Prédire l'aspect linguistique en anglais au moyen de transformers

Eleni Metheniti, Tim van de Cruys, Nabil Hathout

Actes de la 28e Conférence sur le Traitement Automatique des Langues Naturelles
Lille, France, 28 juin au 2 juillet 2021

Prédire l'aspect linguistique en anglais au moyen de *transformers*

Eleni Metheniti[1, 2] Tim van de Cruys[3] Nabil Hathout [1]
(1) CLLE, CNRS & Université Toulouse - Jean Jaurès, France
(2) IRIT, CNRS & Université Toulouse - Paul Sabatier, France
(3) Faculté des Lettres, KU Leuven, Belgique

RÉSUMÉ

L'aspect du verbe décrit la manière dont une action, un événement ou un état exprimé par un verbe est lié au temps ; la télicité est la propriété d'un syntagme verbal qui présente une action ou un événement comme étant mené à son terme ; la durée distingue les verbes qui expriment une action (dynamique) ou un état (statique). Ces caractéristiques essentielles à l'interprétation du langage naturel, sont également difficiles à annoter et à identifier par les méthodes de TAL. Dans ce travail, nous estimons la capacité de différents modèles de type *transformers* pré-entraînés (BERT, RoBERTa, XLNet, ALBERT) à prédire la télicité et la durée. Nos résultats montrent que BERT est le plus performant sur les deux tâches, tandis que les modèles XLNet et ALBERT sont les plus faibles. Par ailleurs, les performances de la plupart des modèles sont améliorées lorsqu'on leur fournit en plus la position des verbes. Globalement, notre étude établit que les modèles de type *transformers* captent en grande partie la télicité et la durée.

ABSTRACT

Classifying Linguistic Aspect in English with Transformers

Verb aspect describes how an action, event, or state of a verb relates to time ; telicity focuses on whether the verb's action or state has an end point 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. In this work, we explore whether different kinds of fine-tuned transformer models (BERT, RoBERTa, XLNet, ALBERT) are successful in the task of binary classification of telicity and duration. Both for telicity and duration, BERT is the most successful, while certain XLNet and ALBERT models completely failed at the classification task. The use of verb position vectors significantly improves performance in most models. The results show that transformers models adequately capture telicity and duration.

MOTS-CLÉS : transformers, apprentissage automatique, aspect lexical, télicité, durée.

KEYWORDS: transformers, machine learning, lexical aspect, telicity, duration.

1 Introduction

L'aspect est une propriété temporelle des actions, des événements et des états décrits par les verbes, au-delà du temps verbal. Il englobe différentes propriétés, telles que la **télicité** et la **durée**. L'action du verbe est dite *télique* si elle a un point final. Lorsque le verbe désigne un état ou lorsque l'accomplissement de l'action du verbe est impossible, qu'il est indéfini ou non pertinent, l'action est dite *atélique*. Une autre propriété aspectuelle est la durée qui distingue les verbes d'état dits *statif* des actions dites

duratives, indépendamment de l'existence d'un point final perçu ou non. Krifka (1998) a établit que la télicité est une propriété de l'ensemble du syntagme verbale et n'est pas une caractéristique du verbe seul. En outre, le contexte est un autre facteur qui détermine la classe aspectuelle d'un syntagme verbal (Siegel, 1998). La télicité n'est donc pas une propriété facile à estimer, en particulier dans les langues qui ont une morphologie flexionnelle pauvre comme l'anglais. Elle n'en demeure pas moins indispensable pour de nombreuses tâches de TAL. L'aspect fournit notamment des informations sur les relations temporelles (Costa & Branco, 2012), sur l'implication textuelle (Hosseini *et al.*, 2018; Kober *et al.*, 2019) et l'ordonnancement des événements (Chambers *et al.*, 2014).

Dans cet article, nous montrons que les architectures de type *transformers* sont capables de déterminer la télicité et la durée lorsqu'on leur applique un *fine-tuning*. L'entraînement est réalisé au moyen d'un jeu de données fournies par Friedrich & Gateva (2017) et des versions pré-entraînées de plusieurs modèles TRANSFORMERS mis à disposition sur Huggingface (Wolf *et al.*, 2020). L'évaluation des modèles est quantitative (jeu de test issu du jeu de données) et qualitative (ensemble de phrases simples et de paires minimales). Les résultats obtenus montrent que les modèles BERT sont les plus performants dans la classification binaire de la télicité et de la durée ; les scores RoBERTa, XLNet et ALBERT sont à l'inverse les plus faibles.

