Self-Supervised Neural Machine Translation and More!

Cristina España-Bonet
DFKI GmbH

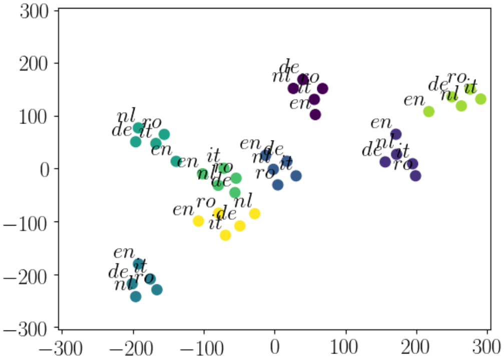
Low-Resource NLP:
Multilinguality and Machine Translation
Webinar Series — Session IV
14th September 2021

Outline

Recap

Multilingual Semantic Space for Context Vectors (easy)

(España-Bonet & van Genabith, 2018)

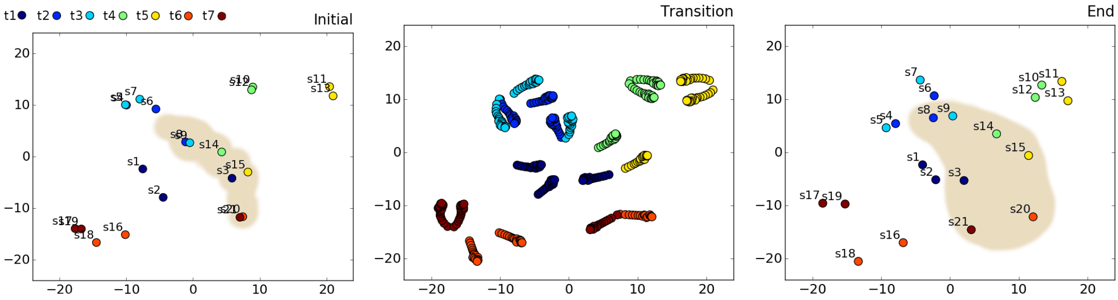


ML-NMT $\{de, en, nl, it, ro\} \rightarrow \{de, en, nl, it, ro\}$ with TED talks (t-SNE projection)

Recap

Evolution of Context Vectors through Training (hard)

(España-Bonet et al., 2017)



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

Recap

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?

Recap

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

Article Talk Read Edit View history Search Wikipedia

Transformers (comics)

There have been three main publishers of the comic book series bearing the name Transformers based on the toy lines of the same name.

The first series was produced by Marvel Comics from 1984 to 1991, which ran for 80 issues and produced four spin-off miniseries.

This was followed by a second volume titled *Transformers: Generation 2*, which ran for 12 issues starting in 1993.

The third series is currently being produced by IDW Publishing starting with an issue #0 in October 2005 and a regular series starting in January 2006.

There are also several limited series being produced by IDW as well.

In addition to these three main publishers, there have also been several other smaller publishers with varying degrees of success.

Artículo Discusión Leer Editar Ver historial Buscar en Wikipedia

Transformers (cómics)

Ha habido tres editores principales en la serie de cómics de Transformers, basados en las líneas de juguetes del mismo nombre.

La primera serie fue producida por Marvel Comics desde 1984 hasta 1991, para ayudar en las ventas de la línea de juguetes de Hasbro.

Desarrolló 80 tomos y produjo cuatro miniseries de spin-off.

Esto fue seguido por un segundo volumen titulado Transformers: Generación 2, que tuvo 12 ediciones a partir de 1993.

La segunda gran serie fue producida por Producciones Dreamwave en 2002 a 2004 con series limitadas, hasta que el compañía se quedó en bancarrota en 2005.

Además de estos tres editores principales, también ha habido varias otras editoriales más pequeñas con diferentes grados de éxito.

Por favor, véase la lista de los cómics de Transformers menores para obtener más información.

En 1984, Marvel comenzó a publicar cómics de Transformers para ayudar en las ventas de la línea de juguetes de Hasbro.

Article Talk Read Edit View history Search Wikipedia

Transformers (comics)

There have been three main publishers of the comic book series bearing the name Transformers based on the toy lines of the same name.

The first series was produced by Marvel Comics from 1984 to 1991, which ran for 80 issues and produced four spin-off miniseries.

This was followed by a second volume titled *Transformers: Generation 2*, which ran for 12 issues starting in 1993.

The third series is currently being produced by IDW Publishing starting with an issue #0 in October 2005 and a regular series starting in January 2006.

There are also several limited series being produced by IDW as well.

In addition to these three main publishers, there have also been several other smaller publishers with varying degrees of success.

Artículo Discusión Leer Editar Ver historial Buscar en Wikipedia

Transformers (cómics)

Ha habido tres editores principales en la serie de cómics de Transformers, basados en las líneas de juguetes del mismo nombre.

La primera serie fue producida por Marvel Comics desde 1984 hasta 1991, para ayudar en las ventas de la línea de juguetes de Hasbro.

Desarrolló 80 tomos y produjo cuatro miniseries de spin-off.

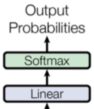
Esto fue seguido por un segundo volumen titulado Transformers: Generación 2, que tuvo 12 ediciones a partir de 1993.

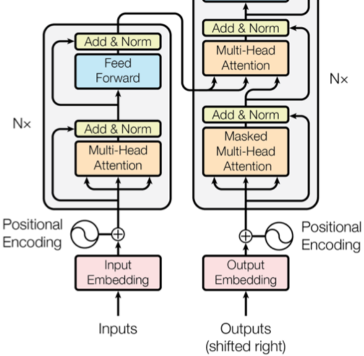
La segunda gran serie fue producida por Producciones Dreamwave en 2002 a 2004 con series limitadas, hasta que el compañía se quedó en bancarrota en 2005.

Además de estos tres editores principales, también ha habido varias otras editoriales más pequeñas con diferentes grados de éxito.

Por favor, véase la lista de los cómics de Transformers menores para obtener más información.

En 1984, Marvel comenzó a publicar cómics de Transformers para ayudar en las ventas de la línea de juguetes de Hasbro.





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

LASER & parallel sentence extraction

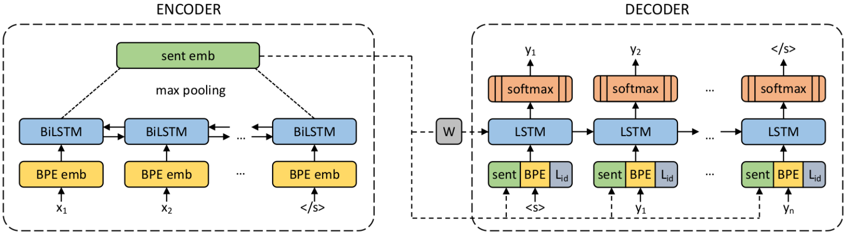
Digression

L*anguage* **A***gnostic* **SE***ntence* **R***epresentations, LASER*

1. Training with (multilingual) parallel corpora, MT task with seq2seq
2. Sentence embeddings from the language agnostic encoder
3. Extract most similar pairs according to semantic similarity
4. Use the parallel sentences to train a supervised NMT system

