Neural Machine Translation for Sumerian–English

Student: Rachit Bansal
Mentors: Dr Niko Schenk, Ravneet Punia, Dr Jacob L. Dahl

What?

- To translate the entire Sumerian Ur-III corpora by making use of monolingual text across Semi-Supervised and Unsupervised techniques.
- Improve the previous work done for Sumerian-English Machine Translation.

Why?

- 1.5M monolingual sentences v/s ~8k parallel.
- Trained on sparse and irregular data, resulting in lack of contextual understanding by the models.

#atf: lang sux	#tr.en
pisan-dub-ba	Basket-of-tablets:
dub gid2-da	long-tablets,
sze erin2 gi-zi	barley of the (labor-)troops
ba-zi dumu na-silim	Bazi, son of Nasilim,
i3-gal2	are here;

What?

- To translate the entire Sumerian Ur-III corpora by making use of monolingual text across Semi-Supervised and Unsupervised techniques.

- Improve the previous work done for Sumerian-English Machine Translation.

Why?

- 1.5M monolingual sentences v/s ~8k parallel.

- Trained on sparse and irregular data, resulting in lack of contextual understanding by the models.

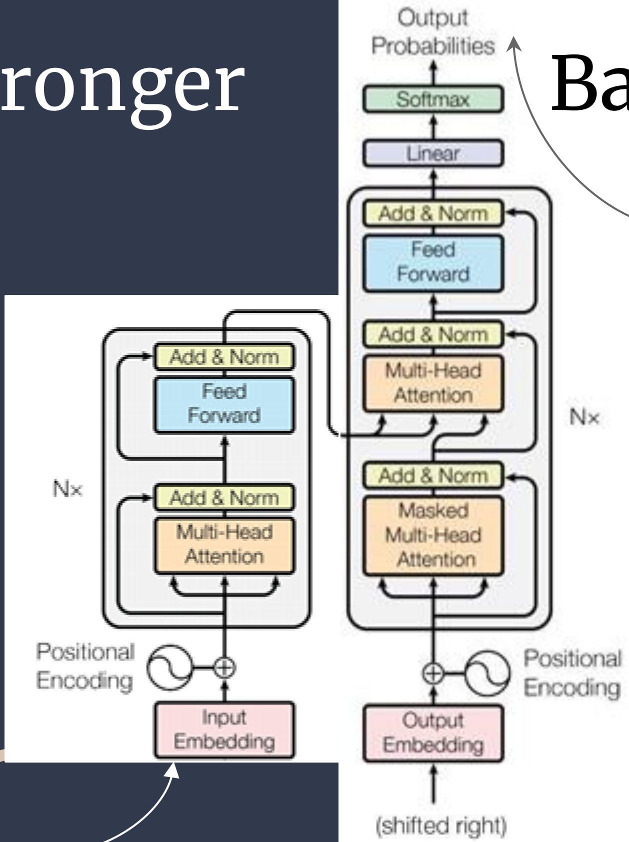
#atf: lang sux

#tr.en

pisan-dub-ba dub gid2-da sze erin2
gi-zi ba-zi dumu na-silim i3-gal2

Basket-of-tablets: long-tablets, barley of the
(labor-)troops Bazi, son of Nasilim, are here;

Stronger Baselines

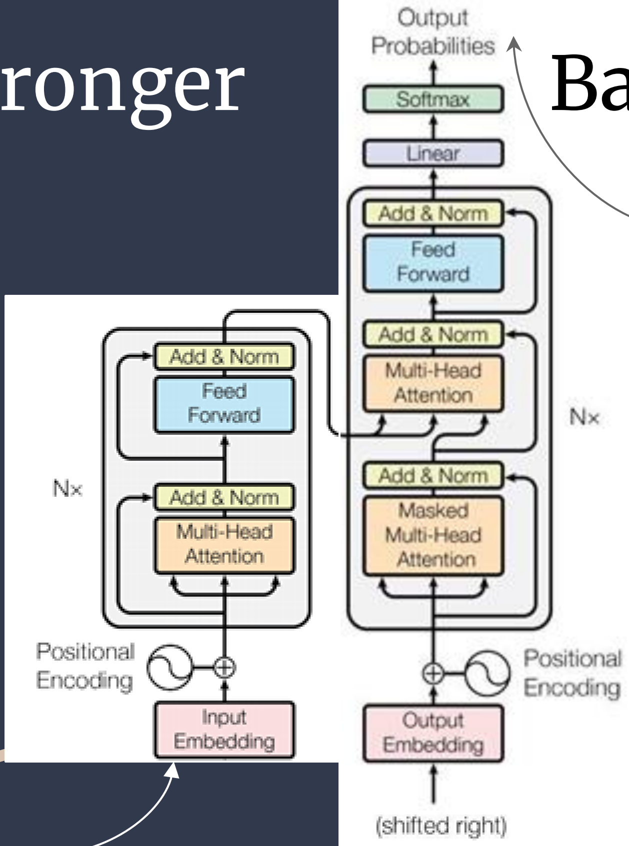


UrIII-Comp:
Complete Sentences from the Ur-III Corpora only.

pisan-dub-ba dub gid2-da sze erin2 gi-zi ba-zi dumu na-silim i3-gal2

Basket-of-tablets: long-tablets, barley of the (labor-)troops Bazi, son of Nasilim, are here;