2 État de l'art

Siegel & McKeown (2000) ont mis au point plusieurs méthodes de classification aspectuelle fondés sur l'identification de marqueurs linguistiques et ont observé que les méthodes par apprentissage supervisé permettent d'obtenir les meilleurs résultats. Friedrich & Palmer (2014) utilisent pour leur part une approche semi-supervisée d'apprentissage de l'aspect lexical, combinant des caractéristiques linguistiques et distributionnelles, afin de prédire la stativité et la durée. Friedrich & Pinkal (2015) reprennent la même approche pour classer l'aspect lexical verbal en plusieurs catégories duratives (habituel, épisodique, statique) puis Friedrich *et al.* (2016) étendent leurs jeux de données et leurs catégories et atteignent une précision de 76% pour la classification supervisée, se rapprochant ainsi des performances humaines estimées à 80%. Plus récemment, Friedrich & Gateva (2017) font état d'une amélioration significative de la classification automatique de la télicité avec un modèle de régression logistique supervisée.

Loáiciga & Grisot (2016) exploitent la télicité pour améliorer la traduction automatique français-anglais. Falk & Martin (2016) prédisent l'aspect verbal dans différents types de contexte par des méthodes d'apprentissage automatique. Pour leur part, Peng (2018) utilise deux modèles compositionnels PLF et LSA pour classer l'aspect en considérant l'ensemble de la proposition et pas seulement le verbe, sans recourir à des données annotées. L'auteur met en évidence l'importance du syntagme verbale et des dépendants du verbe dans l'interprétation de la télicité. Kober *et al.* (2020) utilisent des modèles distributionnels compositionnels pour déterminer l'aspect des verbes anglais en contexte. Leur travail confirme que le contexte du verbe et les mots grammaticaux qui expriment le temps sont des caractéristiques déterminantes pour la classification aspectuelle.

3 Expériences

Nos expérimentations sont basées sur le *fine-tuning* de modèles *transformers* pour classer des séquences relativement à leur télicité et leur durée (séparément). L'exactitude des modèles spécialisés (*fine-tuned*) est testée en prédisant la télicité et la durée de phrases annotées manuellement. Le *fine-tuning* est une méthode qui consiste à adapter un modèle à une tâche spécifique, en ajoutant une couche supplémentaire dédiée à la tâche en question. Il est ainsi possible d'exploiter les connaissances existantes du modèle et de le spécialiser sur une tâche spécifique sans disposer d'importantes ressources spécialisées, sans avoir recours à une grande puissance de calcul ni à un entraînement long.

Les annotations de télicité et de durée que nous utilisons étant basées sur le verbe principal de la phrase, nous affinons chaque modèle de deux manières : (*i*) en l'entraînant uniquement sur les entrées et les étiquettes de télicité ou de durée ; (*ii*) en fournissant en plus un vecteur `token_type_ids` qui indique la position du verbe dans la séquence d'entrée comme illustré en (1).

(1)

tokens	He	**worked**	well	and	earned	much	.	[SEP]
token_type_ids	0	1	0	0	0	0	0	0

Les modèles sont affinés en utilisant les jeux de données *gold* et *silver* développés et distribués par Friedrich & Gateva (2017). Les annotations *gold* sont basées sur le jeu de données MASC (Ide *et al.*, 2008), tandis que les annotations *silver* ont été construites en utilisant le corpus parallèle InterCorp anglais-tchèque (Čermák & Rosen, 2012) qui permet de déterminer la télicité et la durée des phrases anglaise en exploitant les marqueurs morphologiques dans les phrases tchèques correspondantes. Nous avons extrait 6 354 phrases annotées pour la télicité (3 220 téliques, 3 134 atéliques) et 5 119 phrases pour la durée (1 861 statiques, 3 258 dynamiques) des jeux de données de Friedrich & Gateva. Nous avons par ailleurs augmenté le nombre d'annotations initiales en considérant que les phrases annotées comme statives (pour la durée) sont atéliques, et que les phrases annotées comme duratives sont téliques. Les phrases ont été pré-traitées : séparation des tokens, conversion en minuscule, troncation à 128 mots, remplissage (*padding*), comme cela est recommandé pour le *fine-tuning*. La troncation n'a pas posé de problèmes car une seule phrase dépasse ce seuil et parce que le verbe cible se trouve toujours dans les 128 premiers mots. Afin de mener une évaluation qualitative, nous avons préparé un deuxième ensemble de données de test composé de 40 phrases annotées pour la télicité et 40 phrases annotées pour la durée, réparties de manière égale dans les quatre catégories télique, atélique, statique et durative. Nous avons également construit un 3e jeu de données composé de « paires minimales » de phrases téliques et atéliques qui soit partagent le même verbe soit ne diffèrent que par le verbe (2).

