# Multilingual Neural Machine Translation

**Cristina España-Bonet**
DFKI GmbH

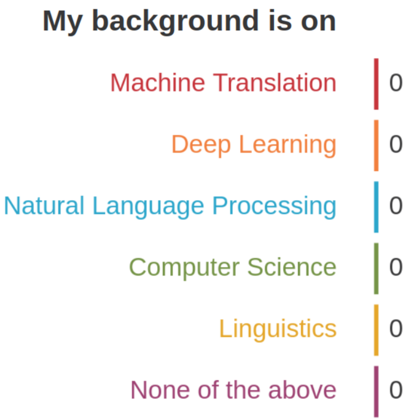
11th Advanced Summer School on NLP
IIIT Hyderabad
23rd June 2022

# Today we'll talk about

1. Neural Machine Translation (Tomorrow!)
   - Multilingual Neural Machine Translation
2. Basics of ML-NMT
   - Behaviour
3. Self-Supervised NMT
4. Evaluating (Large Scale) ML-NMT
   - WMT 2021 Shared Tasks
   - DeltaLM

# Today we'll talk about

## *Before Starting... what do we Know?*



**The Encoder–Decoder Model (with attention)**

1. encodes a sequence of word vectors into a fixed-sized context vector
2. decodes the fixed-sized vector back into a variable-length sequence

Several NLP tasks use nowadays enc–dec architectures:

- Machine translation, but also...
- text summarisation, question answering, chatbots, speech recognition...

**The Encoder–Decoder Model (with attention)**

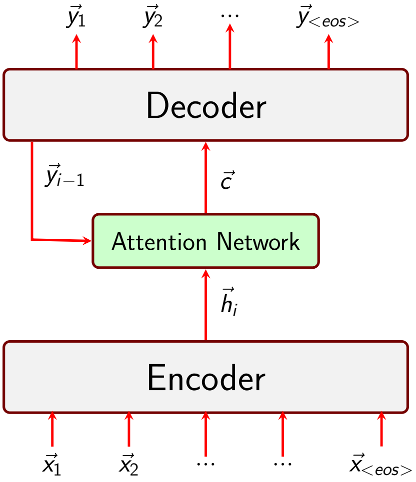
1. encodes a sequence of word vectors into a fixed-sized context vector
2. decodes the fixed-sized vector back into a variable-length sequence

Several NLP tasks use nowadays enc–dec architectures:

- Machine translation, but also...
- text summarisation, question answering, chatbots, speech recognition...

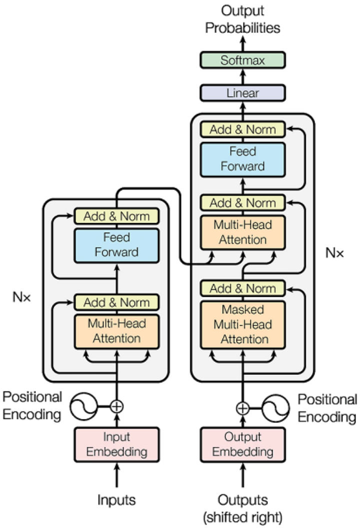
# Neural Machine Translation

## *Basic NMT Model*



# Neural Machine Translation

## *A Transformer to Rule them All!*




(Vaswani et al., 2017)

# Multilingual Machine Translation

## Why?

# Multilingual Neural Machine Translation

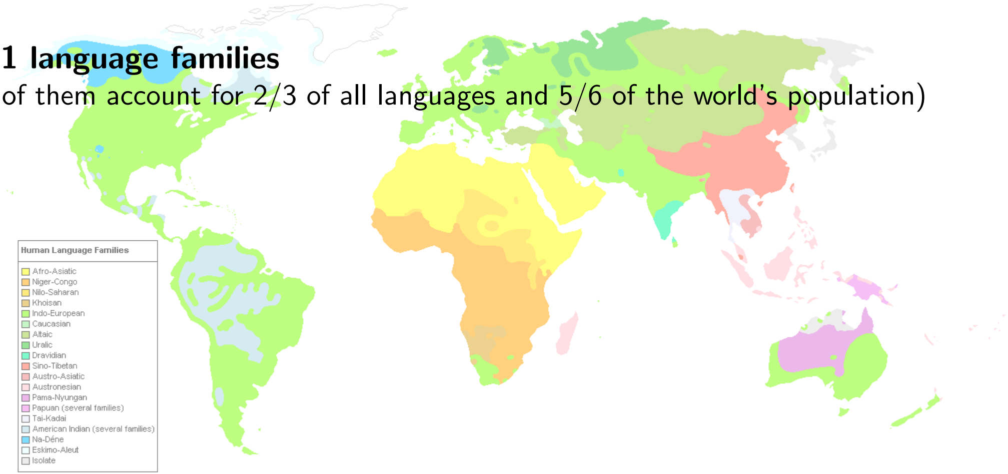
## *Why?*

- There are >7000 languages in the world
  - Do we want/need 7000x7000 MT systems?
  - Do we want 1 MT system to translate from 7000 into 7000 languages?
- Languages share features

# Multilingual Neural Machine Translation

## *Language Relatedness*

- **141 language families**
  (6 of them account for 2/3 of all languages and 5/6 of the world's population)



# Multilingual Neural Machine Translation

## *Why?*

- There are >7000 languages in the world
  - Do we want/need 7000x7000 MT systems?
  - Do we want 1 MT system to translate from 7000 into 7000 languages?
- Languages share features
  - Commonalities among languages can help
  - Main motivation: low-resource languages, but...

# Multilingual Neural Machine Translation

## *Why?*

- There are >7000 languages in the world
  - Do we want/need 7000x7000 MT systems?
  - Do we want 1 MT system to translate from 7000 into 7000 languages?
- Languages share features
  - Commonalities among languages can help
  - **Main motivation: low-resource languages, but...**

# Multilingual Neural Machine Translation

## *ML-NMT can be Convenient and Simple*

- Easier to deploy and mantain (1 system instead of N)
  - Can put together several high-resource languages (capacity!)
  - Help ambiguity?
- Can put together several related languages
  - Can add low-resourced languages to benefit from high-resourced
  - Even zero-shot!
- Code-switching can be dealt almost by construction
  - Bidirectional NMT?

