapping Approaches: Isomorphism (and Other!) Assumption

We talked about:

- Isomorphism:
  spaces should be isomorphic for (linear) mappings to be effective



(Figure from *Conneau et al., 2017)*

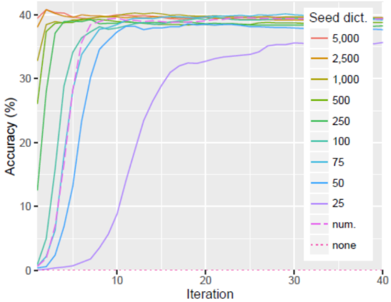
- (Solving the) Procrustes Problem
- Hubness and margin-based similarity measures

(Supervised) Cross-Lingual Embeddings by Mapping

1. We have monolingual embeddings
2. We have a (small) dictionary
3. We solve the **Procrustes problem** to find the projection matrix W
4. Given a word in L1 and W, the equivalent word in L2 can be found by its nearest neighbours according to a **margin-based similarity** measure

Recap, Unsupervised MT

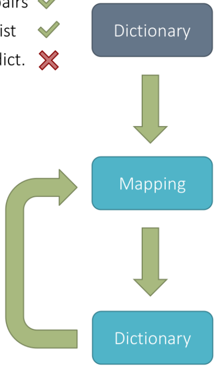
Self-Learning (Mikel Artetxe Slide)



$$W^* = \underset{W \in O(n)}{\arg\min} \sum_i \min_j \left\| X_{i*} W - Z_{j*} \right\|^2$$

- 25 word pairs ✓
- Numeral list ✓
- Random dict. ✗

<u>Self-learning</u>

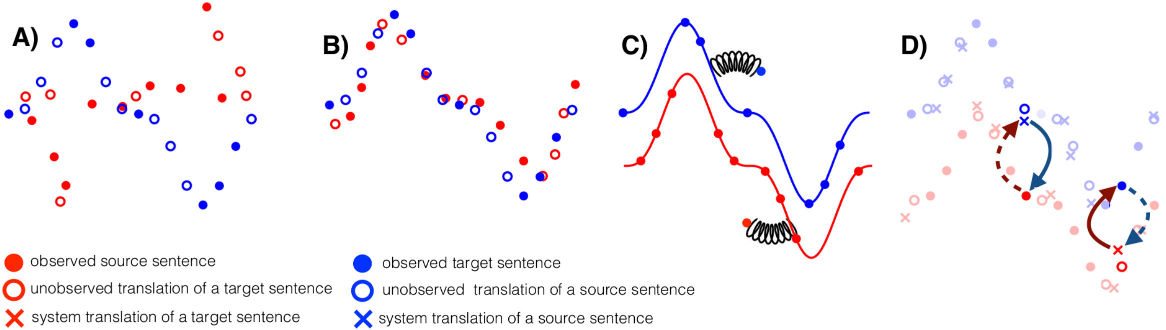


The difference between supervised and unsupervised is the (induction of) the **seed dictionary**

Recap, Unsupervised MT

The Three Principles (from Lample et al., ICLR, 2018)

Initialisation **Denoising (LM)** **Backtranslation**



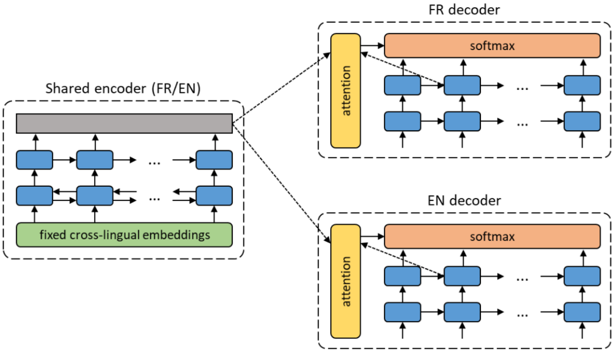
Recap, Unsupervised MT

Basics with Principles (Slides from Mikel Artetxe)

Training

- Supervised

Une fusillade a eu lieu à l'aéroport international de Los Angeles.



There was a shooting in Los Angeles International Airport.

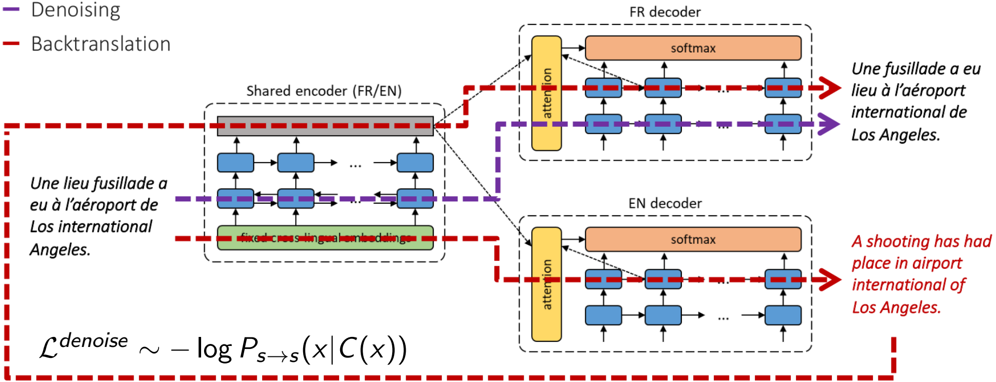
Recap, Unsupervised MT

Basics with Principles (Slides from Mikel Artetxe)

<u>Training</u>

- ~~Supervised~~
- Denoising
- Backtranslation

$$\mathcal{L}^{back} \sim -\log P_{s\to t}(y|u^*(y)) - \log P_{t\to s}(x|v^*(x))$$



$$\mathcal{L}^{denoise} \sim -\log P_{s\to s}(x|C(x)) - \log P_{t\to t}(y|C(y))$$

Recap, Unsupervised MT

Evaluation with BLEU

		newstest2014					newstest2016	
		fr-en	en-fr	de-en	en-de		de-en	en-de
Supervised	*Vaswani et al. (2017)*	-	41.0	-	28.4	-	-	-
	Edunov et al. (2018)	-	45.6	-	35.0	-	-	-
NMT	*Artetxe et al. (2018)*	15.6	15.1	10.2	6.6		-	-
	Lample et al. (2018a)	14.3	15.1	-	-		13.3	9.6
	Lample et al. (2018b)	24.2	25.1	-	-		21.0	17.2
SMT	*Artetxe et al. (2018)*	25.9	26.2	17.4	14.1		23.1	18.2
	Lample et al. (2018b)	27.2	28.1	-	-		22.9	17.9
	Artetxe et al. (2019)	28.4	30.1	20.1	15.8		25.4	19.7
SMT+ NMT	*Lample et al. (2018b)*	27.7	27.6	-	-		25.2	20.2
	Artetxe et al. (2019)	**33.5**	**36.2**	**27.0**	**22.5**		**34.4**	**26.9**
Leaderboard	*Unsupervised*	GPT-3	MASS	GPT-3	GPT-3		Artetxe19	Artetxe19

Unsupervised MT for Low-Resource

An Approach for Low-Resource MT?