Digression

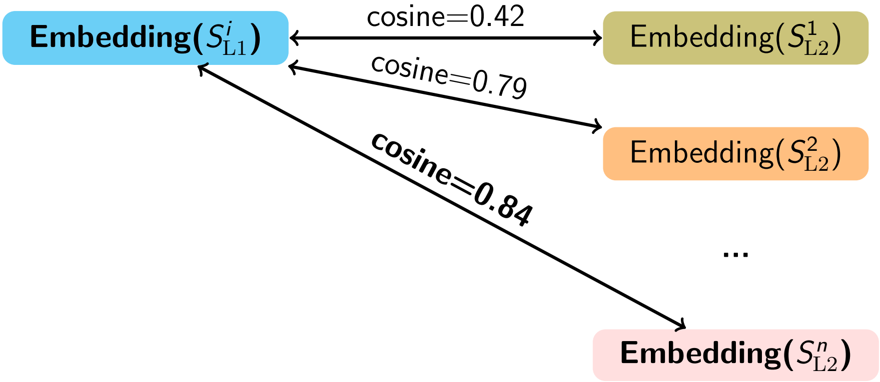
Architecture (based on Schwenk 2018)



- Training with (multilingual) parallel corpora, MT task
- Sentence embeddings from the language agnostic encoder
- **L**anguage **A**gnostic **SE**ntence **R**epresentations: 1024-dim embeddings

Digression

The Key Point: Margin-based Similarity for Scoring Pairs



Threshold=0.80 ($\forall i$)

Digression

The Key Point: Margin-based Similarity for Scoring Pairs

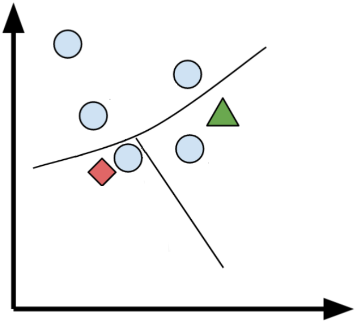
(A)	*Les produits agricoles sont constitués de thé, de riz, de sucre, de tabac, de camphre, de fruits et de soie.*
0.818	Main crops include wheat, sugar beets, potatoes, cotton, tobacco, vegetables, and fruit.
0.817	The fertile soil supports wheat, corn, barley, tobacco, sugar beet, and soybeans.
0.814	Main agricultural products include grains, cotton, oil, pigs, poultry, fruits, vegetables, and edible fungus.
0.808	The important crops grown are cotton, jowar, groundnut, rice, sunflower and cereals.

(B)	*Mais dans le contexte actuel, nous pourrons les ignorer sans risque.*
0.737	But, in view of the current situation, we can safely ignore these.
0.499	But without the living language, it risks becoming an empty shell.
0.498	While the risk to those working in ceramics is now much reduced, it can still not be ignored.
0.488	But now they have discovered they are not free to speak their minds.

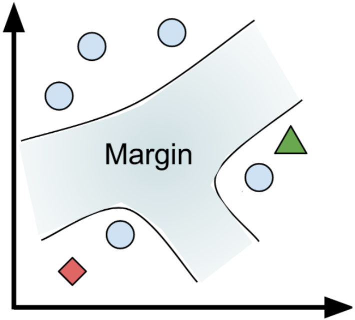
- Cosine similarity has a different scale per sentence

Digression

The Key Point: Margin-based Similarity for Scoring Pairs



Cosine accepted pairs



Margin accepted pairs

(Adapted from Yang et al, 2019)

Digression

The Key Point: Margin-based Similarity for Scoring Pairs

$$\mathrm{margin}_{\mathrm{LASER}}(S_{\mathrm{L1}}, S_{\mathrm{L2}}) = \frac{\cos(S_{\mathrm{L1}}, S_{\mathrm{L2}})}{\mathrm{avr}_{\mathrm{kNN}}(S_{\mathrm{L1}}, P_k)/2 + \mathrm{avr}_{\mathrm{kNN}}(S_{\mathrm{L2}}, Q_k)/2}$$

where $\quad \mathrm{avr}_{\mathrm{kNN}}(X, Y_k) = \sum_{Y \in k\mathrm{NN}(X)} \frac{\cos(X,Y)}{k} \quad$ (average similarity)

Digression

The Key Point: Margin-based Similarity for Scoring Pairs

Artetxe et al.

$$\mathrm{margin}_{\mathrm{LASER}}(S_{\mathrm{L1}}, S_{\mathrm{L2}}) = \frac{\cos(S_{\mathrm{L1}}, S_{\mathrm{L2}})}{\mathrm{avr}_{\mathrm{kNN}}(S_{\mathrm{L1}}, P_k)/2 + \mathrm{avr}_{\mathrm{kNN}}(S_{\mathrm{L2}}, Q_k)/2}$$

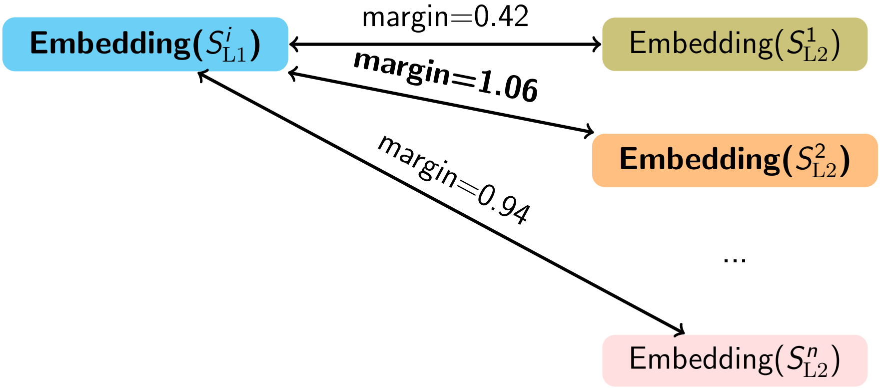
Conneau et al., 2018

$$\mathrm{margin}_{\mathrm{CSLS}}(S_{\mathrm{L1}}, S_{\mathrm{L2}}) = \cos(S_{\mathrm{L1}}, S_{\mathrm{L2}}) - \mathrm{avr}_{\mathrm{kNN}}(S_{\mathrm{L1}}, P_k)/2 - \mathrm{avr}_{\mathrm{kNN}}(S_{\mathrm{L2}}, Q_k)/2$$

where $\mathrm{avr}_{\mathrm{kNN}}(X, Y_k) = \sum_{Y \in k\mathrm{NN}(X)} \frac{\cos(X,Y)}{k}$ (average similarity)