**Simple?**

# Multilingual Neural Machine Translation

## *ML-NMT can be Convenient and Simple*

- Easier to deploy and mantain (1 system instead of N)
  - Can put together several high-resource languages (capacity!)
  - Help ambiguity?
- Can put together several related languages
  - Can add low-resourced languages to benefit from high-resourced
  - Even zero-shot!
- Code-switching can be dealt almost by construction
  - Bidirectional NMT?

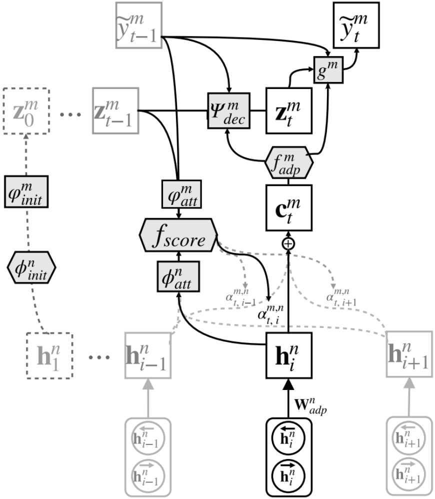
**Simple?**

- ML-NMT can be as **simple** as we want
- ML-NMT can be as **complicated** as we want :-)

- ML-NMT can be as **simple** as we want
- ML-NMT can be as **complicated** as we want :-)

# Multilingual Neural Machine Translation

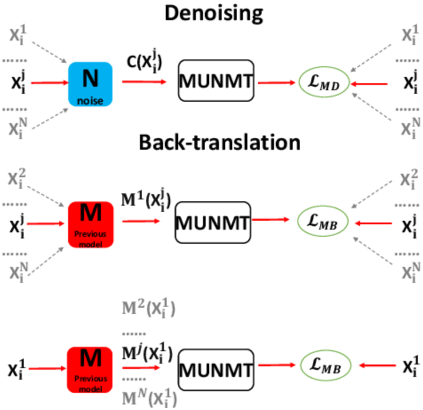
## *Multi-Way, ML-NMT with a Shared Attention Mechanism (Firat et al. 2016)*



Attention-based encoder-decoder that admits a shared attention mechanism with multiple encoders and decoders

# Multilingual Neural Machine Translation

## *Knowledge Distillation for ML-Unsupervised NMT (Sun et al. 2020)*




Single encoder and a single decoder, making use of multilingual data

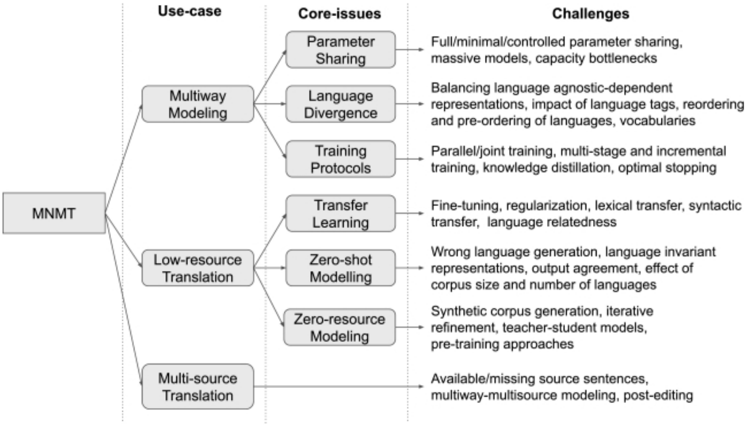
# Multilingual Neural Machine Translation

## *Approaches*

- Multiple encoders and/or decoders
- One encoder, one decoder, joint vocabulary, mixed data in all language pairs
- Any combination you can think of :-)

# Multilingual Neural Machine Translation

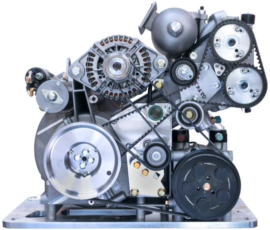
## *A Survey of Multilingual Neural Machine Translation (Dabre et al., 2020)*



# Multilingual Neural Machine Translation

*One Encoder, one Decoder. Easy-peasy!*

traveling around the world



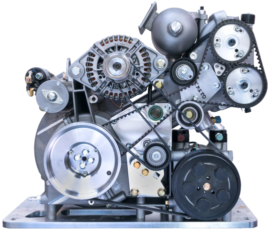
um die Welt reisen

NMT Brain
en2de

# Multilingual Neural Machine Translation

## *One Encoder, one Decoder. Easy-peasy!*

I like hummus



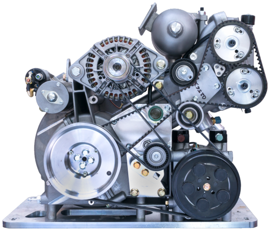
ich mag Hummus

NMT Brain
en2de

# Multilingual Neural Machine Translation

*One Encoder, one Decoder. Easy-peasy!*

m'agrada l'hummus



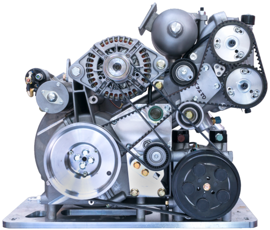
ich mag Hummus

NMT Brain
{en,ca}2de

# Multilingual Neural Machine Translation

*One Encoder, one Decoder. Easy-peasy!*

<2en> m'agrada l'hummus → 



→ I like hummus

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

## *Remember! Basic NMT Model*

**The Encoder–Decoder Model (with attention)**

1. encodes a sequence of word vectors into a fixed-sized context vector
2. decodes the fixed-sized vector back into a variable-length sequence

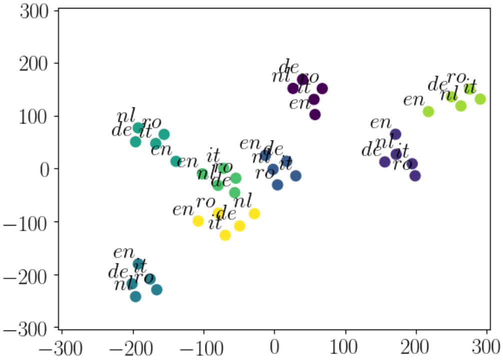
**The Encoder–Decoder Model (with attention)**

