COGNITION, LANGUES, LANGAGE, ERGONOMIE UMR 5263

How Relevant Are Selectional Preferences for Transformer-based Language Models?

Eleni Metheniti (CLLE-CNRS|IRIT)

02.11.2020

How ***Relevant*** Are **Selectional Preferences** for **Transformer-based Language Models**?

Part 0:

How do computers learn (human) language?

How do computers learn language?

With machine learning language models!

- **Character vectors:**

  ~~🍎~~ → *apple* → `a p p l e` → `[1, 16, 16, 11, 5]`

- **Sub-word vectors:**

  e.g. Byte-pair encoding (BPE): ~~🍎~~ → *apple* → `app le` → `[165, 436]`

- **Word-level vectors:**

  e.g. One-hot encoding: 🍎 → *apple* → `25` → `[1, 0, 0, 0, ...]`

How do computers learn language?

With machine learning language models!

- **Character vectors:**

  🍎 → *apple* → `a p p l e` → `[1, 16, 16, 11, 5]`

- **Sub-word vectors:**

  e.g. [Byte-pair encoding (BPE)]: 🍎 → *apple* → `app le` → `[165, 436]`

- **Word-level vectors:**

  e.g. [One-hot encoding]: 🍎 → *apple* → `25` → `[1, 0, 0, 0, …]`

Words aren't random values!

Language models: Word Embeddings

Vectors → Word Embeddings! ↴

	1	2	3	4	5	6	7	...	N
🍎	**1**	0	0	0	0	0	0	0	0
🍊	0	**1**	0	0	0	0	0	0	0
🍎🥧	0	0	**1**	0	0	0	0	0	0

Can we improve them?

Language models: Word Embeddings

Vectors → Word Embeddings! ↴

	1	**2**	**3**	**4**	**5**	**6**	**7**	**...**	**N**
🍎	**1**	0	0	0	0	0	0	0	0
🍊	0	**1**	0	0	0	0	0	0	0
🍎🥧	0	0	**1**	0	0	0	0	0	0

Can we improve them? Yes!↴

	Food	**Fruit**	**Apple**	**Sweet**	**...**
🍎	**1**	**1**	**1**	**0.5**	0
🍊	**1**	**1**	0	**0.5**	0
🍎🥧	**1**	0	**1**	**1**	0