Digression

The Key Point: Margin-based Similarity for Scoring Pairs



Threshold=1.04 **($\forall i$)**

Digression

Parallel Sentence Extraction

$$\cos(S_{\mathrm{L1}}, S_{\mathrm{L2}})$$

$$\mathrm{margin}_{\mathrm{CSLS}}(S_{\mathrm{L1}}, S_{\mathrm{L2}})$$

$$\mathrm{margin}_{\mathrm{LASER}}(S_{\mathrm{L1}}, S_{\mathrm{L2}})$$

Func.	**Retrieval**	**EN-DE**			**EN-FR**		
		P	**R**	**F1**	**P**	**R**	**F1**
Abs. (cos)	Forward	78.9	75.1	77.0	82.1	74.2	77.9
	Backward	79.0	73.1	75.9	77.2	72.2	74.7
	Intersection	84.9	80.8	82.8	83.6	78.3	80.9
	Max. score	83.1	77.2	80.1	80.9	77.5	79.2
Dist.	Forward	94.8	94.1	94.4	91.1	**91.8**	91.4
	Backward	94.8	94.1	94.4	91.5	91.4	91.4
	Intersection	94.9	94.1	94.5	91.2	**91.8**	91.5
	Max. score	94.9	94.1	94.5	91.2	**91.8**	91.5
Ratio	Forward	95.2	**94.4**	**94.8**	**92.4**	91.3	91.8
	Backward	95.2	**94.4**	**94.8**	92.3	91.3	91.8
	Intersection	**95.3**	**94.4**	**94.8**	**92.4**	91.3	**91.9**
	Max. score	**95.3**	**94.4**	**94.8**	**92.4**	91.3	**91.9**

Table 2: BUCC results (precision, recall and F1) on the training set, used to optimize the filtering threshold.

Mining of parallel corpora

- **WikiMatrix**: Mining 135M Parallel Sent. in 1620 Language Pairs from WP
- **CCMatrix**: Mining Billions of High-Quality Parallel Sentences on the WEB
- `https://github.com/facebookresearch/LASER`

Mining of parallel corpora

- **WikiMatrix**: Mining 135M Parallel Sent. in 1620 Language Pairs from WP
- **CCMatrix**: Mining Billions of High-Quality Parallel Sentences on the WEB
- `https://github.com/facebookresearch/LASER`

Others

- Cross-lingual Natural Language Inference (XNLI)
- Cross-lingual text classification
- Cross-lingual similarity search

Digression

Let's join the main path again

Self-Supervised NMT

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

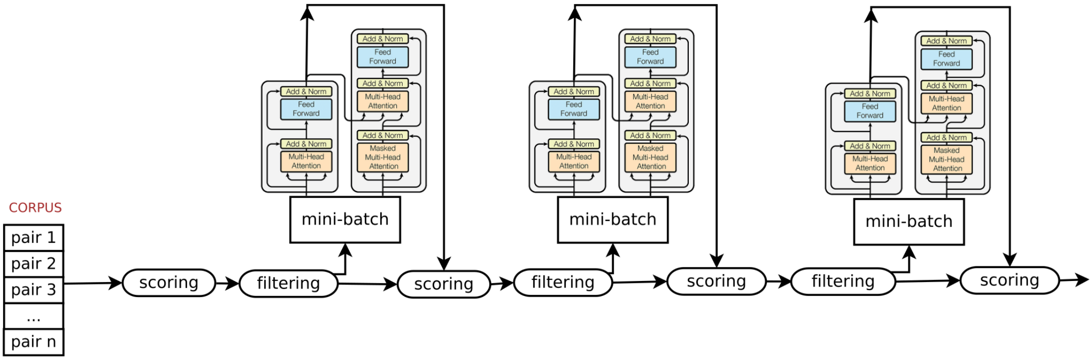
Self-Supervised NMT

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



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

Joint Training: Key Points

1. **Sentence Representation**

2. **Scoring function**

Self-Supervised NMT

Joint Training: Key Points

1. **Sentence Representation**
   the sum of word embeddings (E_w) and the hidden states in an RNN or the encoder outputs in a transformer (E_h):

$$E_w = \sum_{t=1}^{T} e_t, \qquad E_h = \sum_{t=1}^{T} h_t$$

2. **Scoring function**

Joint Training: Key Points

1. **Sentence Representation**

   S_{L1} and S_{L2} vector representations for each sentence of a pair (E_w or E_h)

2. **Scoring function**

   cosine similarility: $$\cos(S_{\mathrm{L1}}, S_{\mathrm{L2}}) = \frac{S_{\mathrm{L1}} \cdot S_{\mathrm{L2}}}{\|S_{\mathrm{L1}}\| \, \|S_{\mathrm{L2}}\|}$$

   margin-based score:

$$\mathrm{margin}(S_{\mathrm{L1}}, S_{\mathrm{L2}}) = \frac{\cos(S_{\mathrm{L1}}, S_{\mathrm{L2}})}{\mathrm{avr}_{\mathrm{kNN}}(S_{\mathrm{L1}}, P_k)/2 + \mathrm{avr}_{\mathrm{kNN}}(S_{\mathrm{L2}}, Q_k)/2}$$

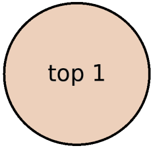
where $\mathrm{avr}_{\mathrm{kNN}}(X, Y_k) = \sum_{Y \in k\mathrm{NN}(X)} \frac{\cos(X,Y)}{k}$ (average similarity)

Self-Supervised NMT

Joint Training: Sentence Selection (Filtering)

1. Input a lot (e.g. set of WP article pairs, web pages, etc)
2. Score all sentence pairs
3. Keep the top one pairs (with constraints!)

E_h src2tgt

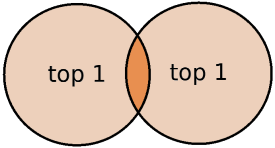


Self-Supervised NMT

Joint Training: Sentence Selection (Filtering)

1. Input a lot (e.g. set of WP article pairs, web pages, etc)
2. Score all sentence pairs
3. Keep the top one pairs (with constraints!)

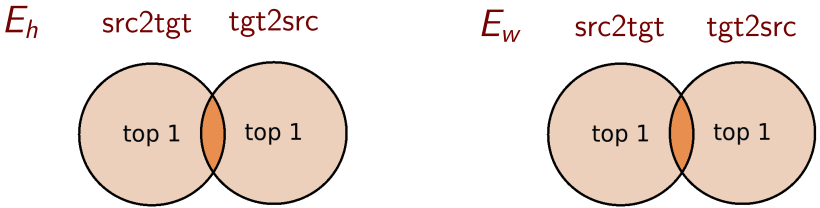
E_h src2tgt tgt2src



Self-Supervised NMT

Joint Training: Sentence Selection (Filtering)

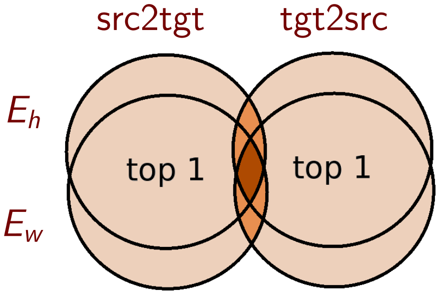
1. Input a lot (e.g. set of WP article pairs, web pages, etc)
2. Score all sentence pairs
3. Keep the top one pairs (with constraints!)



Self-Supervised NMT

Joint Training: Sentence Selection (Filtering)

Intersection of intersection of intersection...

