About Time: Do *Transformers* Learn Temporal Verbal Aspect?

Anonymous ACL submission

Abstract

Aspect is a linguistic concept that describes how an action, event, or state of a verb phrase is situated in time. In this paper, we explore whether different transformer models are capable of identifying aspectual features. We focus on two specific aspectual features: telicity and duration. Telicity marks whether the verb's action or state has an endpoint or not (telic/atelic), and duration denotes whether a verb expresses an action (dynamic) or a state (stative). These features are integral to the interpretation of natural language, but also hard to annotate and identify with NLP methods. Our results show that transformer models adequately capture information on telicity and duration in their vectors, even in their pretrained forms, but are somewhat biased with regard to verb tense and word order.

1 Introduction

Aspect is a linguistic concept that characterizes how an action, event, or state (expressed by a verb phrase) relates to time, beyond the scope of the verb's tense; via aspect, information such as frequency, duration, and completion is conveyed. Languages may express aspect in various ways, e.g. by using grammatical verb tense (incomplete actions with continuous/progressive, perfect progressive and imperfect, complete actions with perfect), morphemes (e.g. Finnish, Czech) or with aspect markers (e.g. Mandarin Chinese). However, certain aspectual features are more complex, and cannot simply be deduced from morphosyntax. In this paper, we focus on two of these aspectual features: telicity and duration. **Telicity** is related to the goal-oriented nature of the verb phrase. The verb's action is said to be *telic* if it has an endpoint; when the verb denotes a state, or when the completion of the verb's action is either indefinite, impossible or irrelevant, then the verb phrase is characterized as *atelic*. **Duration** is another aspectual feature, different from telicity: it distinguishes between verbs that describe a state (*stative*) or an action (*durative*) regardless of whether they have a perceived endpoint or not. The perception of telicity and duration is the outcome of the entire verbal phrase, and not solely the verb's features (Krifka, 1998). Besides, the context can also place constraints on the aspectual class of a verb (Siegel, 1998). Therefore, making sound judgments on aspectual features such as telicity and duration, especially in a morphologically-poor language like English, is not always an easy task—our datasets in Section 3.1 and Appendix B provide some examples of sentences where these features are hard to assess, even for a human. Nevertheless, correctly identifying them is indispensable to many natural language processing tasks.

In recent years, transformer-based models have shown great success in NLP tasks which traditionally require in-depth language analysis and complex strategies on capturing dependencies, semantic information, and world knowledge. However, it remains unclear whether the success of these models is due to a genuine capability to accurately model linguistic meaning, or whether the models are just very good at picking up statistical correlations, but fail to capture fine-grained semantic distinctions (Ettinger, 2020). With this research question in mind, our goal is to investigate whether transformer-based architectures (both with and without fine-tuning) are able to capture the semantic information related to telicity and duration. To do so, we make use of two datasets annotated for telicity and duration (Friedrich and Gateva, 2017; Alikhani and Stone, 2019), and we conduct a range of experiments using several pretrained transformer architectures in two languages (English and French). We aim to explore the capabilities of transformer architectures in classifying aspect beyond mere quantitative evaluation: we made custom qualitative datasets in order to observe how complex context, verb tense and prepositional phrases affect classification. We find that classification with fine-tuned models is very successful—both for telicity and duration—but this success can be largely attributed to the knowledge built up during pre-training, as contextual word embeddings by themselves are already quite capable of capturing this information. We noticed that complex cases where the context was conflicting with the verbal aspect were harder for the models to classify, and we provide evidence that misclassification in complex sentences is related to verb tense and word order. Finally, comparing the two languages we investigate, even though the French models show lower accuracy, they were more successful in classifying more difficult cases of telicity and duration, because of the properties of verbal tense in French.

2 Previous Work

Siegel and McKeown (2000) were the first to propose natural language processing methods for aspectual classification; they used decision trees, genetic programming, and logistic regression to locate linguistic indicators of stativity and completeness, and observed that there was an improvement on the classification of these features, especially with supervised methods, compared to uninformed classification.

Friedrich and Palmer (2014) use a semi-supervised approach for learning lexical aspect, combining linguistic and distributional features, in order to predict a verb's stativity/duration, and also released two datasets of annotated sentences for stativity. Friedrich and Pinkal (2015) extended this approach by classifying verbal lexical aspect into multiple categories of duration, habitual/episodic/static, and Friedrich et al. (2016) expanded their datasets and categories, achieving 76% accuracy on supervised classification compared to the 80% of their human baseline. In their most recent work, Friedrich and Gateva (2017) have released two datasets in English with gold and silver annotations of telicity and duration (gold is human annotated; silver is obtained from parallel English–Czech corpora where aspectual features were extracted from Czech morphological markers). With these datasets and a L1-regularized multi-class logistic regression model, they report significant improvement on automatic telicity classification.

Loáiciga and Grisot (2016) exploit telicity in order to improve on French–English machine translation; they are using verb classification of telicity (defined as *boundedness*) and notice improvement on the translation of tense. Falk and Martin (2016) also use a machine learning approach, alongside morpho-syntactic and semantic annotations, to predict the aspect of French verbs in different contexts (*verb readings*). Moving away from hard-coded annotations and lexical aspect, Peng (2018) uses two different compositional models to classify aspect, exploring the entire clause and not only the verb, with the use of distributional vectors and without annotated linguistic features, and highlights the importance of the verbal phrase and the verb's dependents in the interpretation of telicity. Kober et al. (2020) propose modeling aspect of English verbs in context, with the use of compositional distributional models, and confirm that a verb's context and closed-class words of tense are strong features for aspect classification.

3 Methodology

3.1 Datasets

Telicity and duration-annotated sentences will be used as two separate datasets for our experiments. The two datasets from which we are sourcing sentences are constructed by Friedrich and Gateva (2017) and by Alikhani and Stone (2019), who have created datasets in the scope of their work.

Friedrich and Gateva's dataset[1] includes gold- and silver-annotations of telicity (telic/atelic) and duration (stative/durative). The gold annotations are based on the MASC dataset (Ide et al., 2008), while the silver annotations were crafted on the basis of the InterCorp parallel corpus of English and Czech (Čermák and Rosen, 2012), extracting the annotations from the Czech morphological markers of telicity and duration and applying them to the English translations. Each annotation corresponds to a specific verb in each sentence and not the entire clause.