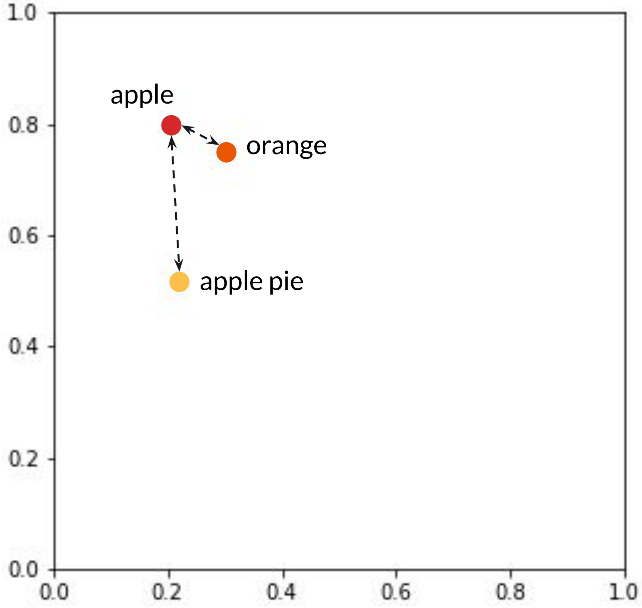
Language models: Word Embeddings

Vectors → Word Embeddings! ↴

	1	2	3	4	5	6	7	...	N
🍎	1	0	0	0	0	0	0	0	0
🍊	0	1	0	0	0	0	0	0	0
🍎🥧	0	0	1	0	0	0	0	0	0

Can we improve them? Yes!↴

	Food	Fruit	Apple	Sweet	...
🍎	1	1	1	0.5	0
🍊	1	1	0	0.5	0
🍎🥧	1	0	1	1	0

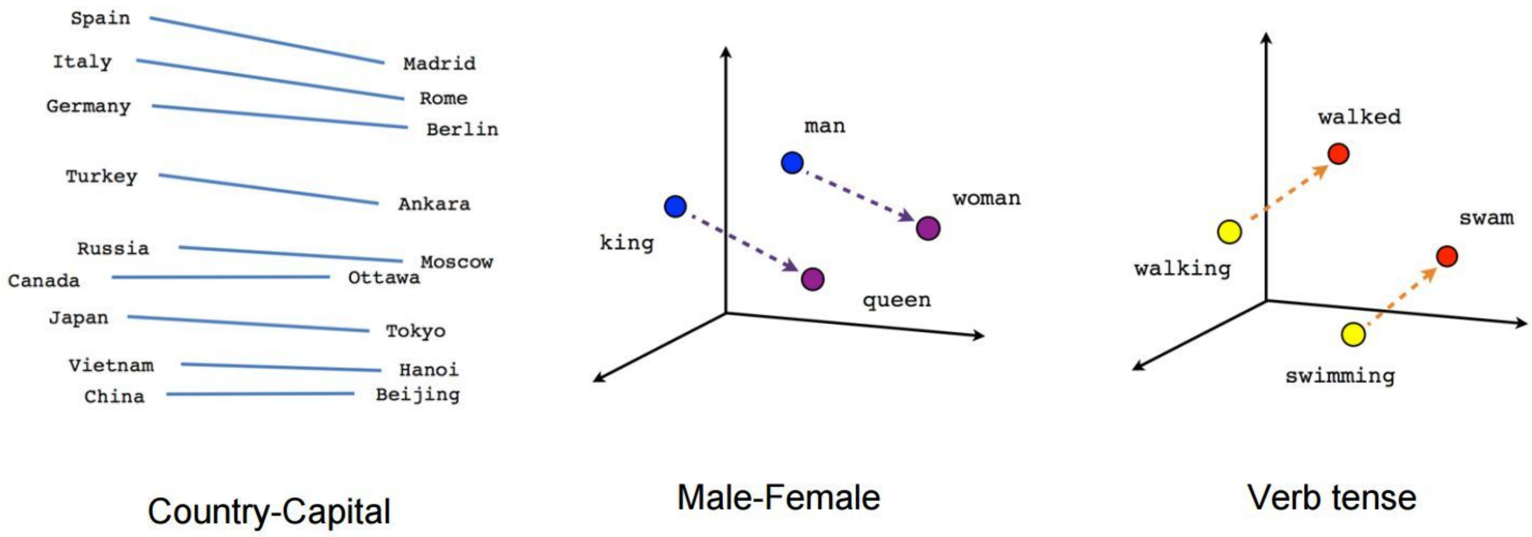


Language models: Word Embeddings

- Text → **Algorithms → (Unsupervised) Word embedding models:**
  word2vec (2013), GloVe (2014), fastText (2015)...

Language models: Word Embeddings

- Text → **Algorithms → (Unsupervised) Word embedding models:** [word2vec](#) (2013), [GloVe](#) (2014), [fastText](#) (2015)...



graphic by Shane Lynn @ https://www.shanelynn.ie/get-busy-with-word-embeddings-introduction

Is one embedding enough?

- Sub-word information? OOV words? Multilingual connections?
- 🍎🥧 ≠ 🍎📱
- 🍎 → `[0.5, 1, 0, 0, 0 …]` AND `[0, 0, 0, 1, 1 …]`

Is one embedding enough?