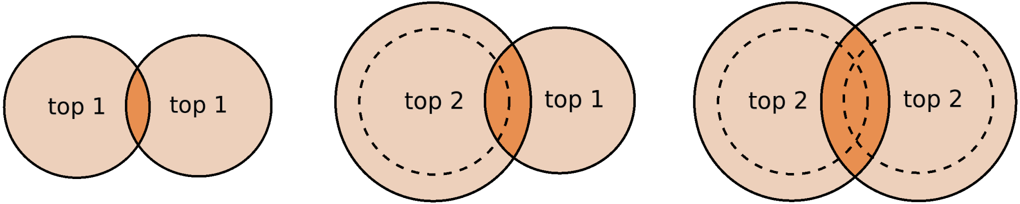


to avoid the need for a threshold
(as compared to LASER bitext mining approach)

Self-Supervised NMT

Sentence Selection: Precision or Recall?

low permissibility | medium permissibility | high permissibility



high precision mode

high recall mode

cosP: E_w, E_h in high precision mode and $\cos(S_{\mathrm{L1}}, S_{\mathrm{L2}})$ are used.

margP: E_w, E_h in high precision mode and $\mathrm{margin}(S_{\mathrm{L1}}, S_{\mathrm{L2}})$ are used.

cosP: E_w, E_h in high precision mode and $\cos(S_{\mathrm{L1}}, S_{\mathrm{L2}})$ are used.

margP: E_w, E_h in high precision mode and $\mathrm{margin}(S_{\mathrm{L1}}, S_{\mathrm{L2}})$ are used.

margR: As **margP** but E_w and E_h are used in the high recall mode.

Self-Supervised NMT

Evaluation, Selected Models

cosP: E_w, E_h in high precision mode and $\cos(S_{\mathrm{L1}}, S_{\mathrm{L2}})$ are used.

margP: E_w, E_h in high precision mode and $\mathrm{margin}(S_{\mathrm{L1}}, S_{\mathrm{L2}})$ are used.

margR: As **margP** but E_w and E_h are used in the high recall mode.

margH: As **margP** with E_h as only representation.
A hard threshold of 1.01 is used.

margE: As **margP** with E_w as only representation.
A hard threshold of 1.00 is used.

Self-Supervised NMT

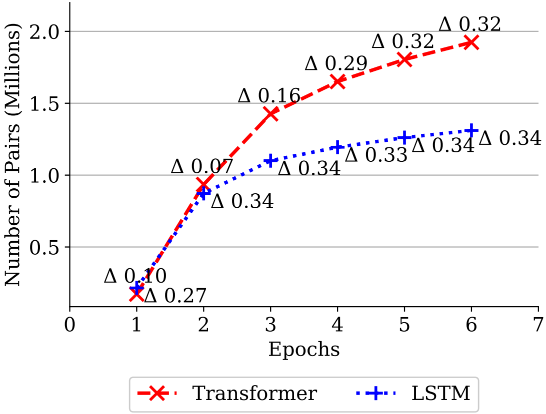
Automatic Evaluation (Transformer; en–fr, en–de, en–es)

Model	Corpus, *en*+*fr* sent. (in millions)	BLEU *en2fr* (newstest2014)	BLEU *fr2en* (newstest2014)
cosP	Wikipedia, 12+8	25.21	24.96
margE	Wikipedia, 12+8	27.33	25.87
margH	Wikipedia, 12+8	24.45	23.83
margP	Wikipedia, 12+8	**29.21**	**27.36**
margR	Wikipedia, 12+8	28.01	26.78

margP: E_w, E_h in high precision mode and $\mathrm{margin}(S_{\mathrm{L1}}, S_{\mathrm{L2}})$

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

Built-In Curriculum (Ruiter et al., EMNLP, 2020)

	#Pairs$_{enfr}$	*en2fr*	*fr2en*	#Pairs$_{ende}$	*en2de*	*de2en*	#Pairs$_{enes}$	*en2es*	*es2en*
NMT$_{init}$	2.14M	21.8±.6	21.1±.5	0.32M	3.4±.3	4.7±.3	2.51M	27.0±.7	25.0±.7
NMT$_{mid}$	3.14M	29.0±.6	26.6±.6	1.13M	11.2±.4	15.0±.6	3.96M	28.3±.7	26.1±.7
NMT$_{end}$	3.17M	28.8±.6	26.5±.6	1.18M	11.9±.5	15.3±.5	3.99M	28.3±.7	26.2±.7
NMT$_{all}$	5.38M	26.8±.7	25.2±.6	2.21M	11.6±.5	15.0±.6	5.41M	27.9±.6	25.9±.8
SS-NMT	5.38M	29.5±.6	27.7±.6	2.21M	14.4±.6	18.1±.6	5.41M	28.6±.7	28.4±.7

Supervised NMT systems trained on the unique pairs collected by SS-NMT in the first (NMT$_{init}$), intermediate (NMT$_{mid}$), final (NMT$_{end}$) and all (NMT$_{all}$) epochs of training

Learning Process in SS-NMT

What's going on? — Built-In Curriculum Learning

Input Documents

Article | Talk | Read | Edit | View history | Search Wikipedia

Transformers (comics)

There have been three main publishers of the comic book series bearing the name Transformers based on the toy lines of the same name.

The first series was produced by Marvel Comics from 1984 to 1991, which ran for 80 issues and produced four spin-off miniseries.

This was followed by a second volume titled *Transformers: Generation 2*, which ran for 12 issues starting in 1993.

The third series is currently being produced by IDW Publishing starting with an issue #0 in October 2005 and a regular series starting in January 2006.

There are also several limited series being produced by IDW as well.

In addition to these three main publishers, there have also been several other smaller publishers with varying degrees of success.

Artículo | Discusión | Leer | Editar | Ver historial | Buscar en Wikipedia

Transformers (cómics)

Ha habido tres editores principales en la serie de cómics de Transformers, basados en las líneas de juguetes del mismo nombre.

La primera serie fue producida por Marvel Comics desde 1984 hasta 1991, para ayudar en las ventas de la línea de juguetes de Hasbro.

Desarrolló 80 tomos y produjo cuatro miniseries de spin-off.

Esto fue seguido por un segundo volumen titulado Transformers: Generación 2, que tuvo 12 ediciones a partir de 1993.

La segunda gran serie fue producida por Producciones Dreamwave en 2002 a 2004 con series limitadas, hasta que el compañía se quedó en bancarrota en 2005.

Además de estos tres editores principales, también ha habido varias otras editoriales más pequeñas con diferentes grados de éxito.

Por favor, véase la lista de los cómics de Transformers menores para obtener más información.

En 1984, Marvel comenzó a publicar cómics de Transformers para ayudar en las ventas de la línea de juguetes de Hasbro.

Learning Process in SS-NMT

Built-In Curriculum Learning

Sentence selection through epochs: Epoch 1

Article Talk Read Edit View history Search Wikipedia

Transformers (comics)

There have been three main publishers of the comic book series bearing the name Transformers based on the toy lines of the same name.

The first series was produced by Marvel Comics from 1984 to 1991, which ran for 80 issues and produced four spin-off miniseries.

This was followed by a second volume titled *Transformers: Generation 2*, which ran for 12 issues starting in 1993.

The third series is currently being produced by IDW Publishing starting with an issue #0 in October 2005 and a regular series starting in January 2006.

There are also several limited series being produced by IDW as well.

In addition to these three main publishers, there have also been several other smaller publishers with varying degrees of success.