Stronger Baselines

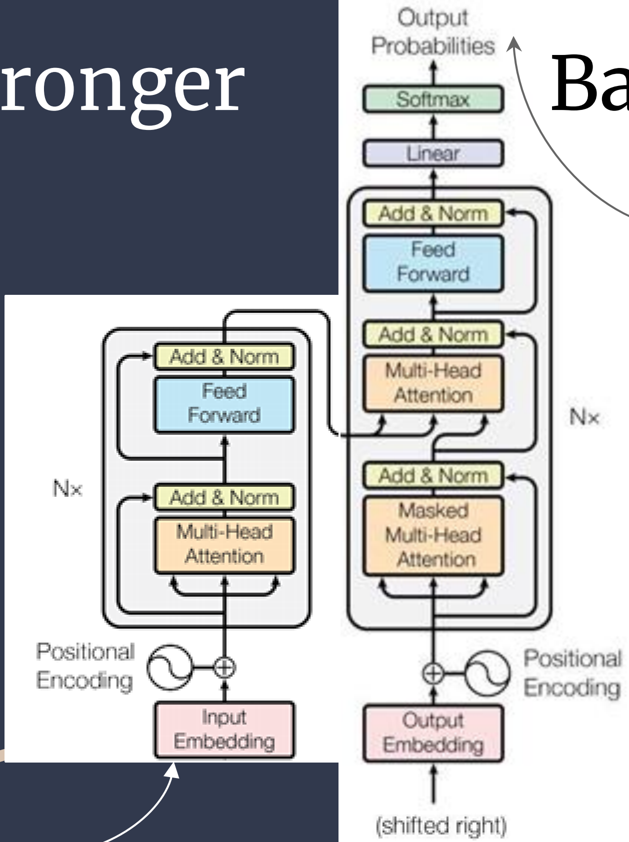


UrIII-Seg:
Partial phrases/segments
from the Ur-III Corpora only.

pisan-dub-ba

Basket-of-tablets:

Stronger Baselines

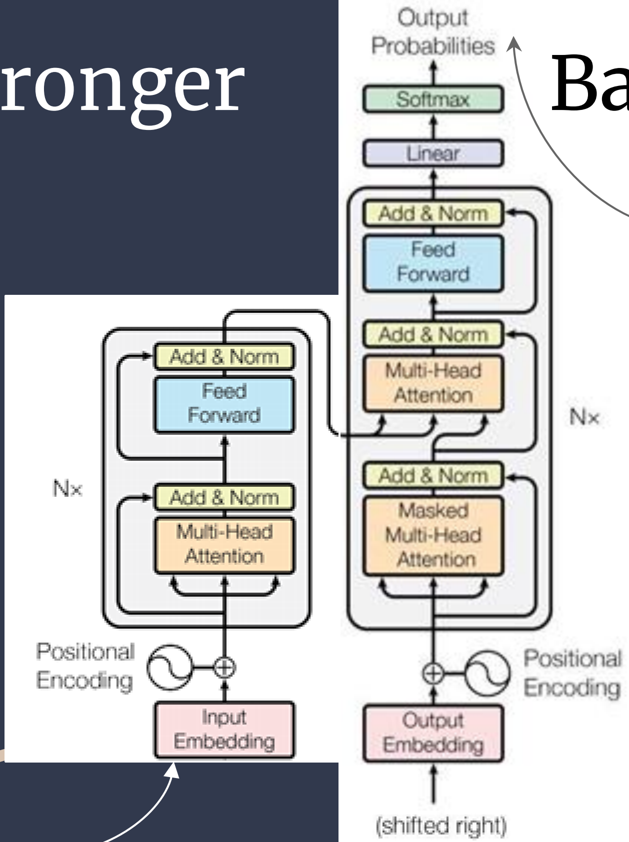


All-Seg:
Partial phrases/segments across all genres of Sumerian texts.

1(asz@c) sze guru7

1 barley-silo,

Stronger Baselines



All-Comp:
Complete Sentences across all genres of Sumerian texts.

1(asz@c) sze guru7 sila3
7(asz@c)1(asz) lu2 szu ba-ti lu2-bi

1 barley-silo, sila3: 7
did one man receive, the men:

Methods

Self-Supervision + Fine-Tuning

Initializing the model parameters after pre-training the encoder on the source and target side monolingual data using **MLM** and **TLM** tasks. [1][2]

Data Augmentation

Expanding the target-side monolingual data for pre-training by using **BERT** Embeddings, **CharSwap** and **WordNet** Synonyms. [3]

Back Translation

Iterative re-training by using the periodic model to translate source or target monolingual data to synthetically expand the parallel data. [4][5]

[1] Cross-lingual Language Model Pretraining, Guillaume Lample and Alexis Conneau, 2019

[2] MASS: Masked Sequence to Sequence Pre-training for Language Generation, Kaitao Song and Xu Tan and Tao Qin and Jianfeng Lu and Tie-Yan Liu, 2019

[3] TextAttack: A Framework for Adversarial Attacks, Data Augmentation, and Adversarial Training in NLP} John X. Morris and Eli Lifland and Jin Yong Yoo and Jake Grigsby and Di Jin and Yanjun Qi, 2020

[4] Understanding Back-Translation at Scale, Edunov, Sergey and Ott, Myle and Auli, Michael and Grangier, David, 2018

[5] Investigating Backtranslation in Neural Machine Translation, Alberto Poncelas and D. Shterionov and A. Way and G. M. D. B. Wenniger and Peyman Passban, 2018

Methods

Self-Supervision + Fine-Tuning

Initializing the model parameters after pre-training the encoder on the source and target side monolingual data using **MLM** and **TLM** tasks. [1][2]

Data Augmentation

Expanding the target-side monolingual data for pre-training by using **BERT** Embeddings, **CharSwap** and **WordNet** Synonyms. [3]

Back Translation

Iterative re-training by using the periodic model to translate source or target monolingual data to synthetically expand the parallel data. [4][5]

[1] Cross-lingual Language Model Pretraining, Guillaume Lample and Alexis Conneau, 2019