- **Sub-word information? OOV words? Multilingual connections?**
- 🍎🥧 ≠ 🍎📱
- 🍎 → `[0.5, 1, 0, 0, 0 …]` **AND** `[0, 0, 0, 1, 1 …]`

> Text →**Neural Network** →hidden state + word2vec embeddings ⇒ **embedding** information + text **dependencies** learned by the NN

Deep contextualised word representation

- [TagLM](#) (2017): [Recurrent Neural Network](#) (RNN)
- [ELMo](#) (2018): [Bidirectional Long Short Term Memory](#) (bi-LSTM) NN

Part 1:

Fine-tuned, deep contextualised word representation:
Transformer-based Language models

The path to Transformers

seq2seq models

↓

learning input serially

The path to Transformers

seq2seq models → **seq2seq + attention**

↓ learning input serially

↓ learning input serially & learning important "shortcuts" of context

The path to Transformers

seq2seq models → seq2seq + attention → Transformers + self-attention

learning input serially

learning input serially & learning important “shortcuts” of context

parallelized learning & use of self-attention for word representations

collects attention from the entire input, creates **representations**

(+ **multi-headed**, i.e. many subspaces!)

Transformer spotlight: BERT

BERT-base
BERT-large

Bidirectional → [MASK]ed predictions both ways

Encoder → encodes NN input

Representations → **word embeddings!**

from Transformers

- Truly Bidirectional: self-attention context from both sides of the word

```
I love eat ##ing [MASK] pie
```

- Pre-train with a **large** amount of data
- Fine-tune with data specific to an NLP task

Petite pause ~~café~~ questions!

What do BERT's embeddings know?

Rogers, A., Kovaleva, O., & Rumshisky, A. (2020). A primer in bertology: What we know about how BERT works. arXiv preprint arXiv:2002.12327.

What do BERT’s embeddings know?

- Do they behave like **traditional embeddings** (distribution, transformations)?
  - Yes… maybe in the higher layers

Rogers, A., Kovaleva, O., & Rumshisky, A. (2020). A primer in bertology: What we know about how BERT works. arXiv preprint arXiv:2002.12327.

What do BERT’s embeddings know?

- Do they behave like **traditional embeddings** (distribution, transformations)?
  - Yes… maybe in the higher layers
- Do they have **syntactic information**?
  - Hierarchical, tree-like structure
  - Bidirectionality really helped!
  - Parts of speech, syntactic chunks and roles, but not distant relations
  - (Probably) No full syntactic trees, but syntactic transformations and dependencies
  - Bad with negation and with “bad” input
  - Does it really understand syntax?

Rogers, A., Kovaleva, O., & Rumshisky, A. (2020). A primer in bertology: What we know about how BERT works. arXiv preprint arXiv:2002.12327.

What do BERT's embeddings know?

- Do they have **semantic information**?
  - Some knowledge of semantic roles, entity types, relations, proto-roles
  - Can't generalize!
- Do they have **world knowledge**?
  - Fills the blanks successfully, but not enough!
  - Bad at inference, bias?!

Rogers, A., Kovaleva, O., & Rumshisky, A. (2020). A primer in bertology: What we know about how BERT works. arXiv preprint arXiv:2002.12327.

Question:

Do BERT encodings capture linguistic information; specifically, the selectional preferences of a verb for its predicates?

Part 2:

Selectional Preferences

What are selectional preferences?

The athlete runs a marathon = (🏃 + ~~🎽~~) + (~~🎽~~ + 🏟️) -> ✔️✔️

The trumpet runs a banana = (🎺 + ~~🎽~~) + (~~🎽~~ + 🍌) -> ❌❌

We can tell the difference... But can BERT?

Methodology

1. Use BERT-base Masked Language Model (MLM)
2. Create sentences with [MASK]ed dependent word
    - Sentences with (in)felicitous head-dependent pairings
3. Retrieve the **probability** assigned to dependent word
    - Use different scenarios: attention can only access certain words!
4. Is the **probability correlated** to the **degree of felicity**?

