A Primer in BERTology:

What we know about how BERT works

16/03/2020

Based on

A Primer in BERTology: What we know about how BERT works,
Anna Rogers, Olga Kovaleva, Anna Rumshisky, arXiv 2020

Presented by: Mojtaba Valipour

PhD student of Computer Science at Data Analytics Lab

CS 886 – Deep Learning for NLP – Ming Li

UNIVERSITY OF WATERLOO

DATA ANALYTICS LAB

Outline

IMAGE CREDIT: http://www.avengingforce.com/?p=871

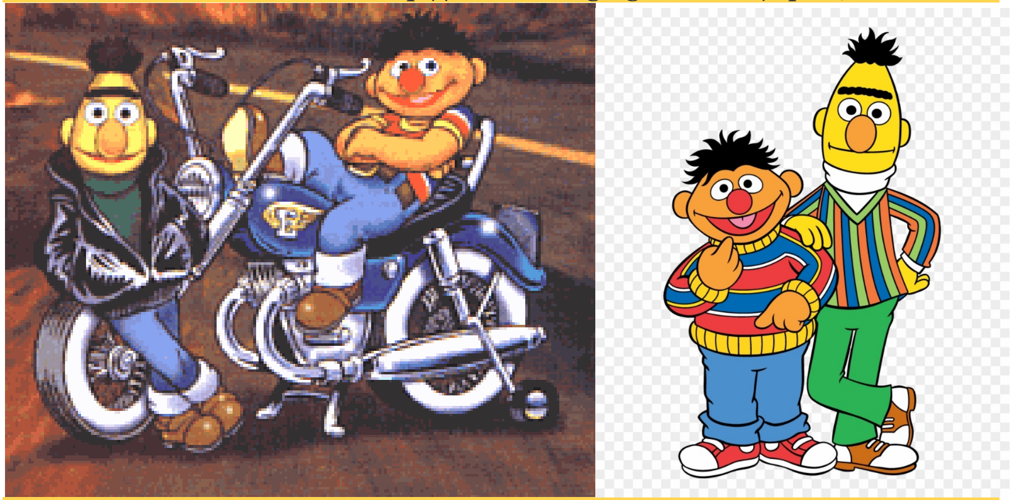


IMAGE CREDIT: www.picgifs.com, www.pinclipart.com

- BERT
- Language Knowledge
- Localizing Knowledge
- Conclusions

Ref:
1- A Primer in BERTology: What we know about how BERT works, A. Rogers, et. al., 2020
2- https://www.youtube.com/watch?v=3VZZbKoXDVM

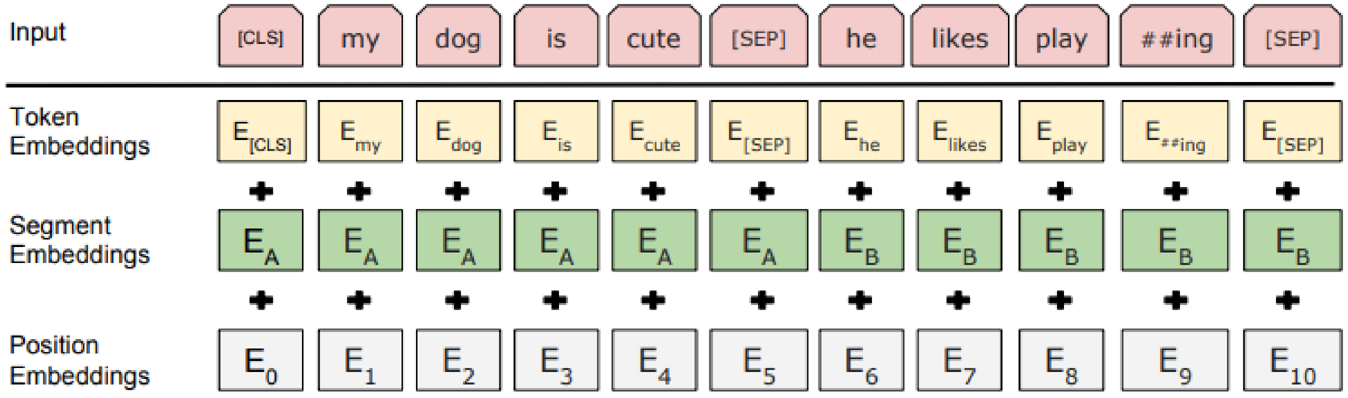
UNIVERSITY OF WATERLOO

WHAT IS BERT?

A Stack of Transformer Encoder Layers

BERT ARCHITECTURE

BERT Base: L=12, H=768, A=12, T: 110M
BERT Large: L=24,H=1024,A=16,T:340M



- *Task 1: Masked Language Modeling (MLM)*
- *Task 2: Next Sentence Prediction (NSP)*

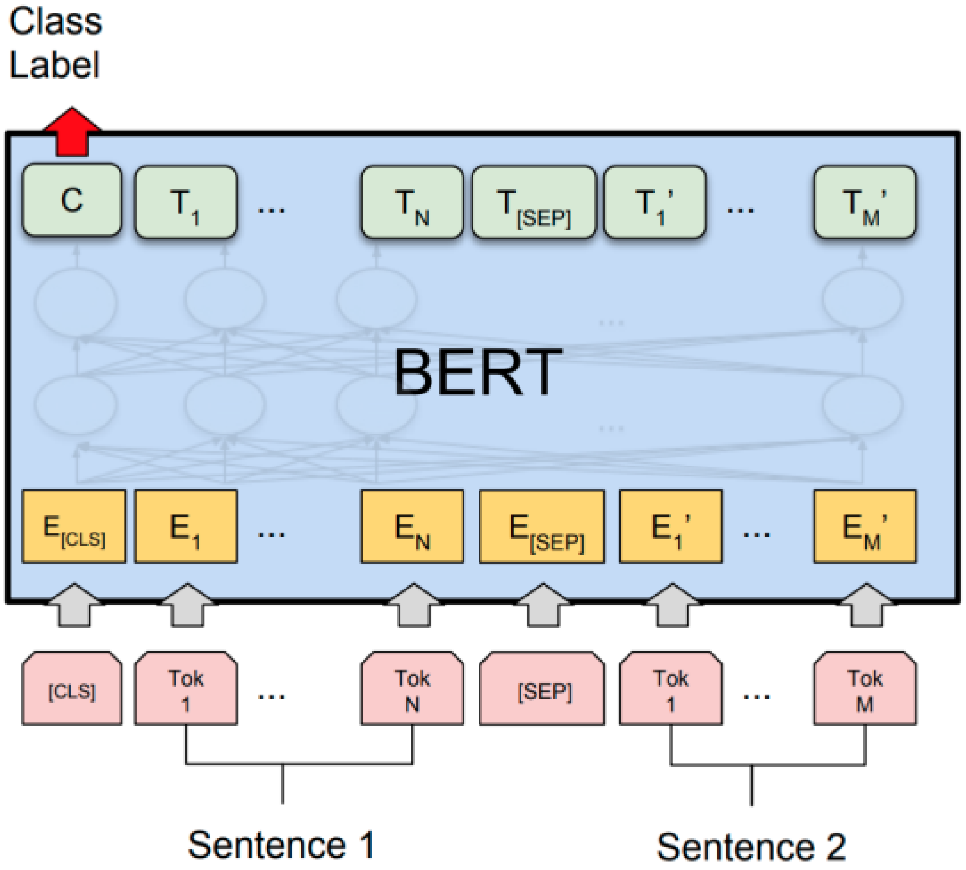


Figure 1: BERT fine-tuning (Devlin et al., 2019).

Ref:
1- A Primer in BERTology: What we know about how BERT works, A. Rogers, et. al., 2020
2- BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, J. Dave et al. 2019

UNIVERSITY OF WATERLOO

BERT ARCHITECTURE

BERT Base: L=12, H=768, A=12, T: 110M
BERT Large: L=24,H=1024,A=16,T:340M

- *Task 1: Masked Language Modeling (MLM)*

- 80% of the time: Replace the word with the `[MASK]` token, e.g., `my dog is hairy` → `my dog is [MASK]`
- 10% of the time: Replace the word with a random word, e.g., `my dog is hairy` → `my dog is apple`
- 10% of the time: Keep the word unchanged, e.g., `my dog is hairy` → `my dog is hairy`. The purpose of this is to bias the representation towards the actual observed word.

