Do pretrained transformers infer telicity like humans?

Yiyun Zhao, Jian Gang Ngui, Lucy Hall Hartley
University of Arizona
Department of Linguistics
Tucson, AZ 85721, USA

Steven Bethard
University of Arizona
School of Information
Tucson, AZ 85721, USA

{yiyunzhao,jgngui,lucyhallhartley,bethard}@email.arizona.edu

Abstract

Pretrained transformer-based language models achieve state-of-the-art performance in many NLP tasks, but it is an open question whether the knowledge acquired by the models during pretraining resembles the linguistic knowledge of humans. We present both humans and pretrained transformers with descriptions of events, and measure their preference for telic interpretations (the event has a natural endpoint) or atelic interpretations (the event does not have a natural endpoint). To measure these preferences and determine what factors influence them, we design an English test and a novel-word test that include a variety of linguistic cues (noun phrase quantity, resultative structure, contextual information, temporal units) that bias toward certain interpretations. We find that humans' choice of telicity interpretation is reliably influenced by theoretically-motivated cues, transformer models (BERT and RoBERTa) are influenced by some (though not all) of the cues, and transformer models often rely more heavily on temporal units than humans do.

1 Introduction

Large pretrained-language models (ELMo: Peters et al., 2018a, BERT: Devlin et al., 2019, RoBERTa Liu et al., 2019, etc.) keep achieving new states of the art in a variety of NLP tasks, leading to a growing interest in exploring what has been acquired by the pretraining objectives.

Many recent works utilize *probes*: shallow, usually supervised classifiers that try to determine which linguistic phenomena are predictable from the pretrained representations. The linguistic properties studied include syntactic relationships (Hewitt and Manning, 2019), morphological information (Belinkov et al., 2017), and semantic knowledge and entailment (Peters et al., 2018b; Goodwin et al., 2020), and the representations studied include sentence embeddings (Adi et al., 2017; Conneau et al., 2018; Ettinger et al., 2018) and token/word-level embeddings (Kim et al., 2019; Ethayarajh, 2019). Recent work has also explored unsupervised analysis of attention heads instead of supervised training (Clark et al., 2019; Kovaleva et al., 2019; Zhao and Bethard, 2020).

Another approach to understanding pretrained language models is to test their behavior on psycholinguistic tasks. Stimuli in psycholinguistic tasks are typically designed to reveal linguistic bias in human behaviors (e.g., grammatical judgments, reading speeds, neural responses). Evaluating pretrained language models on such tasks can thus provide insights on the linguistic biases acquired by the models. The linguistic properties studied include subject-verb agreement (Linzen et al., 2016; Gulordava et al., 2018), filler-gap depedencies (Wilcox et al., 2018), garden-path effects (Futrell et al., 2019), other types of syntactic awareness (Marvin and Linzen, 2018; Hu et al., 2020), and variations of grammatical judgment based on availability of contexts (Lau et al., 2020). Most works focus on syntactic knowledge (though see Ettinger (2020)'s work on semantic and pragmatic aspect).

Our work contributes to this line of research. While most previous studies probe syntactic properties, we investigate a semantic property: telicity (whether an event has reached an end point or not). Telicity encoding is covert in English and needs to be inferred from information such as inherent telicity of the main verb, the argument structure, the quantity of noun phrases, and other constituents (Dowty, 1986; Verkuyl, 2013). For instance, the description "John read the book" might allow both telic (he finished reading the book) or atelic (he did not finish reading the book) interpretations, while the description "John reached the park" has only a telic interpretation. There is a rich literature on what linguistic information biases listeners toward which types of interpretations. Thus, our study aims first to test, given a description of an event,

Proceedings of the 25th Conference on Computational Natural Language Learning (CoNLL), pages 72–81
November 10–11, 2021. ©2021 Association for Computational Linguistics

which linguistic structures influence interpretation preference, both for humans and pretrained language models. We test telicity preference using an adverbial phrase task frequently used in semantics that is also well-aligned with the masking task used during model pretraining. This allows us to probe telicity preference without fine-tuning on telicity and is directly comparable to the task we assign to human subjects. It also avoids tokenization issues since the prepositions are single word-pieces. We also extend these English sentence tests to novel-word tests where we replace content words such as verbs and nouns with made-up words guaranteed to be in neither the vocabulary of native English speakers nor the vocabulary of the pretrained models. Novel-word tests have deep roots in psycholinguistic studies and are often used to evaluate linguistic generalizability (Berko, 1958). In our experiments, they allow us to test to what extent linguistic features such as plurality still affect the telicity preferences in novel contexts.

Our findings suggest that humans attend to the cues predicted by linguistic literature in both real and novel conditions, and that pretrained transformers have acquired some but not all of these cues:

1. Humans reliably followed linguistic theory in their attention to the inherent telicity of verbs, quantized noun phrases, resultative structures, and the telicity of the surrounding context.
2. Transformers followed linguistic theory in their attention to the inherent telicity of verbs and to number and non-quantized noun phrases, but were inconsistent for other cues.
3. Humans had stronger effects of the cues, more often passing both our strict decision and soft ranking measures, while transformers mostly passed only our soft ranking measure.
4. Temporal units in the adverbial phrase affected both human and transformers' telicity preferences, with humans relying more on standard linguistic cues of telicity, and transformers relying more on temporal units.