SP-10K: Selectional Preference Corpus

- Pairs of **head word** + **dependent word**, **score** of plausibility (felicity) 0-10
- 10K pairs, 2500 words, 5 categories:
  - nsubj = verb + noun
  - dobj = verb + noun
  - amod = noun + adjective

    one-hop relation

  - nsubj_amod = verb + adjective (+ noun)
  - dobj_amod = verb + adjective (+noun)

    two-hop relation

invest money

invest e-mail

- Combined with ukWaC corpus

Hongming Zhang, Hantian Ding, and Yangqiu Song. 2019. SP-10K: A large-scale evaluation set for selectional preference acquisition.
Adriano Ferraresi, Eros Zanchetta, Marco Baroni, and Silvia Bernardini. 2008. Introducing and evaluating ukWaC, a very large web-derived corpus of English.

Our corpus

Type	Word pairs in ukWaC	Final sents	Avg. plausibility score
nsubj	958 / 2,000	30,526	6.64
dobj	980 / 2,000	56,777	7.39
amod	1,030 / 2,000	23,110	7.62
nsubj_amod	956 / 2,061	12,911	5.75
dobj_amod	922 / 2,063	21,839	6.32
TOTAL	4846 / 10,124	145,163	

- Short sentences (4-15), distance of pair < 5
- Problems with BERT tokenizer, problems with SP-10K
- Too low plausibility -> impossible to find!

Prediction process

dobj

amod

require additional (dobj_amod) → if you **require additional** information please contact residential and catering services

Prediction process

require additional (dobj_amod)