1. encodes a sequence of word vectors into a **fixed-sized context vector**
2. decodes the fixed-sized vector back into a variable-length sequence

# Multilingual Neural Machine Translation

## *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 Neural Machine Translation

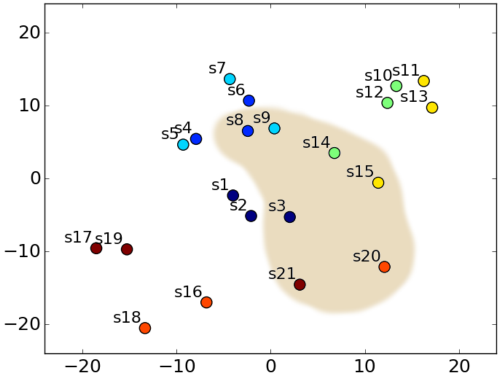
## *Multilingual Semantic Space for Context Vectors (easy)*

- 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 Neural Machine Translation

## *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 Neural Machine Translation

## *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 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 Neural Machine Translation

## 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 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 Neural Machine Translation

## 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 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 Neural Machine Translation

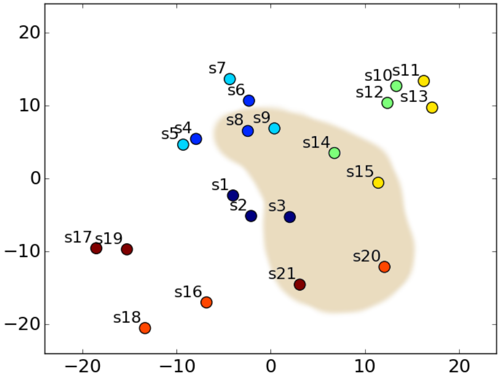
## *Multilingual Semantic Space for Context Vectors (hard)*

| | |
|---|---|
| $s1$:$t1$ | 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 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 Neural Machine Translation

## *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 Semantic Space for Context Vectors*

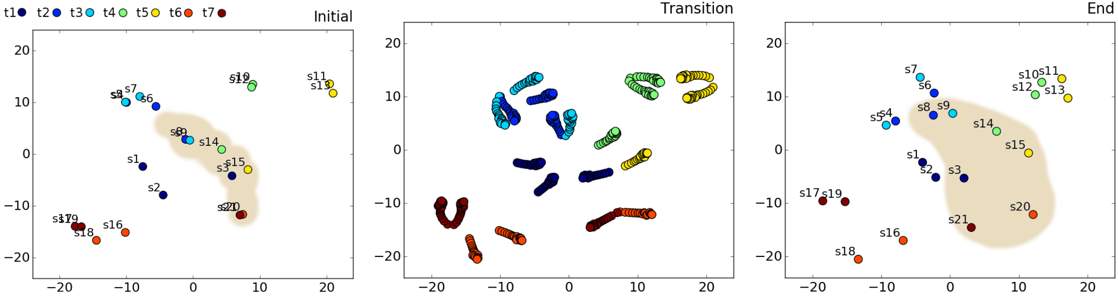
- Related languages cluster better together
  (for distant languages there might not even exist a mapping)
- The nature of the corpus also affects the clustering
  (corpus in different domains per language make the learning more difficult)
- These trends are common in several NLP tasks
- **What happens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

- 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 Neural Machine Translation

## *Evolution of Context Vectors through Training (hard)*

*(España-Bonet et al., 2017)*




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

# How are you doing? Need a Break?

## *Already a Long Way!*

## Where were we?

- 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
- ML word (or context) vectors lie in the same space

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

**Can we profit from it?**

# 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

## *Question*

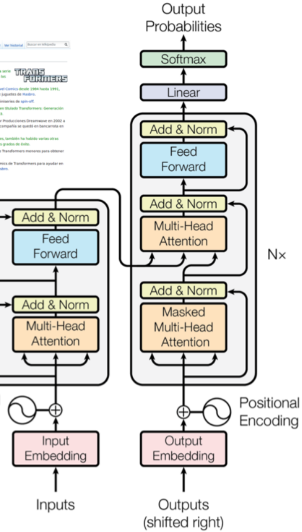
- NMT embeddings differentiate translations from non-translations very soon
- In a standard NMT, all training sentences are (should be) translations
- Can we feed the system with any kind of sentence pair and let itself decide if it is useful or not?
- **Yes, we can!**

# 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-supervision?**

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

# 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

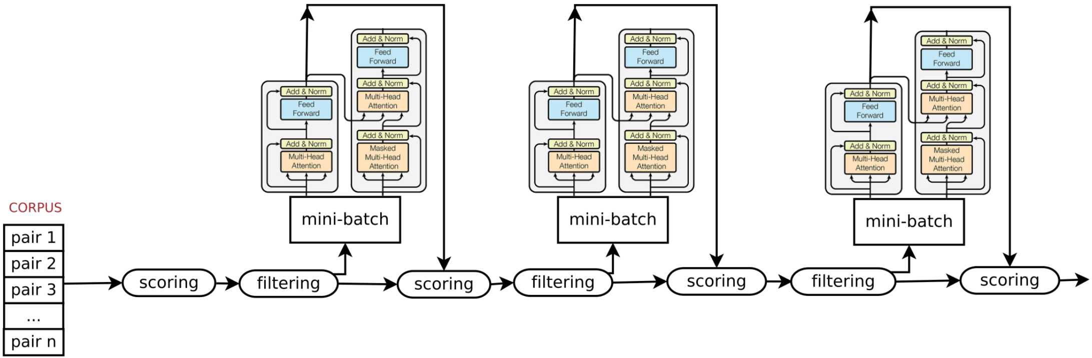
# Self-Supervised NMT

## *Main Idea III (Ruiter et al., 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*



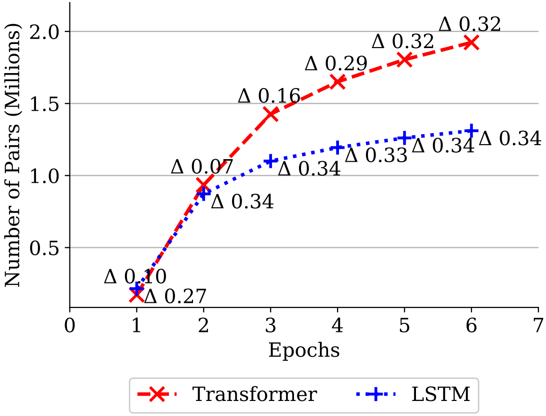
# Self-Supervised NMT

## *Algorithm Description*

1. Internal NMT **representation**: $E_w$ (words); $E_h$ (sentence)
2. **Score** all sentence pairs in a lot (i.e. WP article)
3. **Filter** options
4. Add filtered sentences into a mini-batch
5. Train system when mini-batch is complete
6. Update weights and continue with more data and go again to 1