- No need for parallel data, only monolingual, **but**
- News Crawl 2007–2013: 749 million tokens in *fr*, 1606 in *de*, 2109 in *en*

Unsupervised MT for Low-Resource

An Approach for Low-Resource MT?

- No need for parallel data, only monolingual, **but**
- News Crawl 2007–2013: 749 million tokens in *fr*, 1606 in *de*, 2109 in *en*

When Does Unsupervised Machine Translation Work?

Kelly Marchisio, Kevin Duhand and Philipp Koehn, WMT 2020

- on different scripts and between dissimilar languages?
- with imperfect domain alignment between source and target corpora?
- with a domain mismatch between training data and the test set?
- on the low-quality data of real low-resource languages?

Unsupervised MT for Low-Resource

When Does Unsupervised Machine Translation Work? Marchisio et al. (2020)

Corpus	**Supervised** A / A	**Parallel** A / A	**Disjoint** A / B	**Diff. Dom.** A / CC*
Ru-En	26.9	23.7 *(-3.2)*	21.2 *(-5.7)*	0.7 *(-26.2)*
Fr-En	29.9	27.6 *(-2.3)*	27.0 *(-2.9)*	3.9 *(-26.0)*

- A, B disjoint parts of UN corpus, CC (Common Crawl)
- SacreBLEU on newstest2019 (Ru-En) and newstest2014 (Fr-En)
- Different domain even more crucial than distant languages
- Why?

Unsupervised MT for Low-Resource

When Does Unsupervised Machine Translation Work? Marchisio et al. (2020)

	Condition	**Min**	**Max**	μ	σ
Fr-En	Parallel	48.00	50.20	49.09	0.69
	Disjoint	37.88	39.09	38.47	0.37
	Diff. Dom.	**0.00**	**17.27**	**7.97**	**7.95**
	News	25.86	28.10	26.97	0.56
	CC	25.87	27.60	26.90	0.51
Ru-En	Parallel	32.24	34.04	32.95	0.47
	Disjoint	25.08	26.96	25.79	0.58
	Diff. Dom.	0.00	0.10	0.01	0.03
	News	22.19	23.77	23.10	0.44
	CC	**0.00**	**24.69**	**12.61**	**11.45**

- Accuracies (%) of induced dictionaries on 10-11 runs. Bold experiments were unstable

Unsupervised MT for Low-Resource

When Does Unsupervised Machine Translation NOT Work? Ruiter et al. (2021)

	English	**Afrikaans**	**Nepali**	**Kannada**	**Yorúbà**	**Swahili**	**Burmese**
Typology	fusional	fusional	fusional	agglutinative	analytic	agglutinative	analytic
Word Order	SVO	SOV,SVO	SOV	SOV	SOV,SVO	SVO	SOV
Script	Latin	Latin	Brahmic	Brahmic	Latin	Latin	Brahmic
sim(L–en)	1.000	0.822	0.605	0.602	0.599	0.456	0.419

- We have seen different domains (src vs. tgt, train vs. test). But also...
- When the word order is very different, different typology, different script
- All this makes mapping word embeddings a challenge

Unsupervised MT for Low-Resource

When Does Unsupervised Machine Translation NOT Work? Ruiter et al. (2021)

Pair	Init.	Config.	Best	UMT	USMT+NMT	LASER	TSS	#P (k)
en2af	WE	B+BT	51.2±.9	27.9±.8	**44.2±.9**	52.1±1.0	35.3	37
af2en	WE	B+BT	52.2±.9	1.4±.1	**0.7±.1**	52.9±.9	–	
en2kn	DAE	B+BT+WT+N	0.3±.0	0.0±.0	**0.0±.0**	–	21.3	397
kn2en	DAE	B+BT+WT+N	0.9±.1	0.0±.0	**0.0±.0**	–	40.3	397
en2my	DAE	B(+BT+WT)	0.1±.0	0.1±.0	**0.0±.0**	0.0±.0	39.3	223
my2en	DAE	B(+BT+WT)	0.7±.1	0.0±.0	**0.0±.0**	0.1±.0	38.6	223
en2ne	DAE	B+BT+WT+N	0.3±.0	0.1±.0	**0.0±.0**	0.5±.1	8.8	–
ne2en	DAE	B+BT+WT(+N)	0.5±.0	0.0±.0	**0.0±.0**	0.2±.0	21.5	–
en2sw	WE	B+BT+WT+N	7.7±.3	3.6±.2	**0.2±.0**	10.0±.3	14.8	995
sw2en	DAE	B+BT	6.8±.2	0.3±.0	**0.0±.0**	8.4±.3	19.7	995
en2yo	WE	B+BT+WT	2.9±.1	1.0±.1	**0.3±.1**	–	12.3	501
yo2en	DAE	B+BT+WT	3.1±.1	0.6±.0	**0.0±.0**	–	22.4	501

BLEU on heterogeneous test sets

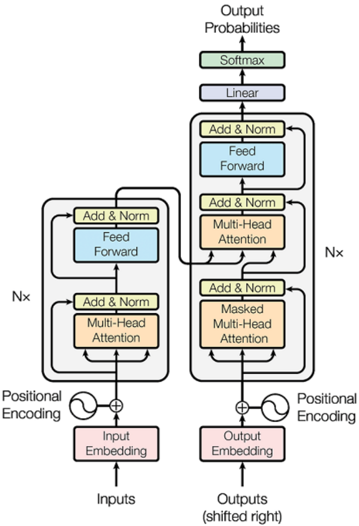
Session III

Outline

1. Unsupervised MT
   - Recap on Basics & Cross-Lingual Embeddings
   - The Low-Resource Setting
2. Supervised NMT
   - Basics
   - The Low-Resource Setting
   - Multilingual Neural Machine Translation
3. Self-Supervised NMT
   - Basics
   - The Low Resource Setting (Session IV)