- *Task 2: Next Sentence Prediction (NSP)*

Input = `[CLS] the man [MASK] to the store [SEP] penguin [MASK] are flight ##less birds [SEP]`
Label = `NotNext`

Input = `[CLS] the man went to [MASK] store [SEP] he bought a gallon [MASK] milk [SEP]`
Label = `IsNext`

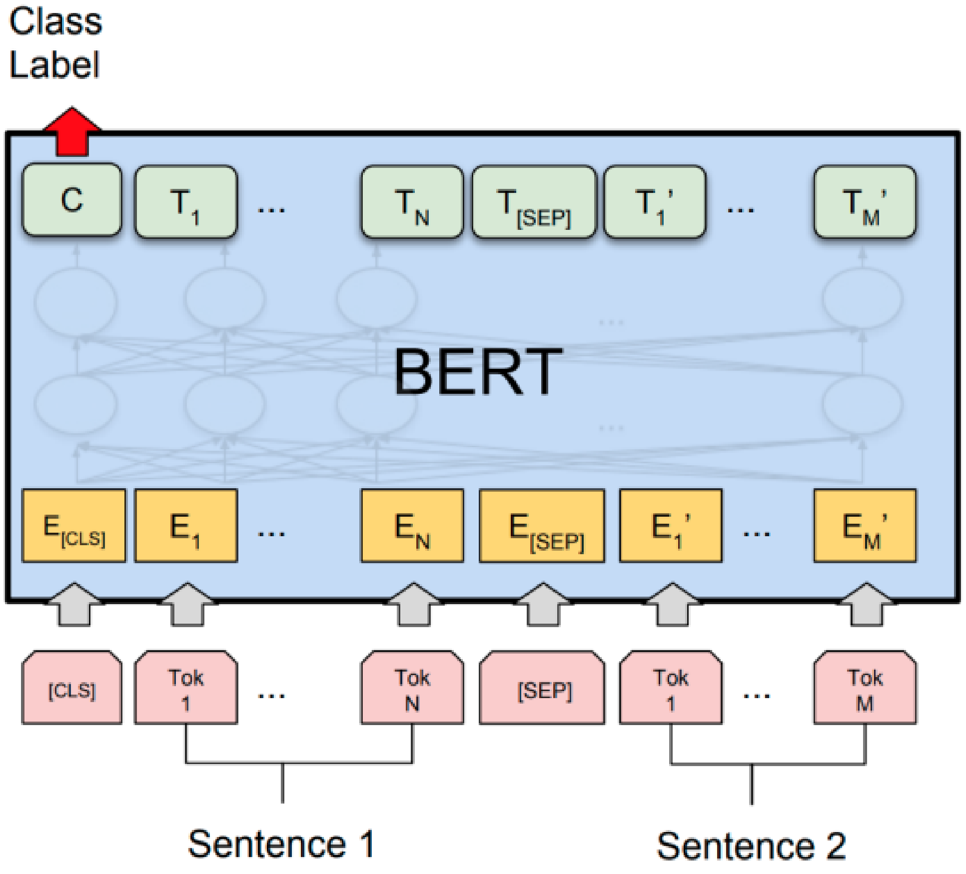
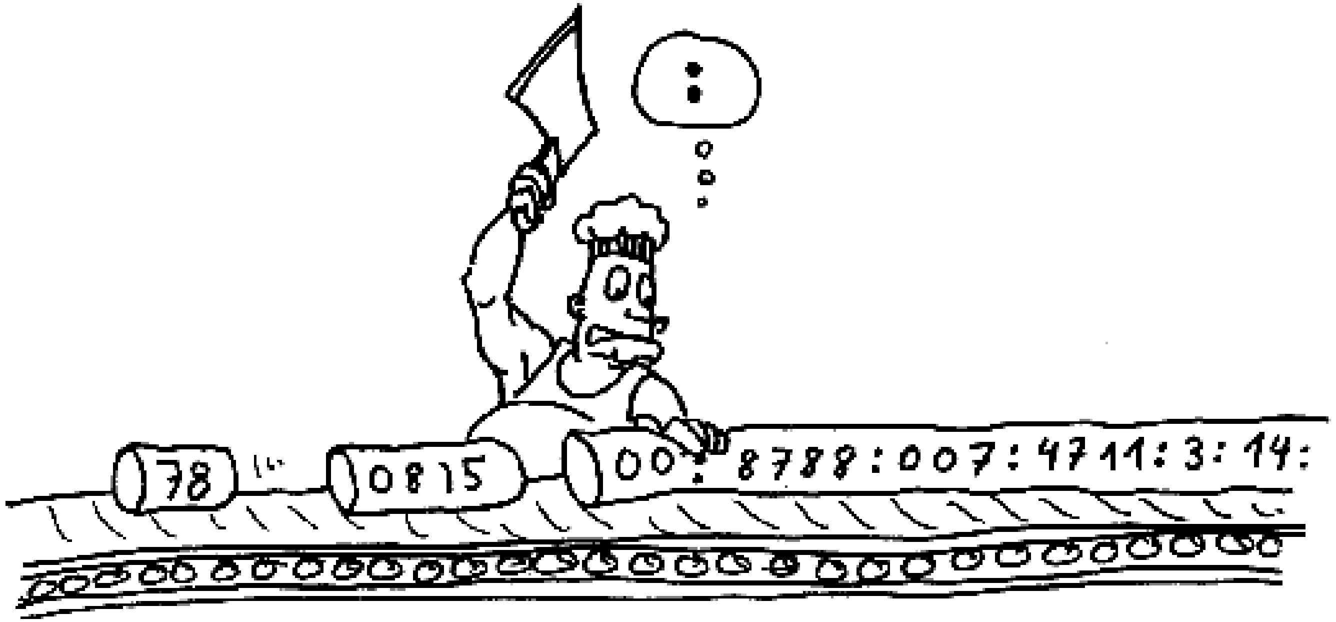


Figure 1: BERT fine-tuning (Devlin et al., 2019).

Ref:
1- A Primer in BERTology: What we know about how BERT works, A. Rogers, et. al., 2020
2- BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, J. Dave et al. 2019

UNIVERSITY OF WATERLOO

BERT TOKENIZER



Ref:
1- A Primer in BERTology: What we know about how BERT works, A. Rogers, et. al., 2020
2- BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, J. Dave et al. 2019
3- https://medium.com/@makcedward/how-subword-helps-on-your-nlp-model-83dd1b836f46
4- https://dvrtechnopark.wordpress.com/2014/07/21/string-tokenizer-in-java/

BPE (INFORMATION THEORY, 1994)

Tokenizers in NLP

Problem: NMT models typically operate with a fixed vocabulary, but translation is an open-vocabulary problem.
Intuition: Various word classes are translatable via smaller units than words
Solution: Encoding rare and unknown word classes as sequences of **subword units**

Named entities:
Barack Obama (English; German)
Барак Обама (Russian)
バラク・オバマ (ba-ra-ku o-ba-ma) (Japanese)

Cognates and loanwords:
claustrophobia (English)
Klaustrophobie (German)
Клаустрофобия (Klaustrofobiâ) (Russian)

Morphologically complex words:
solar system (English)
Sonnensystem (Sonne + System) (German)
Naprendszer (Nap + Rendszer) (Hungarian)

Hypothesis: A segmentation of rare words into appropriate subword units is sufficient to allow for the neural translation network to learn transparent translations.

Ref:
1- Neural Machine Translation of Rare Words with Subword Units, Sennrich, et al., 2015

BPE (INFORMATION THEORY, 1994)

Tokenizer in NLP

Problem: NMT models typically operate with a fixed vocabulary, but translation is an open-vocabulary problem.
Intuition: Various word classes are translatable via smaller units than words
Solution: Encoding rare and unknown word classes as sequences of subword units

BPE Data Compression:

Aaabdaaabac

Ref:
1- Neural Machine Translation of Rare Words with Subword Units, Sennrich, et al., 2015
2- A New Algorithm for Data Compression, Philip Gage, 1994
3- https://en.wikipedia.org/wiki/Byte_pair_encoding
4- https://leimao.github.io/blog/Byte-Pair-Encoding/

BPE (INFORMATION THEORY, 1994)

Tokenizer in NLP

Problem: NMT models typically operate with a fixed vocabulary, but translation is an open-vocabulary problem.
Intuition: Various word classes are translatable via smaller units than words
Solution: Encoding rare and unknown word classes as sequences of subword units

BPE Data Compression:

Aaabdaaabac

ZabdZabac
Z=aa

Ref:
1- Neural Machine Translation of Rare Words with Subword Units, Sennrich, et al., 2015
2- A New Algorithm for Data Compression, Philip Gage, 1994
3- https://en.wikipedia.org/wiki/Byte_pair_encoding
4- https://leimao.github.io/blog/Byte-Pair-Encoding/

BPE (INFORMATION THEORY, 1994)

Tokenizer in NLP

Problem: NMT models typically operate with a fixed vocabulary, but translation is an open-vocabulary problem.
Intuition: Various word classes are translatable via smaller units than words
Solution: Encoding rare and unknown word classes as sequences of subword units

BPE Data Compression:

Aaabdaaabac

ZabdZabac
Z=aa

ZYdZYac
Y=ab
Z=aa

Ref:
1- Neural Machine Translation of Rare Words with Subword Units, Sennrich, et al., 2015
2- A New Algorithm for Data Compression, Philip Gage, 1994
3- https://en.wikipedia.org/wiki/Byte_pair_encoding
4- https://leimao.github.io/blog/Byte-Pair-Encoding/

BPE (INFORMATION THEORY, 1994)

Tokenizer in NLP

Problem: NMT models typically operate with a fixed vocabulary, but translation is an open-vocabulary problem.
Intuition: Various word classes are translatable via smaller units than words
Solution: Encoding rare and unknown word classes as sequences of subword units

BPE Data Compression:

Aaabdaaabac

ZabdZabac
Z=aa

ZYdZYac
Y=ab
Z=aa

BPE Segmentation Init:

Vocabulary:
{l, o, w, e, r, n, s, t, i, d}

Dictionary:
{5: l o w
2: l o w e r
6: n e w e s t
3: w i d e s t}

Ref:
1- Neural Machine Translation of Rare Words with Subword Units, Sennrich, et al., 2015
2- A New Algorithm for Data Compression, Philip Gage, 1994
3- https://en.wikipedia.org/wiki/Byte_pair_encoding
4- https://leimao.github.io/blog/Byte-Pair-Encoding/

BPE (INFORMATION THEORY, 1994)

Tokenizer in NLP

Problem: NMT models typically operate with a fixed vocabulary, but translation is an open-vocabulary problem.
Intuition: Various word classes are translatable via smaller units than words
Solution: Encoding rare and unknown word classes as sequences of subword units

BPE Data Compression:	BPE Segmentation Init:	Iter2:
Aabdaaabac ---------------- ZabdZabac Z=aa ---------------- ZYdZYac Y=ab Z=aa	Vocabulary: {l, o, w, e, r, n, s, t, i, d} Dictionary: {5: l o w 2: l o w e r 6: n e w e s t 3: w i d e s t}	Vocabulary: {l, o, w, e, r, n, s, t, i, d, es} Dictionary: {5: l o w 2: l o w e r 6: n e w es t 3: w i d es t}