# Self-Supervised NMT

## *What's going on? —* margP *models*




- The mean difference in similarity between accepted and rejected pairs increases (Δ)
- The number of extracted sentences increases with Δ
- Changes are more prominent at the beginning of the training

# Self-Supervised NMT

## *Open Problems*

1. **Distant Languages** (no/few homographs)
2. **Low-resourced languages**

Similar issues in unsupervised NMT, bilingual embeddings, etc.

Same "solutions"?

**Additions** (Unsupervised NMT-inspired?)

- Initalisation
  - Word embeddings (bilingual word2vec-like embeddings, BWE)
  - Sentence embeddings (BART-style training, Denoising Autoencoder DAE)

- Data augmentation
  - *Online* back-translation
  - Word by word translation (nearest neighbour in BWE)
  - Noise (token deletion, substitution and permutation)

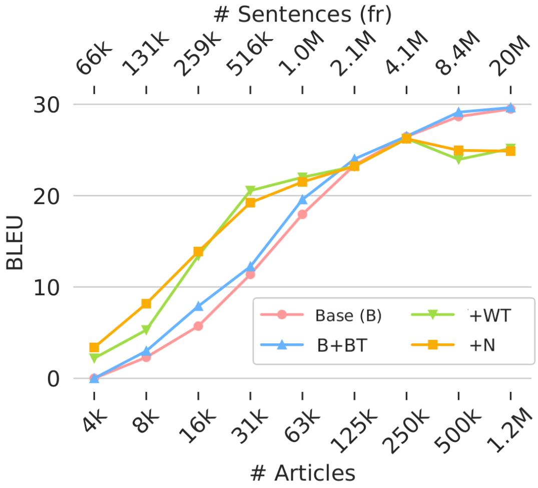
# Self-Supervised NMT

## *Algorithm Description*

1. System initialisation
2. Extract pairs as usual (scoring, filtering)
3. On-line back-translation of rejected pairs
    1. SS-NMT filtering to remove low-quality back-translations
    2. Word translation for rejected back-translations
4. Add noise

# Self-Supervised NMT

## *Data Augmentation vs. Corpus Size*




- WT and N damage high-resource setting
- Significant improvements mid-resource setting
- Small improvements in the low-resource simulated setting

(English & French Wikipedias)

# Self-Supervised NMT

## *But, is this Real Low Resource?*

- Artificial low-resourced setting 👍 (lots of mono data, few comparable)
- Real setting 👎 (few mono data, few comparable, distant languages)

| | **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 |

# Self-Supervised NMT

## *Mmmm... What else?*

- Multilinguality
- Fine-tuning

# Self-Supervised NMT

## *Mmmm... What else?*

- Multilingualit
  - Multilingual comparable corpora
  - Multilingual denosing autoencoder, MDAE
- Fine-tuning
  - Bilingual comparable corpora

# Self-Supervised NMT

## *Automatic Evaluation (BLEU scores on Different Sets)*

Language (L)

**Latin**

| Initialization | | yo B | yo +BT | yo +WT | yo +N | af B | af +BT | af +WT | af +N | sw B | sw +BT | sw +WT | sw +N |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| en2L | none | 0.3±0.1 | 0.3±0.1 | 2.2±0.1 | 0.0±0.0 | 48.1±0.9 | 49.0±1.0 | 1.1±0.1 | 37.1±0.8 | 4.2±0.2 | 6.1±0.2 | 0.9±0.1 | 5.6±0.2 |
| en2L | WE | 0.5±0.1 | 0.4±0.1 | 2.9±0.1 | 0.9±0.0 | 48.1±0.9 | 51.2±0.9 | 8.4±0.5 | 41.7±0.9 | 4.4±0.2 | 5.1±0.2 | 3.0±0.2 | 7.7±0.3 |
| en2L | DAE | 2.0±0.1 | 2.3±0.1 | 2.8±0.1 | 1.2±0.1 | 44.8±0.9 | 48.6±0.9 | 42.3±0.9 | 38.9±0.9 | 5.3±0.2 | 7.2±0.3 | 4.7±0.2 | 4.7±0.2 |
| en2L | MDAE | 1.7±0.1 | 1.5±0.1 | 1.1±0.1 | 2.0±0.1 | 42.1±0.9 | 42.1±0.9 | 36.6±0.9 | 30.3±0.7 | 6.5±0.3 | 7.4±0.3 | 3.3±0.2 | 3.4±0.2 |
| L2en | none | 0.5±0.1 | 0.6±0.1 | 2.7±0.1 | 0.2±0.0 | 47.9±0.9 | 51.3±0.9 | 0.7±0.1 | 38.6±0.9 | 3.6±0.2 | 5.5±0.3 | 0.4±0.0 | 5.0±0.2 |
| L2en | WE | 0.6±0.1 | 0.5±0.1 | 2.5±0.1 | 0.0±0.0 | 48.6±0.9 | 52.2±0.9 | 5.8±0.4 | 43.7±0.9 | 3.6±0.2 | 4.2±0.2 | 2.1±0.1 | 6.3±0.2 |
| L2en | DAE | 2.6±0.1 | 3.0±0.1 | 3.1±0.1 | 2.0±0.1 | 46.2±0.9 | 50.4±0.9 | 43.1±0.9 | 39.5±0.8 | 4.8±0.2 | 6.8±0.2 | 5.6±0.2 | 5.9±0.2 |
| L2en | MDAE | 4.6±0.1 | 4.7±0.1 | 3.9±0.1 | 3.5±0.1 | 43.1±0.9 | 42.5±0.9 | 38.4±0.9 | 31.9±0.8 | 6.8±0.2 | 7.9±0.3 | 4.0±0.2 | 3.5±0.2 |