Supervised NMT

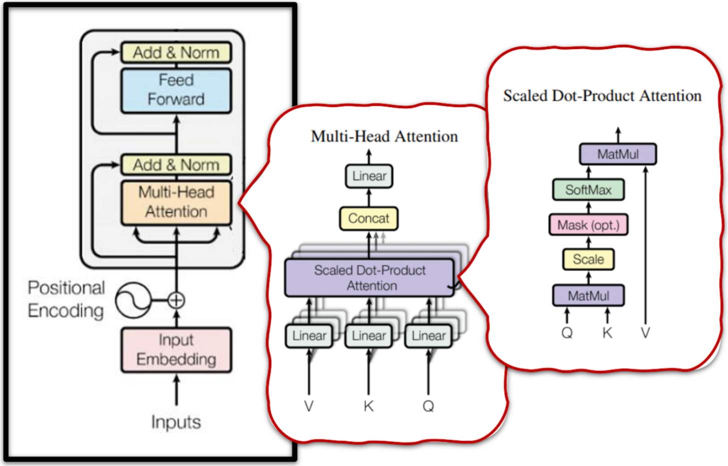
The Transformer, a Seq2Seq Architecture



(Vaswani et al., 2017)

Supervised NMT

NLP 2020 Summary: Transformer Blocks



Supervised NMT

Neural Machine Translation, Results

- Papers *fight* for a $+1$ BLEU improvement
- Several evaluation campaigns, traditional and general: WMT and IWSLT
- Automatic (from BLEU to COMET...) vs manual (DA) evaluations
- Super-human performance vs. fair evaluations
- 2021 campaign being evaluated right now

Supervised NMT

WMT 2020: High-Resource, Close Languages (Direct Assessments)

English→German

Ave.	Ave. z	System
90.5	0.569	HUMAN-B
87.4	0.495	OPPO
88.6	0.468	Tohoku-AIP-NTT
85.7	0.446	HUMAN-A
84.5	0.416	Online-B
84.3	0.385	Tencent-Translation
84.6	0.326	VolcTrans
85.3	0.322	Online-A
82.5	0.312	eTranslation
84.2	0.299	HUMAN-paraphrase
82.2	0.260	AFRL
81.0	0.251	UEDIN
79.3	0.247	PROMT-NMT
77.7	0.126	Online-Z
73.9	−0.120	Online-G
68.1	−0.278	zlabs-nlp
65.5	−0.338	WMTBiomedBaseline

German→English

Ave.	Ave. z	System
82.6	0.228	VolcTrans
84.6	0.220	OPPO
82.2	0.186	HUMAN
81.5	0.179	Tohoku-AIP-NTT
81.3	0.179	Online-A
81.5	0.172	Online-G
79.8	0.171	PROMT-NMT
82.1	0.167	Online-B
78.5	0.131	UEDIN
78.8	0.085	Online-Z
74.2	−0.079	WMTBiomedBaseline
71.1	−0.106	zlabs-nlp
20.5	−1.618	yolo

Supervised NMT

WMT 2020: High-Resource, Distant Languages (Direct Assessments)

English→Japanese

Ave.	Ave. z	System
79.7	0.576	HUMAN
77.7	0.502	NiuTrans
76.1	0.496	Tohoku-AIP-NTT
75.8	0.496	OPPO
75.9	0.492	ENMT
71.8	0.375	NICT-Kyoto
71.3	0.349	Online-A
70.2	0.335	Online-B
63.9	0.159	zlabs-nlp
59.8	0.032	Online-Z
53.9	−0.132	SJTU-NICT
52.8	−0.164	Online-G

Japanese→English

Ave.	Ave. z	System
75.1	0.184	Tohoku-AIP-NTT
76.4	0.147	NiuTrans
74.1	0.088	OPPO
75.2	0.084	NICT-Kyoto
73.3	0.068	Online-B
70.9	0.026	Online-A
71.1	0.019	eTranslation
64.1	−0.208	zlabs-nlp
66.0	−0.220	Online-G
61.7	−0.240	Online-Z

Supervised NMT

WMT 2020: Lower-Resource, Distant Languages (Direct Assessments)

English→Khmer

Ave.	Ave. z	System
77.4	0.478	GTCOM
76.1	0.435	Online-B
74.6	0.386	Huawei-TSC
73.3	0.349	HUMAN
71.1	0.266	VolcTrans
63.8	0.059	Online-Z
60.9	−0.061	OPPO
57.0	−0.164	Online-Z

Khmer→English

Ave.	Ave. z	System
69.0	0.168	Online-B
69.4	0.146	GTCOM
68.5	0.136	Huawei-TSC
62.6	−0.047	VolcTrans
58.1	−0.210	OPPO
56.9	−0.222	Online-Z
55.5	−0.282	Online-G

English→Pashto

Ave.	Ave. z	System
73.0	0.244	GTCOM
71.9	0.180	Huawei-TSC
70.4	0.162	OPPO
69.7	0.158	Online-B
68.8	0.092	HUMAN
67.7	0.055	Online-Z
66.9	−0.029	VolcTrans

Pashto→English

Ave.	Ave. z	System
67.3	0.032	Online-B
66.7	0.024	GTCOM
65.5	−0.016	Huawei-TSC
62.7	−0.106	VolcTrans
62.1	−0.164	OPPO
61.0	−0.195	Online-Z

Supervised NMT

The Low-Resource Setting

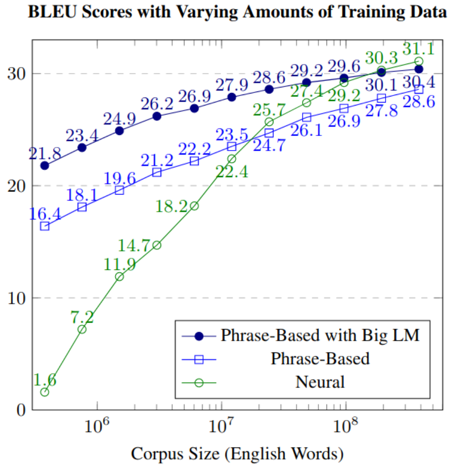
- Deep learning needs a huge amount of data
- As any machine learning problem, **parameter tuning** is crucial...
- but it is also extremely slow for NMT
- Initial belief that SMT is better than NMT
- **Nope!** Tune your system... and use a network you can *fill*
  - small network, fewer layers, larger dropout, less vocabulary...