The "Captions" dataset[2] by Alikhani and Stone (2019) was created from five image–text corpora, in order to study inferential connections in sentences. It has been annotated for telicity (telic/atelic) and duration (stative/durative/punctual) based on the verb's aspect. Even though the focus of the original work was on the head verb of each sentence, the verbs were not separately annotated, therefore we used dependency parsing with spaCy (Honnibal

[1] https://github.com/annefried/telicity

[2] https://github.com/malihealikhani/Captions

et al., 2020) in order to extract the verb and its position for our experiments. We noticed some inconsistencies in annotation, which we corrected, and we also excluded the sentences annotated with the *punctual* label, since this label did not exist in Friedrich and Gateva's dataset.

In Table 1 we present the sizes of the datasets and our final dataset. We split this dataset in training, validation and test sets with a ratio of 80-10-10%.

We also created some smaller datasets for testing purposes, in order to observe specific phenomena in our models. First, we created forty sentences annotated for telicity, and forty for duration, a sample of which can be found in Table 2. We also crafted additional sentences on telicity in minimal pairs, where each pair includes the same verb but in a context that has a different degree of telicity (see examples in Table 3). We also created variations for some of these sentences, moving prepositional phrases to different positions in the sentence or changing the verb tense without changing the meaning or the degree of telicity, in order to test whether the models are sensitive not only to specific verbs but also word position and tenses (see Table 4). For the sake of transparency and reproducibility, these datasets are presented in full in Appendix B.

3.2 Verb position

Aspect is generally attributed to the verb; we therefore wanted to mark the position of the verb in the sentence. To do so, we made use of `token_type_ids` vectors to specify the position of the verb form without auxiliaries (or multiple positions, when the verb is split into subwords by the model tokenizer). An example is shown in Table 5. Unfortunately, RoBERTa based models (`roberta` and `camembert`) do not support the use of `token_type_ids` vectors, therefore they will only be used without explicit verb position.

Type	Label	Friedrich	Captions	Current	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	-	

Table 1: Number of sentences and annotations in each dataset, and our final dataset sizes.

label	sentence
telic	I **ate** a fish for lunch .
telic	John **built** a house in a year .
telic	The cat **drank** all the milk .
atelic	John **watched** TV .
atelic	I always **spill** milk when I pour it in my mug .
atelic	Cork **floats** on water.
stative	Bread **consists** of flour, water and yeast.
stative	This box **contains** a cake.
stative	I have **disliked** mushrooms for years.
durative	She **plays** tennis every Friday.
durative	The snow **melts** every spring.
durative	The boxer is **hitting** his opponent.

Table 2: A sample from our qualitative dataset.

label	sentence
telic	I will **receive** new stock on Friday.
atelic	I will **receive** new stock on Fridays.
telic	The boy is **eating** an apple.
atelic	The boy is **eating** apples.
telic	I **drank** the whole bottle.
atelic	I **drank** juice.
telic	The Prime Minister **made** that declaration yesterday.
atelic	The Prime Minister **made** that declaration for months.

Table 3: A sample of minimal pairs for telicity.

label	sentence
telic	John **built** a house in a year.
telic	John had **built** a house in a year.
telic	In a year, John had **built** a house.
atelic	We **swim** in the lake in the afternoons.
atelic	We **swim** in the lake each afternoon.
atelic	In the afternoons , we **swim** in the lake.
atelic	Each afternoon , we **swim** in the lake .

Table 4: A sample of sentence variations for specific phenomena.

tokens	He	**worked**	well	and	earned	much	.		
vector	`0`	`1`	`0`	`0`	`0`	`0`	`0`		
tokens	He	**work**	**###ed**	well	and	earn	###ed	much	.
vector	`0`	`1`	`1`	`0`	`0`	`0`	`0`	`0`	`0`

Table 5: Sentence tokens and the corresponding `token_type_ids` vectors, depending on tokenization. Each sequence also includes the model's special tokens and padding.

3.3 Transformer models

Transformers are neural network models which assign weighted attention to the different parts of the input with a sequence of alternating neural feedforward layers and self-attention layers. These models have proven to be very successful in a variety of NLP tasks, and they have been shown to implicitly capture syntactic and semantic information and dependencies.

BERT (Devlin et al., 2019) is a transformer-based bi-directional encoder, which is trained by randomly masking words in the input sequence and learning to fill the word in the masked position, while also learning to predict the next sentence given the first sentence.

RoBERTa (Liu et al., 2019) has the same model architecture as BERT, but focuses only on the language masking modeling objective, and expands BERT's use of subwords from unseen words to almost all tokens. The model modifies key hyper-parameters in BERT, has been trained with much larger mini-batches and learning rates, and has improved results on the masked language modeling objective and on downstream task performance.

XLNet (Yang et al., 2019) is an auto-regressive pretraining model which introduces permutation language modeling, where all tokens are predicted but in random order (unlike BERT, which predicts only the masked tokens). This method allows the model to better learn dependencies and relations between words. XLNet reportedly outperforms BERT on tasks such as question answering, natural language inference, sentiment analysis, and document ranking.

ALBERT (Lan et al., 2019) is a transformer architecture, based on BERT but using fewer parameters more efficiently; the vocabulary is decomposed into two small matrices and the size of the hidden layer embeddings (which learn context-dependent representations) is separated from the vocabulary embeddings (which learn context-independent representations). ALBERT has managed to outperform BERT on tasks such as reading comprehension, proving that better exploitation of contextual representations could be more beneficial than larger training and parameter sizes.

In Table 6 we are listing the pretrained models we used. We made use of the implementations provided by the `transformers` library (Wolf et al., 2020).

3.4 Fine-tuning

One of our experiments explores the process of fine-tuning a transformer model for binary sequence classification of telicity and duration (separately), and testing the fine-tuned model's accuracy on predicting the telicity or duration annotated label of a sentence. Fine-tuning is the strategy of adapting a pretrained model to a specific task, by adding an extra layer on top of the existing ones and specializing it on the given task. Thus, we can exploit the existing model's knowledge from its contextual word embeddings, and further specialize the model on a specific task without the need for large specialized resources, large computational power and long training times; in many tasks, fine-tuned transformer models have consistently provided state-of-the-art results (Sun et al., 2019).

Model	Lang.	Layers	Embed.	Hidden	Heads	Param.
bert-base-cased	EN	12	-	768	12	109M
bert-base-uncased	EN	12	-	768	12	110M
bert-large-cased	EN	24	-	1024	16	335M
bert-large-uncased	EN	24	-	1024	16	336M
roberta-base	EN	12	-	768	12	125M
roberta-large	EN	24	-	1024	16	355M
xlnet-base-cased	EN	12	-	768	12	110M
xlnet-large-cased	EN	24	-	1024	16	340M
albert-base-v2	EN	12	128	768	12	11M
albert-large-v2	EN	24	128	1024	16	17M
camembert-base	FR	12	-	768	12	110M
camembert-large	FR	24	-	1024	16	335M
flaubert-small-cased	FR	6	-	512	8	54M
flaubert-base-uncased	FR	12	-	768	12	137M
flaubert-base-cased	FR	12	-	768	12	138M
flaubert-large-cased	FR	24	-	1024	16	373M

Table 6: The pretrained models we used in our experiments.