(2)

télique : The girl **walked** a kilometer yesterday.	atélique : The girl **walked** yesterday.
télique : She **noticed** him.	atélique : She **looked** at him.

Nous utilisons des modèles pré-entraînés de la bibliothèque Python `transformers` (Wolf *et al.*, 2020), et en particulier les modèles de classification de séquences. La mise en œuvre reprend les recommandations de l'équipe qui a développée la bibliothèque de *fine-tuning* de ces modèles. Les architectures utilisées sont : BERT, RoBERTa, XLNet et ALBERT, dans les versions de `base`, `large` (taille grande), `cased` (avec les majuscules) et `uncased` (en minuscules) disponibles.

En outre, deux autres modèles de classification binaire sont utilisés comme méthodes de base : un modèle de régression logistique simple implémenté en Python (Celik, 2021) avec les paramètres par défaut, une lemmatisation et la suppression des mots vides et un modèle à réseau de neurones convolutionnel (CNN) implémenté avec Keras (Schapira, 2019). Ce deuxième modèle est largement utilisé comme méthode de base pour les tâches de classification de texte (Kim, 2014).

4 Résultats

4.1 Évaluation quantificative

Au cours du *fine-tuning*, nous avons déterminé les performances des modèles dans la prédiction des étiquettes binaires sur un jeu de données de validation (10% des phrases). La précision et la matrice de confusion ont été calculées en utilisant la bibliothèque Python `scikit-learn` (Pedregosa *et al.*, 2011). Les résultats pour les données de test (10% des phrases) sont présentés en Table 1 pour la télicité et en Table 2 pour la durée. Les modèles les plus performants pour la **télicité** sont `bert-base-cased` et `bert-large-cased`. Globalement, les modèles BERT sont significativement meilleurs que les autres architectures, les modèles RoBERTa étant modérément performants tandis que XLNet et ALBERT (`xlnet-large-cased`, `albert-base-v2`, `albert-large-v2`) tendent à prédire la même étiquette pour toutes les phrases. Par ailleurs, on observe une amélioration nette pour les modèles les plus performants lorsque la position du verbe dans la phrase est fournie lors de l'entraînement. La précision augmente par exemple de 11% pour `bert-base-cased` (65% → 76%) et pour `bert-large-cased` (68% → 79%) . En revanche, elle baisse pour les modèles les moins performants. Les modèles qui obtiennent les meilleurs résultats surpassent aussi très largement nos deux méthodes de base (CNN et régression logistique), les performances de ces dernières s'avérant proches de celles des modèles les moins performants.