The Low-Resource Setting

Revisiting Low-Resource Neural Machine Translation

Koehn and Knowles, 2017

- 6 challenges for NMT
  - Amounts of training data
- BLEU scores for English–Spanish systems



Sennrich and Zhang, ACL, 2019

German→English IWSLT results, BLEU

ID	system	100k words	3.2M words
1	phrase-based SMT	15.87 ± 0.19	26.60 ± 0.00
2	NMT baseline	0.00 ± 0.00	25.70 ± 0.33

Sennrich and Zhang, ACL, 2019

German→English IWSLT results, BLEU

ID	system	100k words	3.2M words
1	phrase-based SMT	15.87 ± 0.19	26.60 ± 0.00
2	NMT baseline	0.00 ± 0.00	25.70 ± 0.33
3	2 + "mainstream improvements" (dropout, tied embeddings, layer normalization, bideep RNN, label smoothing)	7.20 ± 0.62	31.93 ± 0.05

Sennrich and Zhang, ACL, 2019

German→English IWSLT results, BLEU

ID	system	100k words	3.2M words
1	phrase-based SMT	15.87 ± 0.19	26.60 ± 0.00
2	NMT baseline	0.00 ± 0.00	25.70 ± 0.33
3	2 + "mainstream improvements" (dropout, tied embeddings, layer normalization, bideep RNN, label smoothing)	7.20 ± 0.62	31.93 ± 0.05
4	3 + **reduce BPE vocabulary** (14k → 2k symbols)	12.10 ± 0.16	-
5	4 + reduce batch size (4k → 1k tokens)	12.40 ± 0.08	31.97 ± 0.26
6	5 + lexical model	13.03 ± 0.49	31.80 ± 0.22

The Low-Resource Setting

Revisiting Low-Resource Neural Machine Translation

Sennrich and Zhang, ACL, 2019

German→English IWSLT results, BLEU

ID	system	100k words	3.2M words
1	phrase-based SMT	15.87 ± 0.19	26.60 ± 0.00
2	NMT baseline	0.00 ± 0.00	25.70 ± 0.33
3	2 + "mainstream improvements" (dropout, tied embeddings, layer normalization, bideep RNN, label smoothing)	7.20 ± 0.62	31.93 ± 0.05
4	3 + **reduce BPE vocabulary** (14k → 2k symbols)	12.10 ± 0.16	-
5	4 + reduce batch size (4k → 1k tokens)	12.40 ± 0.08	31.97 ± 0.26
6	5 + lexical model	13.03 ± 0.49	31.80 ± 0.22
7	5 + **aggressive (word) dropout**	15.87 ± 0.09	$\mathbf{33.60} \pm 0.14$
8	7 + other hyperparameter tuning (learning rate, model depth, label smoothing rate)	$\mathbf{16.57} \pm 0.26$	32.80 ± 0.08
9	8 + lexical model	16.10 ± 0.29	33.30 ± 0.08

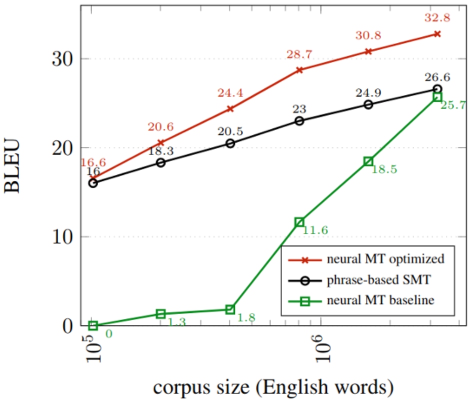
The Low-Resource Setting

Revisiting Low-Resource Neural Machine Translation

Sennrich and Zhang, ACL, 2019

- German→English learning curve
- Beginning of Koehn & Knowles graph

BLEU Scores with Varying Amounts of Training Data



So, clever hyper-parameter tuning is important, but this does not exclude other techniques

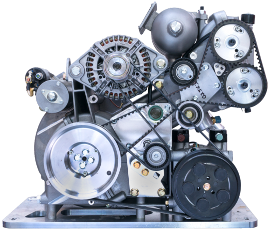
- Data augmentation
- Pre-training
- **Multilinguality**

- Machine translation is at least a bilingual task
- Neural machine translation encodes semantics in vectors (WE)
- Straightforward extension of NMT to multilingual NMT (ML-NMT)
- Simple architecture for ML-NMT: shared encoder & shared decoder
- ML word (or context) vectors lie in the same space (CL-WE)

Multilingual NMT

Basics: Mix the Corpus

traveling around the world



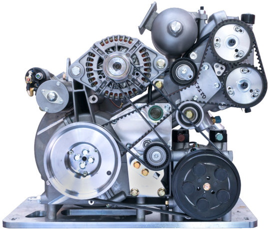
um die Welt reisen

NMT Brain
en2de

Multilingual NMT

Basics: Mix the Corpus

I like hummus



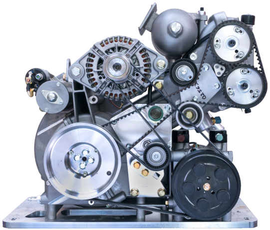
ich mag Hummus

NMT Brain
en2de

Multilingual NMT

Basics: Mix the Corpus

m'agrada l'hummus



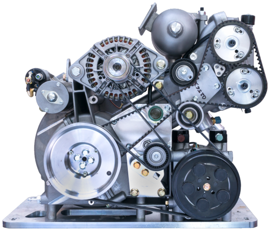
ich mag Hummus

NMT Brain
{en,ca}2de

Multilingual NMT

Basics: Mix the Corpus

<2en> m'agrada l'hummus



I like hummus

NMT Brain
{en,ca}2{en,de}

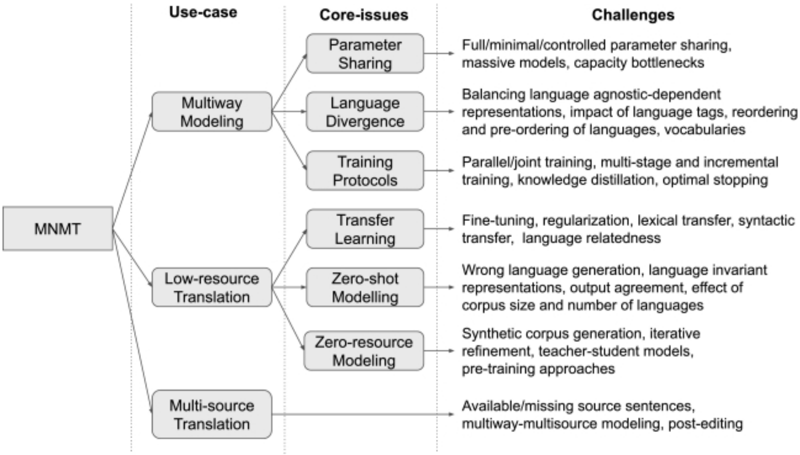
Multilingual NMT

Why should I go Multilingual?