dobj

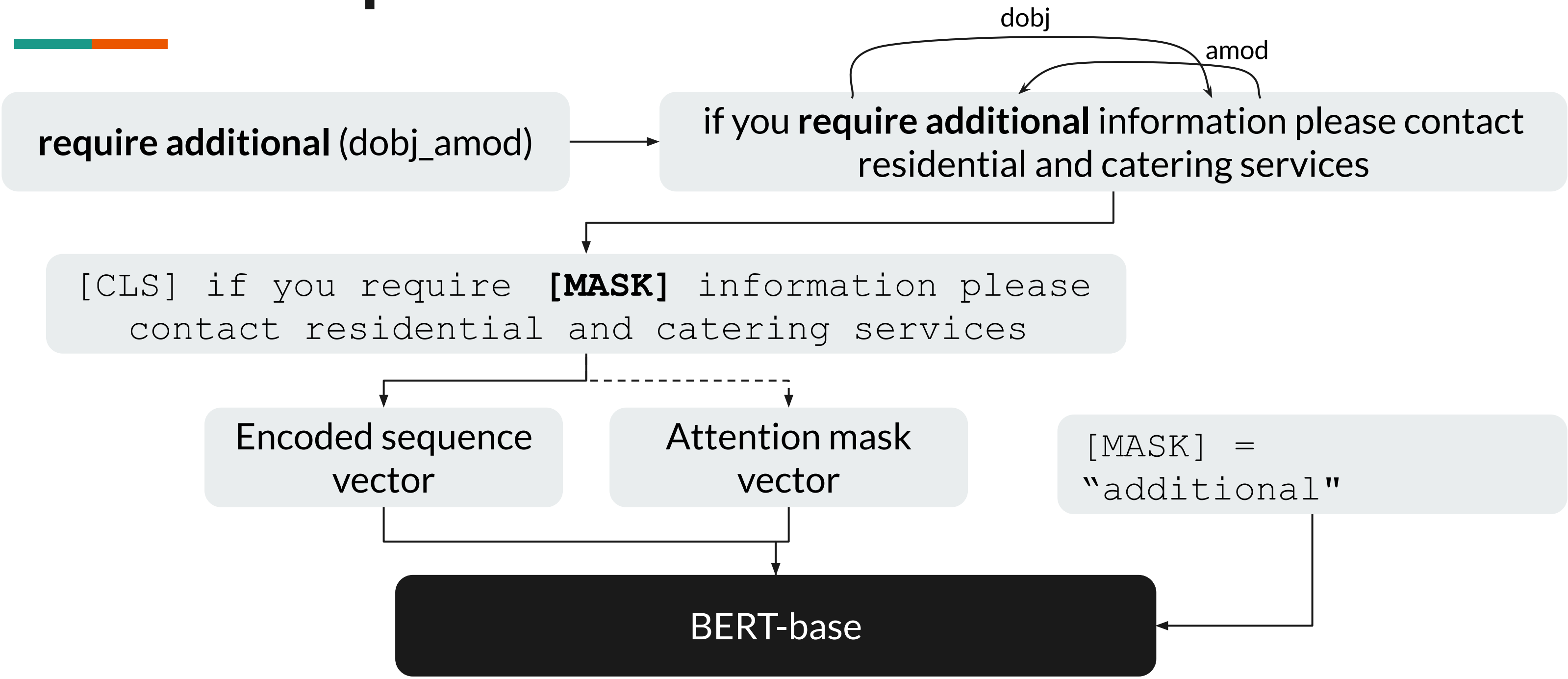
amod

if you **require additional** information please contact residential and catering services

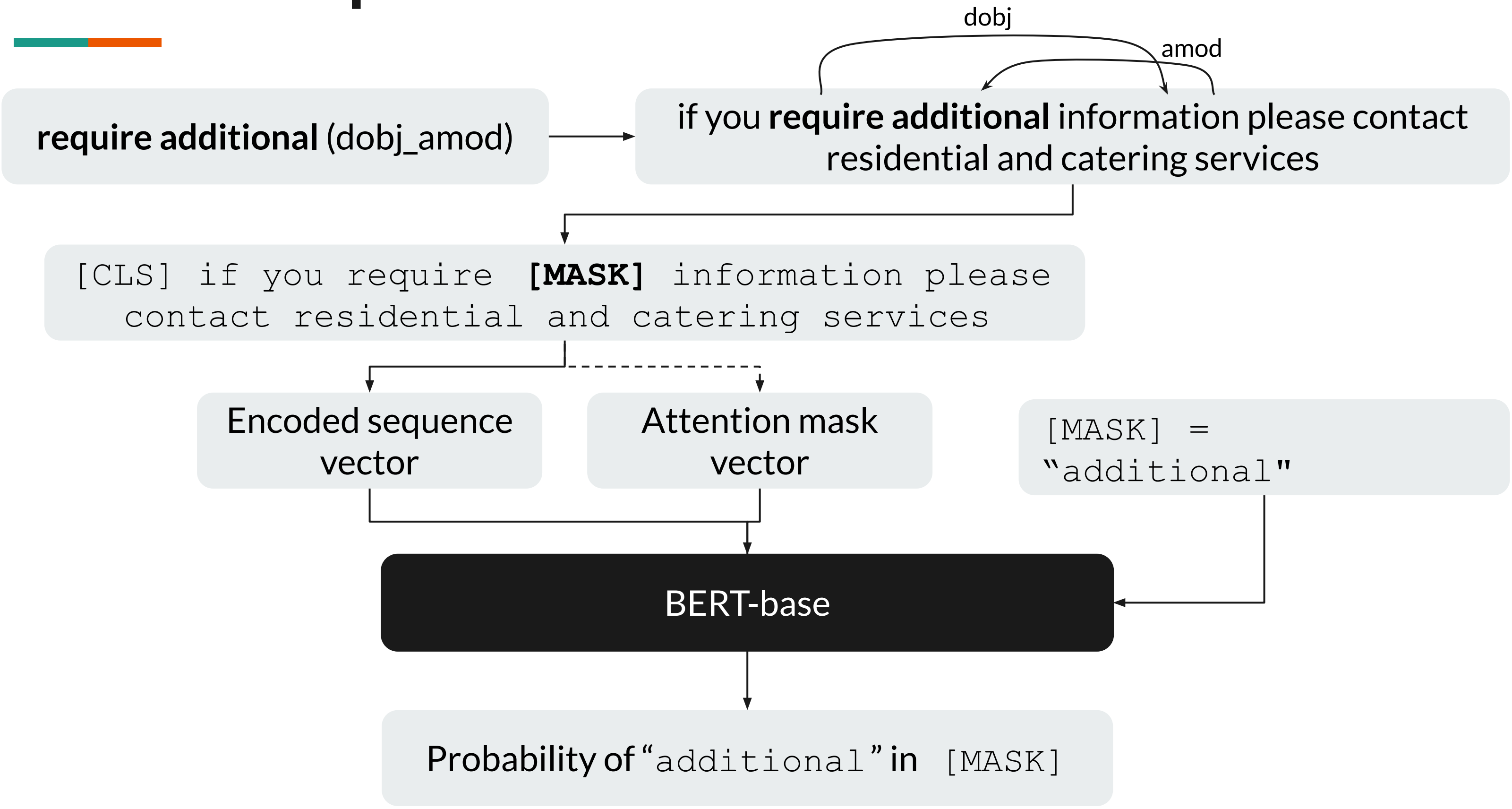
`[CLS] if you require` **`[MASK]`** `information please contact residential and catering services`