2 Background on Telicity

Telicity is an important semantic property that indicates whether an event has an inherent endpoint (*telos*) or boundary. For instance, the description 'John solved the problem' typically elicits a *telic* interpretation where the event reaches the endpoint when the problem gets resolved. In contrast, the description 'John gazed at the sunset' typically elicits an *atelic* interpretation where there is no clear boundary when the event is completed. The property also plays an important role in temporal inference. For example, when a punctual temporal adverbial phrase modifies an atelic event (e.g., 'John ran at 8 am'), the timestamp is inferred as the inception time of the event, but when it modifies a telic event (e.g., 'John arrived at 8 am'), the timestamp is inferred as the endpoint of the event.

English encodes telicity with both overt grammatical aspect (e.g., progressives or perfectives) and covert situational (lexical) aspect. We focus on the situational aspect of telicity because it is considered the *fundamental aspectual class* (Siegel and McKeown, 2000) and its coding is non-transparent, i.e., it must be constructed by the human or model.

2.1 Time Adverbial Test

One common way to probe telicity preference is to use time adverbial judgments as telicity of events select specific adverbial phrases (Dowty, 1986; Vendler, 1957; Rothstein, 2008). Specifically, readers are presented with sentences (e.g., "John read the book") and asked whether they prefer 'for' or 'in' when adding a time adverbial (e.g., "John read the book for/in 20 minutes"). Here, we consider a preference for "for" as a preference for an atelic interpretation (e.g., John read the book and did not finish it), and a preference for "in" as a preference for a telic interpretation (e.g., John read the book and finished it). Linguistic literature explores various information (verb, quantity of object noun phrases, resultative structures, and contexts) that bias toward one interpretation or the other.

2.2 Verb Type

Vendler (1957) classified verbs into four categories – activity (e.g., 'run' in its intransitive form, 'drive a car'), state (e.g., 'hate', 'be happy'), achievement (e.g., 'recognize', 'die') and accomplishment (e.g., 'recover'). Verbs of activity or state have no logical endpoint, while verbs of achievement and accomplishment do. Thus, events of activity (e.g., 'John ran') and state (e.g., 'John was married') often bias toward atelic interpretations, and events of achievement (e.g., 'John reached Tucson') and accomplishment (e.g., 'John recovered from illness') often bias toward telic interpretations.

2.3 Noun Phrase Quantity

A verb's semantics may interact with the quantity of the verb's direct object noun phrase

(Krifka, 1989; Verkuyl, 2013) to influence the telicity interpretation of an event. Quantized noun phrases in English with overt determiners such as 'the'/'a'/'three'/'many' bias toward telic interpretations (see example 1) whereas non-quantized noun phrases such as mass nouns and bare plurals bias toward atelic interpretations (see example 2).

1. John ran a lap/the lap/three laps/several laps ~~for~~/in two hours. (telic)[1]
2. John ran laps for/~~in~~ two hours. (atelic)

The bias from quantized noun phrases is not absolute. For example, "John ran the lap for 10 minutes" would be plausible under the interpretation that John did not finish running the lap and activity phrases such as 'push (three) carts' or 'wave (the) flags' are generally insensitive to the change of quantity in the object noun phrase.

2.4 Resultative Structure

The resultative construction transforms atelic activities into accomplishment predicates via an adjunct phrase that specifies the state that the verb phrase obtains (Pustejovsky, 1991)[2]. Thus, description of an event with the resultative structure often biases toward telic interpretation.

3. John cut the log for/~~in~~ two hours. (atelic)
4. John cut the log in half ~~for~~/in two hours. (telic)

Example 4 shows that adding the phrase 'in half' introduces the final state of the activity of 'cut the log', leading to a telic preference.

2.5 Context

Surrounding contexts can also affect the telicity interpretation of an event. Filip (2004) points out that a large class of verbs (e.g. 'read', 'drain') alternate between telic and atelic interpretation based on context (see examples 5 and 6).

5. John read the book for/~~in~~ an hour and he has not finished it. (atelic)
6. John read the book ~~for~~/in an hour but Mary has not finished it. (telic)

The sentence 'John read the book' allows both telic and atelic interpretations, but the context leads to a preference of one interpretation over the other.

3 Method

We adapted the time adverbial test (Dowty, 1986; Vendler, 1957; Rothstein, 2008) into forced-choice fill-in-the-blank tests. For human subjects, we ask participants to select "in" or "for" to fill in the blank of a given sentence, e.g.:

John loved Mary ____ 2 years.

a. in

b. for

For pretrained transformers, we used masked-language modeling. Specifically, we replace the preposition with the masking token, e.g.,

John loved Mary [MASK] 2 years

and then compare the probability at the [MASK] token of predicting 'for' vs. predicting 'in', and select the preposition with the higher probability.

3.1 Materials

We designed two sets of tests, an English set and a novel-word set (see table 1).