[2] MASS: Masked Sequence to Sequence Pre-training for Language Generation, Kaitao Song and Xu Tan and Tao Qin and Jianfeng Lu and Tie-Yan Liu, 2019

[3] TextAttack: A Framework for Adversarial Attacks, Data Augmentation, and Adversarial Training in NLP} John X. Morris and Eli Lifland and Jin Yong Yoo and Jake Grigsby and Di Jin and Yanjun Qi, 2020

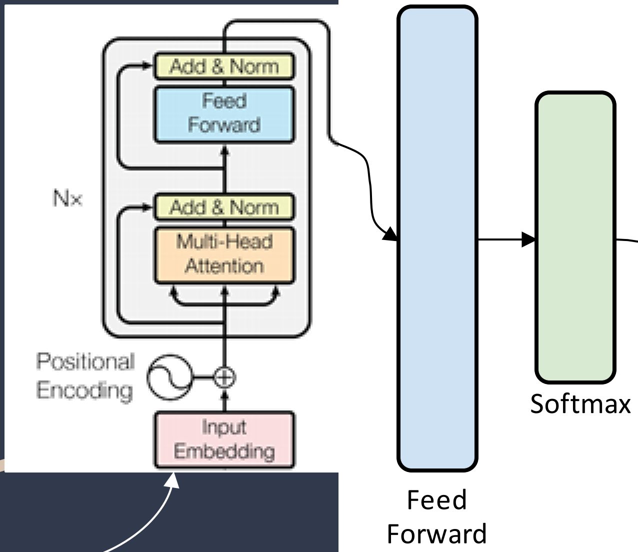
[4] Understanding Back-Translation at Scale, Edunov, Sergey and Ott, Myle and Auli, Michael and Grangier, David, 2018

[5] Investigating Backtranslation in Neural Machine Translation, Alberto Poncelas and D. Shterionov and A. Way and G. M. D. B. Wenniger and Peyman Passban, 2018

Self Supervision

Initializing the model parameters after pre-training the encoder on the source and target side monolingual data using **MLM** and **TLM** tasks.

Masked Language Modelling (MLM)



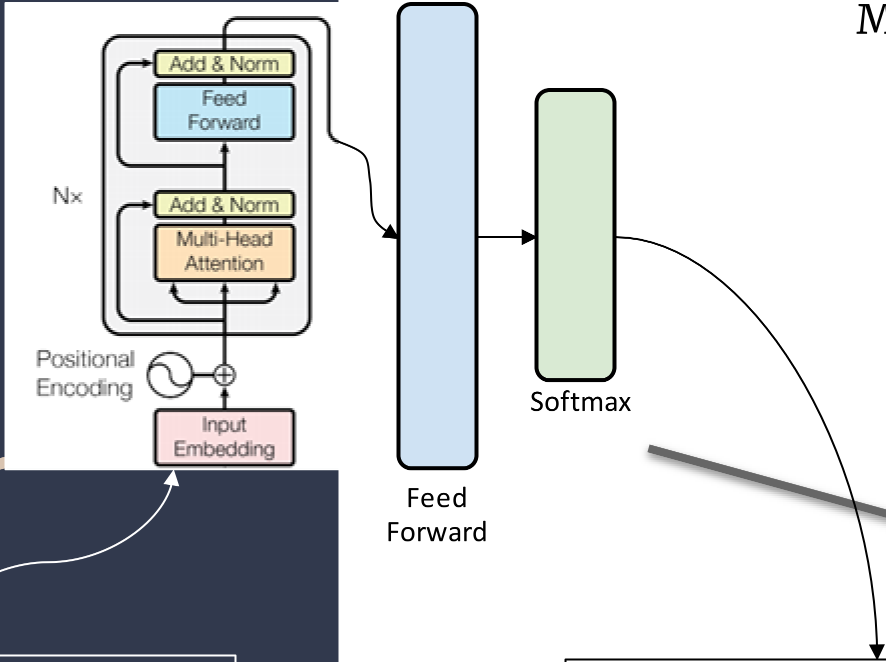
1(asz@c) sze guru7 <**MASK**>
7(asz@c)1(asz) lu2 szu ba-ti lu2-bi

1(asz@c) sze guru7 **sila3**
7(asz@c)1(asz) lu2 szu ba-ti lu2-bi

Self Supervision

Initializing the model parameters after pre-training the encoder on the source and target side monolingual data using **MLM** and **TLM** tasks.

Translation Language Modelling (TLM)



1(asz@c) sze guru7 <**MASK**> 7
did one man <**MASK**>, the men:

1(asz@c) sze guru7 **sila3** 7
did one man **recieve**, the men:

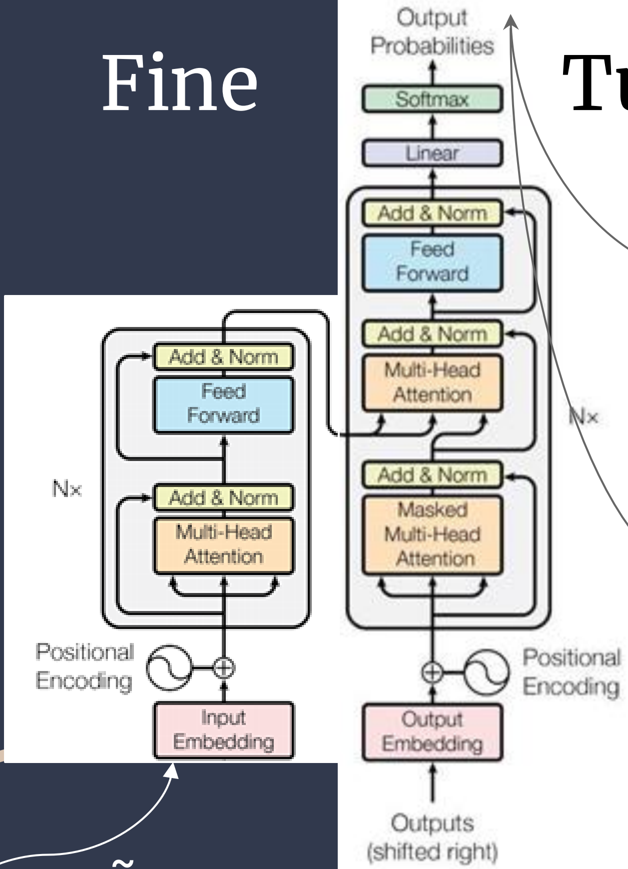
Fine Tuning

Denoising Auto-Encoding

Language modelling task to return the input, provided some noise.