`[MASK] = “additional"`

Prediction process



Prediction process



Attention mask

sentence		the	film	**tells**	the	**story**	of	that	trial
standard	[CLS]	the	film	tells	the	[MASK]	of	that	trial
	[1,	1,	1,	1,	1,	1,	1,	1,	1]
head	[CLS]	the	film		the	[MASK]	of	that	trial
	[1,	1,	1,	0,	1,	1,	1,	1,	1]
context	[CLS]			tells		[MASK]			
	[1,	0,	0,	1,	0,	1,	0,	0,	0]
control	[CLS]					[MASK]			
	[1,	0,	0,	0,	0,	1,	0,	0,	0]

Results

- Kendall τ correlation of probability + plausibility
- Significant correlation: **<-0.4** or **>0.4**

	standard	head	context	control
nsubj	0.03	-0.02	0.16	-0.01
dobj	0.05	-0.07	0.05	-0.05
amod	0.04	-0.06	0.24	-0.04
nsubj_amod	-0.01	-0.13	0.29	0
dobj_amod	0.06	0.01	-0.03	0.02

Micro-averaged

	standard	head	context	control
nsubj	0.19	0.15	0.29	0.08
dobj	0.16	0.04	0.27	0.05
amod	0.15	0.03	0.35	0.03
nsubj_amod	0.01	-0.04	0.22	0.06
dobj_amod	0.14	0.1	0.2	0.07

Macro-averaged

Results: nsubj

- Do we notice some head categories with strong positive/negative correlations? **NO**

  e.g. *kill:* strong positive, *shoot*: strong negative, *strike*: no correlation
- What happens with **attention masks**?
  - head: (slightly) worse than standard
  - context: better than standard, but not strong correlation

Results: dobj

- Do we notice some head categories with strong positive/negative correlations? **NO**
- Do we notice some dependent categories with strong correlations? **No...**

  e.g. *blame customer* < *blame management* (but not with head mask!)
- What happens with **attention masks**?
  - head: worse than standard
  - context: better than standard, but not strong correlation

Results: amod

- NB: Overall highest plausibility scores
- Do we notice some head categories with strong positive/negative correlations? **NO**
- BERT likes high-frequency adjectives, but they are not always the best fit...
- What happens with **attention masks**?
    - head: worse than standard
    - context: better than standard, but not strong correlation

Results: nsubj_amod

- NB: Overall lower plausibility scores
- Do we notice some head categories with strong positive/negative correlations? **NO**
- BERT likes high-frequency adjectives, but they are not always the best fit...
- What happens with **attention masks**?
  - head: worse than standard
  - context: better than standard, but not strong correlation (+0.20 improvement!)