The English set uses sentences and linguistic features frequently discussed in the literature (Vendler, 1957; Krifka, 1989; Pustejovsky, 1991; van Hout, 1999). The English set will not only provide insights about differences between human performance and model performance, but also assess variability among human preferences against linguistic theory, as studies have found that human judgments are more variable than theoretical claims (Gibson and Fedorenko, 2010).

The novel-word set uses sentence templates designed to target linguistic factors of interest, but uses novel words for verbs and nouns. Novel words are taken from the ARC nonword database[3] with parameters as only orthographically exisiting onsests and bodies that has 4-6 letters and more than 1 phoneme. We also asked two native English speakers to go over the list to make sure the novel words do not associate with existing meanings. This will allow us to separate the influence of other structural/semantic cues from the verb's telicity preference (since novel verbs will have no known telicity) and the event's typical duration (since novel verbs and nouns will have no known duration).

3.2 Participants

The surveys were administered on Qualtrics[4]. 120 native speakers of English (based on their language

[1] For all examples, we underline the event, strikeout the dispreferred (but often still possible) preposition, and note the preferred interpretation (atelic or telic) in parentheses.

[2] Not all adjuncts lead to resultative constructions. For instance, 'John sold the house cheap' is not a resultative construction as the adjunct 'cheap' modifies the manner of activity 'selling' rather than denoting a result state.

[3] https://www.cogsci.mq.edu.au/research/resources/nwdb/nwdb.html

[4] https://www.qualtrics.com/

	English sentences		Novel word templates	
Factor	Number	Example	Number	Example
Typical Verb	33	John swam { } 30 minutes. John died { } 10 minutes.		
Noun Phrase	30	John built houses { } 10 months. John built three houses { } 10 months.	12	Mary did not [V] a single [N] but John [V+ed] [N+s]{ } 2 hours. Mary did not [V] a single [N] but John [V+ed] three [N+s] { } 2 hours.
Resultative	18	John hammered the metal { } 2 hours. John hammered the metal flat { } 2 hours.	20	John [V+ed] it again and again. He [V+ed] it empty { } 5 hours. John [V+ed] it again and again. He [V+ed] it { } 5 hours.
Context	18	John baked the cake { } an hour and he is still baking it. John baked the cake{ } an hour which is faster than he expected.	8	John [V+ed] { } it 5 seconds but Mary has not finished it. John [V+ed] it { } 5 seconds but he has not finished it.
Transitivity			12	John [V+ed] it { } 5 years. John [V+ed] him { } 5 years. John [V+ed] { } 5 years.

Table 1: Number of test stimuli and example sentences for English and novel-word telicity tests

profile) were recruited via Prolific[5] for the two sets (60 for each). Participants were paid $2.80 for the 25-minute English survey and $1.80 for the 15-minute novel-word survey. We filtered out participants who failed to achieve 90% accuracy on filler questions (non-ambiguous forced-choice questions such as 'John grew roses in/~~on~~ his garden') or did not complete the survey in time, resulting in 59 participants for the English set and 60 participants for the novel-word set.

We tested our surveys on BERT uncased (base and large) and RoBERTa (base and large) models. All four models are pretrained transformer models and are frequently used in the NLP field.

3.3 Analysis

To investigate whether humans and transformers attend to the aforementioned linguistic cues, we compare their performance under three measures.

Firstly, we used a strict decision measure: we ask whether the use of "for" is significantly different from 50% via Wilcoxon sign-rank tests and visualized responses using violin plots[6] for each category of interest. The error bars indicate the 95% bootstrapped confidence interval over the mean.

Secondly, we used a soft ranking measure: we ask whether humans or transformers showed a theoretically-motivated tendency (i.e., significantly greater preference for "for" in items expected to have an atelic preference than in items expected to have a telic preference) by fitting Bayesian mixed effect logistic regression models[7] to predict telicity preference from linguistic cues and temporal unit features. We included a random subject intercept for human participants. In novel word sets for transformers, a random item intercept is included to obtain stable estimates as we repeated many items with only the novel word changed. Features are sliding contrast coded[8]. This coding schema enables the comparison between the mean of predicted variable on one level to the mean of the previous level so that we can see the ranking between features. For instance, if we contrast coded the time unit factor (second, hour, week, year) then, it means the model would compare seconds to hours, hours to weeks and weeks to years. Tables for these tests include one column of coefficients and p-values for each contrast, i.e., for each pairwise comparison.

Lastly, we used ablations to evaluate which cues (theoretically-motivated or temporal unit cues) contribute more to explaining participants' responses by comparing the residual change after removing related features via ANOVA tests. Tables for these ablations include one column of residuals and p-values for removal of the structural features and one column for removal of the time unit features.

[5] https://www.prolific.co/

[6] For the English set, where each transformer produces only a few scores per category, we plot the model's mean prediction as a letter, overlaid on the human violin plot.

[7] We used the R blme package and added a normal distribution prior when there were linear separation problems.