We fine-tune the models as Devlin et al. (2019) have recommended, with some modifications; we use a batch size of 32 and a learning rate of 2×10^{-5}. We apply dropout with probability p = 0.1 and weight decay with $\lambda = 0.01$. We use the PyTorch's ADAM as our optimizer (AdamW) without bias correction. We fine-tune each model for a maximum of 4 epochs, following the recommendation of Devlin et al. (2019) to train for 2-4 epochs when fine-tuning on a specific task. For `base` models each training epoch took ~3 minutes and for `large` models ~7 minutes, using a single GPU.

As baselines, we make use of two standard binary classification models trained and tested on the same sets: a simple bag-of-words logistic regression model, implemented with the Python library *scikit-learn* (Pedregosa et al., 2011) with default parameters and data scaling, and a one-dimensional convolutional neural network model (CNN) implemented with `Pytorch` (Paszke et al., 2019) and trained for 50 epochs, which is commonly used for text classification tasks (Kim, 2014). The CNN model is trained with the fastText 300-dimensional embeddings (Bojanowski et al., 2017), embedding dimension of 300, filter size of $[3, 4, 5]$, 100 filters per dimension, dropout rate of 0.5, learning rate of 0.01 and the Adadelta optimizer.

Next to a quantitative evaluation, we make use of our qualitative test sets for an in-depth investigation of predictions for specific cases, such as verb tenses and word position, by examining the probability distribution of the predicted labels. We equally visualize which tokens the attention mechanism focuses on in a sentence, in order to observe how the context is interpreted and attended to by the model—based on previous work by Clark et al. (2019) and Subudhi (2019).

3.5 Classification with layer embeddings and logistic regression

Pretrained models already contain linguistic information in their contextualized word embeddings, which we can extract and use with task-specific models for classification. The process of extracting the knowledge of a transformer model's embeddings has been explored since the popularization of contextual word embeddings with ELMo (Peters et al., 2018), since it allows for faster computations with results comparable to fine-tuned transformer models (Tang et al., 2019). We equally conduct an experiment without any finetuning, where we apply a logistic regression to the contextual embeddings of each layer as provided by the pre-trained model. We extract the contextual word embeddings (for the annotated verb) from each layer of a transformer model, and we train a logistic regression model (using `scikit-learn`) to classify telicity and duration, in order to examine how much information relevant to telicity and duration has been learned by each layer.

3.6 Comparing English and French transformer models

We also wanted to examine whether telicity and duration were classifiable in a different language with transformer models. We chose French, as it differs from English in the way verb tenses are formed (conjugation, compound tenses) and used (present continuous is morphologically the same as present simple), but it does not have a dedicated morpheme to expressing telicity such as Finnish and Czech. There are two monolingual French transformer models, FlauBERT (Le et al., 2020) and CamemBERT (Martin et al., 2020) which we can compare to our English models. We translated our datasets of telicity and duration in French, with the DeepL translator[3] and manually reviewed them (with special care for our qualitative test sets). We use the resulting dataset to fine-tune the FlauBERT and CamemBERT models, and assess their abilities on aspectual classification.

4 Results

4.1 Classification accuracy and probabilities

During the fine-tuning process, we were able to identify via validation which models were most and least successful in predicting binary tags. The results for validation are presented in Table 7 for telicity and Table 8 for duration.

On classifying **telicity**, the best performing model was `bert-large-cased`. Overall, BERT models outperformed the other architectures, but all models achieved accuracy of > 0.80. When trained with the extra information of verb position in the sentence, accuracy improved for all models and sets ($+0.01 - 0.04$). Examining the probability distribution of the two labels, we observed that the BERT models, both `base` and `large`, with the use of the verb position, were the most "confident" in assigning a label to a sentence (with the probability of each label being > 0.9) while the `large` versions of other models were the ones whose probability distribution included more cases with lower label probability. In Figure 2 (Appendix A.1) we are comparing the probability distributions for the most and least successful model in terms of accuracy.

Our findings on classifying **duration** were similar to the ones on telicity, with the models performing overall better on this classification task despite the dataset being smaller. The BERT models were the most successful ones, achieving accuracy of up to 0.96, however all models achieved accuracy of > 0.93. The effect of the use of the verb position information is not apparent in this classification task, since we notice an improvement or deterioration of 0.01 in most models. Examining the probability distribution of the two labels, all models were very confident in classifying sentences, regardless of their accuracy. In Figure 3 (Appendix A.1) we are comparing the probability distributions for the most and least successful model in terms of accuracy.

In both cases, the fine-tuned transformers models outperformed the baselines we have established.

[3] `https://www.deepl.com/translator`

Model	Verb	Acc.	Prec.	Rec.	F1
bert-base-uncased	yes	0.86	0.86	0.86	0.86
	no	0.81	0.81	0.81	0.81
bert-base-cased	yes	0.87	0.87	0.87	0.87
	no	0.81	0.80	0.80	0.80
bert-large-uncased	yes	0.86	0.86	0.86	0.86
	no	0.81	0.80	0.80	0.80
bert-large-cased	yes	**0.88**	**0.87**	**0.87**	**0.87**
	no	0.81	0.81	0.80	0.80
roberta-base	no	0.84	0.84	0.84	0.84
roberta-large	no	0.80	0.81	0.79	0.79
xlnet-base-cased	yes	0.82	0.82	0.82	0.82
	no	0.81	0.81	0.81	0.80
xlnet-large-cased	yes	0.82	0.82	0.82	0.82
	no	0.80	0.80	0.80	0.80
albert-base-v2	yes	0.84	0.84	0.84	0.84
	no	0.81	0.80	0.80	0.80
albert-large-v2	yes	0.80	0.80	0.80	0.80
	no	0.82	0.81	0.81	0.81
CNN (50 epochs)	no	0.75	0.75	0.75	0.75
Logistic Regression	no	0.61	0.61	0.61	0.61

Table 7: Results of classification accuracy on the telicity test set. 'Verb' refers to training the model with the added information of the verb position.

