About Time : (How) do Transformers learn Temporal Verbal Aspect?

Eleni Metheniti (CLLE-CNRS | IRIT)
Nabil Hathout (CLLE-CNRS)
Tim Van de Cruys (KU Leuven)

29.11.2021
Thématiques actuelles de la recherche en TAL

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

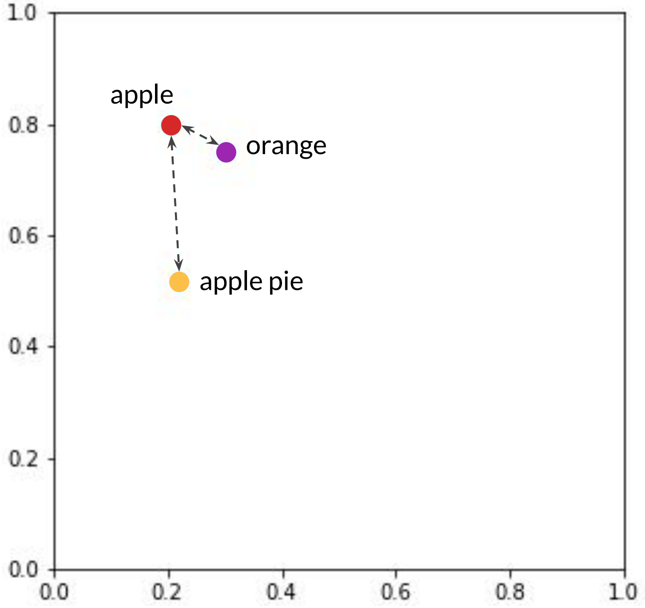
Language models: Word Embeddings

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

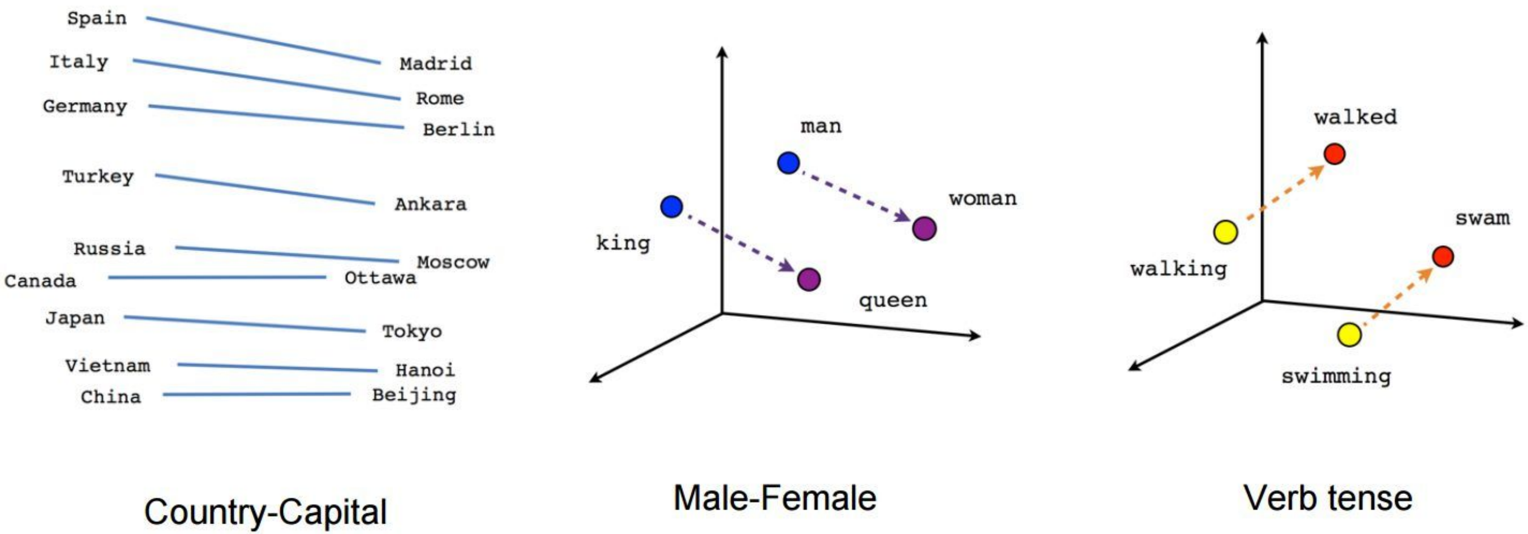


Language models: Word Embeddings

- 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

Fine-tuned, deep contextualised word representations: 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

Even more Transformers!