[8] https://github.com/burchill/zplyr/blob/master/R/sliding_contrast.R

	second vs hour		hour vs week		week vs year	
Subject	Coef	P-value	Coef	P-value	Coef	P-value
Human	0.15	1.45×10^{-9}	0.03	0.237	0.11	2.53×10^{-5}
BERT-base	0.31	$<2.0 \times 10^{-16}$	0.34	$<2.0 \times 10^{-16}$	0.09	5.53×10^{-4}
BERT-large	0.25	$<2.0 \times 10^{-16}$	0.06	3.06×10^{-3}	0.18	$<2.0 \times 10^{-16}$
RoBERTa-base	0.39	$<2.0 \times 10^{-16}$	0.11	1.65×10^{-8}	0.10	2.31×10^{-7}
RoBERTa-large	0.28	$<2.0 \times 10^{-16}$	0.13	4.58×10^{-9}	0.06	0.0138

Table 2: Influence of temporal units on telicity preference among human subjects and transformers: results of the soft ranking measure described in section 3.3.

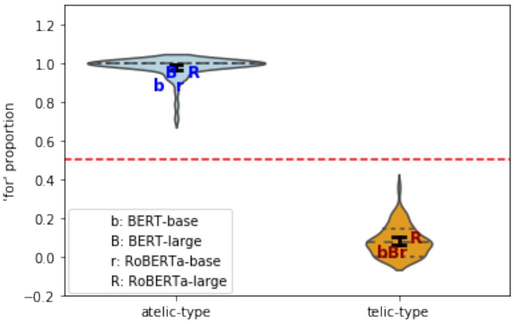

Figure 1: Influence of verb type on telicity preference among human subjects and transformers: visualization of the strict decision measure described in section 3.3.

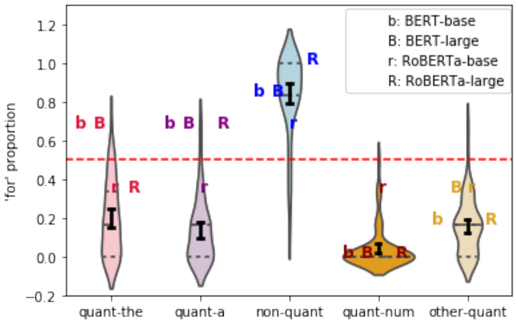

Figure 2: Influence of noun phrase quantity on telicity preference among human subjects and transformers in the English set: visualization of the strict decision measure described in section 3.3.

4 Results

4.1 Verb Type

The linguistics literature suggests that certain verbs have a strong telicity preference (e.g., 'swam' is strongly atelic and 'died' is strongly telic). This implies that descriptions of events whose telicity is dominated by a strongly telic verb should see more use of 'in', whereas those dominated by a strongly atelic verb should see more use of 'for'.

Figure 1 shows that humans follow these predictions under our strict decision measure: preference for 'for' with atelic verbs was higher than 50% (p = 4.65×10^{-13}) and preference for 'for' with telic verbs was lower than 50% (p = 1.38×10^{-11}). Pretrained transformers (marked by letters in fig. 1) have similar preferences, indicating that they can reflect the inherent telicity of common verbs.

4.2 Temporal units

Linguistic literature has generally not considered temporal units (seconds, hours, etc.) to be an important factor for telicity calculation. But temporal units are pragmatically important: achievements happen instantly while states and activities last longer. Preliminary experiments suggested temporal units might impact telicity judgments, so we varied temporal units (seconds, hours, weeks, years) in all novel word sentences.

Table 2 shows that the preference for an atelic interpretation, i.e., the preference for 'for', increases as the temporal units get larger for both humans and transformers as all coefficients are positive and significant for humans and transformers, except that human participants did not show a significant difference between hour and week.

Given that temporal units predict human and transformer judgments of telicity, we include temporal unit as a factor in all mixed effect and ANOVA models in the following sections.

4.3 Noun Phrase Quantity

The linguistics literature suggests that quantized noun phrases lead to a greater preference for telic interpretations than non-quantized noun phrases.

Figure 2 shows that in the English set, humans follow these predictions, consistently passing our strict decision measure: preference for 'for' in all quantized NPs was lower than 50% ('a': p=2.51×10^{-11}; 'the': p=3.12×10^{-10}; cardinal numbers: p=3.50×10^{-13}; other quantifiers: p=1.70×10^{-11}), while non-quantized NPs were significantly higher than 50% (p=1.86×10^{-10}).

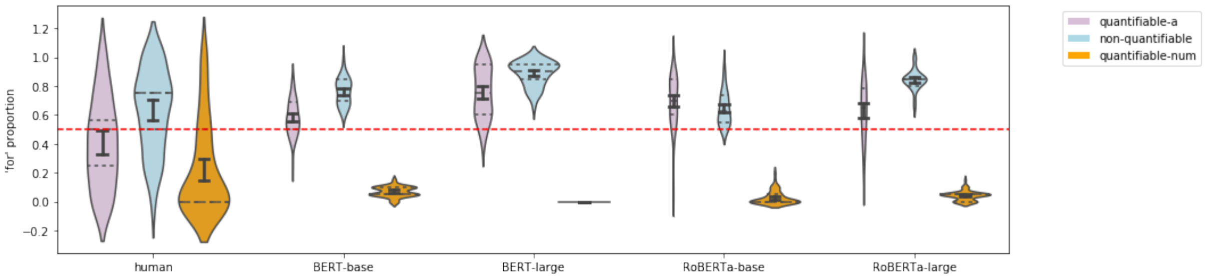

Figure 3: Influence of noun phrase quantity on telicity preference among human subjects and transformers in the novel word set: visualization of the strict decision measure described in section 3.3.