- Shared vocabulary among languages (hummus!)
- Remember dictionaries in supervised mappings for CL-WE?
  (numbers are also shared vocabulary!)
- In the low-resource setting, we use small BPE
  that's a lot of shared vocabulary!
- Very simple to implement (tagging a corpus)
- Simpler to mantain (1 vs. $N(N-1)$ models)

Multilingual NMT

A Survey of Multilingual Neural Machine Translation (Dabre et al., 2020)



Should I go Multilingual?

In general,
multilinguality is good for the low-resource language (if any);
neutral or bad for the high-resource language in the group (if any)

Besides, it has other applications
SS-NMT

Multilingual NMT

Towards Self-Supervised NMT

- Machine translation is at least a bilingual task
- Neural machine translation encodes semantics in vectors

- Machine translation is at least a bilingual task
- Neural machine translation encodes semantics in vectors

- Straightforward extension of NMT to multilingual NMT (ML-NMT)
- Simple architecture for ML-NMT: shared encoder & shared decoder

  `<2en> Es war ein riesiger Erfolg` ‖ `It was a huge success`

  `<2de> È stato un enorme successo` ‖ `Es war ein riesiger Erfolg`

Multilingual NMT

Towards Self-Supervised NMT

- Machine translation is at least a bilingual task
- Neural machine translation encodes semantics in vectors

- Straightforward extension of NMT to multilingual NMT (ML-NMT)
- Simple architecture for ML-NMT: shared encoder & shared decoder

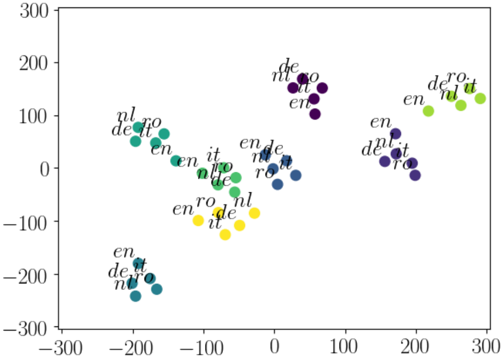
  `<2en> Es war ein riesiger Erfolg` ‖ `It was a huge success`
  `<2de> È stato un enorme successo` ‖ `Es war ein riesiger Erfolg`

- ML word (or context) vectors lie in the same space, **but how?**

Multilingual NMT

Multilingual Semantic Space for Context Vectors (easy)

(España-Bonet & van Genabith, 2018)



ML-NMT $\{de, en, nl, it, ro\} \rightarrow \{de, en, nl, it, ro\}$ with TED talks (t-SNE projection)

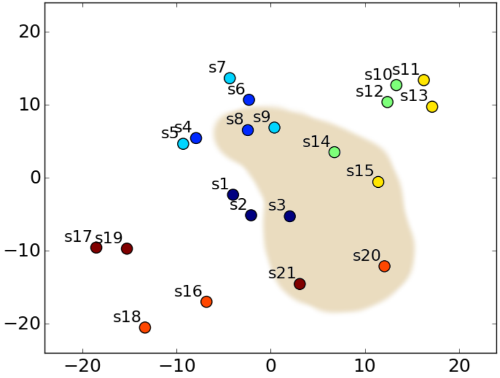
Multilingual Semantic Space for Context Vectors (easy)

- Sentences are clustered according to semantics (not languages)
- Ideal corpus, not a big challenge for NMT
- Let's see something more challenging (for the NMT system!)

Multilingual NMT

Multilingual Semantic Space for Context Vectors (hard)

(España-Bonet et al., 2017)



ML-NMT $\{en, es, ar\} \rightarrow \{en, es, ar\}$ with heterogeneous corpora (t-SNE projection)

Multilingual NMT

Multilingual Semantic Space for Context Vectors (hard)

$s1$:$t1$	Spain princess testifies in historic fraud probe
$s2$:$t1$	Princesa de España testifica en juicio histórico de fraude
$s3$:$t1$	أميرة أسبانيا تدلي بشهادتها في قضية احتيال تاريخي.
$s4$:$t2$	You do not need to worry.
$s5$:$t3$	You don't have to worry.
$s6$:$t2$	No necesitas preocuparte.
$s7$:$t3$	No te tienes por que preocupar.
$s8$:$t2$	لا ينبغي أن تقلق
$s9$:$t3$	لا ينبغي أن تجزع.
$s10$:$t4$	Mandela's condition has 'improved'
$s11$:$t5$	Mandela's condition has 'worsened over past 48 hours'
$s12$:$t4$	La salud de Mandela ha 'mejorado'
$s13$:$t5$	La salud de Mandela 'ha empeorado en las últimas 48 horas'
$s14$:$t4$	لقد تحسّنت حالة مانديلا الصحية.
$s15$:$t5$	ساءت الحالة الصحية لمانديلا خلال ال ٤٨ ساعة الماضية.
$s16$:$t6$	Vector space representation results in the loss of the order which the terms are in the document.
$s17$:$t7$	If a term occurs in the document, the value will be non-zero in the vector.
$s18$:$t6$	La representación en el espacio de vecores implica la pérdida del órden en el que los términos ocurren en el documento.
$s19$:$t7$	Si un término ocurre en el document, el valor en el vector será distinto de cero.
$s20$:$t6$	يؤدي تمثيلُ فضاءِ المتجهِ إلى فقد الترتيب الذي تكون عليه المصطلحات في الوثيقة.
$s21$:$t7$	إذا ما ورد مصطلح في الوثيقة، فالقيمة ستكون غيرصفرية المتجه.

Multilingual NMT

Multilingual Semantic Space for Context Vectors (hard)

$s1{:}t1$	Spain princess testifies in historic fraud probe
$s2{:}t1$	Princesa de España testifica en juicio histórico de fraude
$s3{:}t1$	أميرة أسبانيا تدلى بشهادتها فى قضية احتيال تاريخى.
$s4{:}t2$	You do not need to worry.
$s5{:}t3$	You don’t have to worry.
$s6{:}t2$	No necesitas preocuparte.
$s7{:}t3$	No te tienes por que preocupar.
$s8{:}t2$	لا ينبغي أن تقلق
$s9{:}t3$	لا ينبغي أن تجزع.
$s10{:}t4$	Mandela’s condition has ’improved’
$s11{:}t5$	Mandela’s condition has ’worsened over past 48 hours’
$s12{:}t4$	La salud de Mandela ha ’mejorado’
$s13{:}t5$	La salud de Mandela ’ha empeorado en las últimas 48 horas’
$s14{:}t4$	لقد تحسنّت حالة مانديلا الصحية.
$s15{:}t5$	ساءت الحالة الصحية لمانديلا خلال ال ٨٤ ساعة الماضية.
$s16{:}t6$	Vector space representation results in the loss of the order which the terms are in the document.
$s17{:}t7$	If a term occurs in the document, the value will be non-zero in the vector.
$s18{:}t6$	La representación en el espacio de vecores implica la pérdida del órden en el que los términos ocurren en el documento.
$s19{:}t7$	Si un término ocurre en el document, el valor en el vector será distinto de cero.
$s20{:}t6$	يؤدي تمثيلُ فضاءِ المتجهِ إلى فقد الترتيب الذي تكون عليه المصطلحات في الوثيقة.
$s21{:}t7$	إذا ما ورد مصطلح في الوثيقة، فالقيمة ستكون غيرصفرية المتجه.