- RoBERTa: more subwords, more mini-batches, larger learning rates
- ALBERT: smaller and more efficient, learns context-dependent and context-independent representations
- XLNET: more computations between words, better dependencies and relations

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.

Can *transformers* capture more fine-grained semantic information…?

… specifically, features of lexical aspect?

What is lexical aspect?

- Lexical aspect ≠ Grammatical aspect ≠ Mood ≠ Tense

- Temporal features of a verb's described action, event or state:
    - frequence
    - **duration**: stative, punctual, durative
    - **telicity**: telic, atelic

Telicity and Duration

- **Telicity:** is there an end point to an action?
    - Telic: “I ate a fish.” “The soup cooled in an hour.”
    - Atelic: “John watched TV.” “Nobody laughs at my jokes.”

- **Duration:** is there an action or a state?
    - Stative: “I disagree with you.” “Bread is made of flour.”
    - ~~Punctual~~: “I knocked on the door.”
    - Durative: “I walked.” “I slept all morning.”

Question

Can transformers understand telicity and duration?

- Does providing the **verb position** help with predictions?
- Which architectures are most **successful**?
- When is classification **possible** or unsuccessful?
- How does the **attention** mechanism focus on aspect?
- Differences between English and **French**?

Fine-tuning a transformer

- Transformers + millions of sentences + ~~hours days~~ months of training ⇒ **Pretrained language models**
- Very good… but can be better!

- Pretrained language models + (small) **specialized** data + (reasonable) training ⇒ **Finetuned language models**
- Even better on a specialized task!