	Soft ranking measure				Ablation			
	NP-non vs NP-a		NP-a vs NP-num		− NP quantity		− time unit	
Subject	Coef	P-value	Coef	P-value	Resid	P-value	Resid	P-value
Human	-0.45	7.80×10^{-9}	-0.44	2.00×10^{-7}	**119.78**	$<2.2 \times 10^{-16}$	71.741	1.80×10^{-15}
BERT-base	-0.87	$< 2.0 \times 10^{-16}$	-2.37	$<2.0 \times 10^{-16}$	**2226.7**	$<2.2 \times 10^{-16}$	1756.4	$<2.2 \times 10^{-16}$
BERT-large	-0.56	$<2.0 \times 10^{-16}$	-4.78	6.23×10^{-13}	**3168.2**	$<2.2 \times 10^{-16}$	682.22	$<2.2 \times 10^{-16}$
RoBERTa-base	0.13	5.08×10^{-4}	-2.03	$<2.0 \times 10^{-16}$	**1938.1**	$<2.2 \times 10^{-16}$	774.18	$<2.2 \times 10^{-16}$
RoBERTa-large	-0.65	$<2.0 \times 10^{-16}$	-1.83	$<2.0 \times 10^{-16}$	**2328.3**	$<2.2 \times 10^{-16}$	667.72	$<2.2 \times 10^{-16}$

Table 3: Influence of noun phrase quantity on telicity preference among human subjects and transformers in the novel word set: results of the soft ranking measure and ablation described in section 3.3.

The pretrained transformers (marked by letters in fig. 2) have similar preferences for non-quantized NPs, cardinal numbers, and other quantifiers, but overuse 'for' in the quantized 'the' and 'a'.

Figure 3 shows that in the novel word set where there is no inherent verb telicity, humans also follow the predictions, again consistently passing our strict decision measure: preference for 'for' was lower than 50% with both 'a' and cardinal number quantized NPs ('a': p=0.028; cardinal number: p=3.56×10^{-7}) and was higher than 50% with non-quantized NPs (p=8.39×10^{-4}). Transformers partially follow the predictions in the novel word set, consistently passing our strict decision measure for cardinal number quantized NPs (p<2.2×10^{-16} for all four models) and non-quantized NPs (BERT-base: p<2.2×10^{-16}; BERT-large: p<2.2×10^{-16}; RoBERTa-base: p=4.89×10^{-10}; RoBERTa-large: p<2.2×10^{-16}), but preferring an atelic interpretation with 'a' quantized NPs, where preference for 'for' was higher than 50% (BERT-base: p=9.81×10^{-14}; BERT-large: p<2.2×10^{-16}; RoBERTa-base: p<2.2×10^{-16}; RoBERTa-large: p=1.83×10^{-11}). Transformers do loosely follow the theoretical predictions under our soft ranking measure: Table 3 shows that they rank non-quantized NPs > quantized 'a' NPs > quantized cardinal number NPs (all coefficients of the mixed effect models are negative, except for RoBERTa-base's ranking of non-quantized NPs vs. 'a' NPs).

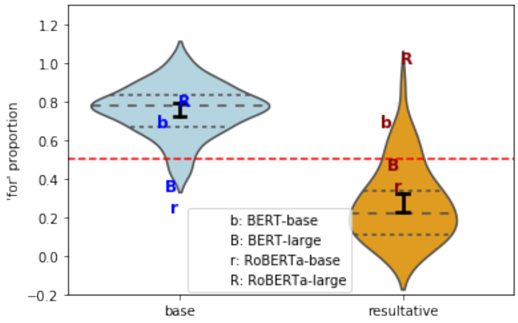

Figure 4: Influence of resultative construction on telicity preference among human subjects and other transformers in the English set: visualization of the strict decision measure described in section 3.3.

Table 3's ablation shows that on the novel word set, both humans and transformers rely more on the noun phrase quantity than the temporal unit: the residuals are greater without the quantity variables than without the time unit variables.

4.4 Resultative Structure

The linguistic literature suggests that the presence of a resultative leads to a greater preference for telic interpretations, i.e., a preference for 'in'.

Figure 4 shows that humans follow these predictions in the English set, consistently passing our

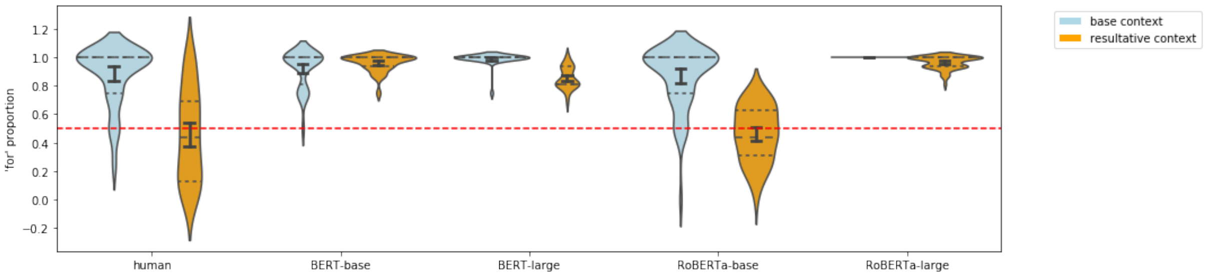

Figure 5: Influence of resultative construction on telicity preference among human subjects and other transformers in the novel word set: visualization of the strict decision measure described in section 3.3.