Multilingual NMT

Multilingual Semantic Space for Context Vectors (hard)

$s1{:}t1$	Spain princess testifies in historic fraud probe
$s2{:}t1$	Princesa de España testifica en juicio histórico de fraude
$s3{:}t1$	أميرة أسبانيا تدلي بشهادتها في قضية احتيال تاريخي.
$s4{:}t2$	You do not need to worry.
$s5{:}t3$	You don't have to worry.
$s6{:}t2$	No necesitas preocuparte.
$s7{:}t3$	No te tienes por que preocupar.
$s8{:}t2$	لا ينبغي أن تقلق
$s9{:}t3$	لا ينبغي أن تجزع.
$s10{:}t4$	Mandela's condition has 'improved'
$s11{:}t5$	Mandela's condition has 'worsened over past 48 hours'
$s12{:}t4$	La salud de Mandela ha 'mejorado'
$s13{:}t5$	La salud de Mandela 'ha empeorado en las últimas 48 horas'
$s14{:}t4$	لقد تحسنّت حالة مانديلا الصحية.
$s15{:}t5$	ساءت الحالة الصحية لمانديلا خلال ال ٨٤ ساعة الماضية.
$s16{:}t6$	Vector space representation results in the loss of the order which the terms are in the document.
$s17{:}t7$	If a term occurs in the document, the value will be non-zero in the vector.
$s18{:}t6$	La representación en el espacio de vecores implica la pérdida del órden en el que los términos ocurren en el documento.
$s19{:}t7$	Si un término ocurre en el document, el valor en el vector será distinto de cero.
$s20{:}t6$	يؤدي تمثيلُ فضاءِ المتجهِ إلى فقد الترتيب الذي تكون عليه المصطلحات في الوثيقة.
$s21{:}t7$	إذا ما ورد مصطلح في الوثيقة، فالقيمة ستكون غيرصفرية المتجه.

Multilingual NMT

Multilingual Semantic Space for Context Vectors (hard)

$s1{:}t1$	Spain princess testifies in historic fraud probe
$s2{:}t1$	Princesa de España testifica en juicio histórico de fraude
$s3{:}t1$	أميرة أسبانيا تدلي بشهادتها في قضية احتيال تاريخي.
$s4{:}t2$	You do not need to worry.
$s5{:}t3$	You don't have to worry.
$s6{:}t2$	No necesitas preocuparte.
$s7{:}t3$	No te tienes por que preocupar.
$s8{:}t2$	لا ينبغي أن تقلق
$s9{:}t3$	لا ينبغي أن تجزع.
$s10{:}t4$	Mandela's condition has 'improved'
$s11{:}t5$	Mandela's condition has 'worsened over past 48 hours'
$s12{:}t4$	La salud de Mandela ha 'mejorado'
$s13{:}t5$	La salud de Mandela 'ha empeorado en las últimas 48 horas'
$s14{:}t4$	لقد تحسّنت حالة مانديلا الصحية.
$s15{:}t5$	ساءت الحالة الصحية لمانديلا خلال ال ٤٨ ساعة الماضية.
$s16{:}t6$	Vector space representation results in the loss of the order which the terms are in the document.
$s17{:}t7$	If a term occurs in the document, the value will be non-zero in the vector.
$s18{:}t6$	La representación en el espacio de vecores implica la pérdida del órden en el que los términos ocurren en el documento.
$s19{:}t7$	Si un término ocurre en el document, el valor en el vector será distinto de cero.
$s20{:}t6$	يؤدي تمثيلُ فضاءِ المتجهِ إلى فقد الترتيب الذي تكون عليه المصطلحات في الوثيقة.
$s21{:}t7$	إذا ما ورد مصطلح في الوثيقة، فالقيمة ستكون غيرصفرية المتجه.

Multilingual NMT

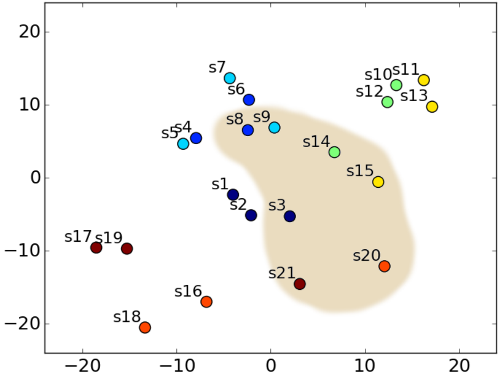
Multilingual Semantic Space for Context Vectors (hard)

$s1$:$t1$	Spain princess testifies in historic fraud probe
$s2$:$t1$	Princesa de España testifica en juicio histórico de fraude
$s3$:$t1$	أميرة أسبانيا تدلي بشهادتها في قضية احتيال تاريخي.
$s4$:$t2$	You do not need to worry.
$s5$:$t3$	You don't have to worry.
$s6$:$t2$	No necesitas preocuparte.
$s7$:$t3$	No te tienes por que preocupar.
$s8$:$t2$	لا ينبغي أن تقلق
$s9$:$t3$	لا ينبغي أن تجزع.
$s10$:$t4$	Mandela's condition has 'improved'
$s11$:$t5$	Mandela's condition has 'worsened over past 48 hours'
$s12$:$t4$	La salud de Mandela ha 'mejorado'
$s13$:$t5$	La salud de Mandela 'ha empeorado en las últimas 48 horas'
$s14$:$t4$	لقد تحسّنت حالة مانديلا الصحية.
$s15$:$t5$	ساءت الحالة الصحية لمانديلا خلال ال ٤٨ ساعة الماضية.
$s16$:$t6$	Vector space representation results in the loss of the order which the terms are in the document.
$s17$:$t7$	If a term occurs in the document, the value will be non-zero in the vector.
$s18$:$t6$	La representación en el espacio de vecores implica la pérdida del órden en el que los términos ocurren en el documento.
$s19$:$t7$	Si un término ocurre en el document, el valor en el vector será distinto de cero.
$s20$:$t6$	يؤدي تمثيلُ فضاءِ المتجهِ إلى فقد الترتيب الذي تكون عليه المصطلحات في الوثيقة.
$s21$:$t7$	إذا ما ورد مصطلح في الوثيقة، فالقيمة ستكون غيرصفرية المتجه.