Artículo Discusión Leer Editar Ver historial Buscar en Wikipedia

Transformers (cómics)

Ha habido tres editores principales en la serie de cómics de Transformers, basados en las líneas de juguetes del mismo nombre.

La primera serie fue producida por Marvel Comics desde 1984 hasta 1991, para ayudar en las ventas de la línea de juguetes de Hasbro.

Desarrolló 80 tomos y produjo cuatro miniseries de spin-off.

Esto fue seguido por un segundo volumen titulado Transformers: Generación 2, que tuvo 12 ediciones a partir de 1993.

La segunda gran serie fue producida por Producciones Dreamwave en 2002 a 2004 con series limitadas, hasta que el compañía se quedó en bancarrota en 2005.

Además de estos tres editores principales, también ha habido varias otras editoriales más pequeñas con diferentes grados de éxito.

Por favor, véase la lista de los cómics de Transformers menores para obtener más información.

En 1984, Marvel comenzó a publicar cómics de Transformers para ayudar en las ventas de la línea de juguetes de Hasbro.

Sentence selection through epochs: Epoch 6

Article Talk Read Edit View history Search Wikipedia

Transformers (comics)

There have been three main publishers of the comic book series bearing the name Transformers based on the toy lines of the same name.

The first series was produced by Marvel Comics from 1984 to 1991, which ran for 80 issues and produced four spin-off miniseries.

This was followed by a second volume titled *Transformers: Generation 2*, which ran for 12 issues starting in 1993.

The third series is currently being produced by IDW Publishing starting with an issue #0 in October 2005 and a regular series starting in January 2006.

There are also several limited series being produced by IDW as well.

In addition to these three main publishers, there have also been several other smaller publishers with varying degrees of success.

Artículo Discusión Leer Editar Ver historial Buscar en Wikipedia

Transformers (cómics)

Ha habido tres editores principales en la serie de cómics de Transformers, basados en las líneas de juguetes del mismo nombre.

La primera serie fue producida por Marvel Comics desde 1984 hasta 1991, para ayudar en las ventas de la línea de juguetes de Hasbro.

Desarrolló 80 tomos y produjo cuatro miniseries de spin-off.

Esto fue seguido por un segundo volumen titulado Transformers: Generación 2, que tuvo 12 ediciones a partir de 1993.

La segunda gran serie fue producida por Producciones Dreamwave en 2002 a 2004 con series limitadas, hasta que el compañía se quedó en bancarrota en 2005.

Además de estos tres editores principales, también ha habido varias otras editoriales más pequeñas con diferentes grados de éxito.

Por favor, véase la lista de los cómics de Transformers menores para obtener más información.

En 1984, Marvel comenzó a publicar cómics de Transformers para ayudar en las ventas de la línea de juguetes de Hasbro.

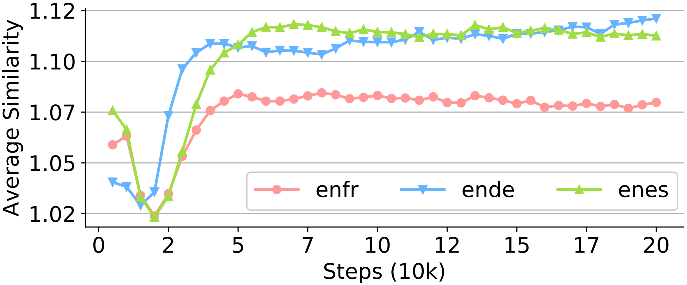
Learning Process in SS-NMT

Self-Induced Curricula

- SS-NMT induces a curriculum when selecting the data to train the MT task
- The order in which sentences are extracted is vital for translation quality (NMTall vs. SS-NMT)
- The data selection shows (at least) 3 curricula:
  1. a task-specific (MT) curriculum
  2. a denoising curriculum
  3. a complexity curriculum

Self-Induced Curricula in SSNMT

Task-specific (MT) Curriculum

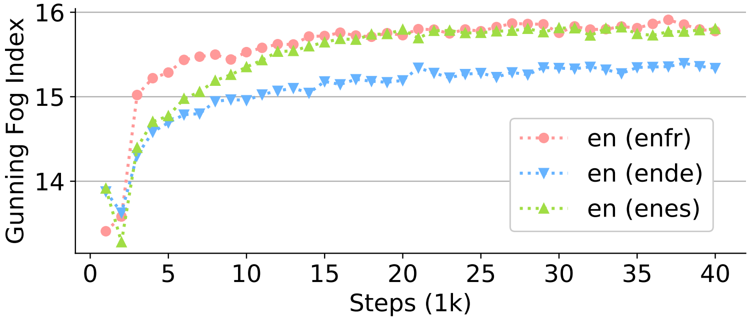


⇝ more cross-lingual similarity → more parallel

⇝ more parallel → closer to MT purpose

Self-Induced Curricula in SSNMT

Complexity Curriculum

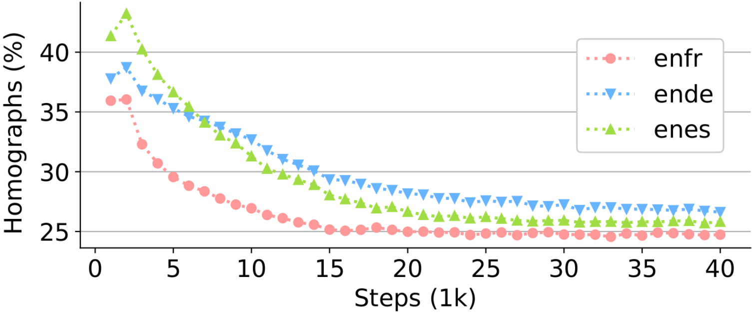


Gunning Fog, readability measure: $\mathrm{GF} = 0.4\left[\left(\frac{w}{s}\right) + 100\left(\frac{c}{w}\right)\right]$

- Increment from GF=11 (high school students) to GF=13 (undergrads)

Self-Induced Curricula in SSNMT

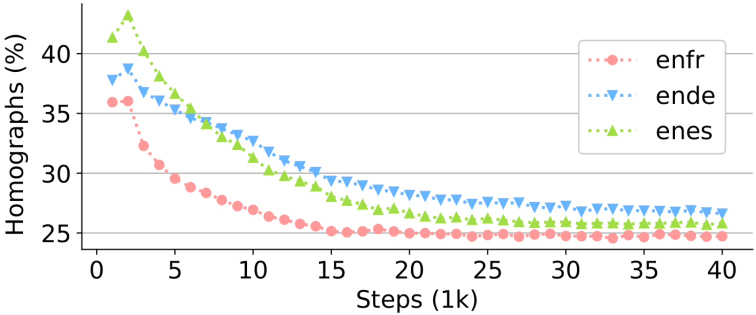
Key Point: Homographs!



- Large % of homographs in the sentences at the beginning of the training less sentences (punctuation, numbers, common BPE), noisier, easier

Self-Induced Curricula in SSNMT

Key Point: Homographs!



- Large % of homographs in the sentences at the beginning of the training less sentences (punctuation, numbers, common BPE), noisier, easier

⇝ What if no homographs?

Self-Induced Curricula in SSNMT

Open Problems

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

Similar issues in unsupervised NMT, bilingual embeddings, etc.