**Brahmic**

| Initialization | | my B | my +BT | my +WT | my +N | ne B | ne +BT | ne +WT | ne +N | kn B | kn +BT | kn +WT | kn +N |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| en2L | none | 0.0±0.0 | 0.0±0.0 | 0.1±0.0 | 0.1±0.0 | 0.0±0.0 | 0.0±0.0 | 0.2±0.0 | 0.1±0.0 | 0.0±0.0 | 0.0±0.0 | 0.2±0.0 | 0.1±0.0 |
| en2L | WE | 0.0±0.0 | 0.0±0.0 | 0.1±0.0 | 0.1±0.0 | 0.0±0.0 | 0.0±0.0 | 0.2±0.0 | 0.1±0.0 | 0.0±0.0 | 0.0±0.0 | 0.2±0.0 | 0.2±0.0 |
| en2L | DAE | 0.1±0.0 | 0.1±0.0 | 0.1±0.0 | 0.0±0.0 | 0.1±0.0 | 0.2±0.0 | 0.1±0.0 | 0.3±0.0 | 0.0±0.0 | 0.0±0.0 | 0.2±0.0 | 0.3±0.0 |
| en2L | MDAE | 0.1±0.0 | 0.1±0.0 | 0.1±0.0 | 0.1±0.0 | 0.9±0.1 | 1.0±0.1 | 0.3±0.1 | 0.3±0.1 | 3.3±0.1 | 3.1±0.1 | 0.8±0.1 | 0.5±0.1 |
| L2en | none | 0.0±0.0 | 0.0±0.0 | 0.1±0.0 | 0.2±0.1 | 0.0±0.0 | 0.0±0.0 | 0.2±0.0 | 0.1±0.0 | 0.0±0.0 | 0.0±0.0 | 0.2±0.0 | 0.7±0.1 |
| L2en | WE | 0.1±0.0 | 0.0±0.0 | 0.2±0.0 | 0.4±0.0 | 0.1±0.0 | 0.0±0.0 | 0.1±0.0 | 0.4±0.1 | 0.0±0.0 | 0.0±0.0 | 0.2±0.0 | 0.2±0.0 |
| L2en | DAE | 0.7±0.1 | 0.6±0.0 | 0.7±0.1 | 0.4±0.1 | 0.3±0.1 | 0.3±0.1 | 0.5±0.1 | 0.5±0.0 | 0.0±0.0 | 0.0±0.0 | 0.7±0.1 | 0.9±0.1 |
| L2en | MDAE | 1.5±0.1 | 1.7±0.1 | 0.8±0.1 | 0.5±0.1 | 3.2±0.1 | 3.3±0.1 | 0.8±0.1 | 0.6±0.1 | 5.2±0.1 | 5.3±0.1 | 1.9±0.1 | 1.4±0.1 |

# Self-Supervised NMT

## *Data Augmentation vs. Multilinguality vs. Fine-tuning*

BLEU scores on different test sets per language

| | *en–af* | | *en–kn* | | *en–my* | | *en–ne* | | *en–sw* | | *en–yo* | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | → | ← | → | ← | → | ← | → | ← | → | ← | → | ← |
| Baseline | 48.1 | 48.6 | 0.0 | 0.0 | 0.0 | 0.1 | 0.0 | 0.1 | 4.4 | 3.6 | 0.5 | 0.6 |
| Best Bilingual | **51.2** | **52.2** | 0.3 | 0.9 | 0.1 | 0.7 | 0.3 | 0.5 | 7.7 | 6.8 | **2.9** | 3.1 |
| MDAE | 42.5 | 42.5 | 3.1 | 5.3 | 0.1 | 1.7 | 1.0 | 3.3 | 7.4 | 7.9 | 1.5 | 4.7 |
| MDAE+F | 46.3 | 50.2 | **5.0** | **9.0** | **0.2** | **2.8** | **2.3** | **5.7** | **11.6** | **11.2** | **2.9** | **5.8** |
| Typology *L* | fusional | | agglutinative | | analytic | | fusional | | agglutinative | | analytic | |
| Word Order *L* | SOV,SVO | | SOV | | SOV | | SOV | | SVO | | SOV,SVO | |
| Word Overlap | 7.1% | | 1.4% | | 2.1% | | 0.6% | | 6.5% | | 5.7% | |
| Tokens *L* | 27.6 M | | 30.0 M | | 15.3 M | | 7.5 M | | 8.7 M | | 0.5 M | |

# Self-Supervised NMT

## *Data Augmentation vs. Multilinguality vs. Fine-tuning*

BLEU scores on different test sets per language

| | *en–af* | | *en–kn* | | *en–my* | | *en–ne* | | *en–sw* | | *en–yo* | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | → | ← | → | ← | → | ← | → | ← | → | ← | → | ← |
| Baseline | 48.1 | 48.6 | 0.0 | 0.0 | 0.0 | 0.1 | 0.0 | 0.1 | 4.4 | 3.6 | 0.5 | 0.6 |
| Best Bilingual | **51.2** | **52.2** | 0.3 | 0.9 | 0.1 | 0.7 | 0.3 | 0.5 | 7.7 | 6.8 | **2.9** | 3.1 |
| MDAE | 42.5 | 42.5 | 3.1 | 5.3 | 0.1 | 1.7 | 1.0 | 3.3 | 7.4 | 7.9 | 1.5 | 4.7 |
| MDAE+F | 46.3 | 50.2 | **5.0** | **9.0** | **0.2** | **2.8** | **2.3** | **5.7** | **11.6** | **11.2** | **2.9** | **5.8** |
| Typology *L* | fusional | | agglutinative | | analytic | | fusional | | agglutinative | | analytic | |
| Word Order *L* | SOV,SVO | | SOV | | SOV | | SOV | | SVO | | SOV,SVO | |
| Word Overlap | 7.1% | | 1.4% | | 2.1% | | 0.6% | | 6.5% | | 5.7% | |
| Tokens *L* | 27.6 M | | 30.0 M | | 15.3 M | | 7.5 M | | 8.7 M | | 0.5 M | |

# 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

- The term multilinguality is usually related to low-resource (LR) settings
- Even if it helps the most in LR settings, HR are currently also improved
- It might imply additional work (adapters, etc)
- In 2021, a multilingual system won WMT for the first time