						télique			atélique		
modèle	**posit. verbe**	**exact.**	**précis.**	**rappel**	**F1-score**	**précis.**	**rappel**	**F1-score**	**précis.**	**rappel**	**F1-score**
bert-base-uncased	oui	0.72	0.72	0.72	0.72	0.73	0.71	0.72	0.72	0.69	0.71
	non	0.66	0.66	0.66	0.66	0.66	0.66	0.66	0.65	0.65	0.65
bert-base-cased	oui	0.76	0.76	0.76	0.76	0.75	0.79	0.77	0.77	0.73	0.75
	non	0.65	0.65	0.65	0.65	0.67	0.61	0.64	0.63	0.69	0.66
bert-large-uncased	oui	0.64	0.64	0.64	0.64	0.66	0.61	0.63	0.63	0.67	0.65
	non	0.66	0.66	0.66	0.66	0.67	0.64	0.65	0.65	0.67	0.66
bert-large-cased	oui	**0.79**	**0.79**	**0.79**	**0.79**	**0.8**	**0.8**	**0.8**	**0.79**	**0.79**	**0.79**
	non	0.68	0.68	0.68	0.68	0.69	0.67	0.68	0.67	0.69	0.68
roberta-base	non	0.64	0.64	0.64	0.64	0.64	0.67	0.65	0.64	0.61	0.63
roberta-large	non	0.66	0.66	0.66	0.66	0.66	0.70	0.68	0.67	0.62	0.64
xlnet-base-cased	oui	0.59	0.59	0.59	0.59	0.59	0.62	0.61	0.59	0.56	0.58
	non	0.61	0.61	0.61	0.61	0.62	0.59	0.60	0.60	0.62	0.61
xlnet-large-cased	oui	0.59	0.59	0.59	0.59	0.59	0.62	0.61	0.59	0.56	0.58
	non	0.51	0.26	0.51	0.34	0.51	1.00	0.67	0.00	0.00	0.00
albert-base-v2	oui	0.49	0.24	0.49	0.33	0.00	0.00	0.00	0.49	1.00	0.66
	non	0.60	0.60	0.60	0.60	0.61	0.60	0.61	0.60	0.60	0.60
albert-large-v2	oui	0.49	0.24	0.49	0.33	0.00	0.00	0.00	0.49	1.00	0.66
	non	0.49	0.24	0.49	0.33	0.00	0.00	0.00	0.49	1.00	0.66
CNN (50 epochs)	non	0.6	0.6	0.6	0.6	0.6	0.62	0.61	0.6	0.58	0.59
Régression logistique	non	0.53	0.63	0.53	0.42	0.52	0.97	0.68	0.74	0.08	0.15

TABLE 1 – Résultats pour la classification de la télicité pour le jeu de données de Friedrich & Gateva.

Les résultats obtenus pour la classification de la **durée** sont globalement meilleurs que ceux de la télicité bien que le jeu de données soit déséquilibré et plus petit. Les modèles `bert-base` l'emportent sur les modèles `bert-large`, tandis que les modèles `roberta-large`, `xlnet-large-case` et `albert-large-v2` s'avèrent incapables de réaliser cette tâche. Nous constatons à nouveau une amélioration significative de la précision lorsque les informations sur la position du verbe sont fournies aux modèles, en particulier pour les plus performants : 70% → 86% pour `bert-base-cased`, 71% → 86% pour `bert-base-uncased`. Là encore, les deux méthodes de base obtiennent des résultats nettement inférieurs à ceux des meilleurs modèles, du même ordre que ceux des modèles

modèle	posit. verbe	exact.	précis.	rappel	F1-score	stative précis.	stative rappel	stative F1-score	durative précis.	durative rappel	durative F1-score
bert-base-uncased	oui	0.86	0.85	0.86	0.85	0.83	0.76	0.79	0.87	0.91	0.89
	non	0.71	0.71	0.71	0.71	0.62	0.57	0.59	0.76	0.80	0.78
bert-base-cased	oui	**0.86**	**0.86**	**0.86**	**0.86**	**0.82**	**0.8**	**0.81**	**0.89**	**0.9**	**0.89**
	non	0.70	0.70	0.70	0.70	0.59	0.55	0.57	0.75	0.78	0.77
bert-large-uncased	oui	0.77	0.77	0.77	0.77	0.70	0.65	0.67	0.81	0.84	0.82
	non	0.70	0.69	0.70	0.70	0.6	0.53	0.56	0.75	0.79	0.77
bert-large-cased	oui	0.74	0.73	0.74	0.73	0.66	0.58	0.62	0.78	0.83	0.80
	non	0.71	0.71	0.71	0.71	0.6	0.58	0.59	0.77	0.78	0.77
roberta-base	non	0.72	0.71	0.72	0.71	0.65	0.49	0.56	0.74	0.85	0.79
roberta-large	non	0.64	0.41	0.64	0.50	0.00	0.00	0.00	0.64	1.00	0.78
xlnet-base-cased	oui	0.70	0.69	0.70	0.68	0.65	0.39	0.49	0.72	0.88	0.79
	non	0.71	0.70	0.71	0.69	0.65	0.43	0.52	0.73	0.87	0.79
xlnet-large-cased	oui	0.64	0.41	0.64	0.50	0.00	0.00	0.00	0.64	1.00	0.78
	non	0.64	0.41	0.64	0.50	0.00	0.00	0.00	0.64	1.00	0.78
albert-base-v2	oui	0.80	0.80	0.80	0.78	0.84	0.54	0.66	0.78	0.94	0.86
	non	0.68	0.66	0.68	0.66	0.59	0.37	0.46	0.70	0.86	0.77
albert-large-v2	oui	0.64	0.41	0.64	0.50	0.00	0.00	0.00	0.64	1.00	0.78
	non	0.64	0.41	0.64	0.50	0.00	0.00	0.00	0.64	1.00	0.78
CNN (50 epochs)	non	0.65	0.64	0.65	0.64	0.54	0.39	0.45	0.7	0.81	0.75
Régression logistique	non	0.64	0.41	0.64	0.50	0.00	0.00	0.00	0.64	1.00	0.78