Results: dobj_amod

- NB: Overall lower plausibility scores
- Do we notice some head categories with strong positive/negative correlations? **NO**

  Harder to make assumptions with two-hop relations
- Do we notice some dependent categories with strong correlations? **NO**
- What happens with **attention masks**?
  - head: worse than standard
  - context: (not) better than standard, but not strong correlation (smallest)

Discussion

- Problems with plausibility scores in the SP-10K corpus
- Never found all word pairs in corpus…
- **No strong correlations** but… **BERT did capture preferences and constraints**!
- Attention: Head word was more important than the entire sequence!

Future directions

- Current work: BERT and Telicity:
    - Telic/Atelic verbs + Prepositional phrases: does BERT see a link?

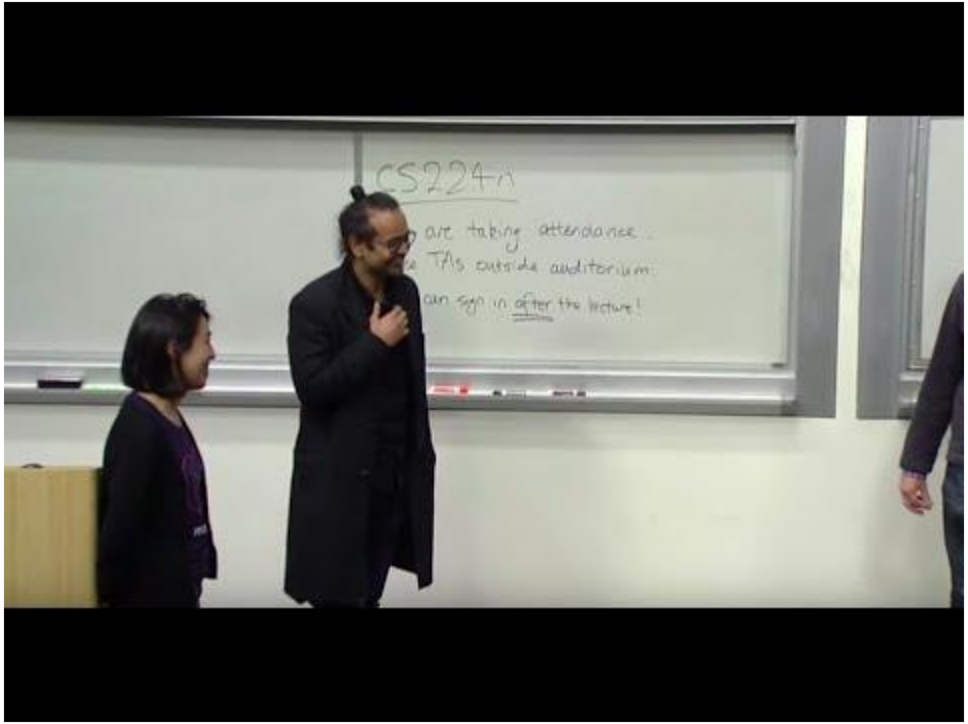
- Future work: Multiple layers and heads, which one of them is best?

Merci pour votre attention!
Y a-t-il des questions?

(Even more) Neural Network resources

- 3Blue1Brown 4-video series (avec sous-titres!): Neural Networks
- Jason Brownlee's blog: Machine Learning Mastery
- Jay Allamar's blog: visualizations of neural networks & videos, very up-to-date
- Stanford University's CS224n: Natural Language Processing with Deep Learning: full lectures in video, slides, special guests
- BERT for dummies: article + some code to get started!
- Rasa YouTube Channel, NLP for Developers

Talk on Transformers (by its creators)



Stanford CS224N:
NLP with Deep Learning
Winter 2019
[Lecture 14 – Transformers and Self-Attention]

Chris Manning,
Ashish Vaswani,
Anna Huang

Merci pour votre attention!
Y a-t-il des questions?