# Evaluating (Large Scale) ML-NMT

**EMNLP 2021**
**SIXTH CONFERENCE ON**
**MACHINE TRANSLATION (WMT21)**

**November 10-11, 2021**
**Punta Cana (Dominican Republic) and Online**

### Home

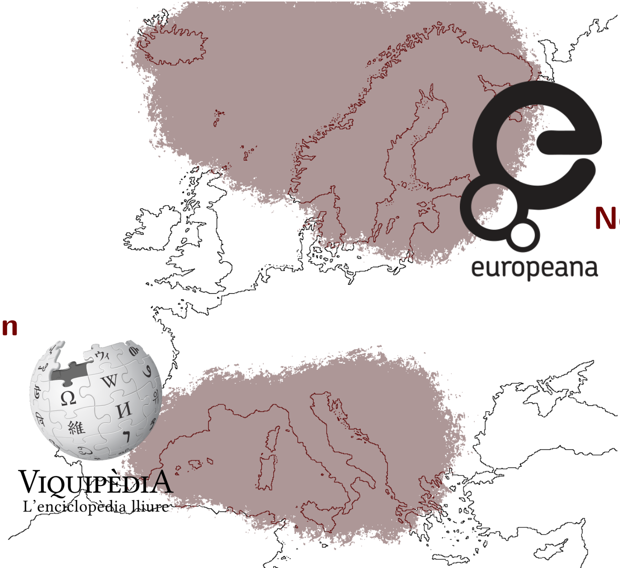
[HOME] [SCHEDULE] [RESULTS]
TRANSLATION TASKS: [NEWS] [SIMILAR LANGUAGES] [BIOMEDICAL] [EUROPEAN LOW RES MULTILINGUAL] [LARGE-SCALE MULTILINGUAL] [TRIANGULAR MT] [EFFICIENCY] [TERMINOLOGY] [UNSUP AND VERY LOW RES] [LIFELONG LEARNING]
EVALUATION TASKS: [QUALITY ESTIMATION] [METRICS]
OTHER TASKS: [AUTOMATIC POST-EDITING]

This conference builds on a series of annual workshops and conferences on statistical machine translation, going back to 2006:

- the NAACL-2006 Workshop on Statistical Machine Translation,
- the ACL-2007 Workshop on Statistical Machine Translation,
- the ACL-2008 Workshop on Statistical Machine Translation,
- the EACL-2009 Workshop on Statistical Machine Translation,
- the ACL-2010 Workshop on Statistical Machine Translation
- the EMNLP-2011 Workshop on Statistical Machine Translation,
- the NAACL-2012 Workshop on Statistical Machine Translation,
- the ACL-2013 Workshop on Statistical Machine Translation,
- the ACL-2014 Workshop on Statistical Machine Translation,
- the EMNLP-2015 Workshop on Statistical Machine Translation,
- the First Conference on Machine Translation (at ACL-2016),
- the Second Conference on Machine Translation (at EMNLP-2017),
- the Third Conference on Machine Translation (at EMNLP-2018),
- the Fourth Conference on Machine Translation (at ACL-2019),
- the Sixth Conference on Machine Translation (at EMNLP-2020).

# Multilingual LR Translation for Indo-European Languages

## *Two Subtasks, two Indo-European Families*




**Icelandic**

**Swedish**

**Norwegian Bokmål**

**Catalan**

**Romanian**

**Occitan**

**Italian**

## Shared Task Challenges

**C1** Multilinguality

**C2** Limited data but related languages

**C3** Specific vocabulary (cultural heritage, NEs)

**C4** Document-level translation

# Multilingual LR Translation for Indo-European Languages

## *Automatic Evaluation, Task 1: North Germanic Languages*

| | Average Ranking | BLEU | TER | chrF | COMET | BertScore |
|---|---|---|---|---|---|---|
| **M2M-100 (baseline)** | 1.0±0.0 | 31.5 | 0.54 | 0.55 | 0.399 | 0.862 |
| **EdinSaar-Contrastive** | 2.2±0.4 | 27.1 | 0.57 | 0.54 | 0.283 | 0.856 |
| **EdinSaar-Primary** | 2.8±0.4 | 27.5 | 0.58 | 0.52 | 0.276 | 0.849 |
| **UBCNLP-Primary** | 4.0±0.0 | 24.9 | 0.60 | 0.50 | 0.076 | 0.847 |
| **UBCNLP-Contrastive** | 5.0±0.0 | 24.0 | 0.61 | 0.49 | -0.068 | 0.837 |
| **mT5-devFinetuned (baseline)** | 6.0±0.0 | 18.5 | 0.78 | 0.42 | -0.102 | 0.810 |

# Multilingual LR Translation for Indo-European Languages

## *Automatic Evaluation, Task 2: Romance Languages*

| | Average Ranking | BLEU | TER | chrF | COMET | BertScore |
|---|---|---|---|---|---|---|
| **CUNI-Primary** | 1.2±0.4 | 50.1 | 0.401 | 0.694 | 0.566 | 0.901 |
| **CUNI-Contrastive** | 1.6±0.5 | 49.5 | 0.404 | 0.693 | 0.569 | 0.901 |
| **TenTrans-Contrastive** | 3.0±0.0 | 43.5 | 0.460 | 0.670 | 0.444 | 0.894 |
| **TenTrans-Primary** | 3.8±0.4 | 43.3 | 0.462 | 0.668 | 0.442 | 0.894 |
| **BSC-Primary** | 5.0±0.7 | 41.3 | 0.402 | 0.647 | 0.363 | 0.884 |
| **M2M-100 (baseline)** | 5.8±0.4 | 40.0 | 0.478 | 0.634 | 0.414 | 0.878 |
| **UBCNLP-Primary** | 7.2±0.4 | 35.4 | 0.528 | 0.588 | 0.007 | 0.854 |
| **mT5-devFinetuned (baseline)** | 8.0±0.7 | 29.3 | 0.592 | 0.553 | 0.059 | 0.850 |
| **UBCNLP-Contrastive** | 8.6±0.5 | 28.5 | 0.591 | 0.529 | -0.374 | 0.825 |