Ref:
1- Neural Machine Translation of Rare Words with Subword Units, Sennrich, et al., 2015
2- A New Algorithm for Data Compression, Philip Gage, 1994
3- https://en.wikipedia.org/wiki/Byte_pair_encoding
4- https://leimao.github.io/blog/Byte-Pair-Encoding/

BPE (INFORMATION THEORY, 1994)

Tokenizer in NLP

Problem: NMT models typically operate with a fixed vocabulary, but translation is an open-vocabulary problem.
Intuition: Various word classes are translatable via smaller units than words
Solution: Encoding rare and unknown word classes as sequences of subword units

BPE Data Compression:	BPE Segmentation Init:	Iter2:	Iter3:
Aaabdaaabac ---------------- ZabdZabac Z=aa ---------------- ZYdZYac Y=ab Z=aa	Vocabulary: {l, o, w, e, r, n, s, t, i, d} Dictionary: {5: l o w 2: l o w e r 6: n e w e s t 3: w i d e s t}	Vocabulary: {l, o, w, e, r, n, s, t, i, d, es} Dictionary: {5: l o w 2: l o w e r 6: n e w es t 3: w i d es t}	Vocabulary: {l, o, w, e, r, n, s, t, i, d, es, est} Dictionary: {5: l o w 2: l o w e r 6: n e w est 3: w i d est}

Ref:
1- Neural Machine Translation of Rare Words with Subword Units, Sennrich, et al., 2015
2- A New Algorithm for Data Compression, Philip Gage, 1994
3- https://en.wikipedia.org/wiki/Byte_pair_encoding
4- https://leimao.github.io/blog/Byte-Pair-Encoding/

BPE (INFORMATION THEORY, 1994)

Tokenizer in NLP

Problem: NMT models typically operate with a fixed vocabulary, but translation is an open-vocabulary problem.
Intuition: Various word classes are translatable via smaller units than words
Solution: Encoding rare and unknown word classes as sequences of subword units

BPE Data Compression:

Aaabdaaabac

ZabdZabac
Z=aa

ZYdZYac
Y=ab
Z=aa

BPE Segmentation Init:

Vocabulary:
{l, o, w, e, r, n, s, t, i, d}

Dictionary:
{5: l o w
2: l o w e r
6: n e w e s t
3: w i d e s t}

Iter3:

Vocabulary:
{l, o, w, e, r, n, s, t, i, d, es, est}

Dictionary:
{5: l o w
2: l o w e r
6: n e w est
3: w i d est}

Iter4:

Vocabulary:
{l, o, w, e, r, n, s, t, i, d, es, est,lo}

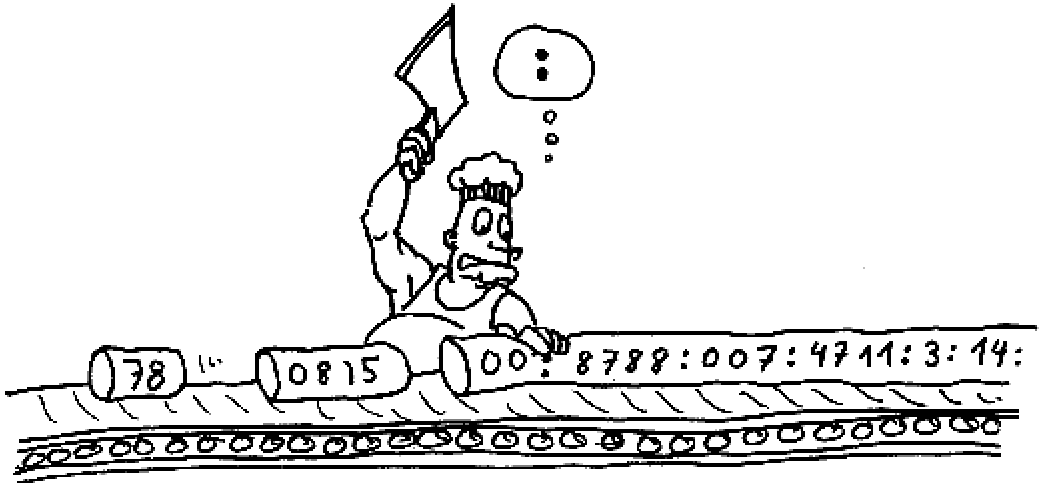
Dictionary:
{5: lo w
2: lo w e r
6: n e w est
3: w i d est}

Ref:
1- Neural Machine Translation of Rare Words with Subword Units, Sennrich, et al., 2015
2- A New Algorithm for Data Compression, Philip Gage, 1994
3- https://en.wikipedia.org/wiki/Byte_pair_encoding
4- https://leimao.github.io/blog/Byte-Pair-Encoding/

BERT TOKENIZER

Wordpiece Algorithm: Wordpiece model uses a likelihood instead of frequency.

1. Prepare a large enough training data (i.e. corpus)
2. Define a desired subword vocabulary size
3. Split word to sequence of characters
4. Build a languages model based on step 3 data
5. Choose the new word unit out of all the possible ones that increases the likelihood on the training data the most when added to the model.
6. Repeating step 5 until reaching subword vocabulary size which is defined in step 2 or the likelihood increase falls below a certain threshold.

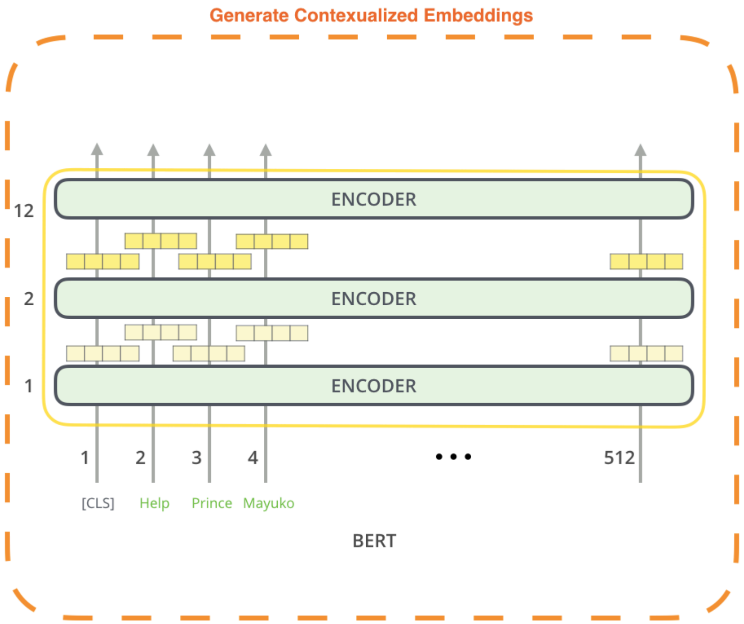


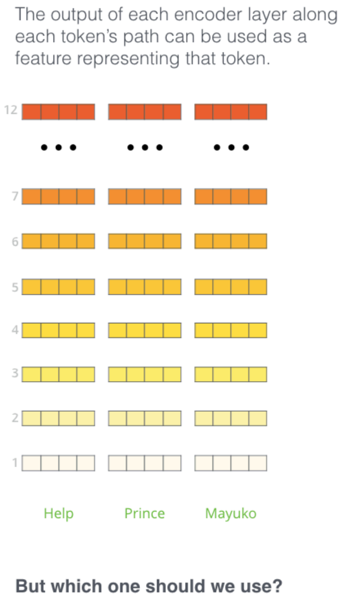
Ref:
1- A Primer in BERTology: What we know about how BERT works, A. Rogers, et. al., 2020
2- BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, J. Dave et al. 2019
3- https://medium.com/@makcedward/how-subword-helps-on-your-nlp-model-83dd1b836f46
4- https://dvrtechnopark.wordpress.com/2014/07/21/string-tokenizer-in-java/
5- Mike Schuster and Kaisuke Nakajima. 2012. Japanese and korean voice search. In Proc. of ICASSP

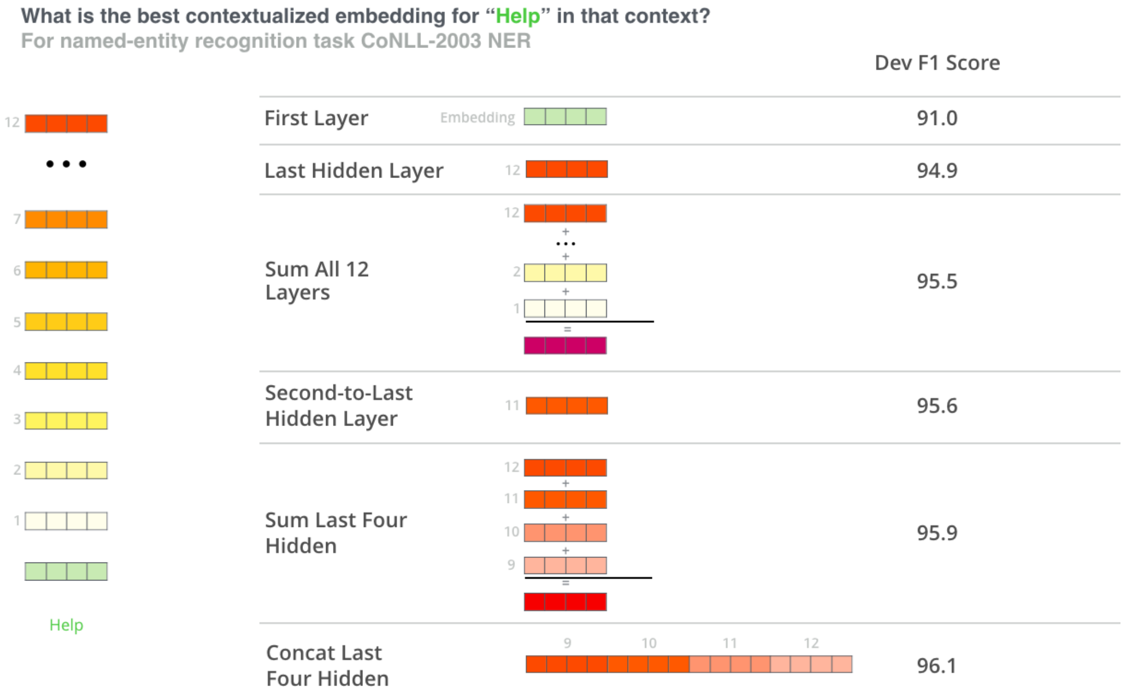
LANGUAGE KNOWLEDGE

What is inside BERT?