TABLE 2 – Résultats pour la classification de la durée pour le jeu de données de Friedrich & Gateva.

sous-performants.

4.2 Évaluation qualitative

Les modèles ont également été testés sur deux jeux de données plus petits que nous avons créés et annotés pour la télicité et la durée afin de tester l'exactitude de la classification sur des phrases qui ne proviennent pas du jeu de données de Friedrich & Gateva.

La précision obtenue avec le premier jeu est meilleure que celle obtenue pour les jeux de test originaux. Elle est notamment nettement plus élevée pour les modèles les plus performants : pour la **télicité**, la précision de `bert-base-uncased` la plus élevée est de 75% sans les positions des verbes et de 85% avec ; `bert-large-cased` s'avère en revanche moins efficace avec une précision de 69% sans les positions des verbes et 70% avec. Nous avons également réalisé une évaluation plus qualitative en examinant les mauvaises prédictions des modèles BERT, ces modèles étant globalement les plus performants. Pour presque tous, les erreurs de classification concernent seulement certaines phrases spécifiques dans lesquelles le syntagme verbal définit un aspect temporel inverse de celui qui est spécifié par une partie du contexte : un syntagme prépositionnel comme dans *I eat a fish for lunch on Fridays* ou le temps grammatical comme dans *The inspectors are always checking every document very carefully*. Dans ces deux exemples, l'action est perçue comme ayant un point final mais le temps continu et la présence de l'adverbe *always* rendent la phrase atélique.

Pour la classification de la **durée**, les résultats du premier jeu de données sont encore meilleurs, avec `bert-base-cased` atteignant une précision de 98% (et 92% sans les vecteurs de position) et `bert-large-cased` une précision de 95% avec ou sans vecteurs de position. Cette amélioration n'est pas surprenante, les modèles étant tous plus performants sur la tâche de classification de la durée, mais aussi parce qu'il est difficile de construire des phrases dans lesquelles le contexte et le verbe expriment des valeurs de durée inverses. La phrase *durative* qui a été le plus souvent mal classée par les modèles est *She's playing tennis right now* ; cette erreur est inattendu, car *play* est toujours un verbe d'action. À l'inverse, *Do you hear music ?* est classée comme *durative* par certains modèles parce qu'ils n'ont probablement pas réussi à capter les connaissances du monde nécessaires à son

interprétation (Rogers *et al.*, 2021).

Les résultats des tests que nous venons de présenter montrent tous que les modèles BERT sont les plus performants. Cependant, certaines questions restent sans réponse. Nous avons donc réalisé des tests supplémentaires sur les modèles de classification au moyen du deuxième jeu de test composé de couples de phrases téliques et atéliques qui partagent le même verbe. Le modèle `bert-base-uncased` obtient les meilleurs résultats avec une précision de 81% pour la classification de la télicité lorsque la position des verbes est fournie. En examinant les mauvaises prédictions, nous observons que certaines phrases sont mal classées par tous les modèles. Ce résultat est attendu, car les paires minimales ont des valeurs de télicité opposées tout en ayant le même verbe et parce que les verbes présentent des affinités fortes avec l'une ou l'autre de ces valeurs. Par exemple, la phrase *The boy is eating an apple* est considérée comme télique, car l'action a un terme perçu (l'objet étant au singulier, l'action du verbe est télique) mais la présence d'un temps continu conduit les modèles à classer incorrectement la phrase comme atélique. De même, la phrase *The Prime Minister made that declaration for months* serait télique, sans la présence du complément de temps *for months*. Ces exemples suggèrent que les modèles accordent trop d'importance au verbe et ne tiennent pas suffisamment compte du temps grammatical et du contexte, notamment lorsque ce dernier contient des compléments de temps.