Multilingual NMT

Multilingual Semantic Space for Context Vectors (hard)

(España-Bonet et al., 2017)



ML-NMT $\{en, es, ar\} \rightarrow \{en, es, ar\}$ with heterogeneous corpora

Multilingual NMT

How Close are Sentences Together?

Cosine similarities between the internal representations of the sentences in STS2017 and newstest2013 when translated from L1 into different languages L2, L3, L4.

L1	{L2, L3, L4}	<2L2–2L3>	<2L2–2L4>	<2L3–2L4>
ar	$\{en,es,\phi\}$	0.97(5)	–	–
en	$\{es,ar,\phi\}$	0.94(5)	–	–
es	$\{ar,en,\phi\}$	0.91(5)	–	–
de	$\{fr,en,es\}$	*0.97(2)	*0.98(2)	*0.96(2)
fr	$\{en,es,de\}$	0.96(2)	*0.96(2)	*0.97(2)
en	$\{es,de,fr\}$	0.96(2)	0.98(2)	0.96(2)
es	$\{de,fr,es\}$	*0.97(2)	*0.96(2)	0.97(2)

Multilingual Semantic Space for Context Vectors

- Related languages cluster better together
  (for distant languages there might not even exist a mapping)
- The nature of the corpus also affects the clustering
  (corpus in different domains per language make the learning more difficult)
- These trends are common in several NLP tasks

Multilingual NMT

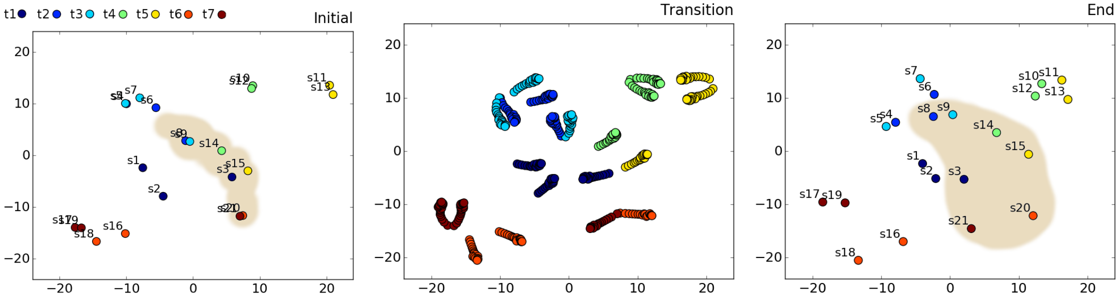
Multilingual Semantic Space for Context Vectors

- Related languages cluster better together
  (for distant languages there might not even exist a mapping)
- The nature of the corpus also affects the clustering
  (corpus in different domains per language make the learning more difficult)
- These trends are common in several NLP tasks
- **What happens during training?**

Multilingual NMT

Evolution of Context Vectors through Training (hard)

(España-Bonet et al., 2017)



ML-NMT $\{en, es, ar\} \rightarrow \{en, es, ar\}$ with heterogeneous corpora

Multilingual NMT

Evolution According to the Similarity: from Translations to Unrelated Sentences

		ar–ar	*en–en*	*ar–en*	*ar–es*	*en–es*
0.1 EPOCHS ($4 \cdot 10^6$sent.)	*trad*					
	semrel					
	unrel					
	$\Delta_{\mathrm{tr-ur}}$					
0.5 EPOCHS ($28 \cdot 10^6$sent.)	*trad*					
	semrel					
	unrel					
	$\Delta_{\mathrm{tr-ur}}$					
1.0 EPOCHS ($56 \cdot 10^6$sent.)	*trad*					
	semrel					
	unrel					
	$\Delta_{\mathrm{tr-ur}}$					
2.0 EPOCHS ($112 \cdot 10^6$sent.)	*trad*	–	–	0.59(07)	0.62(07)	0.71(07)
	semrel	0.80(10)	0.83(08)	0.54(08)	0.60(08)	0.67(08)
	unrel	0.37(12)	0.34(11)	0.26(09)	0.30(10)	0.29(10)
	$\Delta_{\mathrm{tr-ur}}$	–	–	0.33(12)	0.32(12)	0.42(12)

Cosine similarities between the obtained representations of the sentences in the STS2017 test set

trad: sim 5
semrel: sim 4
unrel: sim 0

Multilingual NMT

Evolution According to the Similarity: from Translations to Unrelated Sentences

		ar–ar	*en–en*	*ar–en*	*ar–es*	*en–es*
0.1 EPOCHS ($4 \cdot 10^6$sent.)	*trad*	–	–	0.26(10)	0.76(05)	0.40(09)
	semrel	0.92(03)	0.93(01)	0.24(10)	0.75(06)	0.38(09)
	unrel	0.65(13)	0.66(13)	0.06(09)	0.53(11)	0.14(10)
	$\Delta_{\text{tr}-\text{ur}}$	–	–	0.20(13)	0.23(12)	0.26(13)
0.5 EPOCHS ($28 \cdot 10^6$sent.)	*trad*	–	–	0.61(07)	0.67(06)	0.76(06)
	semrel	0.86(07)	0.87(06)	0.58(08)	0.65(07)	0.73(07)
	unrel	0.48(12)	0.43(12)	0.30(10)	0.37(11)	0.37(11)
	$\Delta_{\text{tr}-\text{ur}}$	–	–	0.32(12)	0.30(12)	0.39(12)
1.0 EPOCHS ($56 \cdot 10^6$sent.)	*trad*	–	–	0.61(08)	0.65(07)	0.74(06)
	semrel	0.83(09)	0.85(07)	0.57(08)	0.63(08)	0.70(08)
	unrel	0.41(12)	0.37(11)	0.27(10)	0.32(11)	0.31(10)
	$\Delta_{\text{tr}-\text{ur}}$	–	–	0.34(12)	0.33(13)	0.43(12)
2.0 EPOCHS ($112 \cdot 10^6$sent.)	*trad*	–	–	0.59(07)	0.62(07)	0.71(07)
	semrel	0.80(10)	0.83(08)	0.54(08)	0.60(08)	0.67(08)
	unrel	0.37(12)	0.34(11)	0.26(09)	0.30(10)	0.29(10)
	$\Delta_{\text{tr}-\text{ur}}$	–	–	0.33(12)	0.32(12)	0.42(12)

Cosine similarities between the obtained representations of the sentences in the STS2017 test set

trad: sim 5
semrel: sim 4
unrel: sim 0

This is a fact. ML-NMT behaves this way.