BERT EMBEDDING







KEY INSIGHTS:

- ***BERT Representation are contextualized,*** *form distinct and clear word senses clusters (Wiedemann et al. 2019)*
- ***Representation of the same words varies*** *depending on position of the sentence likely due to* ***NSP objective****. (Mickus et al., 2019)*
- ***Later BERT layers produce more context specific representations*** *(Ethayarajh et al. 2019)*

Ref:
1- A Primer in BERTology: What we know about how BERT works, A. Rogers, et. al., 2020
2- http://jalammar.github.io/illustrated-bert/

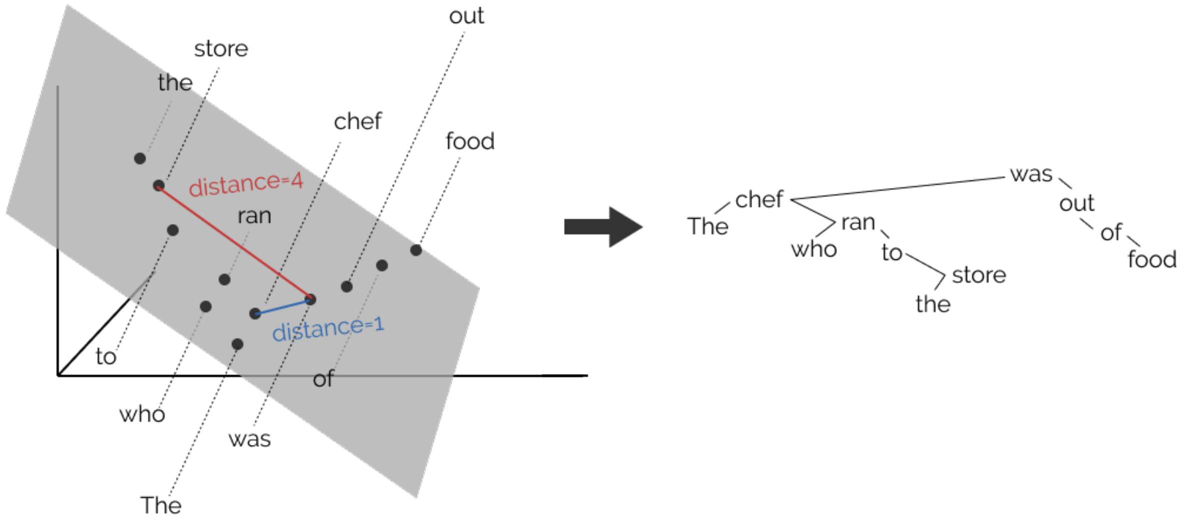
BERT KNOWLEDGE

KEY INSIGHTS:

- ***BERT Representation are hierarchical rather than linear*** *(Lin et. al., 2019)*
- *BERT Embeddings encode information about parts of speech, syntactic chunks and roles. (Tenney et. al. 2019 and Liu et al. 2019)*

- ***BERT knowledge of syntax is partial*** *(since probing not works for long distant parent nodes). (Liu et al. 2019)*

- ***Syntactic structure is not directly encoded in self-attention weights, but they can be transformed to reflect it.***



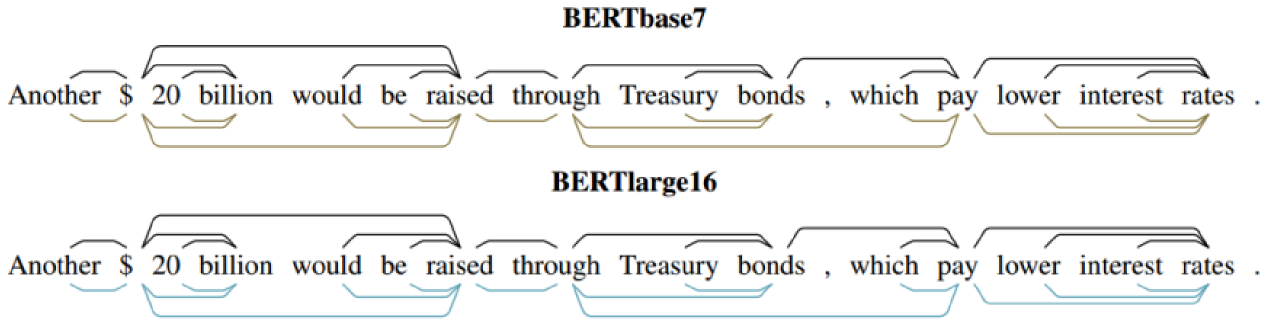


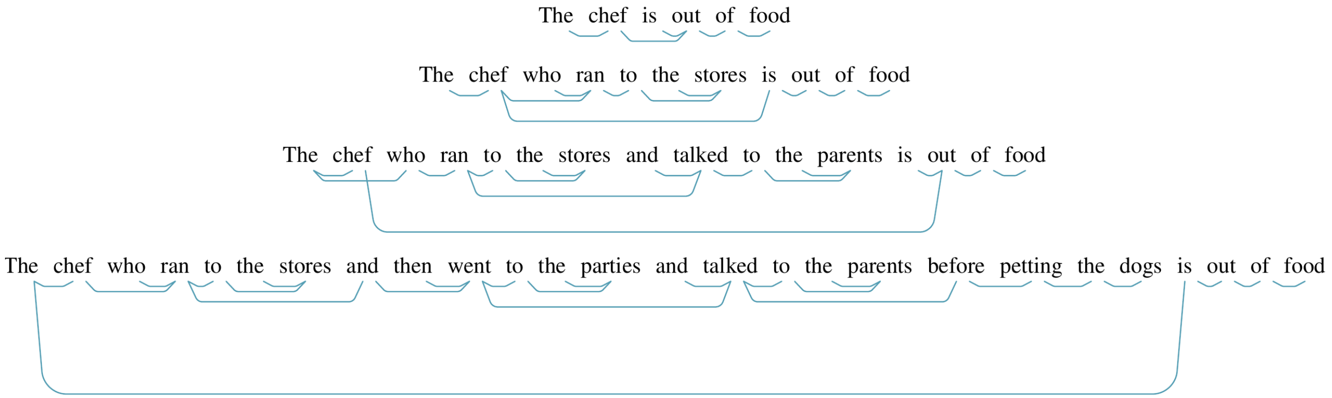
Figure 2: Parse trees recovered from BERT representations by Hewitt et al. (2019)

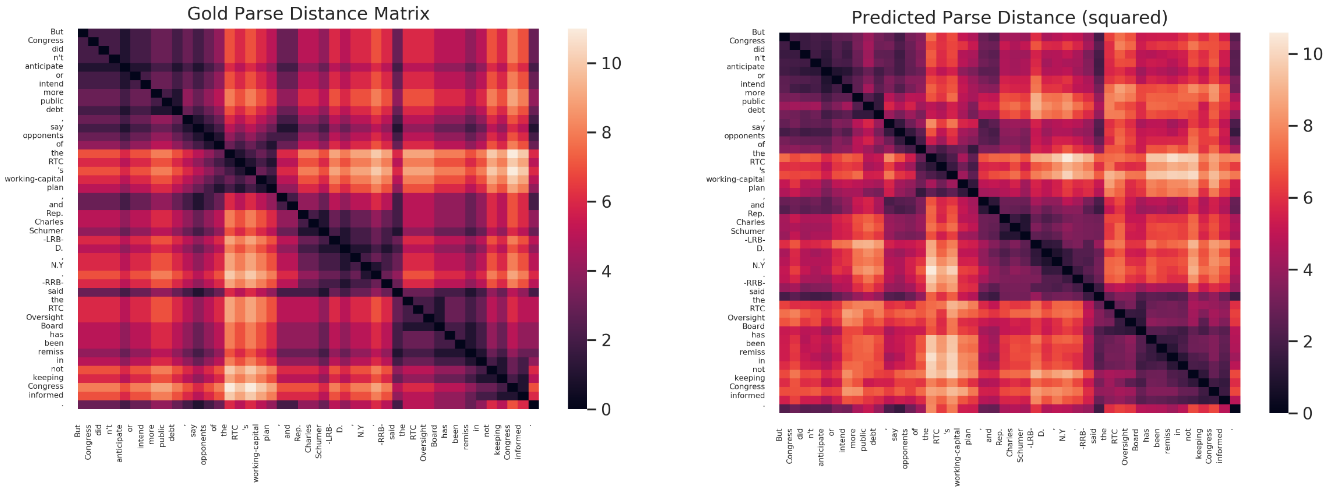
Ref:
1- A Primer in BERTology: What we know about how BERT works, A. Rogers, et. al.
2- https://www.youtube.com/watch?v=3VZZbKoXDVM
3- https://github.com/john-hewitt/structural-probes

BERT KNOWLEDGE

KEY INSIGHTS:

- ***BERT Representation are hierarchical rather than linear*** *(Lin et. al., 2019)*
- *BERT Embeddings encode information about parts of speech, syntactic chunks and roles. (Tenney et. al. 2019 and Liu et al. 2019)*
- ***BERT knowledge of syntax is partial*** *(since probing not works for long distant parent nodes). (Liu et al. 2019)*
- ***Syntactic structure is not directly encoded in self-attention weights, but they can be transformed to reflect it.***



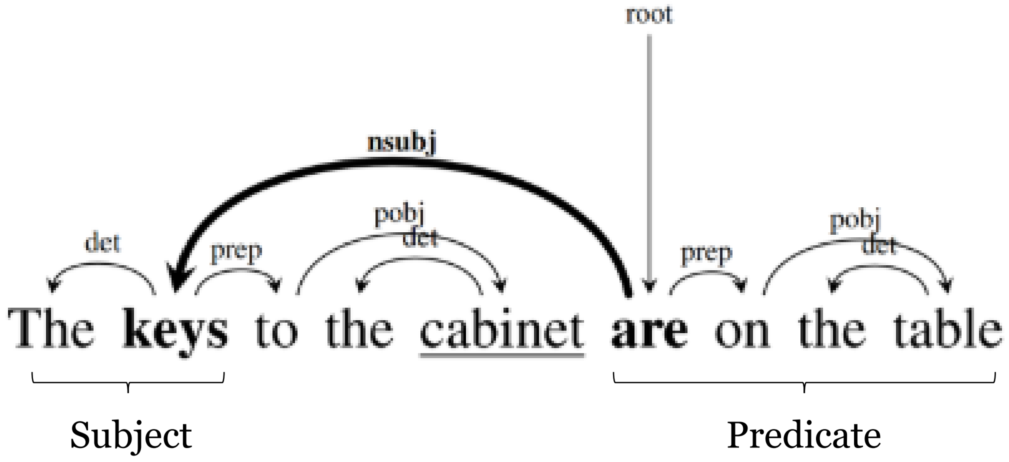


Ref:
1- A Primer in BERTology: What we know about how BERT works, A. Rogers, et. al., 2020
2- https://www.youtube.com/watch?v=3VZZbKoXDVM
3- https://github.com/john-hewitt/structural-probes

BERT KNOWLEDGE

KEY INSIGHTS:

- ***BERT takes subject-predicate agreement into account when performing the cloze task.*** *(Goldberg 2019)*
- ***BERT doesn't understand negation and is insensitive to malformed input.*** *No predictions change even with shuffled word order, truncated sentences, removed subjects and objects. (Ettinger, 2019)*
- ***BERT encodes information about entity types, relations, semantic roles, and proto-roles.*** *(Tenney et al. 2019)*
- ***BERT struggles with representation of numbers*** *maybe because of wordpiece tokenization. (Wallace et al. 2019)*

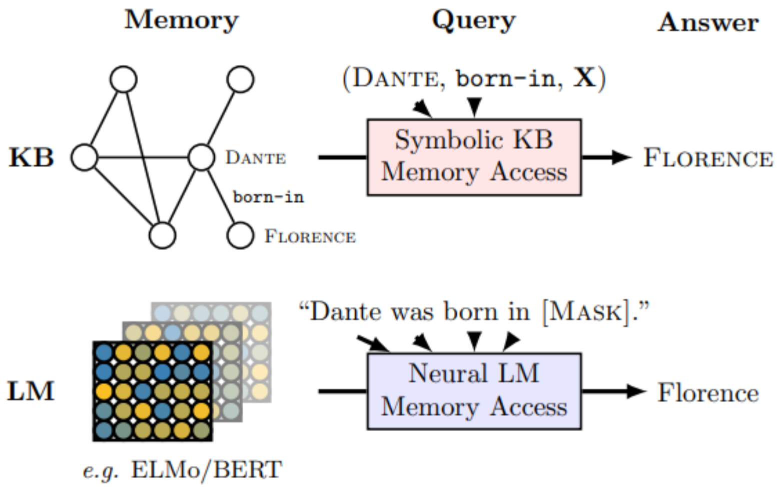


POS	The important thing about Disney is that it is a global [brand]$_1$. → NN (Noun)
Constit.	The important thing about Disney is that it [is a global brand]$_1$. → VP (Verb Phrase)
Depend.	[Atmosphere]$_1$ is always [fun]$_2$ → nsubj (nominal subject)
Entities	The important thing about [Disney]$_1$ is that it is a global brand. → Organization
SRL	[The important thing about Disney]$_2$ [is]$_1$ that it is a global brand. → Arg1 (Agent)
SPR	[It]$_1$ [endorsed]$_2$ the White House strategy... → {awareness, existed_after, ...}
Coref.^{O}	The important thing about [Disney]$_1$ is that [it]$_2$ is a global brand. → True
Coref.^{W}	[Characters]$_2$ entertain audiences because [they]$_1$ want people to be happy. → True Characters entertain [audiences]$_2$ because [they]$_1$ want people to be happy. → False
Rel.	The [burst]$_1$ has been caused by water hammer [pressure]$_2$. → Cause-Effect(e_2, e_1)

Ref:
1- A Primer in BERTology: What we know about how BERT works, A. Rogers, et. al., 2020
2- https://www.youtube.com/watch?v=3VZZbKoXDVM
3- https://nlp.stanford.edu//~johnhew//structural-probe.html
4- What do you learn from context? Probing for sentence structure in contextualized word representations, Ian Tenney et al., ICLR 2019

UNIVERSITY OF WATERLOO

BERT KNOWLEDGE



Birds have ___.	feathers	wings [-1.8], nests [-3.1], **feathers [-3.2]**, died [-3.7], eggs [-3.9]
Typing requires ___.	speed	patience [-3.5], precision [-3.6], registration [-3.8], accuracy [-4.0], **speed [-4.1]**
Time is ___.	finite	short [-1.7], passing [-1.8], precious [-2.9], irrelevant [-3.2], gone [-4.0]
You would celebrate because you are ___.	alive	happy [-2.4], human [-3.3], **alive [-3.3]**, young [-3.6], free [-3.9]
Skills can be ___.	taught	acquired [-2.5], useful [-2.5], learned [-2.8], combined [-3.9], varied [-3.9]
A pond is for ___.	fish	swimming [-1.3], fishing [-1.4], bathing [-2.0], **fish [-2.8]**, recreation [-3.1]
Francesco Bartolomeo Conti was born in ___.	Florence	Rome [-1.8], **Florence [-1.8]**, Naples [-1.9], Milan [-2.4], Bologna [-2.5]
Adolphe Adam died in ___.	Paris	**Paris [-0.5]**, London [-3.5], Vienna [-3.6], Berlin [-3.8], Brussels [-4.0]
English bulldog is a subclass of ___.	dog	dogs [-0.3], breeds [-2.2], **dog [-2.4]**, cattle [-4.3], sheep [-4.5]
The official language of Mauritius is ___.	English	**English [-0.6]**, French [-0.9], Arabic [-6.2], Tamil [-6.7], Malayalam [-7.0]
Patrick Oboya plays in ___ position.	midfielder	centre [-2.0], center [-2.2], **midfielder [-2.4]**, forward [-2.4], midfield [-2.7]
Hamburg Airport is named after ___.	Hamburg	Hess [-7.0], Hermann [-7.1], Schmidt [-7.1], **Hamburg [-7.5]**, Ludwig [-7.5]
The original language of Mon oncle Benjamin is ___.	French	**French [-0.2]**, Breton [-3.3], English [-3.8], Dutch [-4.2], German [-4.9]
Dani Alves plays with ___.	Barcelona	Santos [-2.4], Porto [-2.5], Sporting [-3.1], Brazil [-3.3], Portugal [-3.7]
Paul Toungui is a ___ by profession .	politician	lawyer [-1.1], journalist [-2.4], teacher [-2.7], doctor [-3.0], physician [-3.7]
Sodium sulfide consists of ___.	sodium	water [-1.2], sulfur [-1.7], **sodium [-2.5]**, zinc [-2.8], salt [-2.9]
Gordon Scholes is a member of the ___ political party.	Labor	Labour [-1.3], Conservative [-1.6], Green [-2.4], Liberal [-2.9], **Labor [-2.9]**
Kenya maintains diplomatic relations with ___.	Uganda	India [-3.0], **Uganda [-3.2]**, Tanzania [-3.5], China [-3.6], Pakistan [-3.6]
iPod Touch is produced by ___.	Apple	**Apple [-1.6]**, Nokia [-1.7], Sony [-2.0], Samsung [-2.6], Intel [-3.1]

KEY INSIGHTS:

- *For some relation types*, vanilla ***BERT is competitive with methods relying on knowledge bases.*** *(Petroni et al. 2019)*
- ***BERT cannot reason based on its word knowledge.*** *It knows that people can walk into houses, and that houses are big, but it cannot infer that houses are bigger than people. (Forbes et al. 2019)*