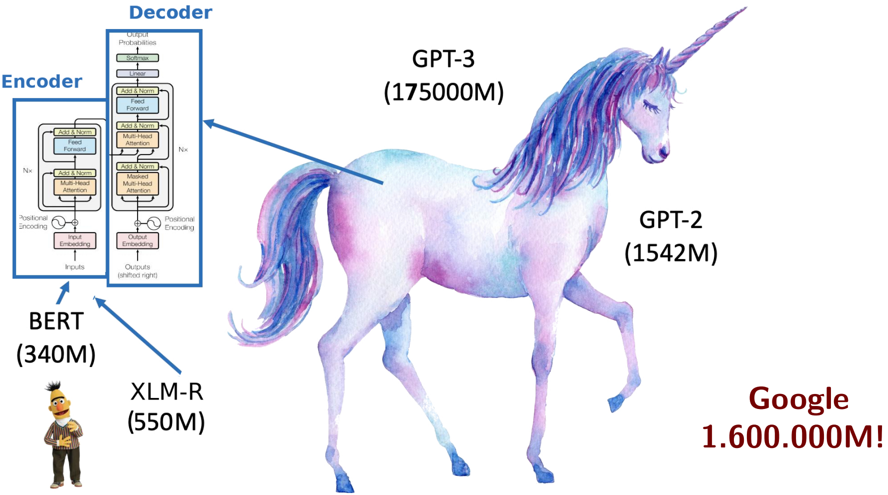
Same "solutions"?

Digression

Pre-trained models for language generation

Pre-trained Models for Language Generation

Transformer Encoder/Decoder for Language Modeling



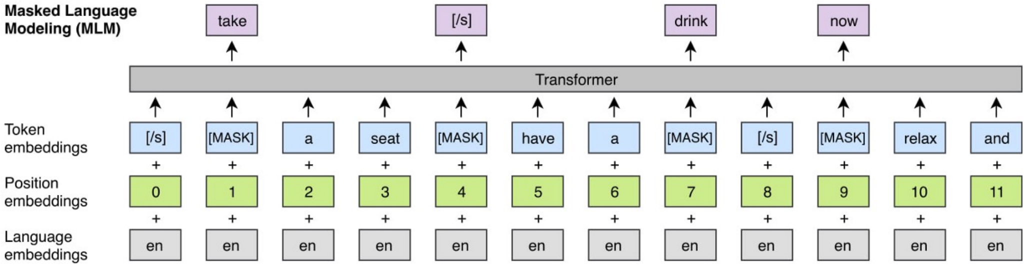
(Adapted from https://www.programmersought.com/article/24793362644/)

Pre-trained Models for Language Generation

Similarities and Differences

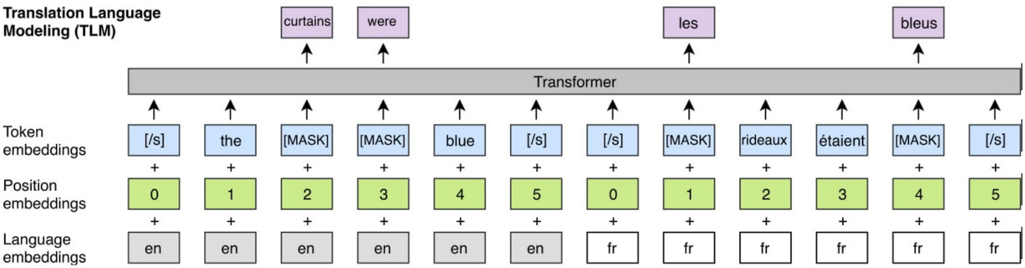
- Encoder vs. decoder vs. both
- Loss function (task)
- Monolingual vs. parallel data
- Monolingual vs. multilingual model
- Noise function (if any)

Masked Language Modeling (MLM) with XLM (Bert-like)



(Conneau and Lample, NIPS 2019)

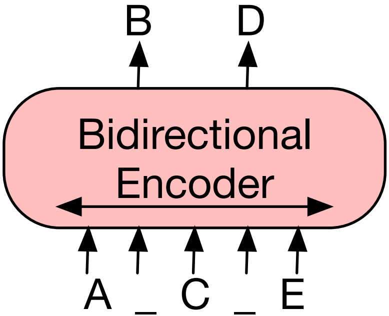
Translation Language Modeling (TLM) with XLM



(Conneau and Lample, NIPS 2019)

Pre-trained Models for Language Generation

Denoising Autoencoders for Language Generation (BERT)

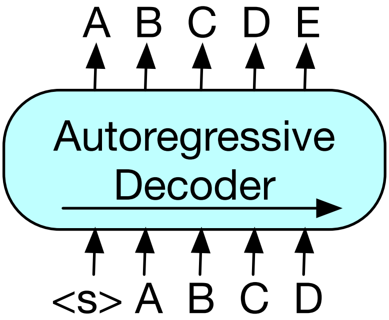


- BERT
- Masked LM

(Images from Lewis et al., ACL 2020)

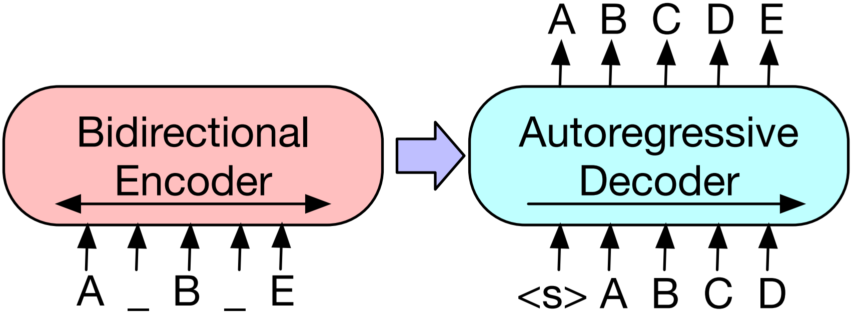
Pre-trained Models for Language Generation

Autoregressive Decoding for Language Generation (GPT-X)



- GPT
- Causal LM
- Good for generation

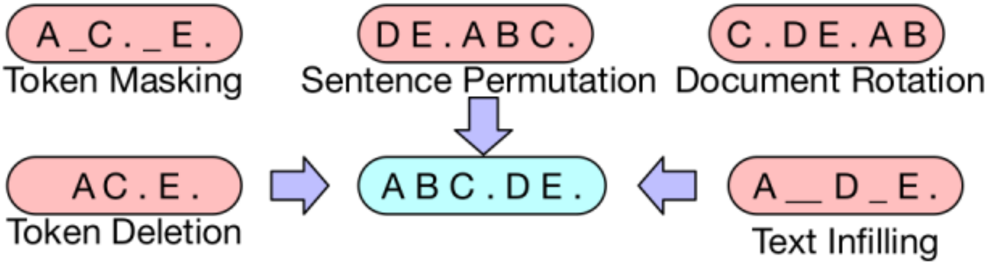
(Image from Lewis et al., ACL 2020)



(Image from Lewis et al., ACL 2020)