Machine Translation

Predicting the target sentence given unabridged source sentence.



1(asz@c) sze guru7 sila3
7(asz@c)1(asz) lu2 szu ba-ti lu2-bi

Unsupervised

1 barley-silo, sila3: 7
did one man receive, the men:

Semi-Supervised

1(asz@c) sze guru7 sila3
7(asz@c)1(asz) lu2 szu ba-ti lu2-bi

Catastrophic Scarcity

- Sumerian text has been translated to English at the level of small words and phrases. This makes the target-side text of our parallel corpora incoherent with the general English text seen in other datasets.

- Obtaining target-side monolingual text which gives the model a true representation of the desired text becomes difficult.

- Semi-supervised techniques relying on monolingual texts suffer.

Methods

Self-Supervision + Fine-Tuning

Initializing the model parameters after pre-training the encoder on the source and target side monolingual data using **MLM** and **TLM** tasks. [1][2]

Data Augmentation

Expanding the target-side monolingual data for pre-training by using **BERT** Embeddings, **CharSwap** and **WordNet** Synonyms. [3]

Back Translation

Iterative re-training by using the periodic model to translate source or target monolingual data to synthetically expand the parallel data. [4][5]

[1] Cross-lingual Language Model Pretraining, Guillaume Lample and Alexis Conneau, 2019

[2] MASS: Masked Sequence to Sequence Pre-training for Language Generation, Kaitao Song and Xu Tan and Tao Qin and Jianfeng Lu and Tie-Yan Liu, 2019

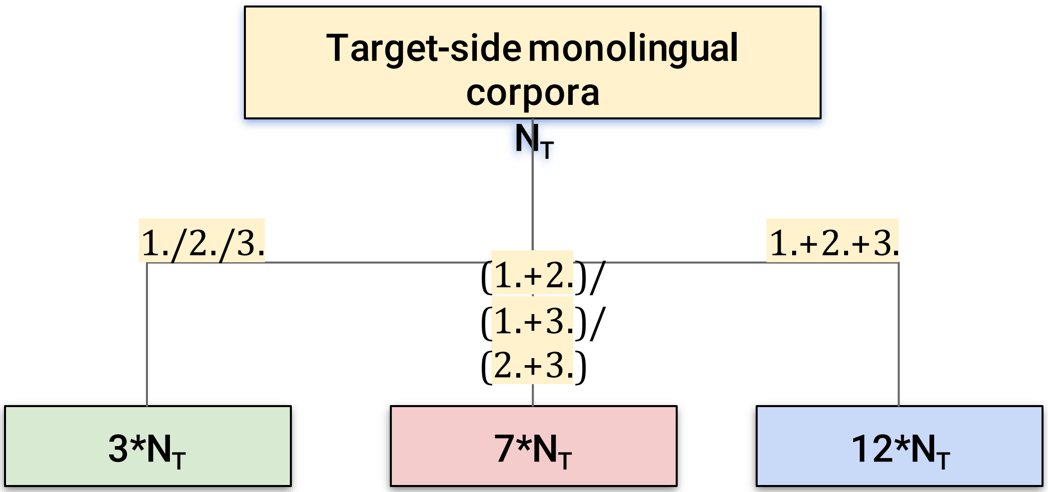
[3] TextAttack: A Framework for Adversarial Attacks, Data Augmentation, and Adversarial Training in NLP} John X. Morris and Eli Lifland and Jin Yong Yoo and Jake Grigsby and Di Jin and Yanjun Qi, 2020

[4] Understanding Back-Translation at Scale, Edunov, Sergey and Ott, Myle and Auli, Michael and Grangier, David, 2018

[5] Investigating Backtranslation in Neural Machine Translation, Alberto Poncelas and D. Shterionov and A. Way and G. M. D. B. Wenniger and Peyman Passban, 2018

Data Augmentation

1. <u>**BERT**</u>: Replacing words by the spatially closest words measured by **Cosine Similarity in BERT Embeddings**, with a threshold of 0.8.
2. <u>**WordNet**</u>: Replacing words with WordNet **synonyms**.
3. <u>**CharSwap**</u>: Introduces certain **character-level perturbations** in the text by substituting, deleting, inserting, and swapping adjacent character tokens.