Model	Verb	Acc.	Prec.	Rec.	F1
bert-base-uncased	yes	**0.96**	**0.96**	**0.96**	**0.96**
	no	0.94	0.94	0.94	0.94
bert-base-cased	yes	**0.96**	**0.96**	**0.96**	**0.96**
	no	0.96	0.95	0.96	0.96
bert-large-uncased	yes	**0.96**	**0.96**	**0.96**	**0.96**
	no	0.95	0.95	0.94	0.94
bert-large-cased	yes	**0.96**	**0.96**	**0.96**	**0.96**
	no	0.95	0.95	0.95	0.95
roberta-base	no	0.95	0.95	0.95	0.95
roberta-large	no	0.95	0.95	0.95	0.95
xlnet-base-cased	yes	0.94	0.94	0.94	0.94
	no	0.95	0.95	0.95	0.95
xlnet-large-cased	yes	0.94	0.94	0.94	0.94
	no	0.95	0.95	0.95	0.95
albert-base-v2	yes	0.95	0.95	0.95	0.95
	no	0.95	0.95	0.95	0.95
albert-large-v2	yes	**0.96**	**0.96**	**0.96**	**0.96**
	no	**0.96**	**0.96**	**0.96**	**0.96**
CNN (50 epochs)	no	0.88	0.88	0.88	0.88
Logistic Regression	no	0.70	0.70	0.69	0.69

Table 8: Results of classification accuracy on the duration test set. 'Verb' refers to training the model with the added information of the verb position.

4.2 Qualitative analysis

As mentioned before, we also created our own annotated datasets of telicity and duration, in order to study aspectual properties beyond the scope of classification metrics. We took a closer look at the correct and incorrect predictions of the models, in order to determine which cases were easier or more difficult for models to classify. For the sake of brevity, we are presenting only a few examples of successes and failures; our goal was to manually examine the strengths and weaknessess of the models in difficult and conflicting cases of classification, hence the smaller qualitative datasets and the presentation of the most interesting examples.

For **telicity**, overall, models were quite successful in classifying the sentences of our qualitative dataset.For example, all models were able to identify that sentences with statements are atelic, such as *Cork floats on water.* and *The Earth revolves around the Sun.*, and sentences with an action were correctly classified almost all the time: *I spilled the milk.* was correctly classified as *telic*, and *I always spill milk when I pour it in my mug.* was also correctly classified as *atelic* (except for the `xlnet` models).

For the majority of the models, the errors in classification could be located in some specific sentences, where the verb or the verbal phrase would be considered (a)telic, but part of the context defines the temporal aspect of the sentence in the opposite way, either a prepositional phrase (e.g. *I eat a fish for lunch on Fridays.*; *eat* with an object would be considered telic, but the prepositional phrase *on Fridays* shows an action without perceived ending) or a grammatical tense (e.g. *The inspectors are always checking every document very carefully.*; even though the action should have a perceived ending, the continuous tense and the presence of the adverb *always* render this sentence atelic).

Moving to our minimal pairs of telic-atelic sentences, we observe that, in most cases, most models are able to classify correctly a sentence based both on the verb action and the context; *I drank the whole bottle.* and *I drank juice.* were correctly classified as *telic* and *atelic* respectively, despite of the presence of the same verb and tense. However, in our qualitative dataset, we noticed that the sentence *The cat drank all the milk.* was incorrectly classified as *atelic* by all the models. Another interesting mistake we noticed was the classification of the pair *The boy is eating an apple.* and *The boy is eating apples.* as both atelic; in the former sentence, the action is telic for pragmatic reasons (one apple that will be finished), but the tense is continuous.

In order to observe specific tenses, word positions and context more extensively, we can examine the variations of a sentence and see whether the models classified them all with the same label or not. The telic sentence *I ate a fish for lunch at noon.*

has confused some of the models, whether the prepositional phrase *at noon* was at the beginning or the end. However, the same sentences regardless of the phrase's position, with past perfect tense *had eaten* is always classified as *telic*. In some complex cases, such as the sentence *The Prime Minister made that declaration for months.* we notice that most models fail to classify it as *atelic* in all its variations, except for when the prepositional phrase is at the start and the tense is present perfect continuous (*has been making*). We noticed that even sentences with a more obvious degree of telicity (*John Wilkes Booth killed Lincoln on 1865.* - *telic*) were sometimes labeled incorrectly, when the prepositional phrase was at the end rather than the start.

Regarding **duration**, the models were less successful at classifying *stative* sentences than *durative*; even some sentences with intransitive verbs, such as *Bread consists of flour, water and yeast.* were classified as *durative*. However, stative sentences with animate subjects such as *I disagree with you.* were correctly classified. Durative sentences, despite of verb tense and context, were always correctly classified, e.g. *She plays tennis every Friday.* and *She's playing tennis right now.*.

4.3 A look at attention

We notice that, out of the models we used in our experiments, BERT models in earlier layers were the ones that showed more "focused" attention to specific tokens; other models had more "diffused" attention from earlier layers. In the final layers, most tokens attended to all tokens or to the special tokens (start and end of sequence). We were specifically interested in comparing the attention from sentences of our qualitative sets, since we had already extensively studied them. In Figure 4 (Appendix A.2), we are comparing a minimal pair of telicity, on layer 3 of the `bert-base-uncased` model (with information on verb position). We selected earlier layers, because later layers specialized on syntactic dependencies (verb attended to subject and object, prepositional phrase attended to its tokens) and the last layers did not focus on any word tokens (in the datasets we examined in this work). In Figure 5 we present the attention that the verb token attributed to the other tokens of the sentence, for all layers and heads of the `bert-base-cased` model. We notice a tendency of the verb "read" to attend to the preposition "for" more than "in", comparing the two sentences (head 4), but overall the verb prefers to attend to its adjacent words and its stronger syntactic dependencies.

4.4 Layer embeddings

By extracting the contextual word embeddings for the verb of each sentence, from each layer, and training a logistic regression model with these embeddings, we were able to examine how much information on telicity and duration is learned by each layer. In Figure 1 we present the accuracy for each layer of the *base* models. Improvement of accuracy is not proportional as we move to higher layers; we notice that for telicity, some models achieve high accuracy in the middle layers, and again in the final layers, with accuracy sometimes dropping in the last layer.

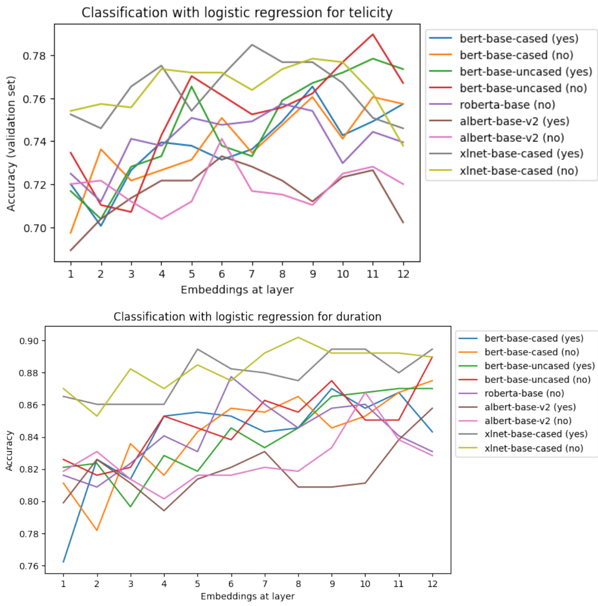


Figure 1: Accuracy of classification of logistic regression, per layer of embeddings, (accuracy on validation set) for `base` models.