Language Generation with (m)BART

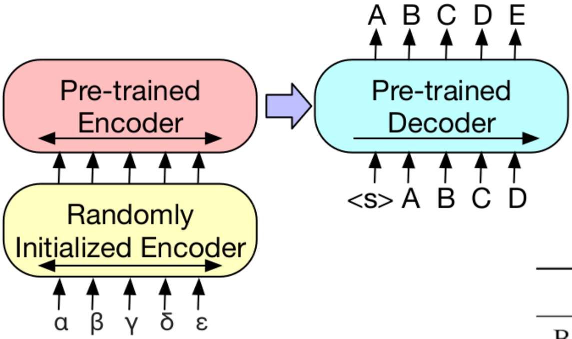
From MLMs to Noise



(Image from Lewis et al., ACL 2020)

Language Generation with (m)BART

BART for Machine Translation

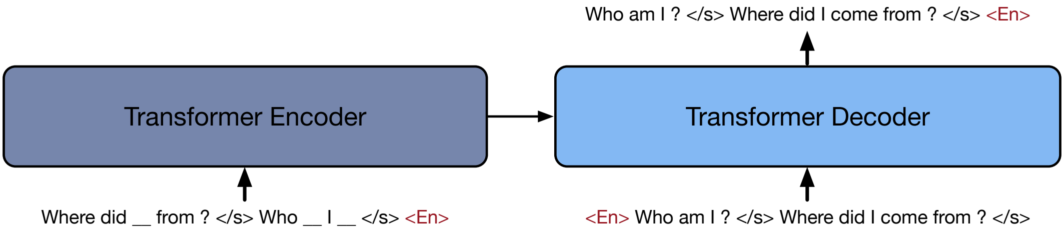


	RO-EN
Baseline	36.80
Fixed BART	36.29
Tuned BART	**37.96**

(Image from Lewis et al., ACL 2020)

Language Generation with (m)BART

Multilingual Denoising Pre-training (mBART)

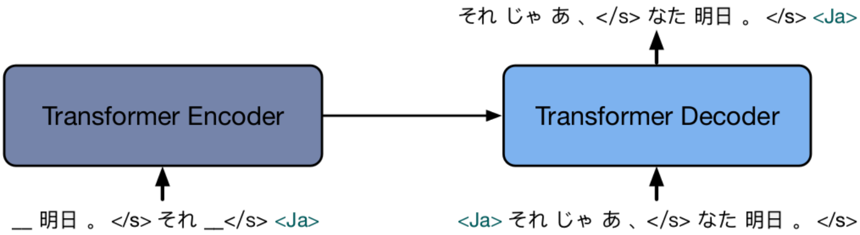


- **Noise**: word-span masking (text infilling) and sentence permutation

(Image from Liu et al., TACL 2020)

Language Generation with (m)BART

Multilingual Denoising Pre-training (mBART)

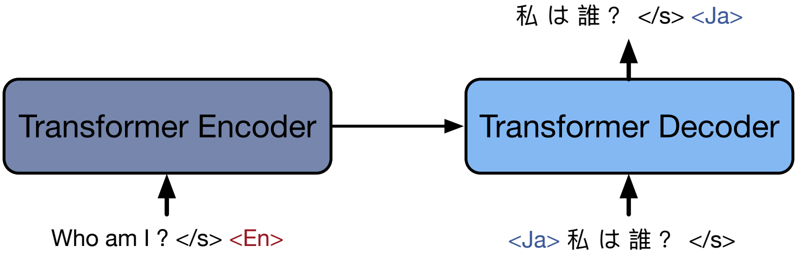


- **Noise**: word-span masking (text infilling) and sentence permutation

(Image from Liu et al., TACL 2020)

Language Generation with (m)BART

mBART: Finetuning for MT



Sentence-level finetuning

(Image from Liu et al., TACL 2020)

Language Generation with (m)BART

mBART: Finetuning for MT, Results

Languages	**En-Gu**		**En-Kk**		**En-Vi**		**En-Tr**		**En-Ja**		**En-Ko**	
Data Source	**WMT19**		**WMT19**		**IWSLT15**		**WMT17**		**IWSLT17**		**IWSLT17**	
Size	10K		91K		133K		207K		223K		230K	
Direction	←	→	←	→	←	→	←	→	←	→	←	→
Random	0.0	0.0	0.8	0.2	23.6	24.8	12.2	9.5	10.4	12.3	15.3	16.3
mBART25	**0.3**	**0.1**	**7.4**	**2.5**	**36.1**	**35.4**	**22.5**	**17.8**	**19.1**	**19.4**	**24.6**	**22.6**

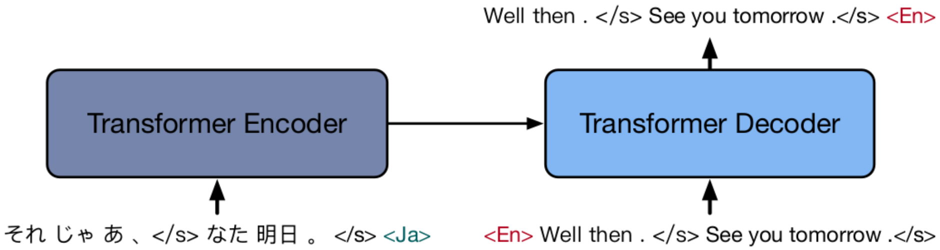
Languages	**En-Nl**		**En-Ar**		**En-It**		**En-My**		**En-Ne**		**En-Ro**	
Data Source	**IWSLT17**		**IWSLT17**		**IWSLT17**		**WAT19**		**FLoRes**		**WMT16**	
Size	237K		250K		250K		259K		564K		608K	
Direction	←	→	←	→	←	→	←	→	←	→	←	→
Random	34.6	29.3	27.5	16.9	31.7	28.0	23.3	34.9	7.6	4.3	34.0	34.3
mBART25	**43.3**	**34.8**	**37.6**	**21.6**	**39.8**	**34.0**	**28.3**	**36.9**	**14.5**	**7.4**	**37.8**	**37.7**

Language Generation with (m)BART

mBART: Finetuning for MT, Results

Languages	**En-Si**		**En-Hi**		**En-Et**		**En-Lt**		**En-Fi**		**En-Lv**	
Data Source	**FLoRes**		**ITTB**		**WMT18**		**WMT19**		**WMT17**		**WMT17**	
Size	647K		1.56M		1.94M		2.11M		2.66M		4.50M	
Direction	←	→	←	→	←	→	←	→	←	→	←	→
Random	7.2	1.2	10.9	14.2	22.6	17.9	18.1	12.1	21.8	20.2	15.6	12.9
mBART25	**13.7**	**3.3**	**23.5**	**20.8**	**27.8**	**21.4**	**22.4**	**15.3**	**28.5**	**22.4**	**19.3**	**15.9**

mBART: Finetuning for MT (II)



Document-level finetuning

(Image from Liu et al., TACL 2020)

mBART: Finetuning for MT (II)

Model	**Random**		**mBART25**	
	s-BLEU	d-BLEU	s-BLEU	d-BLEU
Sent-MT	34.5	35.9	36.4	38.0
Doc-MT	$\times$	7.7	**37.1**	**38.5**

- No document-level data for previous tests
- Results with German–English

Language Generation with (m)BART

mBART: Comparison with Other Pre-training Approaches

Pre-training		Fine-tuning		
Model	**Data**	**En→Ro**	**Ro→En**	**+BT**
Random	None	34.3	34.0	36.8
XLM	En Ro	-	35.6	38.5
MASS	En Ro	-	-	39.1
BART	En	-	-	38.0
XLM-R	CC100	35.6	35.8	-
BART-En	En	36.0	35.8	37.4
BART-Ro	Ro	37.6	36.8	38.1
mBART02	En Ro	**38.5**	**38.5**	**39.9**
mBART25	CC25	37.7	37.8	38.8

Digression

Let's join the main path again

Self-Supervised NMT

SSNMT in the Low Resource Setting

Open Problems

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

Similar issues in unsupervised NMT, bilingual embeddings, etc.