Methods

Self-Supervision + Fine-Tuning

Initialising the model parameters after pre-training the encoder on the source and target side monolingual data using **MLM** and **TLM** tasks.[1][2]

Data Augmentation

Expanding the target-side monolingual data for pre-training by using **BERT** Embeddings, **CharSwap** and **WordNet** Synonyms. [3]

Back Translation

Iterative re-training by using the periodic model to translate source or target monolingual data to synthetically expand the parallel data. [4][5]

[1] Cross-lingual Language Model Pretraining, Guillaume Lample and Alexis Conneau, 2019

[2] MASS: Masked Sequence to Sequence Pre-training for Language Generation, Kaitao Song and Xu Tan and Tao Qin and Jianfeng Lu and Tie-Yan Liu, 2019

[3] TextAttack: A Framework for Adversarial Attacks, Data Augmentation, and Adversarial Training in NLP} John X. Morris and Eli Lifland and Jin Yong Yoo and Jake Grigsby and Di Jin and Yanjun Qi, 2020

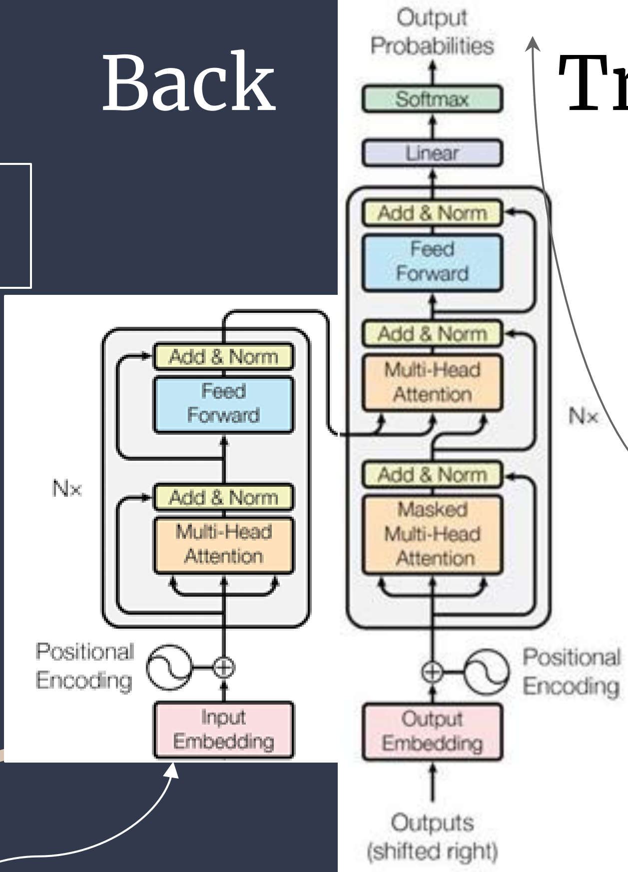
[4] Understanding Back-Translation at Scale, Edunov, Sergey and Ott, Myle and Auli, Michael and Grangier, David, 2018

[5] Investigating Backtranslation in Neural Machine Translation, Alberto Poncelas and D. Shterionov and A. Way and G. M. D. B. Wenniger and Peyman Passban, 2018

Back Translation

Step-1

Train a model using the existing parallel corpora



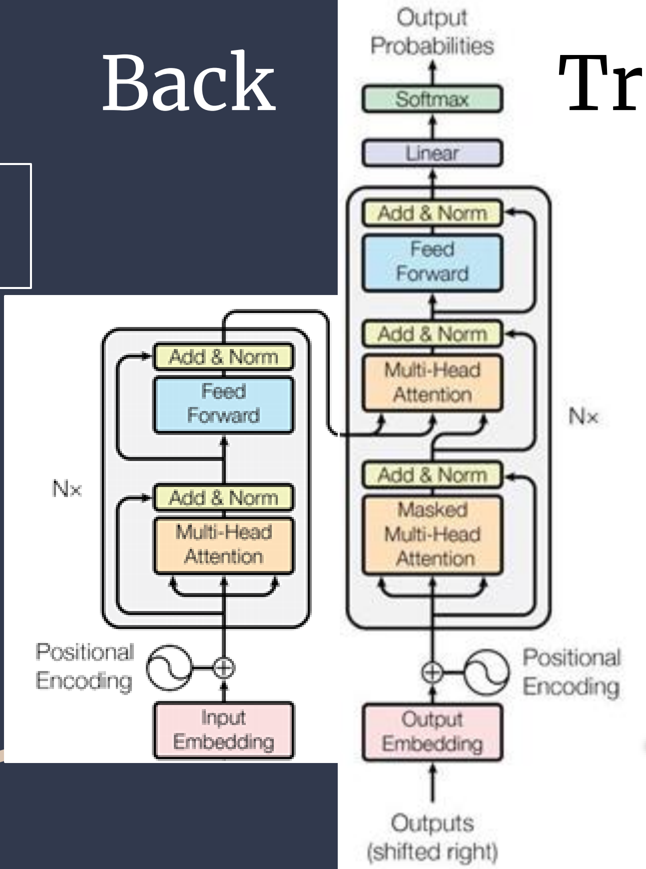
P_S

P_T

Back Translation

Step-1

Train a model using the existing parallel corpora



Step-2

Divide the source-side monolingual data into ‘n’ shards.

P_S	M_S^1	. . .	M_S^n

P_T

Back Translation

Step-1

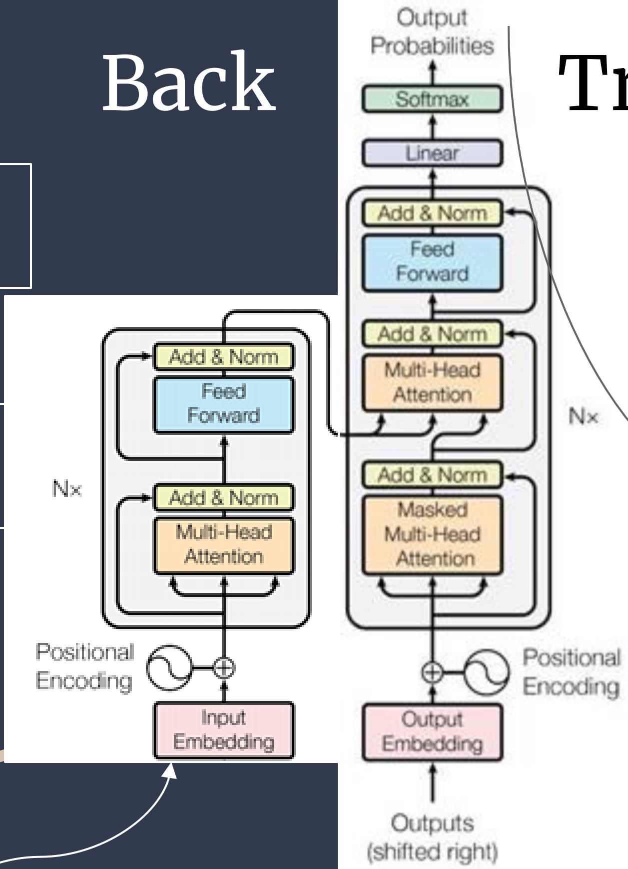
Train a model using the existing parallel corpora