Experimental setup

Pretrained transformer models

EN: BERT, RoBERTa, XLNet, Albert
FR: CamemBERT, FlauBERT

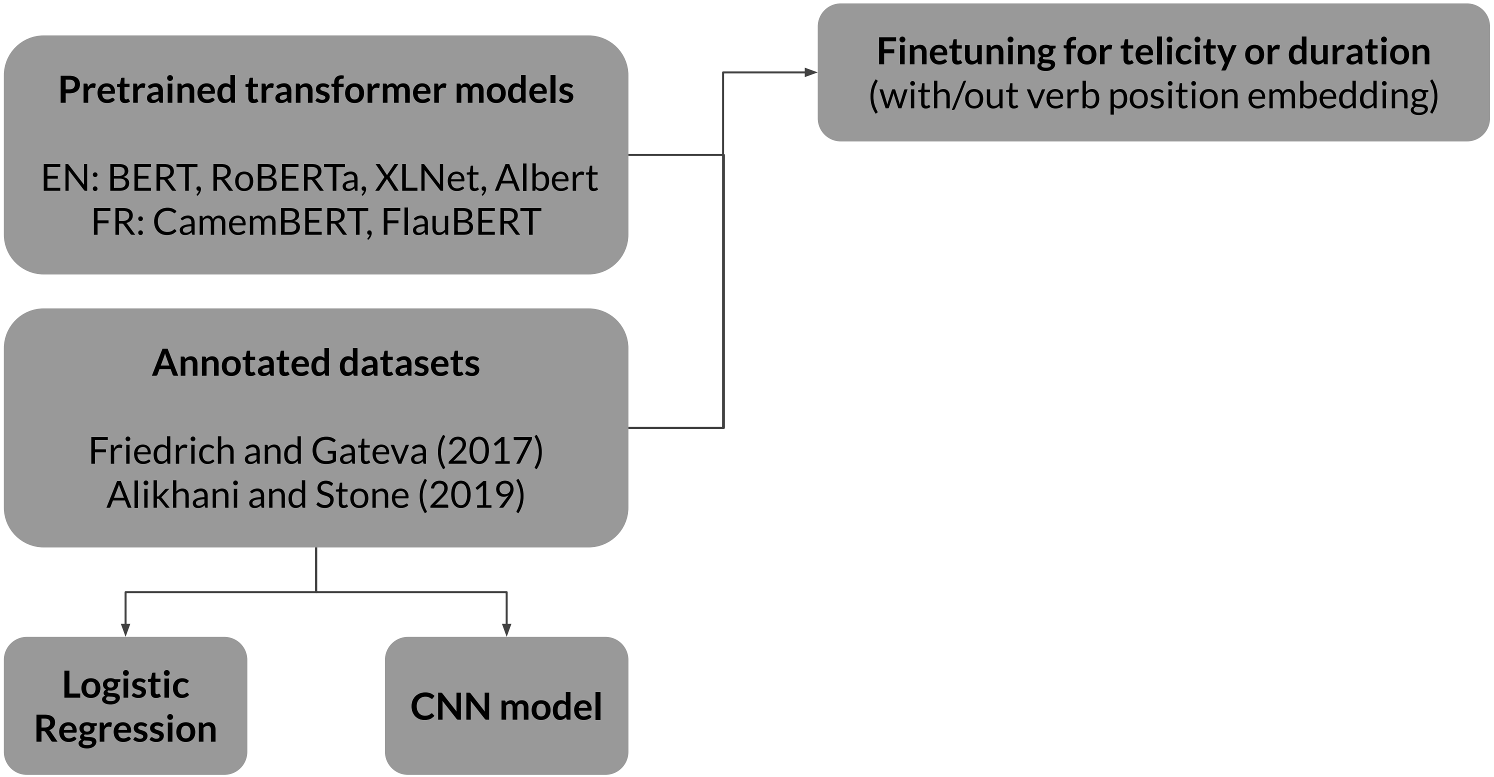
Annotated datasets

Friedrich and Gateva (2017)
Alikhani and Stone (2019)

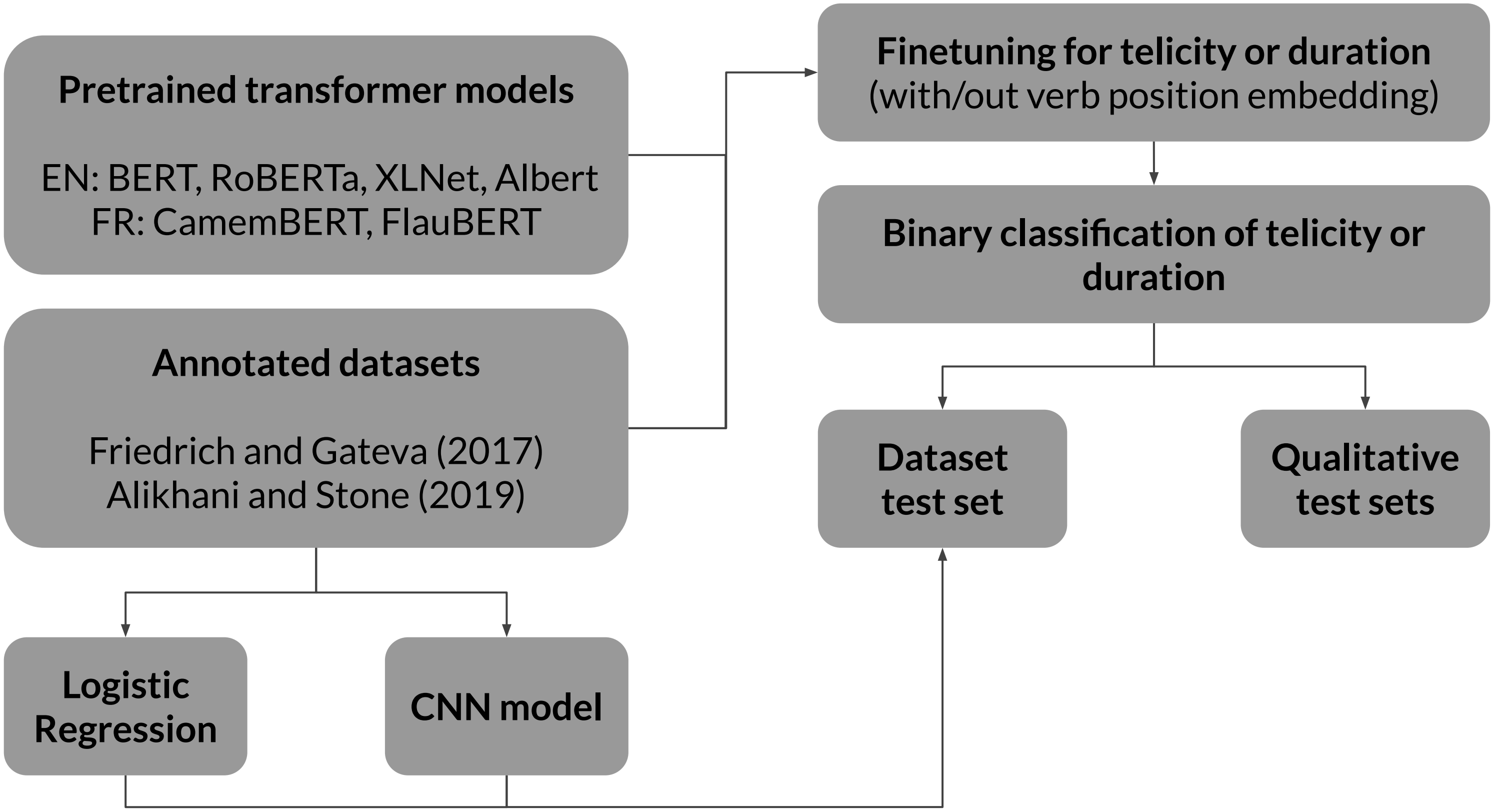
Alikhani, M., & Stone, M. (2019, June). "Caption" as a Coherence Relation: Evidence and Implications. In Proceedings of the Second Workshop on Shortcomings in Vision and Language (pp. 58-67).