Same “solutions”?

SSNMT in the Low Resource Setting

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)

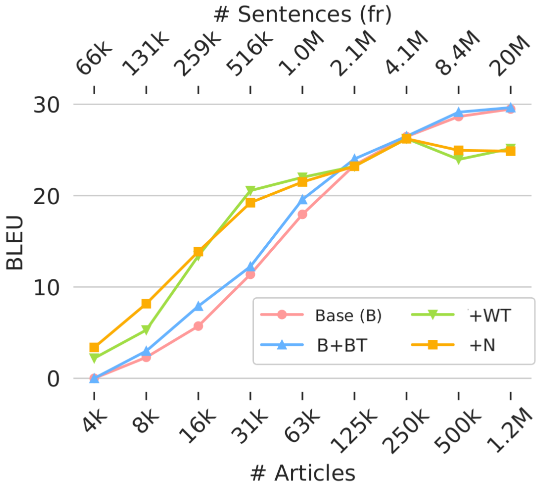
SSNMT in the Low Resource Setting

How does it Work?

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

SSNMT in the Low Resource Setting

A Simulated Setting: Data Augmentation vs. Corpus Size



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

(English & French Wikipedias)

SSNMT in the Low Resource Setting

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

SSNMT in the Low Resource Setting

Automatic Evaluation (BLEU scores on Different Sets)

Language (L)

Latin

yo

Initialization		B	+BT	+WT	+N
en2L	none	0.3±0.1	0.3±0.1	2.2±0.1	0.0±0.0
en2L	WE	0.5±0.1	0.4±0.1	2.9±0.1	0.9±0.0
en2L	DAE	2.0±0.1	2.3±0.1	2.8±0.1	1.2±0.1
L2en	none	0.5±0.1	0.6±0.1	2.7±0.1	0.2±0.0
L2en	WE	0.6±0.1	0.5±0.1	2.5±0.1	0.0±0.0
L2en	DAE	2.6±0.1	3.0±0.1	3.1±0.1	2.0±0.1

af

Initialization		B	+BT	+WT	+N
en2L	none	48.1±0.9	49.0±1.0	1.1±0.1	37.1±0.8
en2L	WE	48.1±0.9	51.2±0.9	8.4±0.5	41.7±0.9
en2L	DAE	44.8±0.9	48.6±0.9	42.3±0.9	38.9±0.9
L2en	none	47.9±0.9	51.3±0.9	0.7±0.1	38.6±0.9
L2en	WE	48.6±0.9	52.2±0.9	5.8±0.4	43.7±0.9
L2en	DAE	46.2±0.9	50.4±0.9	43.1±0.9	39.5±0.8

sw

Initialization		B	+BT	+WT	+N
en2L	none	4.2±0.2	6.1±0.2	0.9±0.1	5.6±0.2
en2L	WE	4.4±0.2	5.1±0.2	3.0±0.2	7.7±0.3
en2L	DAE	5.3±0.2	7.2±0.3	4.7±0.2	4.7±0.2
L2en	none	3.6±0.2	5.5±0.3	0.4±0.0	5.0±0.2
L2en	WE	3.6±0.2	4.2±0.2	2.1±0.1	6.3±0.2
L2en	DAE	4.8±0.2	6.8±0.2	5.6±0.2	5.9±0.2

Brahmic

my

Initialization		B	+BT	+WT	+N
en2L	none	0.0±0.0	0.0±0.0	0.1±0.0	0.1±0.0
en2L	WE	0.0±0.0	0.0±0.0	0.1±0.0	0.1±0.0
en2L	DAE	0.1±0.0	0.1±0.0	0.1±0.0	0.0±0.0
L2en	none	0.0±0.0	0.0±0.0	0.1±0.0	0.2±0.1
L2en	WE	0.1±0.0	0.0±0.0	0.2±0.0	0.4±0.0
L2en	DAE	0.7±0.1	0.6±0.0	0.7±0.1	0.4±0.1

ne

Initialization		B	+BT	+WT	+N
en2L	none	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	0.2±0.0	0.1±0.0
en2L	DAE	0.1±0.0	0.2±0.0	0.1±0.0	0.3±0.0
L2en	none	0.0±0.0	0.0±0.0	0.2±0.0	0.1±0.0
L2en	WE	0.1±0.0	0.0±0.0	0.1±0.0	0.4±0.1
L2en	DAE	0.3±0.1	0.3±0.1	0.5±0.1	0.5±0.0

kn

Initialization		B	+BT	+WT	+N
en2L	none	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	0.2±0.0	0.2±0.0
en2L	DAE	0.0±0.0	0.0±0.0	0.2±0.0	0.3±0.0
L2en	none	0.0±0.0	0.0±0.0	0.2±0.0	0.7±0.1
L2en	WE	0.0±0.0	0.0±0.0	0.2±0.0	0.2±0.0
L2en	DAE	0.0±0.0	0.0±0.0	0.7±0.1	0.9±0.1

SSNMT in the Low Resource Setting

Mmmm... What else?

- Multilinguality
- Fine-tuning

SSNMT in the Low Resource Setting

Mmmm... What else?

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

SSNMT in the Low Resource Setting

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

SSNMT in the Low Resource Setting

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**

SSNMT in the Low Resource Setting

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*	
	$\rightarrow$	$\leftarrow$	$\rightarrow$	$\leftarrow$	$\rightarrow$	$\leftarrow$	$\rightarrow$	$\leftarrow$	$\rightarrow$	$\leftarrow$	$\rightarrow$	$\leftarrow$
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	

SSNMT in the Low Resource Setting

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

Automatic Evaluation in the Low-Resource Setting

Thoughts

- We have seen several ways to approach LR-MT (and we'll see more!)
  - Multilinguality, fine-tuning, UMT, SSNMT, etc.
- What makes MT low-resource?
  - data size, word overlap, typology, word order, and a long etc.
- How can we compare?
  - few standardized data, test sets... of course, low-resource!
- Even more... what is a good metric?
  - BLEU makes sense with small values? Metrics based on multilingual LMs (BertScore, Comet, etc) don't know the language!

Automatic Evaluation in the Low-Resource Setting

As Always, it's Late...

More to come!!

Thanks! And...

wait!

Self-Supervised Neural Machine Translation and More!

Cristina España-Bonet
DFKI GmbH

Low-Resource NLP:
Multilinguality and Machine Translation
Webinar Series — Session IV
14th September 2021