Step-2

Divide the source-side monolingual data into 'n' shards.

Step-3

Translate the adjacent shard using the trained model



P_S	M_S^1	M_S^2	. . .	M_S^n	

P_T	M_T^1

Back Translation

Step-1

Train a model using the existing parallel corpora

Step-2

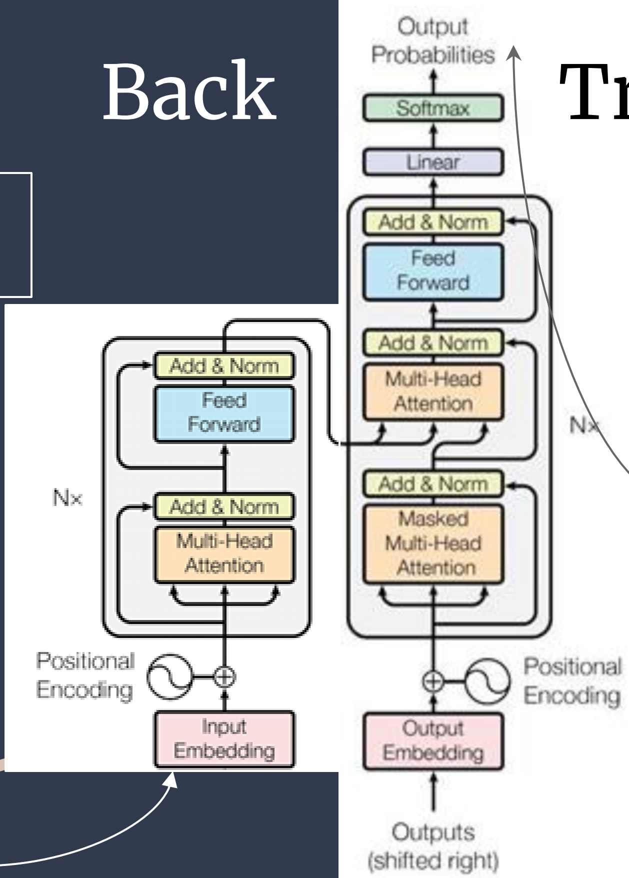
Divide the source-side monolingual data into ‘n’ shards.

Step-3

Translate the adjacent shard using the trained model

Step-4

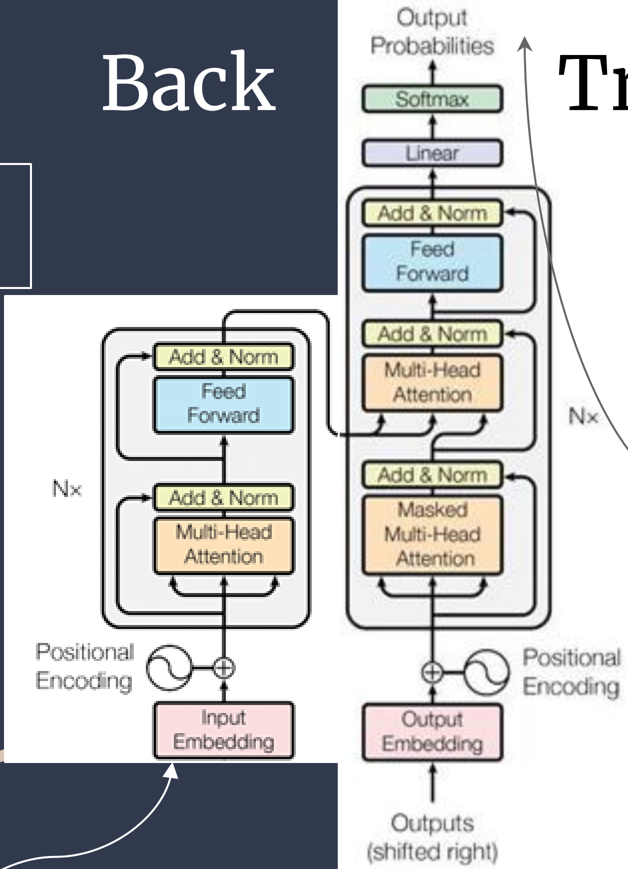
Re-train the model using the stacked data



$P_S + M_S^1$ | M_S^2 | . . . | M_S^n

$P_T + M_T^1$

Back Translation



Step-1

Train a model using the existing parallel corpora

Step-2

Divide the source-side monolingual data into 'n' shards.