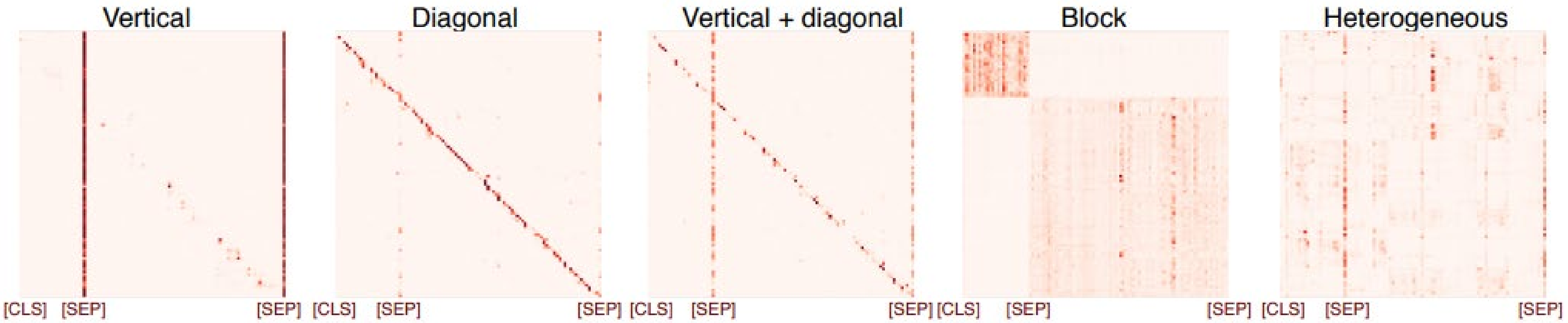
Ref:
1- A Primer in BERTology: What we know about how BERT works, A. Rogers, et. al., 2020
2- https://www.youtube.com/watch?v=3VZZbKoXDVM
3- https://nlp.stanford.edu//~johnhew//structural-probe.html

LOCALIZING BERT LINGUISTIC KNOWLEDGE

<u>Attention weight</u>: How much a particular word will be weighted when computing the next representation for the current word. (Clark et al. 2019)

KEY INSIGHTS:

- **Most self-attention heads do not directly encode any non-trivial linguistic information** since less than half of them had the heterogeneous pattern maybe due to <u>overparametrization</u>. (Kovaleva et al. 2019)



Ref:
1- A Primer in BERTology: What we know about how BERT works, A. Rogers, et. al., 2020
2- https://www.youtube.com/watch?v=3VZZbKoXDVM
3- https://nlp.stanford.edu//~johnhew//structural-probe.html

LOCALIZING BERT LINGUISTIC KNOWLEDGE

Attention weight: How much a particular word will be weighted when computing the next representation for the current word. (Clark et al. 2019)

KEY INSIGHTS:

- **Most self-attention heads do not directly encode any non-trivial linguistic information** since less than half of them had the heterogeneous pattern maybe due to overparameterization. (Kovaleva et al. 2019)
- **Some BERT attention heads seems to specialize in certain types of syntactic relations.** (Htut et al. 2019, Clark et al. 2019) Clark et al. identify a BERT head that can be directly used as a classifier to perform coreference resolution on par with a rule-based system.
- **No single head has the complete syntactic tree information.** (Htut et al. 2019, Clark et al. 2019)
- **Even when attention heads specialize in tracking semantic relations, they do not necessarily contribute to BERT's performance on relevant tasks.** Authors identified two heads of base BERT, in which self-attention maps were closely aligned with annotations of core frame semantic relations. Although such relations should have been instrumental to tasks such as inference, a head ablation study showed that these heads were not essential for BERT's success on GLUE tasks. (Kovaleva et al. 2019, Baker et al. 1998)

Ref:
1- A Primer in BERTology: What we know about how BERT works, A. Rogers, et. al., 2020
2- https://www.youtube.com/watch?v=3VZZbKoXDVM
3- https://nlp.stanford.edu//~johnhew//structural-probe.html

BERT LAYERS

KEY INSIGHTS:

- ***Lower layers have the most linear word order information.*** *Lin et al. (2019) report a decrease in the knowledge of linear word order around layer 4 in BERT-base accompanied by increased knowledge of hierarchical sentence structure.*
- **Syntactic information is the most prominent in the middle BERT layers.** The middle layers of Transformers are overall the best-performing and the most transferable across tasks. (Liu et al. 2019)
- **The final layers of BERT are the most task-specific.**

Each column represents a probing task, and each row represents a contextualizer layer.

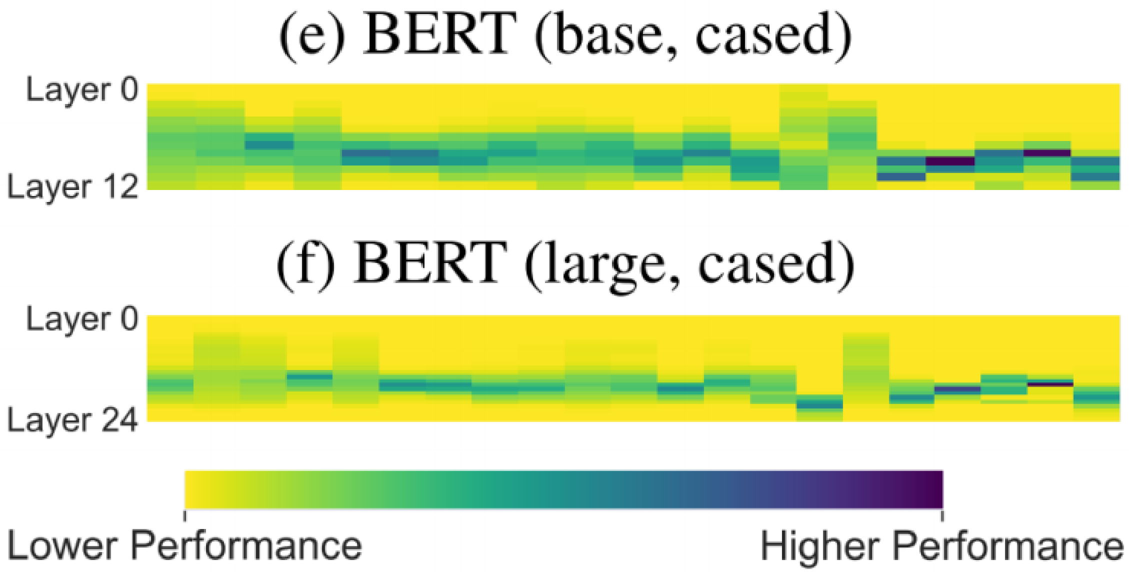


Figure 5: BERT layer transferability (columns correspond to probing tasks) (Liu et al., 2019a).

Ref:
1- A Primer in BERTology: What we know about how BERT works, A. Rogers, et. al., 2020
2- https://www.youtube.com/watch?v=3VZZbKoXDVM
3- https://github.com/john-hewitt/structural-probes

UNIVERSITY OF WATERLOO

BERT PRE-TRAINING

KEY INSIGHTS:

- **Removing NSP does not hurt or slightly improves task performance.** (Liu et al. 2019)
- **Dynamic masking slightly improves on BERT's MLM.** (Liu et al. 2019)
- It is possible to integrate external knowledge into BERT like entity embedding as input as in E-BERT (Poerner et al. 2019) and ERNIE (Zhang et al. 2019)
- SemBERT (Zhang et al. 2020) integrates semantic role information with BERT representations.

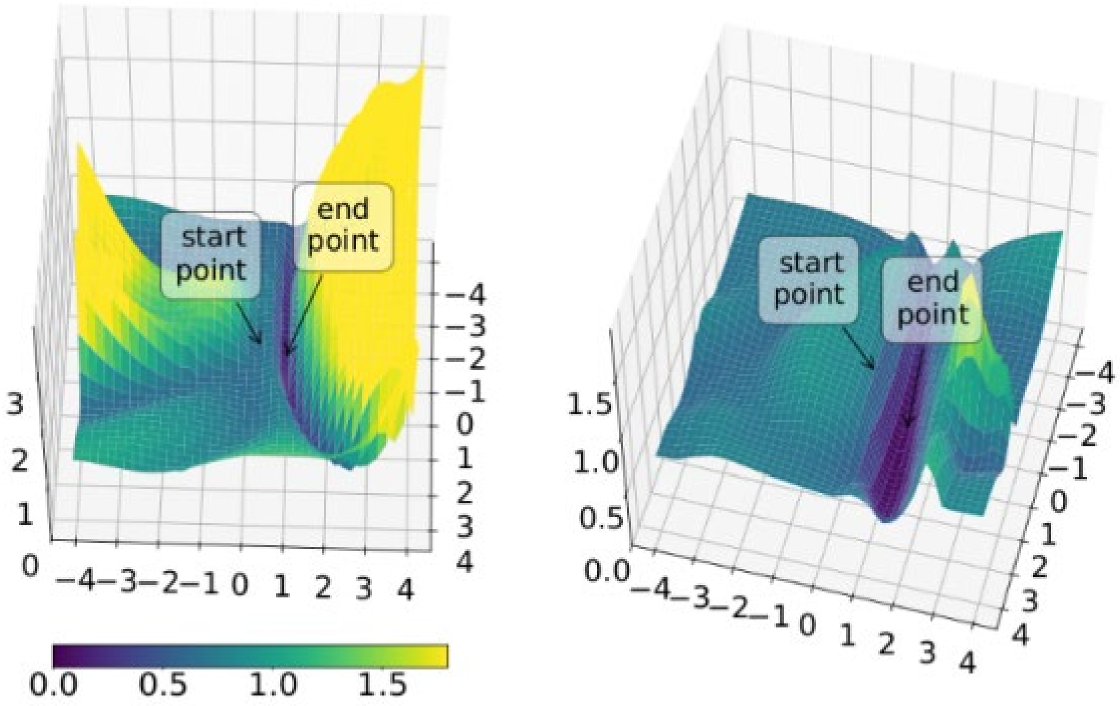


Figure 6: Pre-trained weights help BERT find wider optima in fine-tuning on MRPC (right) than training from scratch (left) (Hao et al., 2019)

Ref:
1- A Primer in BERTology: What we know about how BERT works, A. Rogers, et. al., 2020
2- https://www.youtube.com/watch?v=3VZZbKoXDVM
3- https://github.com/john-hewitt/structural-probes

BERT ARCHITECTURE

KEY INSIGHTS:

- **The number of heads was not as significant as the number of layers**. (Wang et al. 2019)
- **Larger hidden representation size was consistently better**, but the gains varied by setting.
- **Large-batch training (8k/32k) improves** both the language model perplexity and downstream task performance. (Liu et al. 2019, Yo et al. 2019)
- **Normalizing (centered around zero) embedding vector of [CLS] stabilizes the training** leading to a performance gain on text classification task. (Zhou et al. 2019)

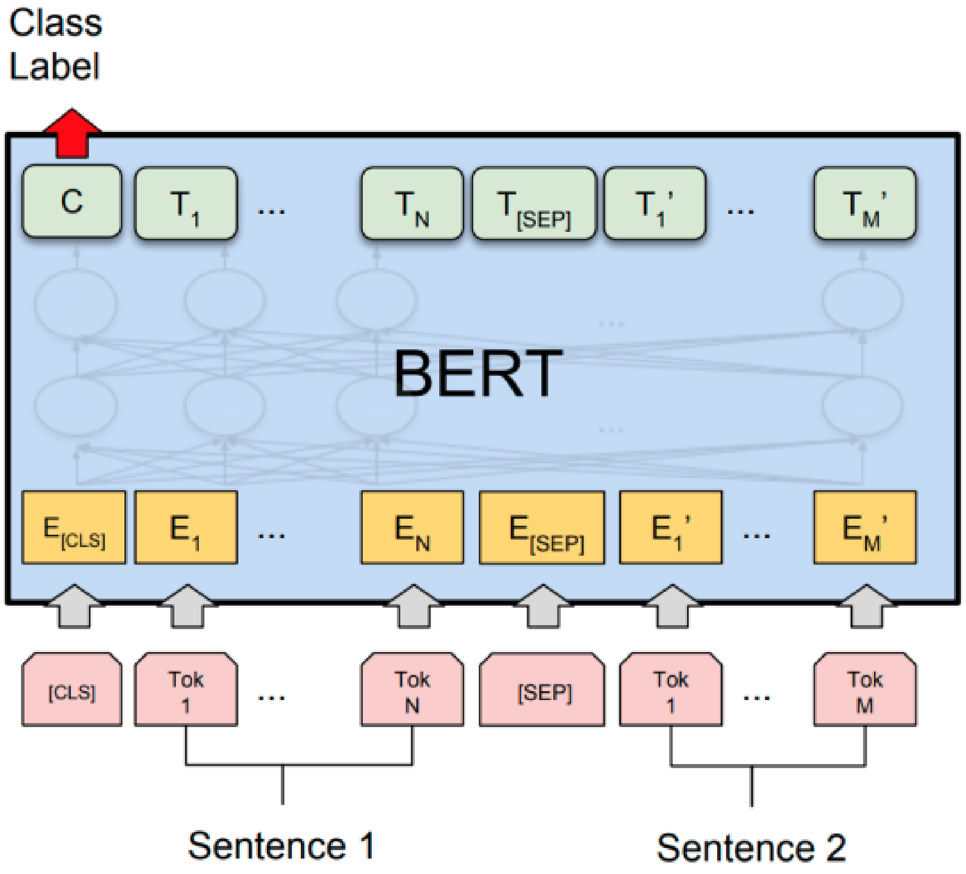


Figure 1: BERT fine-tuning (Devlin et al., 2019).

Ref:
1- A Primer in BERTology: What we know about how BERT works, A. Rogers, et. al., 2020
2- https://www.youtube.com/watch?v=3VZZbKoXDVM
3- https://github.com/john-hewitt/structural-probes

OVERPARAMETRIZATION

Why knowledge distillation and quantization can efficiently compressed BERT

KEY INSIGHTS:

- **All but a few Transformer heads could be pruned without significant losses in performance**. (Voita et al. 2019)
- **Most heads in the same layer of BERT show similar self-attention patterns** perhaps related to the fact that the output of all self-attention heads in a layer passed through the same MLP. (Clark et al. 2019)
- Depending on the task, **some BERT heads/layers are not only useless, but also harmful** to the downstream task performance. (Michel et al. 2019)
- Clark et al. 2019 suggests that one of the possible reason why BERT ends with redundant heads and layers, is the use of attention dropouts.

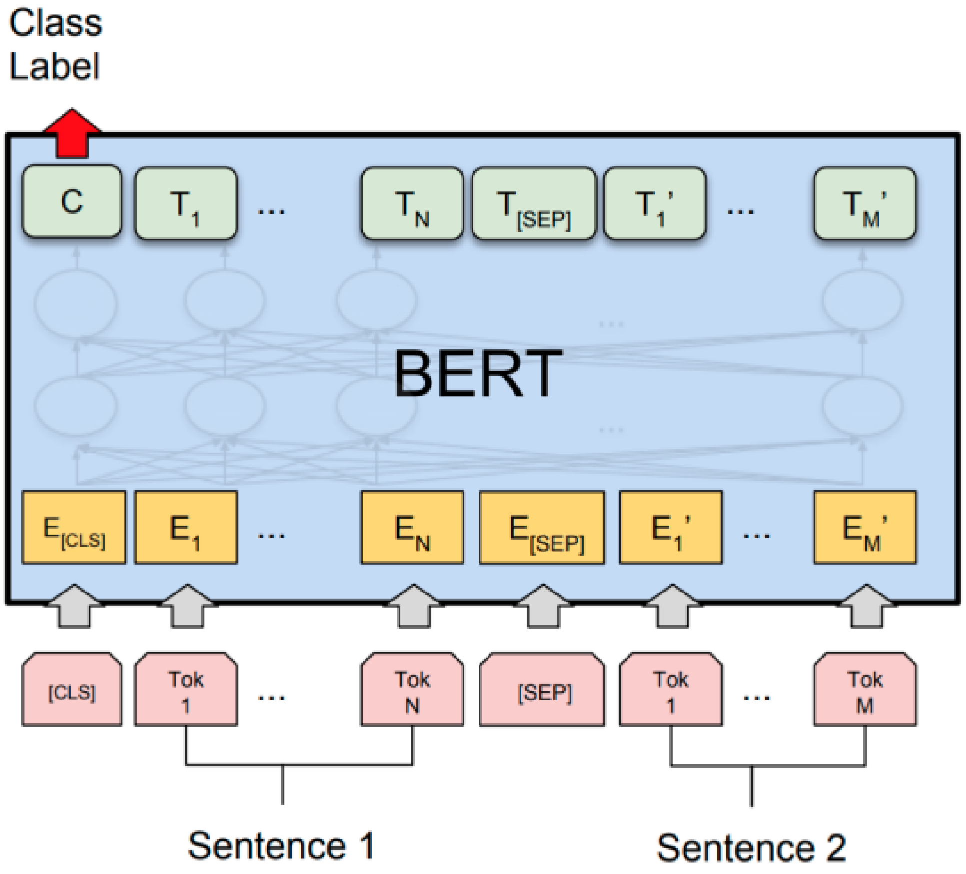


Figure 1: BERT fine-tuning (Devlin et al., 2019).

Ref:
1- A Primer in BERTology: What we know about how BERT works, A. Rogers, et. al., 2020
2- https://www.youtube.com/watch?v=3VZZbKoXDVM
3- https://github.com/john-hewitt/structural-probes

BERT COMPRESSION

		Compression	Performance	Speedup	Model	Evaluation
Distillation	DistilBERT (Sanh et al., 2019)	×2.5	90%	×1.6	$BERT_6$	All GLUE tasks
	$BERT_6$-PKD (Sun et al., 2019a)	×1.6	97%	×1.9	$BERT_6$	No WNLI, CoLA and STS-B
	$BERT_3$-PKD (Sun et al., 2019a)	×2.4	92%	×3.7	$BERT_3$	No WNLI, CoLA and STS-B
	(Aguilar et al., 2019)	×2	94%	-	$BERT_6$	CoLA, MRPC, QQP, RTE
	BERT-48 (Zhao et al., 2019)	×62	87%	×77	$BERT_{12}{}^{*\dagger}$	MNLI, MRPC, SST-2
	BERT-192 (Zhao et al., 2019)	×5.7	94%	×22	$BERT_{12}{}^{*\dagger}$	MNLI, MRPC, SST-2
	TinyBERT (Jiao et al., 2019)	×7.5	96%	×9.4	$BERT_4{}^{*\dagger}$	All GLUE tasks
	MobileBERT (Sun et al.)	×4.3	100%	×4	$BERT_{24}{}^{\dagger}$	No WNLI
	PD (Turc et al., 2019)	×1.6	98%	$\times 2.5^3$	$BERT_6{}^{\dagger}$	No WNLI, CoLA and STS-B
	MiniBERT(Tsai et al., 2019)	$\times 6^{\S}$	98%	$\times 27^{\S}$	$mBERT_3{}^{\dagger}$	CoNLL-2018 POS and morphology
	BiLSTM soft (Tang et al., 2019)	×110	91%	$\times 434^{\ddagger}$	$BiLSTM_1$	MNLI, QQP, SST-2
Quant.	Q-BERT (Shen et al., 2019)	×13	99%	-	$BERT_{12}$	MNLI, SST-2
	Q8BERT (Zafrir et al., 2019)	×4	99%	-	$BERT_{12}$	All GLUE tasks
Other	ALBERT-base (Lan et al., 2019)	×9	97%	×5.6	$BERT_{12}{}^{**}$	MNLI, SST-2
	ALBERT-xxlarge (Lan et al., 2019)	×0.47	107%	×0.3	$BERT_{12}{}^{**}$	MNLI, SST-2
	BERT-of-Theseus (Xu et al., 2020)	×1.6	98%	-	$BERT_6$	No WNLI

Table 1: Comparison of BERT compression studies. Compression, performance retention, and inference time speedup figures are given with respect to $BERT_{base}$, unless indicated otherwise. Performance retention is measured as a ratio of average scores achieved by a given model and by $BERT_{base}$. The subscript in the model description reflects the number of layers used. *Smaller vocabulary used. †The dimensionality of the hidden layers is reduced. **The dimensionality of the embedding layer is reduced. ‡Compared to $BERT_{large}$. §Compared to mBERT.