Friedrich, A., & Gateva, D. (2017, September). Classification of telicity using cross-linguistic annotation projection. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (pp. 2559-2565).

Experimental setup



Experimental setup



Datasets

- Training and quantitative analysis:

Type	Label	F&G	A&S	Ours	1st exp.	2nd exp.
telicity	**telic**	1,831	785	2,885	5,083	6,173
	atelic	2,661	1,256	3,288		
duration	**stative**	1,860	419	2,036	4,095	4,081
	durative	38	1,843	2,045		
	~~**punctual**~~	-	~~355~~	-		

- Qualitative analysis:
  - 40 sentences for telicity, 40 for duration
  - 40 “minimal pairs” of telicity
  - More pairs on telicity, with different word order and tense

First experiment

- Article:
  Eleni Metheniti, Tim van de Cruys, Nabil Hathout. Prédire l'aspect linguistique en anglais au moyen de transformers. *Traitement Automatique des Langues Naturelles* (TALN 2021), 2021, Lille, France. pp.209-218.

- Poster

Telicity results

- All models achieved accuracy of >0.80
- BERT models outperformed the rest: 0.88 (bert-large-cased)
- RoBERTa models quite successful, XL-Net and ALBERT models less successful

- **Verb positions**: very small improvement (+1-5%)

Model	Verb position?	Accuracy
bert-base-uncased	yes	0.86
	no	0.81
bert-base-cased	yes	0.87
	no	0.81
bert-large-uncased	yes	0.86
	no	0.81
bert-large-cased	**yes**	**0.88**
	no	0.81
roberta-base	no	0.84
roberta-large	no	0.8
xlnet-base-cased	yes	0.82
	no	0.81
xlnet-large-cased	yes	0.82
	no	0.8
albert-base-v2	yes	0.84
	no	0.81
albert-large-v2	yes	0.8
	no	0.82
CNN (50 epochs)	no	0.75
Logistic Regression	no	0.61

Duration results

- Very high accuracy, models achieved accuracy of >0.93
- BERT models slightly outperformed the rest (in general)
- All models were very successful

- **Verb positions**: no improvement (±1-2%)

Model	Verb position?	Accuracy
bert-base-uncased	**yes**	**0.96**
	no	0.94
bert-base-cased	yes	0.96
	no	**0.96**
bert-large-uncased	**yes**	**0.96**
	no	0.95
bert-large-cased	**yes**	**0.96**
	no	0.95
roberta-base	no	0.95
roberta-large	no	0.95
xlnet-base-cased	yes	0.94
	no	0.95
xlnet-large-cased	yes	0.94
	no	0.95
albert-base-v2	yes	0.95
	no	0.95
albert-large-v2	**yes**	**0.96**
	no	**0.96**
CNN (50 epochs)	no	0.88
Logistic Regression	no	0.7

Qualitative analysis: telicity

- Correct in most cases and models, but problem with conflicting verb - context
  - ✓ *Cork floats on water.*
  - ✓ *The Earth revolves around the Sun.*
  - ✓ *I spilled the milk.*
  - ✓ *I always spill milk when I pour it in my mug.*