Step-3

Translate the adjacent shard using the trained model

Step-4

Re-train the model using the stacked data

Step-5

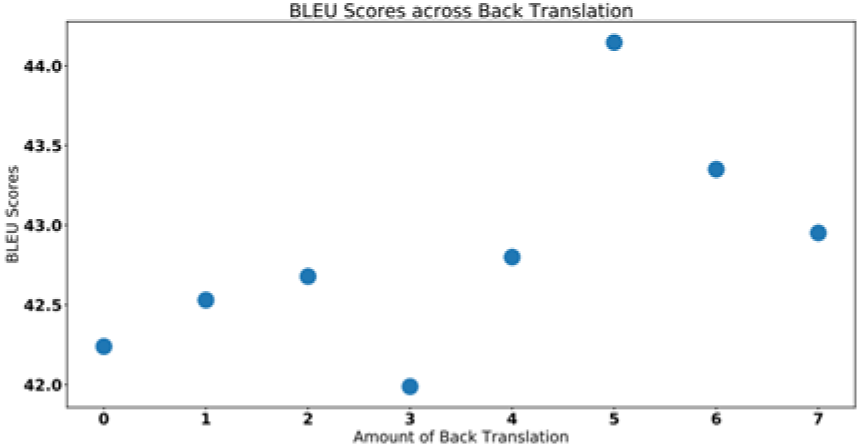
Repeat.

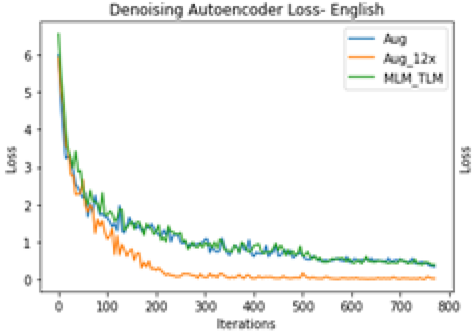
$P_S + M_S^1 + \ldots + M_S^n$

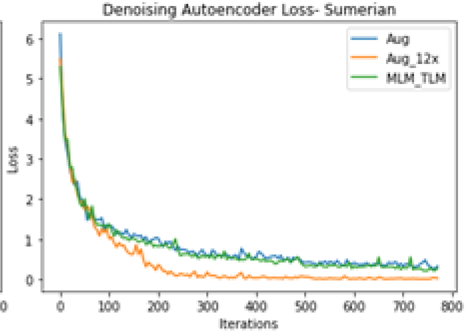
$P_T + M_T^1 + \ldots + M_T^n$

Experimental Results

Technique	Supervised	Un-supervised	Semi-Supervised	Human Evaluation
Vanilla Transformer M-BT				
`UrIIISeg`	36.32			2.202
`UrIIIComp`	33.45			2.242
`AllSeg`	37.01			2.360
`AllComp`	42.23			2.431
+3*M-BT			41.98	2.358
+5*M-BT			**44.14**	**2.504**
+7*M-BT			42.95	2.367
XLM				
MLM, `Orig`		4.49	15.04	
MLM + TLM, `WMT`		0.94	–	
`Mixed`		13.08	21.23	1.104, –
`Orig`		12.73	24.64	1.294, –
XLM + Data Augmentation				
`BERT`		13.06	29.50	1.320, 1.704
`WordNet`		13.08	28.57	1.269, 1.690
`CharSwap`		12.92	29.04	
`BERT+WordNet`		13.34	26.57	1.460, 1.666
`BERT+CharSwap` `+WordNet`		13.23	30.10	– , 1.757







Making sense of the models

Observing ‘why’ certain methods give higher metric scores than others.

Interpreting the results of various models using **gradient-based** and **perturbation-based** algorithms.

A net **attribution** for each output token is obtained, with respect to the spans of input text.

Created a generalisable pipeline for interpreting machine translation models.*

Actual	Human Expert	Model-1	Semi-Supervised DataAug XLM	Model-2	Unsupervised DataAug XLM	Model-3	Unsupervised Orig TLM XLM
Output Word	Visualisations	Output Word	Visualisations	Output Word	Visualisations	Output Word	Visualisations
barley	#s sze-ba geme2 usz-bar kiszib3 ur-{d}asznan ugula #e	barley	#s sze-ba geme2 usz-bar kiszib3 ur-{d}asznan ugula #e	Monthy	#s sze-ba geme2 usz-bar kiszib3 ur-{d}asznan ugula #e	Basketoftablets	#s sze-ba geme2 usz-bar kiszib3 ur-{d}asznan ugula #e
rations	#s sze-ba geme2 usz-bar kiszib3 ur-{d}asznan ugula #e	rations	#s sze-ba geme2 usz-bar kiszib3 ur-{d}asznan ugula #e	rations	#s sze-ba geme2 usz-bar kiszib3 ur-{d}asznan ugula #e	rations	#s sze-ba geme2 usz-bar kiszib3 ur-{d}asznan ugula #e
the	#s sze-ba geme2 usz-bar kiszib3 ur-{d}asznan ugula #e	the	#s sze-ba geme2 usz-bar kiszib3 ur-{d}asznan ugula #e	the	#s sze-ba geme2 usz-bar kiszib3 ur-{d}asznan ugula #e	the	#s sze-ba geme2 usz-bar kiszib3 ur-{d}asznan ugula #e
female	#s sze-ba geme2 usz-bar kiszib3 ur-{d}asznan ugula #e	female	#s sze-ba geme2 usz-bar kiszib3 ur-{d}asznan ugula #e	shephards	#s sze-ba geme2 usz-bar kiszib3 ur-{d}asznan ugula #e	labortroops	#s sze-ba geme2 usz-bar kiszib3 ur-{d}asznan ugula #e
weavers	#s sze-ba geme2 usz-bar kiszib3 ur-{d}asznan ugula #e	weavers	#s sze-ba geme2 usz-bar kiszib3 ur-{d}asznan ugula #e	weavers	#s sze-ba geme2 usz-bar kiszib3 ur-{d}asznan ugula #e	weavers	#s sze-ba geme2 usz-bar kiszib3 ur-{d}asznan ugula #e
under	#s sze-ba geme2 usz-bar kiszib3 ur-{d}asznan ugula #e	under	#s sze-ba geme2 usz-bar kiszib3 ur-{d}asznan ugula #e	from	#s sze-ba geme2 usz-bar kiszib3 ur-{d}asznan ugula #e	255	#s sze-ba geme2 usz-bar kiszib3 ur-{d}asznan ugula #e
seal	#s sze-ba geme2 usz-bar kiszib3 ur-{d}asznan ugula #e	seal	#s sze-ba geme2 usz-bar kiszib3 ur-{d}asznan ugula #e	seal	#s sze-ba geme2 usz-bar kiszib3 ur-{d}asznan ugula #e	seal	#s sze-ba geme2 usz-bar kiszib3 ur-{d}asznan ugula #e
of	#s sze-ba geme2 usz-bar kiszib3 ur-{d}asznan ugula #e	of	#s sze-ba geme2 usz-bar kiszib3 ur-{d}asznan ugula #e	of	#s sze-ba geme2 usz-bar kiszib3 ur-{d}asznan ugula #e	of	#s sze-ba geme2 usz-bar kiszib3 ur-{d}asznan ugula #e
UrAnan	#s sze-ba geme2 usz-bar kiszib3 ur-{d}asznan ugula #e	Lugalniglagare	#s sze-ba geme2 usz-bar kiszib3 ur-{d}asznan ugula #e	Ninlil	#s sze-ba geme2 usz-bar kiszib3 ur-{d}asznan ugula #e	weavers	#s sze-ba geme2 usz-bar kiszib3 ur-{d}asznan ugula #e
foreman	#s sze-ba geme2 usz-bar kiszib3 ur-{d}asznan ugula #e	foreman	#s sze-ba geme2 usz-bar kiszib3 ur-{d}asznan ugula #e	foreman	#s sze-ba geme2 usz-bar kiszib3 ur-{d}asznan ugula #e	female	#s sze-ba geme2 usz-bar kiszib3 ur-{d}asznan ugula #e

