About time:
Do *transformers* learn temporal verbal aspect?

Eleni Metheniti, Tim Van de Cruys, Nabil Hathout

in **Cognitive Modeling and Computational Linguistics (CMCL)**

May 2022

Laboratoire Cognition, Langues, Langage, Ergonomie (CLLE) - UT2J, CNRS
Institut de Recherche en Informatique de Toulouse (IRIT) -UT3
Leuven.AI institute

Lexical aspect and time

What is lexical aspect?

Temporal features of a verb's described action, event or state:

- frequence
- duration: *stative, punctual, durative*
- telicity: *telic, atelic*

Careful! Lexical aspect ≠ Grammatical aspect ≠ Mood ≠ Tense

Telicity & 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 this an action or a state?

- **Stative**: “I disagree with you.” “Bread is made of flour.”
- **Punctual**: “I knocked on the door.”
- **Durative**: “I knocked on the door.” “I walked.” “I slept all morning.”

Lexical aspect and language acquisition

- *Aspect hypothesis* (Shirai, 1991; Shirai and Andersen, 1995):
  - Telicity associated with past and perfectivity
  - Activity -> Accomplishment -> Achievement, not state
- Conflict verb-context: delayed processing (Todorova et al., 2000)
- DO not strong influence, Prepositions important for telicity

- Stativity is difficult! (Rocca, 2002)
- Perfectivity before duration (Wen, 1997)

But will transformers be successful?

Our research questions

- 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?
- Differences between English and French models?

Experiment: Finetuning & Classifying for telicity/duration

Experimental setup

Pretrained transformer models

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

Logistic Regression

CNN model

Experimental setup

Pretrained transformer models

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

Annotated datasets

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

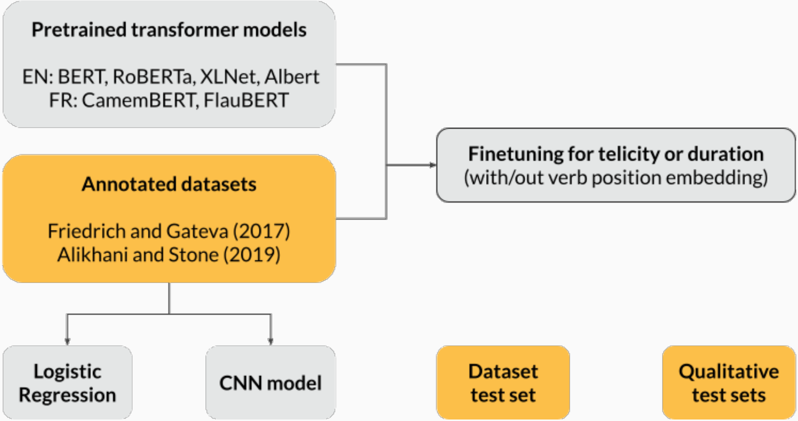
Logistic Regression

CNN model

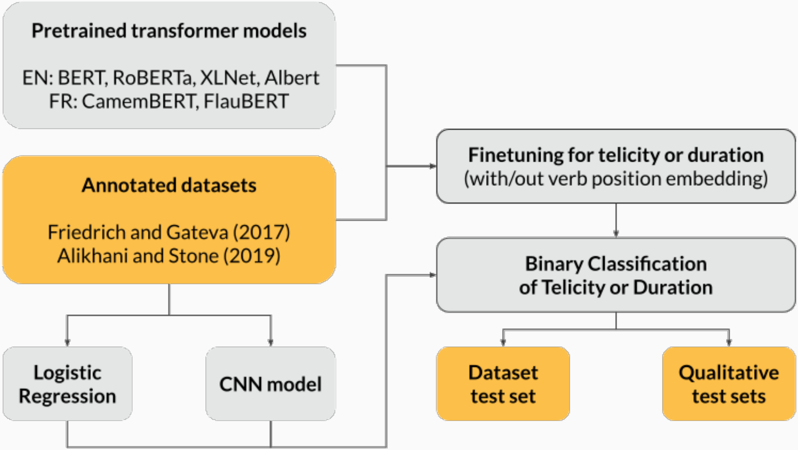
Dataset test set

Qualitative test sets

Experimental setup



Experimental setup



Verb position information

Model input:

- *input_ids*
- *attention_mask*
- *token_type_ids*

Tokens	He	worked	well	and	earned	much	.		
token_type_ids	*[0*	*1*	*0*	*0*	*0*	*0*	*0]*		
Tokens (subwords)	He	work	###ed	well	and	earn	###ed	much	.
token_type_ids	*[0*	*1*	*1*	*0*	*0*	*0*	*0*	*0*	*0]*

Datasets (English)

Training and quantitative analysis:

Type	Label	Friedrich and Gateva	Alikhani and Stone	Ours	Total
telicity	*telic*	1,831	785	2,885	6,173
	atelic	2,661	1,256	3,288	
duration	*stative*	1,860	419	2,036	4,081
	durative	38	1,843	2,045	
	~~punctual~~	-	~~355~~	-	

Qualitative analysis:

- 40 sentences for telicity, 40 for duration
- 40 sentences of “minimal pairs” of telicity
- 80 sentences with variations of word order and tense, for telicity

Datasets (French)

Training/validation/test sets:

- Machine translated from English, partially evaluated for translation and annotation accuracy by us

Qualitative analysis:

- 40 sentences for telicity, 40 for duration
- 40 sentences of "minimal pairs" of telicity
- 40 sentences with variations of word order and tense, for telicity

Results (EN)

Quantitative results: Telicity (EN)

- All models achieved accuracy of >0.80
- BERT models outperformed the rest: 0.88 (bert-large-cased)
- RoBERTa models quite successful, XLNet and ALBERT models less successful
- Verb positions: very small improvement (+1-5%)
- Similar accuracy for sentences with seen/unseen verbs in training (±1-4%)

Model	Verb?	Acc.
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	0.80
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	0.80
	no	0.82
CNN (50 epochs)	no	0.75
Logistic Regression	no	0.61

Quantitative results: Duration (EN)

- Very high accuracy, models achieved accuracy of >0.93
- BERT models slightly outperformed the rest (in general)
- All models were very successful
- Verb position information: no improvement (±1-2%)
- Similar accuracy for sentences with seen/unseen verbs in training (±1-3%)

Model	Verb?	Acc.
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	0.70

Qualitative results: Telicity (EN)

Correct in most cases and models, but problems when there is conflict between verb and 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.