4.5 French classification

The results of the classification for telicity and duration are presented in Tables 9 and 10. Accuracy with these datasets and these models is lower than for English and there is no improvement with the use of verb position.However, we notice that these fine-tuned models performed better on the qualitative sets than their English counterparts, avoiding common mistakes such as classifying the atelic sentence *Je mange un poisson à midi le vendredi.* ("I eat a fish for lunch of Fridays.") as telic. We do notice the same mistake in the duration classification, the models failing to classify sentences of world knowledge such as *Le pain est composé de farine, d'eau et de levure.* ("Bread consists of flour, water and yeast.") as stative.

Model	Verb	Acc.	Prec.	Rec.	F1
camembert-base	no	**0.77**	0.77	0.78	0.77
camembert-large	no	0.76	0.77	0.77	0.77
flaubert-small-cased	yes	0.69	0.70	0.70	0.69
	no	0.73	0.73	0.73	0.72
flaubert-base-uncased	yes	0.74	0.75	0.74	0.72
	no	0.76	0.76	0.76	0.75
flaubert-base-cased	yes	0.76	0.76	0.77	0.76
	no	**0.77**	0.78	0.78	0.78
flaubert-large	yes	0.73	0.74	0.74	0.72
	no	0.75	0.76	0.76	0.74

Table 9: Accuracy metrics for telicity classification with French transformer models.

Model	Verb	Acc.	Prec.	Rec.	F1
camembert-base	no	0.82	0.82	0.82	0.82
camembert-large	no	**0.87**	0.87	0.87	0.87
flaubert-small-cased	yes	0.79	0.79	0.79	0.79
	no	0.81	0.81	0.81	0.8
flaubert-base-uncased	yes	0.80	0.81	0.80	0.80
	no	0.84	0.84	0.84	0.84
flaubert-base-cased	yes	0.81	0.82	0.82	0.81
	no	0.83	0.83	0.83	0.83
flaubert-large	yes	0.81	0.81	0.81	0.80
	no	0.87	0.87	0.87	0.87

Table 10: Accuracy metrics for duration classification with French transformer models.

5 Discussion

Transformer models were quite successful in the classification tasks, outperforming our baselines to a large extent, and they proved to be quite successful even without fine-tuning in our experiment in Section 4.4. Contextual embeddings proved to be an efficient way to encode the aspectual information of a verb and its interaction with its context, and this knowledge is probably already learned in the pretraining process. In addition, BERT's self-attention mechanism on earlier layers demonstrated a certain understanding of a sentence's syntax, with more focused attention between the core elements of a sentence, which probably allowed for better processing of the verb's features and its context, compared to RoBERTa and ALBERT models. XLNet models, despite the architecture's reported improved performance on longer dependencies in other NLP tasks, were not able to attend to context more efficiently than BERT or encode more pertinent information in the encodings.

The superior performance of the duration classification with fine-tuned models did raise a question: from our datasets, most stative questions came from the Friedrich dataset and most durative sentences from the Captions dataset; did the models learn to classify duration or to identify the different corpora? With our qualitative analysis on two languages, we can conclude that the models are indeed able to classify duration and were successful because of the little overlap between stative and durative verbs and contexts. However, the models struggled with sentences for which world knowledge is crucial, which is a known issue (Rogers et al., 2021).

From our experiment with verb tenses and prepositional phrases in Section 4.2, we noticed that perfect and continuous tenses are beneficial to classification by the models, and leading a sentence with a prepositional phrase of time sometimes improved predictions. However, infelicitous context will almost always confuse the models. In addition, our findings on the French datasets showed that, even with lower-performing models, the choices that a language makes in expressing aspect did affect the models' capabilities of classifying aspect.

6 Conclusion

In this study, we conducted several experiments that test the capability of transformer models to grasp aspectual categories, viz. telicity and duration. We tested this capability using a binary classification setting. Using two annotated datasets for telicity and duration (Friedrich and Gateva, 2017; Alikhani and Stone, 2019), we fine-tuned transformers models of different architectures and in two languages and found that transformers models were very successful on the classification of aspect even when trained on small datasets. Providing the verb position as additional information improved performance in both telicity and duration classification for English. The pretained transformer models also proved that they possess knowledge of aspect even without fine-tuning, from our experiment in contextual word embeddings per layer. However, our models revealed weaknesses during our qualitative analysis which were not surprising; for infelicitous sentences, where the verbal aspect contradicted the temporal information in the context (e.g. telic verb with an atelic prepositional phrase, resulting in an overall atelic sentence), the models failed.

References

Malihe Alikhani and Matthew Stone. 2019. "Caption" as a Coherence Relation: Evidence and Implications. In *Proceedings of the Second Workshop on Shortcomings in Vision and Language*, pages 58–67.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

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.

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.

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.

Ingrid Falk and Fabienne Martin. 2016. Automatic identification of aspectual classes across verbal readings. In ** Sem 2016 THE FIFTH JOINT CONFERENCE ON LEXICAL AND COMPUTATIONAL SEMANTICS*.

Annemarie Friedrich and Damyana Gateva. 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.

Annemarie Friedrich and Alexis Palmer. 2014. Automatic prediction of aspectual class of verbs in context. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 517–523.

Annemarie Friedrich, Alexis Palmer, and Manfred Pinkal. 2016. Situation entity types: automatic classification of clause-level aspect. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1757–1768.

Annemarie Friedrich and Manfred Pinkal. 2015. Automatic recognition of habituals: a three-way classification of clausal aspect. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2471–2481.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength Natural Language Processing in Python.

Nancy Ide, Collin Baker, Christiane Fellbaum, Charles Fillmore, and Rebecca Passonneau. 2008. MASC: the Manually Annotated Sub-Corpus of American English. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. European Language Resources Association (ELRA).

Yoon Kim. 2014. Convolutional neural networks for sentence classification. *CoRR*, abs/1408.5882.