  - ✗ *I eat a fish for lunch on Fridays.*
  - ✗ *The inspectors are always checking every document very carefully.*

Qualitative analysis: telicity

- Minimal pairs:
  - ✓ *I drank **the whole bottle**.*
  - ✓ *I drank **juice**.*
  - ✗ *The cat drank **all** the milk.*

  - ✗ *The boy is eating **an apple**.*
  - ✓ *The boy is eating **apples**.*

Qualitative analysis: telicity

- Word order and tenses:
  - ✗ *I **ate** a fish for lunch **at noon**. **At noon** I **ate** a fish for lunch.*
  - ✓ *I **had eaten** a fish for lunch **at noon**. **At noon** I **had eaten** a fish for lunch.*

  - ✗ *The Prime Minister **made** that declaration **for months**.*
  - ✓ ***For months** the Prime Minister **has been making** that declaration.*

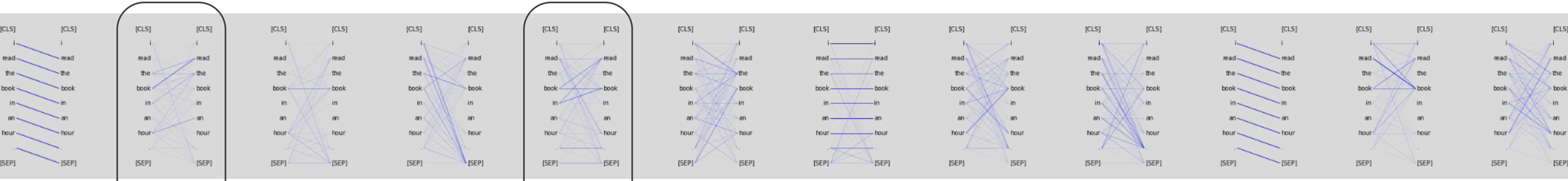
Qualitative analysis: duration

- *Stative* sentences were more difficult than *durative* sentences for the models:
  - ✗ *Bread consists of flour, water and yeast.*
  - ✓ *I disagree with you.*

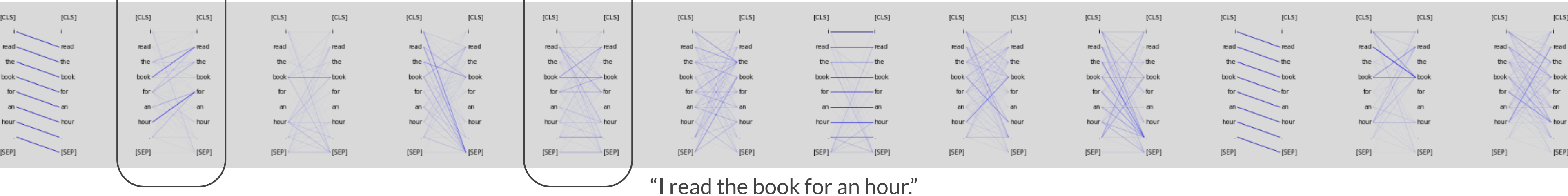
- Durative sentences always correctly classified:
  - ✓ She plays tennis every Friday.
  - ✓ She is playing tennis right now.

Attention mechanism

- BERT models in earlier layers: “focused” attention to specific tokens
- Other models: “diffused” attention early
- bert-base-uncased, layer 3, heads 1-12:



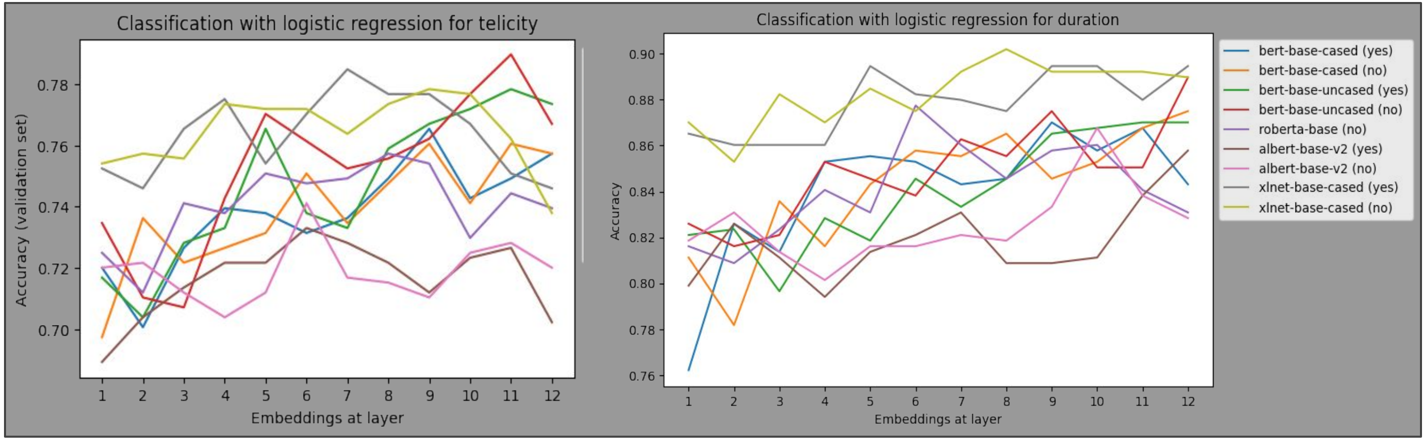
“I read the book in an hour.”



“I read the book for an hour.”

What do pretrained embeddings already know?

- Classification with (contextual) embeddings for verb & logistic regression, per layer
- Higher accuracy in middle layers and final layers, drops in the last



Classification for French

- Same datasets (translated), same procedure of classification
- Telicity: 0.77 (camembert-base, flaubert-base-cased)
- Duration: 0.87 (camembert-large, flaubert-large-cased)
- Verb position deteriorated the results

- Better performance at qualitative sets!
- Telicity:
  - ✓ *Je mange un poisson à midi les vendredis.*
- Duration:
  - ✗ *Le pain est composé de farine, d'eau et de levure.*

Discussion

- Contextual embeddings are good at telicity & duration, even without finetuning!
- Why did BERT models outperform? Probably because of segmentation?

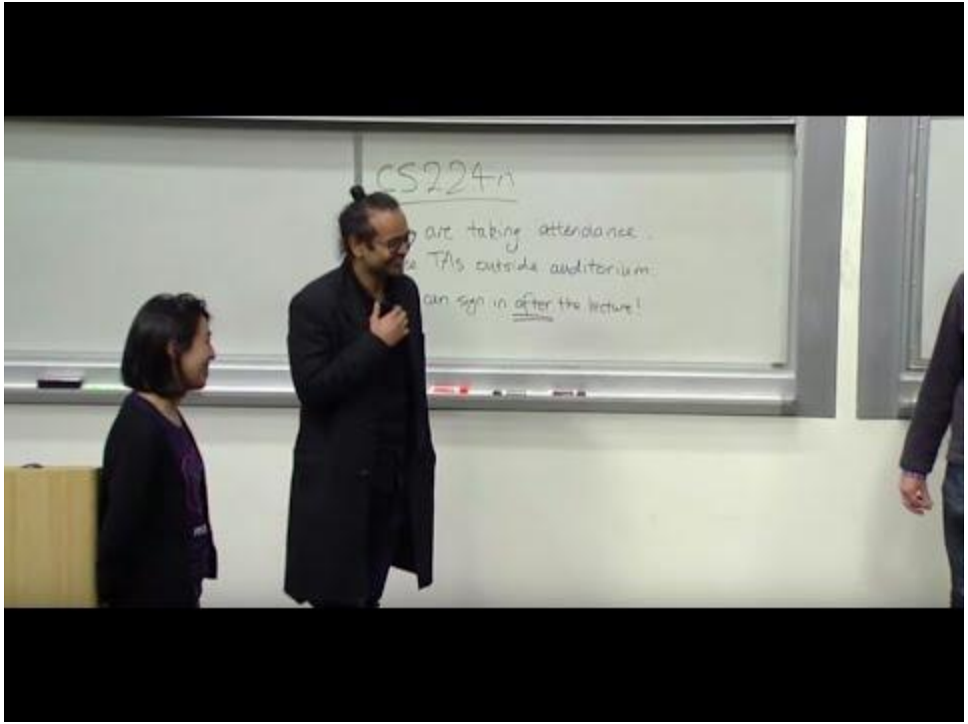
- Qualitative analysis:
  - Verb features > context > infelicitous context
  - Word order, tense were influential (to some degree)
  - French morphosyntax might have been “easier” for the models than English

Merci pour votre (self-)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