	Soft ranking measure		Ablation			
	base vs resultative		− base/resultative		− time unit	
Subject	Coef	P-value	Resid	P-value	Resid	P-value
Human	-1.69	<2.0 × 10^{-16}	**235.17**	<2.2 × 10^{-16}	63.35	1.30 × 10^{-13}
BERT-base	0.46	0.0094	6.17	0.01299	**87.96**	<2.2 × 10^{-16}
BERT-large	-1.45	1.16 × 10^{-5}	50.6	1.13 × 10^{-12}	**97.51**	<2.2 × 10^{-16}
RoBERTa-base	-1.51	<2 × 10^{-16}	195.63	<2.2 × 10^{-16}	**250.60**	<2.2 × 10^{-16}
RoBERTa-large	-1.49	0.0361	13.62	0.0002235	**60.00**	5.87 × 10^{-13}

Table 4: Influence of resultative construction on telicity preference among human subjects and other transformers in the novel word set: results of the soft ranking measure and ablation described in section 3.3.

strict decision measure: preference for 'for' with resultatives was lower than 50% (p=2.98 × 10^{-9}). In contrast, transformers are close to or higher than 50% for these resultatives.

Figure 5 shows that in the novel word set, humans do not pass our strict decision measure: preference for 'for' with resultatives did not differ from 50% (p=0.3271). RoBERTa-base is similar to humans (p=0.1997). All other transformers strongly prefer 'for', the opposite of the theoretical prediction (BERT-base: p=8.33 × 10^{-12}; BERT-large: p=1.72 × 10^{-11}, RoBERTa-large: p=9.82 × 10^{-12}). Humans and transformers except BERT-base do loosely follow the theoretical predictions under our soft ranking measure: Table 4 shows that they have a lower preference for 'for' in a resultative context than in the base context (the coefficients of the mixed effects models are negative).

Table 4's ablation shows that on the novel word set, humans relied more on resultative structure, while transfomers relied more on time unit.

4.5 Context

The linguistics literature suggests that additional context expressing telicity will affect interpretation: with a telic context 'in' should be preferred while with an atelic context 'for' should be preferred.

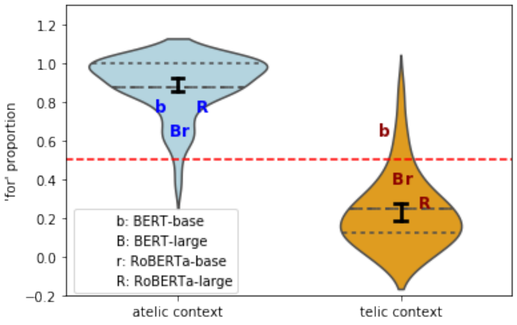

Figure 6: Influence of context information on telicity preference among human subjects and other transformers in the English set: visualization of the strict decision measure described in section 3.3.

Figure 6 shows that in the English set, humans follow these predictions, consistently passing our strict decision measure: preference for 'for' was lower than 50% with telic contexts (p=1.54 × 10^{-9}) and higher than 50% with atelic contexts (p = 1.74 × 10^{-11}). Most pretrained transformers demonstrated a similar tendency.

Figure 7 shows that in the novel word set, humans also follow the predictions, consistently passing our strict decision measure: preference for 'for' was lower than 50% with telic contexts (p=7.58 × 10^{-4}) and higher than 50% with atelic contexts (p=7.42 × 10^{-11}). Transformers did not pass our strict decision measure in the novel word set, with all models either indistiguishable from or higher than 50% (BERT-base: p=2.15 × 10^{-8};

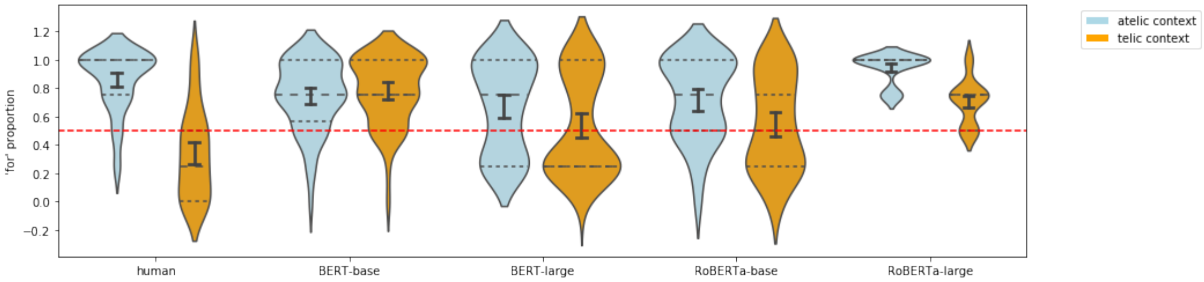

Figure 7: Influence of context information on telicity preference among human subjects and other transformers in the novel word set: visualization of the strict decision measure described in section 3.3.