Thomas Kober, Malihe Alikhani, Matthew Stone, and Mark Steedman. 2020. Aspectuality Across Genre: A Distributional Semantics Approach. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4546–4562, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Manfred Krifka. 1998. The origins of telicity. In *Events and grammar*, pages 197–235. Springer.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. ALBERT: A lite BERT for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*.

Hang Le, Loïc Vial, Jibril Frej, Vincent Segonne, Maximin Coavoux, Benjamin Lecouteux, Alexandre Allauzen, Benoit Crabbé, Laurent Besacier, and Didier Schwab. 2020. Flaubert: Unsupervised language model pre-training for french. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 2479–2490.

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. *arXiv preprint arXiv:1907.11692*.

Sharid Loáiciga and Cristina Grisot. 2016. Predicting and Using a Pragmatic Component of Lexical Aspect of Simple Past Verbal Tenses for Improving english-to-french Machine Translation. In *Linguistic Issues in Language Technology, Volume 13, 2016*. CSLI Publications.

Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric Villemonte De La Clergerie, Djamé Seddah, and Benoît Sagot. 2020. Camembert: a tasty french language model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7203–7219.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward

Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Qiwei Peng. 2018. *Towards aspectual classification of clauses in a large single-domain corpus*. School of Informatics, University of Edinburgh, Edingburgh, UK.

Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of NAACL-HLT*, pages 2227–2237.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2021. A Primer in BERTology: What we know about how BERT works. In *Transactions of the Association for Computational Linguistics*, volume 8, pages 842–866. MIT Press.

Eric V. Siegel and Kathleen R. McKeown. 2000. Learning Methods to Combine Linguistic Indicators:Improving Aspectual Classification and Revealing Linguistic Insights. In *Computational Linguistics*, volume 26, pages 595–627.

Eric Victor Siegel. 1998. *Linguistic Indicators for Language Understanding: Using machine learning methods to combine corpus-based indicators for aspectual classification of clauses*. Columbia University. Ph.D. thesis.

Krishan Subudhi. 2019. Bert attention visualization.

Chi Sun, Xipeng Qiu, Yige Xu, and Xuanjing Huang. 2019. How to fine-tune BERT for text classification? In *China National Conference on Chinese Computational Linguistics*, pages 194–206. Springer.

Raphael Tang, Yao Lu, Linqing Liu, Lili Mou, Olga Vechtomova, and Jimmy Lin. 2019. Distilling task-specific knowledge from bert into simple neural networks. *arXiv preprint arXiv:1903.12136*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-Art Natural Language Processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. XLNet: Generalized autoregressive pretraining for language understanding. *Advances in Neural Information Processing Systems*, 32:5753–5763.

František Čermák and Alexandr Rosen. 2012. The Case of InterCorp, a multilingual parallel corpus. In *International Journal of Corpus Linguistics*, volume 13, pages 411–427.

A Additional figures

A.1 Probability distributions

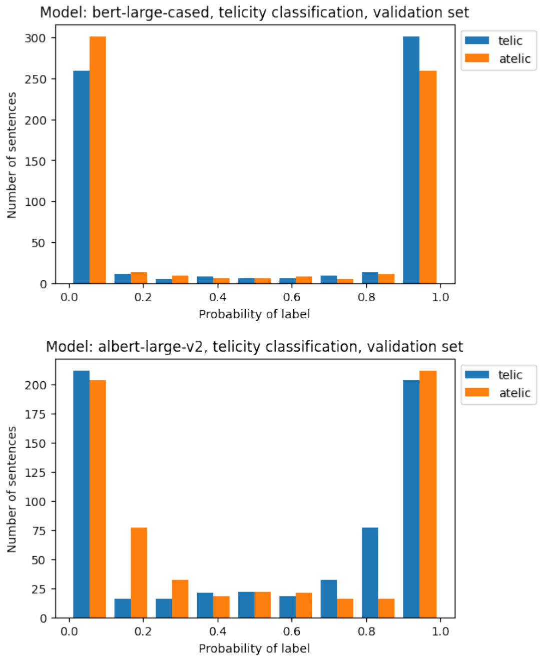


Figure 2: Probability distribution for the telicity labels, for the most successful model (`bert-large-cased` with verb position) and the least successful model (`albert-large-v2` without verb position).

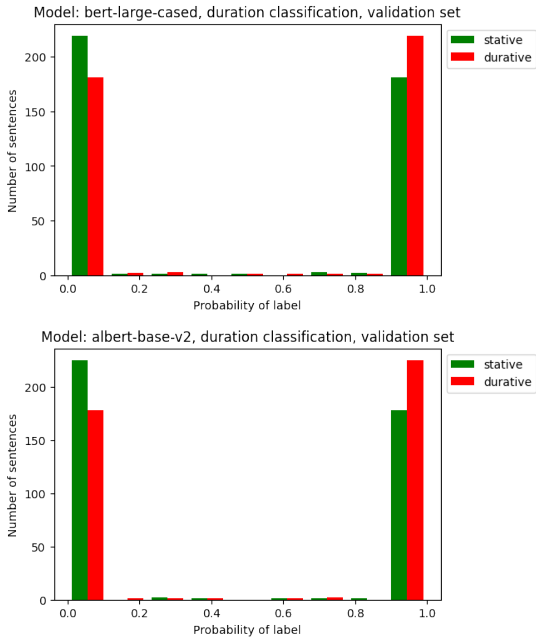


Figure 3: Probability distribution for the duration labels, for the most successful model (`bert-large-cased` with verb position) and the least successful model (`albert-large-v2` without verb position).