* cdli-gh/Sumerian-English-NMT/interpretability

Potential Usage

For Assyriologists to decipher the meaning of phrases and Sumerian text which do not have a clearly defined meaning yet.

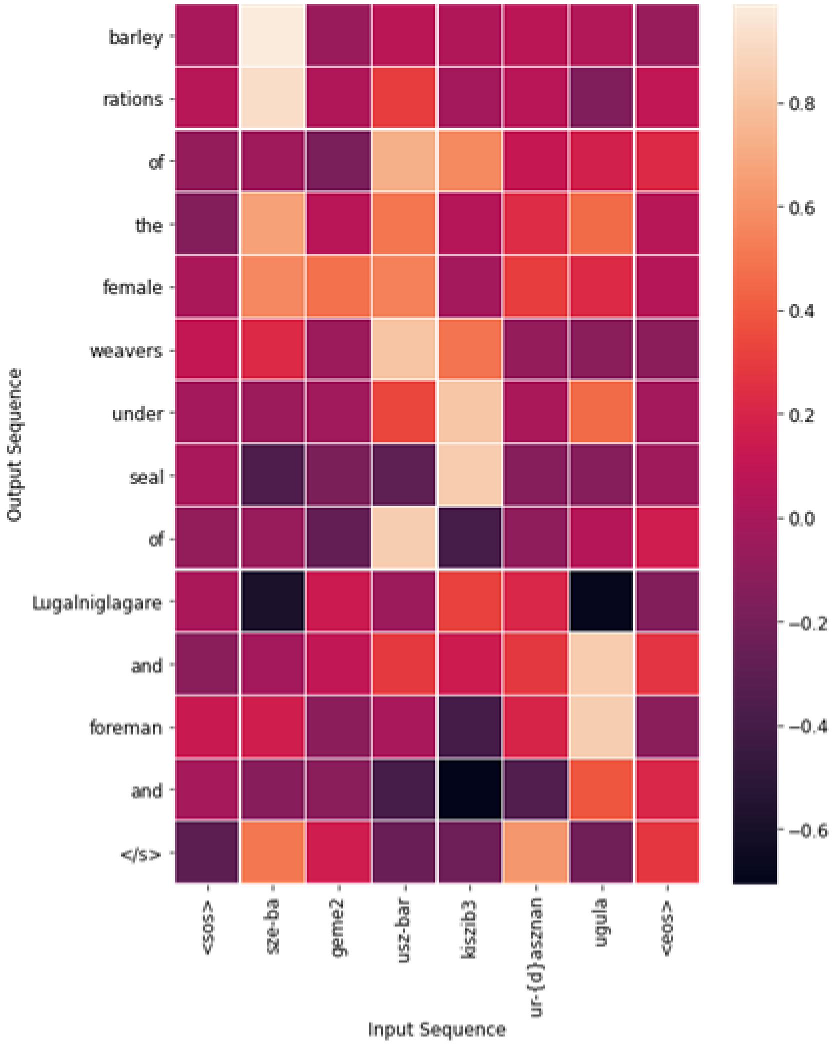
Future Work

Solving the issue of utf-8 encoding in the Sumerian text.

Dissect the problem of Catastrophic Scarcity across other Cuneiform Languages.

Deploying the Sumerian-English Translation pipeline to the CDLI Framework.

Working on the potential applications of the interpretability algorithms, specially for Sumerian.



Special Thanks to...

Émilie Pagé-Perron

Jacob L. Dahl

Ilya Khait

Lafont Bertrand

The entire MTAAC Team

Future Plans

Continue tuning and improving the project.

Inviting and guiding future contributions.

Work on more interesting developments around NLP for CDLI

Contributing as a Google Summer of Code student for 2021.

Thank you.