	Soft ranking measure		Ablation			
	atelic context vs telic context		− context		− time unit	
Subject	Coef	P-value	Resid	P-value	Resid	P-value
Human	-2.87	$<2.0 \times 10^{-16}$	**155.46**	$<2.2 \times 10^{-16}$	7.74	0.05168
BERT-base	2.23	0.0169	9.85	0.001694	**286.99**	$<2.2 \times 10^{-16}$
BERT-large	-2.63	9.41×10^{-7}	35.72	$<2.28 \times 10^{-9}$	**249.19**	$<2.2 \times 10^{-16}$
RoBERTa-base	-2.98	1.15×10^{-8}	53.75	$<2.27 \times 10^{-13}$	**213.5**	$<2.2 \times 10^{-16}$
RoBERTa-large	-4.04	7.04×10^{-10}	86.19	$<2.2 \times 10^{-16}$	**210.55**	$<2.2 \times 10^{-16}$

Table 5: Influence of context information on telicity preference among human subjects and other transformers: results of the soft ranking and ablation measures described in section 3.3.

RoBERTa-large: p=1.93×10^{-9}; BERT-large: p=0.241; RoBERTa-base:p=0.203). Transformers except BERT-base do loosely follow the theoretical predictions under our soft ranking measure: Table 5 shows that they have a higher preference for 'for' in atelic contexts than in telic contexts.

Table 5's ablation shows that on the novel word set, humans relied more on context, while transformers relied more on the time unit.

5 Discussion

We tested how telicity interpretation preferences among human subjects and pre-trained transformers are influenced by a variety of linguistic information—verb type, noun phrase quantity, resultative structure, and contextual information—under the assumption that the preposition choice ('in'/'for') in time adverbial phrases reflects their intepretation preference (telic or atelic) for a given description of an event. Survey results can be found in https://github.com/yiyunzhao/telicity-probing.git

We tested on both natural English and novel-word sets. We found that humans reliably used all theoretically-motivated cues with both natural English and novel predicates, passing our strict decision measure in all cases except resultatives with novel predicates (and there they at least passed our soft ranking measure). We found that transformers followed humans in being highly sensitive to verbs with a strong telicity preference, and to number and non-quantized object noun phrases. However, for most other theoretically motivated cues, transformer sensitivity was weaker, and they passed only our soft ranking measure, not our strict decision measure, with resultatives being especially challenging. We also found that both human and transformers' preferences are affected by temporal units, with humans relying on this cue less than other cues but transformers relying on it heavily.

While time adverbial tests are easy to pose to both humans and pre-trained transformers, these tests have some limitations. Though choice of preposition hints at the participant's interpretation, we do not see the participant's exact mental picture. For example, we found unexpectedly variable responses to some well cited stimuli, e.g., 52.54% of participants preferred 'for' in 'John died in/for 10 minutes', a standard example where 'in' is expected. For human participants, it would be useful to ask those who preferred 'John died for 10 minutes' to describe the specific scene in their mind: perhaps they conceived of a specific scenario where John died but was resuscitated. For transformers, eliciting such descriptions is an open area for future research.

Because the adverbial test aligns so well with the task on which the transformer models were pretrained, one might worry that it overestimates the understanding of the transformer models as they could perhaps solve the fill-in-the-blank problem without really understanding telicity. This is why we individually tested several different types of telicity cues, and indeed found that transformers failed to use many of these cues as well as humans. Still, future work could explore temporal inference and counting tests for additional insights. For example, one could ask "How long does it take John to finish the book" in the sentence "John read a book in an hour" (≤1 hour) vs. "John read a book for an hour" (≥1 hour). Extracting such temporal inferences from pre-trained models is an interesting direction for future research.

Acknowledgements

Research reported in this publication was supported by the National Library of Medicine of the National Institutes of Health under Award Number R01LM010090. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

References

Yossi Adi, Einat Kermany, Yonatan Belinkov, Ofer Lavi, and Yoav Goldberg. 2017. Fine-grained analysis of sentence embeddings using auxiliary prediction tasks. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.

Yonatan Belinkov, Nadir Durrani, Fahim Dalvi, Hassan Sajjad, and James Glass. 2017. What do neural machine translation models learn about morphology? In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 861–872, Vancouver, Canada. Association for Computational Linguistics.

Jean Berko. 1958. The child's learning of english morphology. *Word*, 14(2-3):150–177.

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning. 2019. What does BERT look at? An analysis of BERT's attention. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 276–286, Florence, Italy. Association for Computational Linguistics.

Alexis Conneau, German Kruszewski, Guillaume Lample, Loïc Barrault, and Marco Baroni. 2018. What you can cram into a single $&!#* vector: Probing sentence embeddings for linguistic properties. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2126–2136, Melbourne, Australia. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

David R. Dowty. 1986. The effects of aspectual class on the temporal structure of discourse: Semantics or pragmatics? *Linguistics and Philosophy*, 9(1):37–61.