A.2 Attention plots

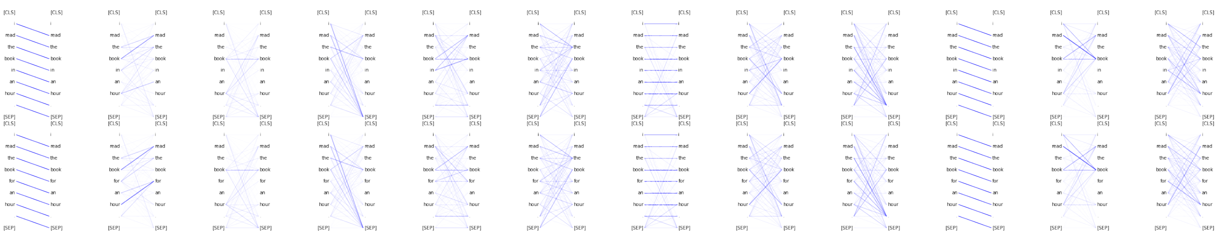
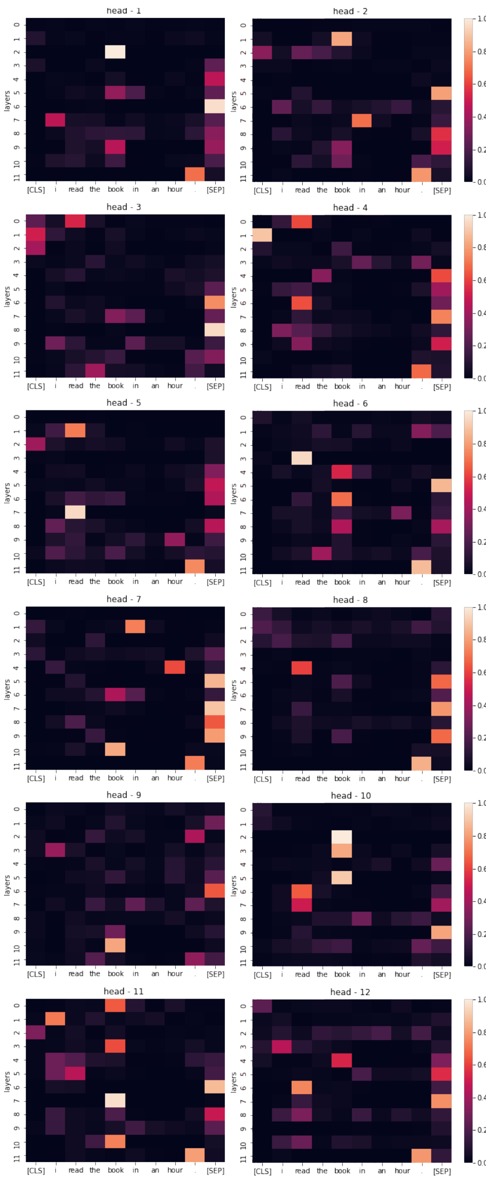
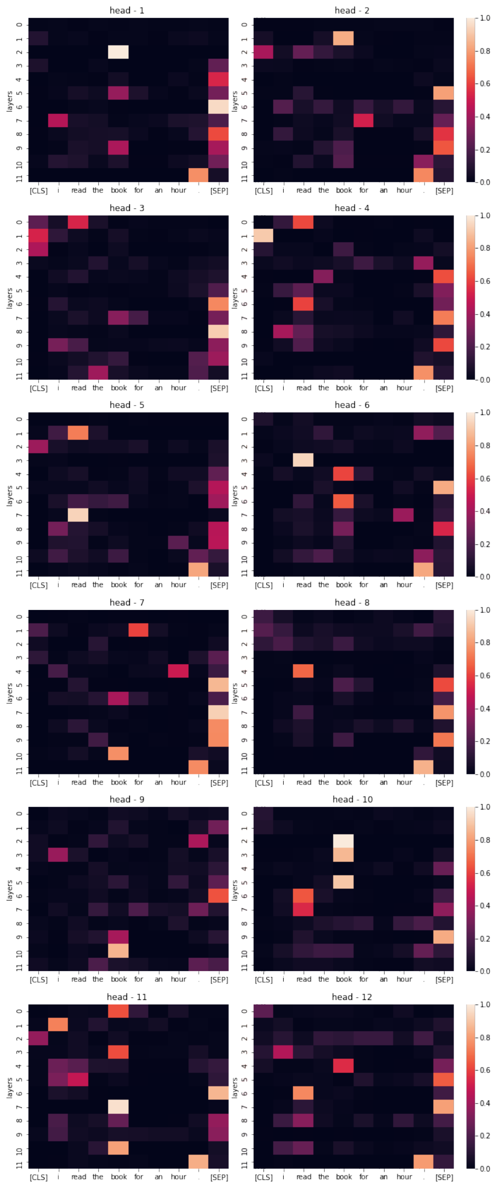


Figure 4: Visualization of attention for the sentences *I read the book in an hour.* (telic, top) and *I read the book for an hour.* (atelic-bottom), from the model `bert-base-uncased` (with verb position information), on the 3rd layer of the model, for all heads (1-12).



I read the book in an hour.



I read the book for an hour.

Figure 5: Visualization of attention of the verb token to all other sentence tokens (x axis), from the model `bert-base-uncased` (with verb position information), on all layers (y axis), for all heads (per plot).

B Qualitative sets

B.1 Telicity test sets

Sentence	Label	Sentence	Label
I ate a fish for lunch .	telic	I eat a fish for lunch on Fridays .	atelic
John built a house in a year .	telic	John is building good houses with his construction company .	atelic
The cat drank all the milk .	telic	John watched TV.	atelic
I spilled the milk .	telic	I always spill milk when I pour it in my mug .	atelic
Yesterday I ran a mile in under 10 minutes .	telic	I 'm running 10 miles every day for my training process .	atelic
The inspector checked our tickets after the first stop .	telic	The inspectors are always checking every document very carefully .	atelic
The classes lasted one hour and took place twice a week over a four-week period .	telic	The damage may last for many years .	atelic
I hung the picture on the wall.	telic	We swim in the lake in the afternoons .	atelic
The vase broke in a million pieces.	telic	In the summer months James sleeps in every morning .	atelic
John kicked the door shut .	telic	Cork floats on water.	atelic
I opened the juice bottle .	telic	My grandfather still lives in his childhood home .	atelic
She opens the door and the dog jumps in her lap .	telic	Nobody laughs at my corny jokes .	atelic
Kim has written a song .	telic	Jenny worked as a doctor her whole life .	atelic
You fell for my trap again .	telic	I am working on a big project now .	atelic
The advancements in technology have changed the world .	telic	Kim is singing .	atelic
Louise made the biggest progress of everyone this year .	telic	Kim is writing a song .	atelic
The dog destroyed the couch .	telic	Grandma is making pancakes for breakfast .	atelic
She cut one single rose from the bush.	telic	He is constantly changing his script .	atelic
The soup cooled in an hour .	telic	We live in a democratic age .	atelic
Jean was born in 1993 in Lyon .	telic	The Earth revolves around the Sun.	atelic

Table 11: 40 sentences with telic and atelic annotations.