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

Qualitative results: Telicity (EN)

Minimal pairs:

✓ I drank the whole bottle.
✓ I drank juice.
X The cat drank all the milk.

X The boy is eating an apple.
✓ The boy is eating apples.

Qualitative results: Telicity (EN)

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 results: Duration (EN)

Stative sentences were more difficult than durative sentences for the models:

- *X* 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.

Results (FR)

Quantitative results (FR)

- Telicity:
  - Best: 0.77 (camembert-base & flaubert-base-cased, without verb)
  - Worst: 0.69 (flaubert-small-cased, with verb)
  - Baselines: 0.71 (CNN), 0.61 (Log. regression)
- Duration:
  - Best: 0.87 (camembert-large & flaubert-large-cased, without verb)
  - Worst: 0.79 (flaubert-small-cased, with verb)
  - Baselines: 0.80 (CNN), 0.68 (Log. regression)
- Verb position deteriorated the results marginally

Qualitative results (FR)

Better performance at qualitative sets than English!

Telicity:

- ✓ Je mange un poisson à midi les vendredis.
- ✗ Le garçon mange une pomme.
- ✓ Le garçon mange des pommes.
- ✗ J'ai bu du jus de fruit.
- ✓ J'ai bu toute la bouteille.

Duration:

- ✗ J'aime le chocolat.
- ✗ Le pain est composé de farine, d'eau et de levure.

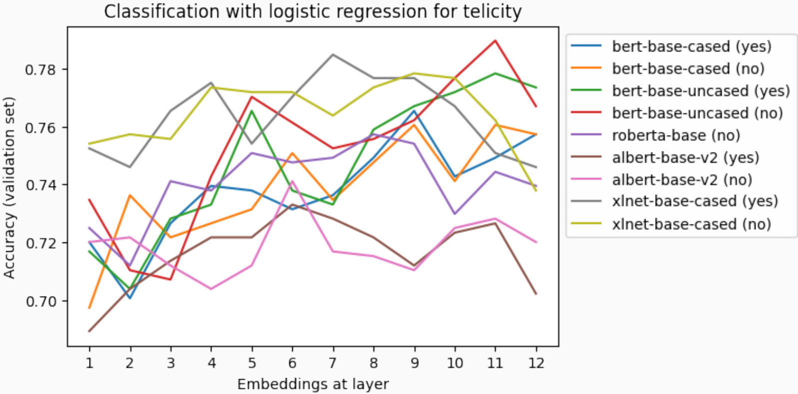
Pretrained vector classification

Additional experiment: Pretrained models and verb vectors

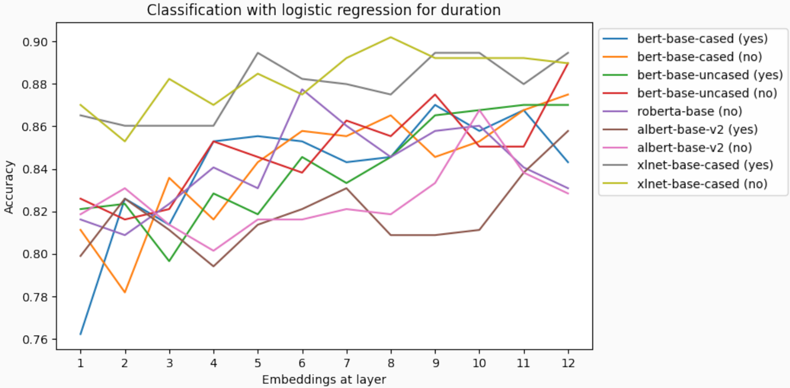
(for English)

- Find verb position in sentence
- Extract its contextual word embeddings, per layer
- Train logistic regression model with verb embeddings
- Predict label on test set

Pretrained models and verb vectors



Pretrained models and verb vectors



Discussion

Discussion

- Contextual embeddings are good at telicity and duration classification, even without finetuning!
- Why did BERT models outperform? Better attention, better semantic representations
- 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

Thank you for your attention!

Selected References

Alikhani, M. and Stone, M. (2019). “Caption” as a Coherence Relation: Evidence and Implications. In *Proceedings of the Second Workshop on Shortcomings in Vision and Language*, pages 58–67.

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

Rocca, S. (2002). Lexical aspect in child second language acquisition of temporal morphology. *The L2 acquisition of tense-aspect morphology*.

Shirai, Y. (1991). Primacy of aspect in language acquisition: Simplified input and prototype.

Shirai, Y. and Andersen, R. W. (1995). The acquisition of tense-aspect morphology: A prototype account. *Language*, 71(4):743–762.

Todorova, M., Straub, K., Badecker, W., and Frank, R. (2000). Aspectual coercion and the online computation of sentential aspect. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 22.

Wen, X. (1997). Acquisition of chinese aspect: An analysis of the interlanguage of learners of chinese as a foreign language. *ITL-International Journal of Applied Linguistics*, 117(1):1–26.