Ref:
1- A Primer in BERTology: What we know about how BERT works, A. Rogers, et. al., 202

MULTILINGUAL BERT

Multilingual BERT (mBERT) was trained on Wikipedia in 104 languages.

KEY INSIGHTS:

- **Except in language generation** (Ronnqvist et al. 2019), **mBERT performs surprisingly well in zero-shot transfer on many tasks.** (Wu and Dredze, 2019; Pires et al. 2019)
- **The model seems to naturally learn high-quality cross-lingual word alignments**. (Libovicky et al. 2019)
- mBERT is simply trained on a multilingual copus, with no language IDs, but **mBERT encodes language identities**. (Wu and Dredze, 2019, Libovicky et al. 2019)
- **It is possible to freeze the Transformer weights and retrain only the input embeddings**. (Artetxe et al. 2019)

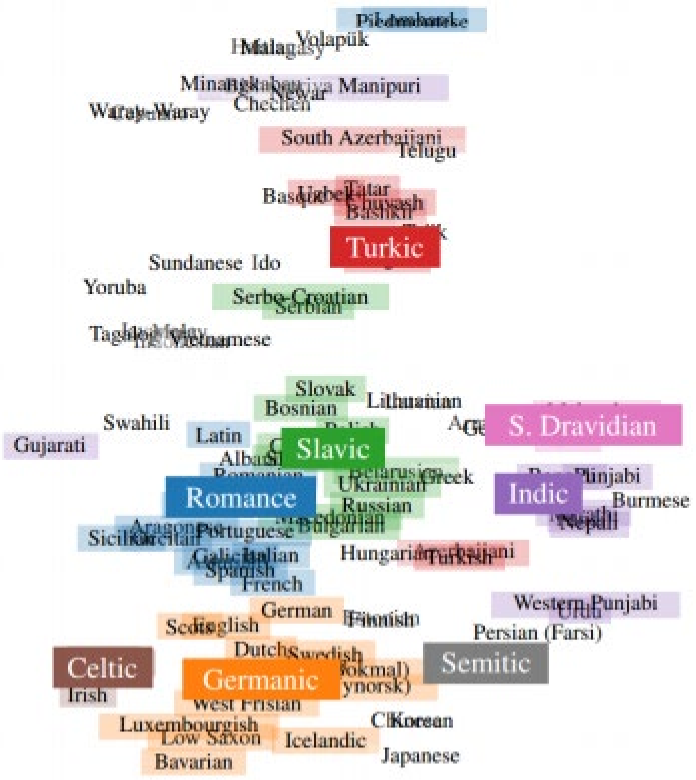


Figure 7: Language centroids of the mean-pooled mBERT representations (Libovický et al., 2019)

Ref:
1- A Primer in BERTology: What we know about how BERT works, A. Rogers, et. al., 2020
2- https://www.youtube.com/watch?v=3VZZbKoXDVM
3- https://github.com/john-hewitt/structural-probes

UNIVERSITY OF WATERLOO

CONCLUSION

KEY INSIGHTS:

- **Benchmark that require verbal reasoning.** While BERT enabled breakthroughs on many NLP benchmarks, a growing list of analysis papers are showing that its verbal reasoning abilities are not as impressive as it seems. (McCoy et al. 2019, Zellers et al. 2019, Si et al. 2019, Rogers et al. 2020, Sugawara et al. 2020)
- **Developing methods to teach reasoning.**
- **Learning what happens at inference time.** At the moment, we know that the knowledge in BERT does not necessarily get used in downstream tasks. (Kovaleva et al. 2019)

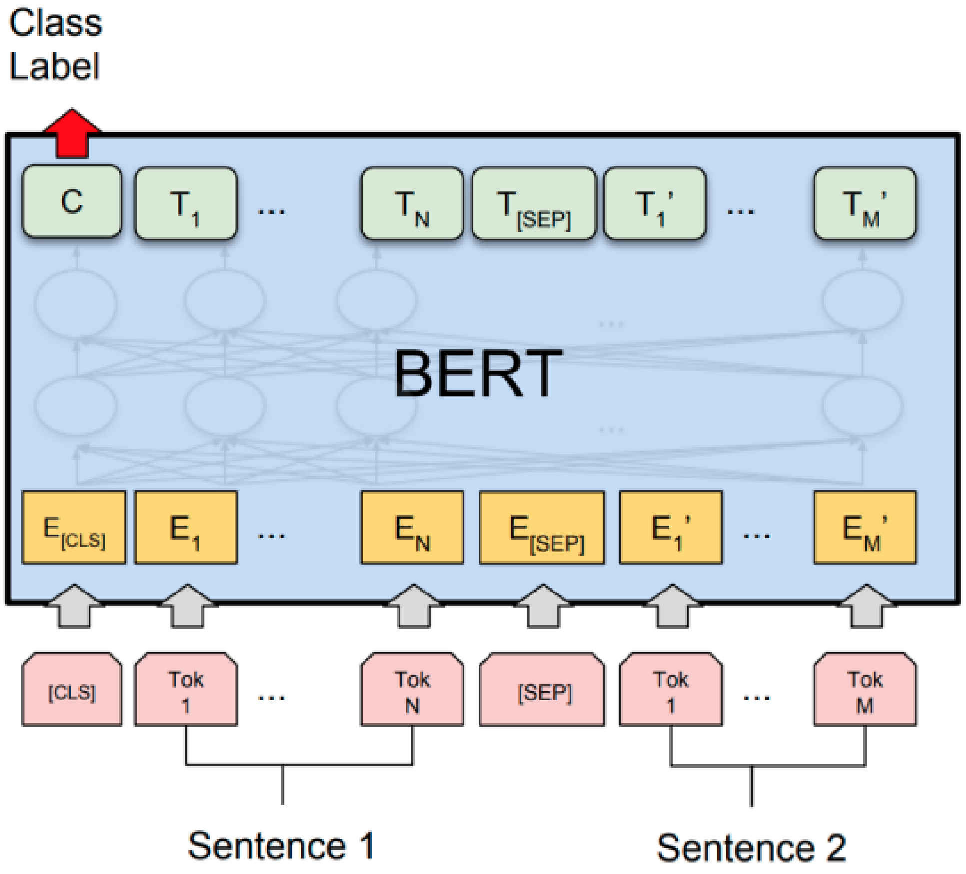
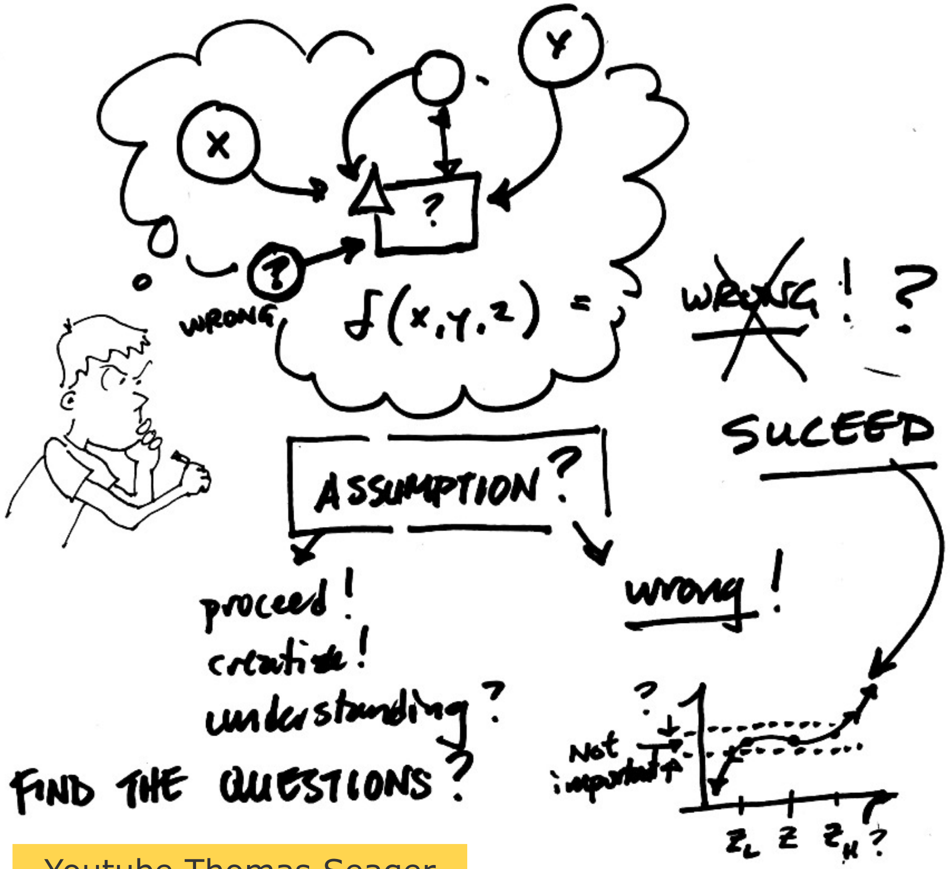


Figure 1: BERT fine-tuning (Devlin et al., 2019).

Ref:
1- A Primer in BERTology: What we know about how BERT works, A. Rogers, et. al., 2020
2- https://www.youtube.com/watch?v=3VZZbKoXDVM
3- https://github.com/john-hewitt/structural-probes

Questions

HikingArtist.com



Youtube Thomas Seager