# Multilingual LR Translation for Indo-European Languages

## *Some Selected Systems*

# Multilingual LR Translation for Indo-European Languages

## *CUNI (Jon et al., 2021)*

- Multilingual supervised machine translation model (primary) enriched with backtranslated data (contrastive)
- 41 M original parallel sentences including all language pairs in the task plus French and English
- Exploration of various subword granularities
- Phonemic representation of texts added via multi-task learning
- Character-level rescoring on the translations *n*-best lists for Catalan–Occitan

# Multilingual LR Translation for Indo-European Languages

## *TenTrans (Yang et al., 2021)*

- 8-to-4 multilingual model with Catalan–Italian–Romanian–Occitan as the target side and Spanish, French, Portuguese and English on the source side.
- In-domain finetuning (data selected using a domain classifier trained with multilingual BERT)
- Knowledge transfer: knowledge distillation of the M2M 1.2B model previously finetuned on the languages of the task
- Primary: ensemble of the in-domain multilingual and the distilled M2M

# Multilingual LR Translation for Indo-European Languages

## *Some Conclusions*

- Systems used direct neural translation, multilingual or bilingual, no translations done through a pivot language
- Multilingual systems trained with additional corpora with the related rich languages as source gave the best performance
- Data augmentation via backtranslations has been beneficial for all the systems
- Few improvements by selecting data close to the domain of the validation set, but the in-domain adaptation was not decisive to win the shared task

# Large-Scale Multilingual Machine Translation

## *Track Details*

**Small Track #1:** 5 Central/East European languages, 30 directions: Croatian, Hungarian, Estonian, Serbian, Macedonian, English

**Small Track #2:** 5 South East Asian languages, 30 directions: Javanese, Indonesian, Malay, Tagalog, Tamil, English

**Large Track:** All Languages, to and from English

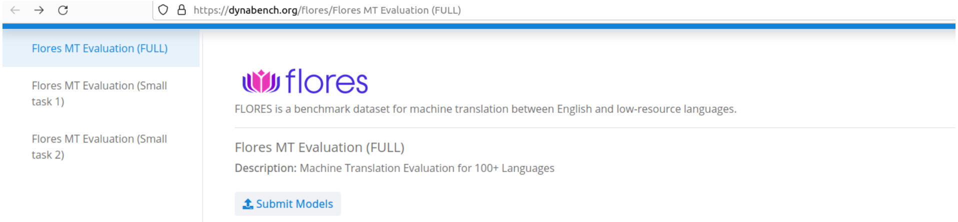
# Large-Scale Multilingual Machine Translation

## *Large Track Languages*

- Afrikaans
- Amharic
- Arabic
- Armenian
- Assamese
- Asturian
- Azerbaijani
- Belarusian
- Bengali
- Bosnian
- Bulgarian
- Burmese
- Catalan
- Cebuano
- Chinese (Simplified)
- Chinese (Traditional)
- Croatian
- Czech
- Danish
- Dutch
- English
- Estonian
- Filipino (Tagalog)
- Finnish
- French
- Fula
- Galician
- Ganda
- Georgian
- German
- Greek
- Gujarati
- Hausa
- Hebrew
- Hindi
- Hungarian
- Icelandic
- Igbo
- Indonesian
- Irish
- Italian
- Japanese
- Javanese
- Kabuverdianu
- Kamba
- Kannada
- Kazakh
- Khmer
- Korean
- Kyrgyz
- Lao
- Latvian
- Lingala
- Lithuanian
- Luo
- Luxembourgish
- Macedonian
- Malay
- Malayalam
- Maltese
- Maori
- Marathi
- Mongolian
- Nepali
- Northern Sotho
- Norwegian
- Nyanja
- Occitan
- Oriya
- Oromo
- Pashto
- Persian
- Polish
- Portuguese
- Punjabi
- Romanian
- Russian
- Serbian
- Shona
- Sindhi
- Slovak
- Slovenian
- Somali
- Sorani Kurdish
- Spanish
- Swahili
- Swedish
- Tajik
- Tamil
- Telugu
- Thai
- Turkish
- Ukrainian
- Umbundu
- Urdu
- Uzbek
- Vietnamese
- Welsh
- Wolof
- Xhosa
- Yoruba
- Zulu

# Large-Scale Multilingual Machine Translation

## *Dynabench Evaluation Platform*




▸ Let's go to Dynabench!

# Large-Scale Multilingual Machine Translation

## *High-Quality Translations*

LANGUAGE-PAIR LEADERBOARD

Dataset

| Source Language | Target Language | Model | BLEU Score |
|---|---|---|---|
| Afrikaans (afr) | English (eng) | DeltaLM+Zcode | 60.86 |
| Welsh (cym) | English (eng) | DeltaLM+Zcode | 60.05 |
| English (eng) | Welsh (cym) | DeltaLM+Zcode | 58.37 |
| English (eng) | Maltese (mlt) | DeltaLM+Zcode | 57.98 |
| Maltese (mlt) | English (eng) | DeltaLM+Zcode | 57.96 |
| Swedish (swe) | English (eng) | DeltaLM+Zcode | 52.63 |
| Danish (dan) | English (eng) | DeltaLM+Zcode | 52.40 |
| Portuguese (Brazil) (por) | English (eng) | DeltaLM+Zcode | 51.29 |
| Welsh (cym) | Maltese (mlt) | DeltaLM+Zcode | 50.15 |
| Afrikaans (afr) | Maltese (mlt) | DeltaLM+Zcode | 49.74 |

Page 1 of 1010

Previous Next

# Large-Scale Multilingual Machine Translation

## *Low-Quality Translations*

LANGUAGE-PAIR LEADERBOARD

Dataset

| Source Language | Target Language | Model | BLEU Score |
|---|---|---|---|
| Lingala (lin) | Fula (ful) | DeltaLM+Zcode | 1.41 |
| Burmese (mya) | Kabuverdianu (kea) | DeltaLM+Zcode | 1.42 |
| Thai (tha) | Umbundu (umb) | DeltaLM+Zcode | 1.42 |
| Igbo (ibo) | Fula (ful) | DeltaLM+Zcode | 1.42 |
| Umbundu (umb) | Khmer (khm) | m2m-124-175m | 1.43 |
| Galician (glg) | Fula (ful) | m2m-124-175m | 1.43 |
| Estonian (est) | Fula (ful) | DeltaLM+Zcode | 1.43 |
| Luo (luo) | Khmer (khm) | 615m | 1.43 |
| Hebrew (heb) | Umbundu (umb) | DeltaLM+Zcode | 1.43 |
| Catalan (cat) | Fula (ful) | m2m-124-175m | 1.44 |