Can we profit from it?

Session III

Outline

1. Unsupervised MT
   - Recap on Basics & Cross-Lingual Embeddings
   - The Low-Resource Setting
2. Supervised NMT
   - Basics
   - The Low-Resource Setting
   - Multilingual Neural Machine Translation
3. Self-Supervised NMT
   - Basics
   - The Low Resource Setting (Session IV)

Self-Supervised NMT

Question

- NMT embeddings differentiate translations from non-translations very soon
- In a standard NMT, all training sentences are (should be) translations
- Can we feed the system with any kind of sentence pair and let itself decide if it is useful or not?

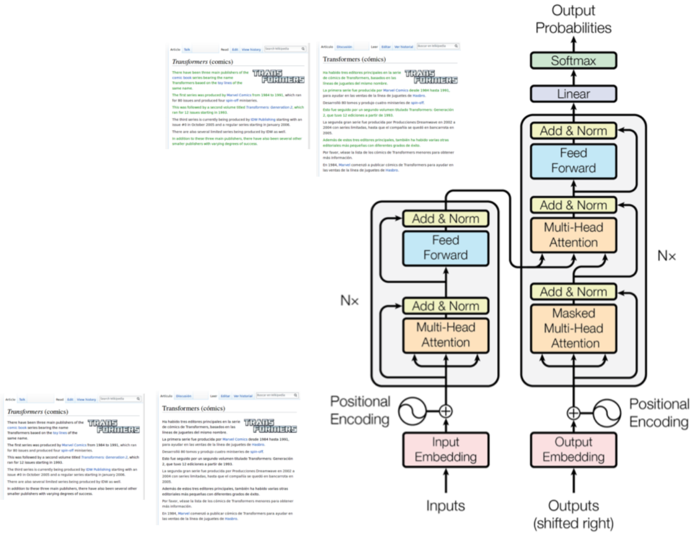
Self-Supervised NMT

Question

- NMT embeddings differentiate translations from non-translations very soon
- In a standard NMT, all training sentences are (should be) translations
- Can we feed the system with any kind of sentence pair and let itself decide if it is useful or not?
- **Yes, we can!**

Self-Supervised NMT

Main Idea I



Self-Supervised NMT

Main Idea II

- Parallel data extraction as an auxiliary task to enable NMT training
- NMT training as an auxiliary task to enhance parallel sentence extraction

Self-Supervised NMT

Main Idea II

- Parallel data extraction as an auxiliary task to enable NMT training
- NMT training as an auxiliary task to enhance parallel sentence extraction

Self-supervision?

Just in a non-standard way, none of the tasks is completely supervised

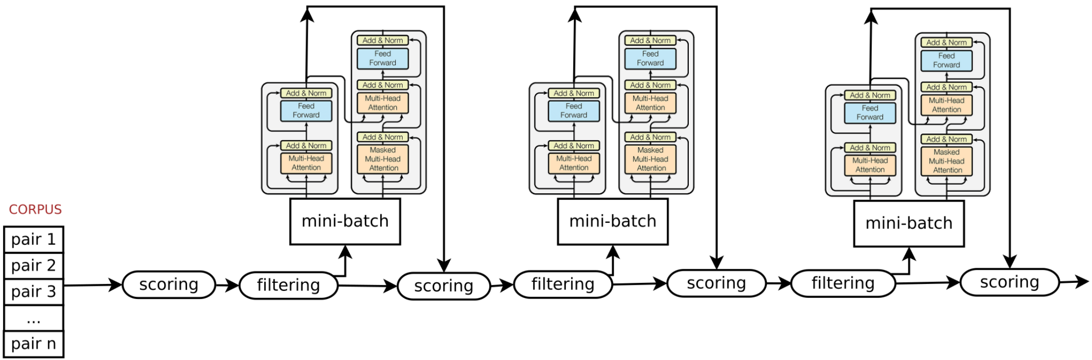
Self-Supervised NMT

Main Idea III (Ruiter et al., ACL, 2019)

- Joint selection of sentences & training NMT
- Uses internal embeddings, i.e., architecture independent
- Bidirectional training $\{$L1, L2$\}\rightarrow\{$L1, L2$\}$ (shared encoder)
- On-line process: embeddings change through epochs, therefore selected sentences change through epochs

Self-Supervised NMT

Training Procedure



More to come!!

Just a spoiler before leaving...

Self-Supervised NMT

SSNMT vs. UMT (vs. NMT)

Pair	Init.	Config.	Best	Base	UMT	UMT+NMT	Laser	TSS	#P (*k*)
en2af	WE	B+BT	**51.2±.9**	48.1±.9	27.9±.8	44.2±.9	**52.1±1.0**	35.3	37
af2en	WE	B+BT	**52.2±.9**	47.9±.9	1.4±.1	0.7±.1	**52.9±.9**	–	–
en2kn	MDAE	B+BT+F	**5.0±.2**	0.0±.0	0.0±.0	0.0±.0	–	21.3	397
kn2en	MDAE	B+BT+F	**9.0±.2**	0.0±.0	0.0±.0	0.0±.0	–	40.3	397
en2my	MDAE	B+BT+F	**0.2±.0**	0.0±.0	0.1±.0	0.0±.0	0.0±.0	39.3	223
my2en	MDAE	B+BT+F	**2.8±.1**	0.0±.0	0.0±.0	0.0±.0	0.1±.0	38.6	223
en2ne	MDAE	B+BT+F	**2.3±.1**	0.0±.0	0.1±.0	0.0±.0	0.5±.1	8.8	–
ne2en	MDAE	B+BT+F	**5.7±.2**	0.0±.0	0.0±.0	0.0±.0	0.2±.0	21.5	–
en2sw	MDAE	B+BT+F	**11.6±.3**	4.2±.2	3.6±.2	0.2±.0	10.0±.3	14.8	995
sw2en	MDAE	B+BT+F	**11.2±.3**	3.6±.2	0.3±.0	0.0±.0	8.4±.3	19.7	995
en2yo	MDAE	B+BT+F	**2.9±.1**	0.3±.1	1.0±.1	0.3±.1	–	12.3	501
yo2en	MDAE	B+BT+F	**5.8±.1**	0.5±.1	0.6±.0	0.0±.0	–	22.4	–

BLEU on heterogeneous test sets

Thanks! And...

wait!

Neural Machine Translation
(Unsupervised, Supervised, Multilingual and Self-Supervised)

Cristina España-Bonet
DFKI GmbH

Low-Resource NLP:
Multilinguality and Machine Translation
Webinar Series — Session III
13th July 2021