Kawin Ethayarajh. 2019. How contextual are contextualized word representations? Comparing the geometry of BERT, ELMo, and GPT-2 embeddings. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 55–65, Hong Kong, China. Association for Computational Linguistics.

Allyson Ettinger. 2020. What BERT is not: Lessons from a new suite of psycholinguistic diagnostics for language models. *Transactions of the Association for Computational Linguistics*, 8:34–48.

Allyson Ettinger, Ahmed Elgohary, Colin Phillips, and Philip Resnik. 2018. Assessing composition in sentence vector representations. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1790–1801, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Hana Filip. 2004. The telicity parameter revisited. In *Semantics and Linguistic Theory*, volume 14, pages 92–109.

Richard Futrell, Ethan Wilcox, Takashi Morita, Peng Qian, Miguel Ballesteros, and Roger Levy. 2019. Neural language models as psycholinguistic subjects: Representations of syntactic state. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 32–42, Minneapolis, Minnesota. Association for Computational Linguistics.

Edward Gibson and Evelina Fedorenko. 2010. Weak quantitative standards in linguistics research. *Trends in Cognitive Sciences*, 14(6):233.

Emily Goodwin, Koustuv Sinha, and Timothy J. O'Donnell. 2020. Probing linguistic systematicity. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1958–1969, Online. Association for Computational Linguistics.

Kristina Gulordava, Piotr Bojanowski, Edouard Grave, Tal Linzen, and Marco Baroni. 2018. Colorless green recurrent networks dream hierarchically. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1195–1205, New Orleans, Louisiana. Association for Computational Linguistics.

John Hewitt and Christopher D. Manning. 2019. A structural probe for finding syntax in word representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4129–4138, Minneapolis, Minnesota. Association for Computational Linguistics.

Angeliek van Hout. 1999. *Event Semantics in the Lexicon-syntax Interface: Verb Frame Alternations in Dutch and their Acquisition*. Utrecht Institute of Linguistics (OTS), Utrecht University.

Jennifer Hu, Jon Gauthier, Peng Qian, Ethan Wilcox, and Roger Levy. 2020. A systematic assessment of syntactic generalization in neural language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1725–1744, Online. Association for Computational Linguistics.

Najoung Kim, Roma Patel, Adam Poliak, Patrick Xia, Alex Wang, Tom McCoy, Ian Tenney, Alexis Ross, Tal Linzen, Benjamin Van Durme, Samuel R. Bowman, and Ellie Pavlick. 2019. Probing what different NLP tasks teach machines about function word comprehension. In *Proceedings of the Eighth Joint Conference on Lexical and Computational Semantics (*SEM 2019)*, pages 235–249, Minneapolis, Minnesota. Association for Computational Linguistics.

Olga Kovaleva, Alexey Romanov, Anna Rogers, and Anna Rumshisky. 2019. Revealing the dark secrets of BERT. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4365–4374, Hong Kong, China. Association for Computational Linguistics.

Manfred Krifka. 1989. Nominal reference, temporal constitution and quantification in event semantics. *Semantics and Contextual Expression*, 75:115.

Jey Han Lau, Carlos Armendariz, Shalom Lappin, Matthew Purver, and Chang Shu. 2020. How furiously can colorless green ideas sleep? Sentence acceptability in context. *Transactions of the Association for Computational Linguistics*, 8:296–310.

Tal Linzen, Emmanuel Dupoux, and Yoav Goldberg. 2016. Assessing the ability of LSTMs to learn syntax-sensitive dependencies. *Transactions of the Association for Computational Linguistics*, 4:521–535.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Rebecca Marvin and Tal Linzen. 2018. Targeted syntactic evaluation of language models. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1192–1202, Brussels, Belgium. Association for Computational Linguistics.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018a. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Matthew Peters, Mark Neumann, Luke Zettlemoyer, and Wen-tau Yih. 2018b. Dissecting contextual word embeddings: Architecture and representation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1499–1509, Brussels, Belgium. Association for Computational Linguistics.

James Pustejovsky. 1991. The syntax of event structure. *Cognition*, 41(1-3):47–81.

Susan Rothstein. 2008. *Structuring Events: A Study in the Semantics of Lexical Aspect*, volume 5. John Wiley & Sons.

Eric V. Siegel and Kathleen R. McKeown. 2000. Learning methods to combine linguistic indicators: Improving aspectual classification and revealing linguistic insights. *Computational Linguistics*, 26(4):595–627.

Zeno Vendler. 1957. Verbs and times. *The philosophical review*, 66(2):143–160.

Hendrik Jacob Verkuyl. 2013. *On the Compositional Nature of the Aspects*, volume 15. Springer Science & Business Media.

Ethan Wilcox, Roger Levy, Takashi Morita, and Richard Futrell. 2018. What do RNN language models learn about filler–gap dependencies? In

Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP, pages 211–221, Brussels, Belgium. Association for Computational Linguistics.

Yiyun Zhao and Steven Bethard. 2020. How does BERT's attention change when you fine-tune? An analysis methodology and a case study in negation scope. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4729–4747, Online. Association for Computational Linguistics.