Page 1 of 1010

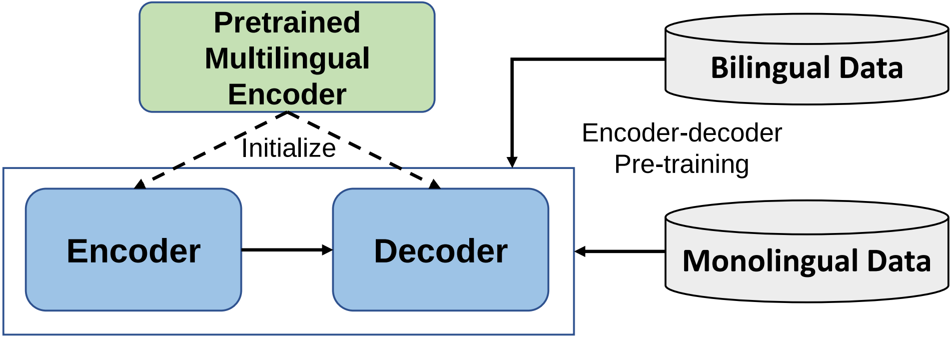
Previous Next

**Main System Characteristics (from the findings paper)**

- Combination of parallel, back-translated and noisy-parallel data
- Based on the pre-trained DeltaLM$_{LARGE}$ (next slides only if soon enough!)
- Mixture of direct and pivoted translation to improve the performance of individual directions
- Progressive learning: starts with a smaller architecture, noisier training data, and later changes to improve performance

# Large-Scale Multilingual Machine Translation

## *DeltaLM: Basic Idea*



# DeltaLM

## *Basic Idea (Ma et al., 2021 —still in arXiv)*

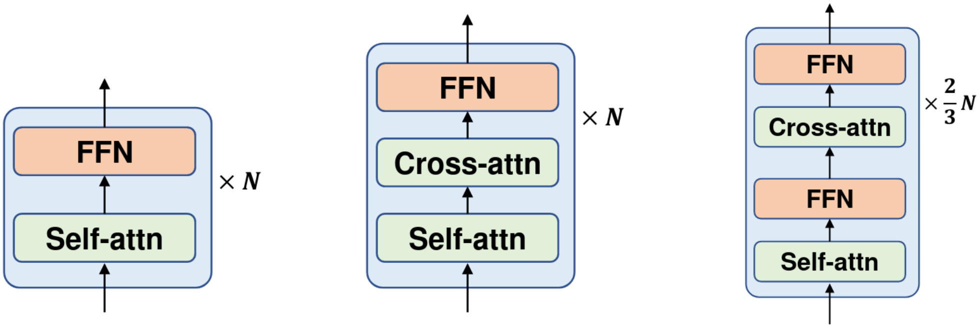
- "The decoder as the task layer of off-the-shelf pre-trained encoders"
- Encoder and the decoder are initialised with the pre-trained multilingual encoder
- Pre-train ΔLM with both monolingual data and bilingual data in a self-supervised way

# DeltaLM

## *Basic Idea (Ma et al., 2021 —still in arXiv)*

- "The decoder as the task layer of off-the-shelf pre-trained encoders"
- Encoder and the decoder are initialised with the pre-trained multilingual encoder
  - **How to initialise a decoder with an encoder??**
- Pre-train ΔLM with both monolingual data and bilingual data in a self-supervised way
  - **What's an appropriate pre-training task??**

## Interleaved Decoder




(a) Vanilla encoder (b) Vanilla decoder (c) Interleaved decoder

# DeltaLM

## *Any Encoder could be Used to Initialise, ΔLM Uses InfoXLM$_{BASE}$*

**InfoXLM (Chi et al., NAACL 2021)**

- 12 layers and 768 hidden states
- Training with large-scale monolingual data and bilingual data
- Tasks: masked language model, translation language model, and cross-lingual contrast objectives
- Shared vocabulary of 250,000 tokens based on the SentencePiece
- By the way... InfoXLM is initialised with XLM-R (550M params)

**DeltaLM (Ma et al., 2021)**

- 24 encoder layers, 12 interleaved decoder layers and 1024 hidden states (360M params)
- Training with large-scale monolingual data and bilingual data
- Tasks: span corruption and translation span corruption
- Shared vocabulary of 250,000 tokens based on the SentencePiece
- Initialised with InfoXLM which is initialised with XLM-R (550M params)

# DeltaLM

## *Pre-training Tasks: Span Corruption*

**Original:**

Thanks for your invitation last week.

**Source**:

Thanks [Mask1] invitation [Mask2].

**Target**:

[Span1] for your [Span2] last week

- Introduced in mT5
- Data: large-scale multilingual corpora in 100 languages (6TB combination of CC100, CC-Net, and Wikipedia)

# DeltaLM

## *Pre-training Tasks: Translation Span Corruption*

**Original:**

Thanks for your invitation last week.
谢谢你上周的邀请。

**Source**:

Thanks [Mask1] invitation [Mask2].
谢谢你上周的[Mask3]。

**Target**:

[Span1] for your [Span2] last week
[Span3] 邀请

- Introduced in mT6
- Data: concatenate two parallel sentences as the input for 77 languages (88GB of bilingual data from CCAligned and OPUS)

# DeltaLM

## *A Comment on the Infrastructure*

- Microsoft's submission trained on 64 NVIDIA V100 or 32 A100 GPUs
- It takes 1 week to train $\Delta$LM with 32 V100 GPUs
- InfoXLM training
- 1.5 Million updates on 500 32GB Nvidia V100 GPUs for XML-R

**Main System Characteristics (from the findings paper)**

- Combination of parallel, back-translated and noisy-parallel data
- Based on the pre-trained DeltaLM
- Mixture of direct and pivoted translation to improve the performance of individual directions
- Progressive learning: starts with a smaller architecture, noisier training data, and later changes to improve performance

# That’s All Folks!

## *Thanks! And...*

# Multilingual Neural Machine Translation

**Cristina España-Bonet**
DFKI GmbH

11th Advanced Summer School on NLP
IIIT Hyderabad
23rd June 2022