Sentence	Label	Sentence	Label
I ate a fish for lunch at noon .	telic	I eat a fish for lunch on Fridays .	atelic
I had eaten a fish for lunch at noon .	telic	I usually eat a fish for lunch of Fridays .	atelic
At noon , I ate a fish for lunch .	telic	On Fridays , I eat a fish for lunch .	atelic
At noon , I had eaten a fish for lunch .	telic	On Fridays , I usually eat a fish for lunch .	atelic
John built a house in a year .	telic	John watched TV .	atelic
John had built a house in a year .	telic	John watched TV all afternoon .	atelic
In a year , John built a house .	telic	John watched TV every afternoon .	atelic
In a year , John had built a house .	telic	John watched TV after finishing his homework .	atelic
I ran a mile in under 10 minutes yesterday .	telic	I 'm running 10 miles every day for my training process .	atelic
I had run a mile in under 10 minutes yesterday .	telic	Every day I 'm running 10 miles for my training process .	atelic
I ran a mile yesterday in under 10 minutes .	telic	We swim in the lake in the afternoons .	atelic
I had run a mile yesterday in under 10 minutes .	telic	We swim in the lake each afternoon .	atelic
Yesterday I ran a mile in under 10 minutes .	telic	In the afternoons , we swim in the lake .	atelic
Yesterday I had run a mile in under 10 minutes .	telic	Each afternoon , we swim in the lake .	atelic
The inspector checked our tickets after the first stop .	telic	Kim is singing .	atelic
The inspector had checked our tickets after the first stop .	telic	Kim is singing a song .	atelic
After the first stop , the inspector checked our tickets .	telic	Kim is writing .	atelic
After the first stop , the inspector had checked our tickets .	telic	Kim is writing a song .	atelic
The classes lasted one hour and took place twice a week over a four-week period .	telic	In the summer months James sleeps in every morning .	atelic
The classes lasted one hour and had taken place twice a week over a four-week period .	telic	James sleeps in every morning in the summer months .	atelic
The classes took place twice a week over a four-week period and lasted one hour .	telic	Grandma is making pancakes for breakfast .	atelic
The classes had taken place twice a week over a four-week period and lasted one hour .	telic	Grandma is making pancakes whenever we visit her .	atelic
Over a four-week period , the classes lasted one hour and took place twice a week .	telic	For breakfast , grandma is making pancakes .	atelic
Over a four-week period , the classes lasted one hour and had taken place twice a week .	telic	Whenever we visit her , grandma is making pancakes .	atelic
Louise made the biggest progress out of everyone this year .	telic	I will receive new stock on Fridays .	atelic
Louise had made the biggest progress out of everyone this year .	telic	I receive new stock on Fridays .	atelic
Out of everyone this year , Louise made the biggest progress .	telic	On Fridays , I will receive new stock ,	atelic
Out of everyone this year , Louise had made the biggest progress .	telic	On Fridays , I receive new stock .	atelic
This year , Louise had made the biggest progress out of everyone .	telic	I read the book for an hour .	atelic
This year , Louise made the biggest progress out of everyone .	telic	I have been reading the book for an hour .	atelic
The soup cooled in an hour .	telic	The Prime Minister made that declaration for months .	atelic
The soup had cooled in an hour .	telic	The Prime Minister has been making that declaration for months .	atelic
In an hour , the soup cooled .	telic	For months the Prime Minister made that declaration .	atelic
In an hour , the soup had cooled .	telic	For months the Prime Minister has been making that declaration .	atelic
John Wilkes Booth killed Lincoln on 1865 .	telic	The workers painted the house for an hour .	atelic
On 1865 , John Wilkes Booth killed Lincoln .	telic	The workers have been painting the house for an hour .	atelic
Lincoln was killed by John Wilkes Booth on 1865 .	telic	The workers painted the house since 8 am .	atelic
On 1865 , Lincoln was killed by John Wilkes Booth .	telic	The workers have been painting the house since 8 am .	atelic
John Wilkes Booth had killed Lincoln before the play ended .	telic	The workers had been painting the house for an hour .	atelic
Before the play ended , John Wilkes Booth had killed Lincoln .	telic	The workers had been painting the house since 8 am .	atelic

Table 12: Test sets on word position and tense variations.

Sentence	Label	Sentence	Label
The girl walked a kilometer yesterday .	telic	The hunter occupied the mountain hut .	atelic
The girl walked yesterday .	atelic	The hunter reached the mountain hut .	telic
I will receive new stock on Friday .	telic	I put on my red dress .	telic
I will receive new stock on Fridays .	atelic	I wore my red dress .	atelic
The boy is eating an apple .	telic	The artist draws a painting .	telic
The boy is eating apples .	atelic	The artist studies a painting .	atelic
I drank the whole bottle .	telic	The policemen entered the church .	telic
I drank juice .	atelic	The policemen watched the church .	atelic
I read the book in an hour .	telic	They caught the boar .	telic
I read the book for an hour .	atelic	They hunted the boar .	atelic
The Prime Minister made that declaration yesterday .	telic	She fell asleep at 8 pm .	telic
The Prime Minister made that declaration for months .	atelic	She slept at 8 pm .	atelic
The workers painted the house in an hour .	telic	She noticed him .	telic
The workers painted the house for an hour .	atelic	She looked at him .	atelic
The hunters chased the deer away .	telic	The people died from starvation .	telic
The hunters chased the deer .	atelic	The people suffered from starvation .	atelic
I finished reading the book at 5 pm .	telic	They built the house .	telic
I stopped reading the book at 5 pm .	atelic	They have been building the house .	atelic
The pond is freezing over .	telic	She ate that sandwich .	telic
It 's freezing outside .	atelic	She has been eating that sandwich .	atelic

Table 13: “Minimal pairs" of telicity.

Sentence	Label	Sentence	Label
She didn’t agree with us .	stative	She plays tennis every Friday .	durative
I don’t believe the news .	stative	She’s playing tennis right now .	durative
Bread consists of flour, water and yeast .	stative	The snow melts every spring .	durative
This box contains a cake .	stative	The snow is melting right now .	durative
I disagree with you .	stative	The boxer hits his opponent .	durative
I have disliked mushrooms for years .	stative	The boxer is hitting his opponent .	durative
This shirt fits me well .	stative	They ate their dinner in silence .	durative
Julie 's always hated dogs .	stative	I walked past the barn .	durative
Do you hear music ?	stative	We learned to make pasta .	durative
This cookbook includes a recipe for bread .	stative	He grew potatoes in his farm .	durative
I 've known Julie for ten years .	stative	I slept all morning .	durative
I like reading detective stories .	stative	We talked for hours on our trips .	durative
I love chocolate .	stative	I will write you a letter tomorrow .	durative
I prefer chocolate ice cream .	stative	She runs ten kilometers a day .	durative
I didn’t realise the problem .	stative	He read a fairytale to his kids .	durative
I didn’t recognise my old friend .	stative	The boy kicked the ball hard .	durative
He didn’t remember my name .	stative	We will go soon .	durative
Your idea sounds great .	stative	He screamed for help .	durative
I suppose John will be late .	stative	The dogs bark all night .	durative
The noise surprised me .	stative	She closed the door .	durative

Table 14: 40 sentences with stative and durative annotations.