Afin de tester davantage encore les modèles, nous avons étudié l'effet d'un masque d'attention sur le contexte et la façon dont il affecte la classification. Cette méthode a été notamment utilisée par Metheniti *et al.* (2020). Les masques d'attention imposent aux modèles de prédire la télicité ou la durée en ne considérant que le verbe. Ils ont été appliqués aux phrases d'entrée lors de la phase de test, sans procéder à un nouveau *fine-tuning* des modèles. La capacité des modèles à déterminer la télicité diminue de manière significative, lorsque la prédiction est réalisée uniquement sur la base du verbe, $79\% \rightarrow 51\%$ par exemple pour `bert-large-cased` sur le jeu de données de Friedrich & Gateva. La baisse rend compte de l'importance du contexte et des dépendants du verbe dans la prédiction de la télicité et de la durée. Ces résultats sont conformes aux prédictions de théories linguistiques comme celle de Krifka (1998) : l'aspect ne dépend pas uniquement du verbe ; il est également déterminé par le contexte.

5 Discussion

Le *fine-tuning* des modèles *transformers* produit des résultats à l'état de l'art pour de nombreuses applications de TALN et dans de nombreux domaines. Ces modèles font cependant l'objet de critiques car il s'agir de « boîte noire » tant au niveau de leur création que de leur déploiement. Les critiques concernent également les stratégies sous-optimales de *fine-tuning* devenues courantes. Dodge *et al.* (2020) soulignent notamment que l'initialisation du processus de *fine-tuning* avec une amorce aléatoire peut produire des résultats sensiblement différents, même avec les mêmes hyperparamètres.

Bien que nos jeux de données soient relativement petits (6K pour la télicité et 4K pour la durée), nous ne les avons pas mélangés avant de les diviser en données d'entraînement, de test et de validation suivant en cela la proposition de Dodge *et al.* (2020). Par ailleurs, nous avons utilisé l'optimisateur ADAM de PyTorch comme cela est recommandé par Zhang *et al.* (2020) au lieu de BERTADAM (Devlin *et al.*, 2019; Wolf *et al.*, 2020). À l'inverse, nous avons suivi les recommandations de Devlin *et al.* (2019) et McCormick & Ryan (2019) de réduire le nombre d'époques d'entraînement et de choisir la meilleure époque en fonction des résultats de validation, la proposition de Dodge *et al.*

(2020) et Mosbach *et al.* (2020) de multiplier les époques d'entraînement pour toutes les tâches s'avérant en définitive contre-productive (Zhang *et al.*, 2020). Signalons également que nous avons répété l'entraînement des modèles avec 75%, 80% et 90% des jeux de données sans observer aucune différence significative dans leur comportement ni aucune baisse de performance.

Notons également que nous nous attendions à ce que les modèles RoBERTa soient moins performants sur nos tâches, car ils exploitent mal les informations contextuelles de mots entiers et ne tiennent pas suffisamment compte des vecteurs de position des verbes. Cependant, nous ne nous attendions pas à ce que les modèles XLNet et ALBERT soient aussi peu performants. XLNet dépasse en effet les modèles BERT sur les tâches de la compréhension et de la classification des textes (Yang *et al.*, 2019) ; ALBERT dépasse BERT lui aussi sur plusieurs tâches, l'une des innovations de cette architecture étant justement ses représentations dépendantes du contexte (Wright, 2019). XLNet et ALBERT obtiennent globalement de mauvais résultats dans toutes les expériences que nous avons réalisées. Cela suggère qu'ils ne sont pas adaptés au *fine-tuning* au moyen de jeux de d'entraînement de petite taille ni à la classification binaire de séquences courtes, ces dernières ne leurs permettant pas de tirer profit de leurs points forts comme la meilleure prise en compte des relations à longue distance.

Une autre question intéressante qui émerge de nos expériences concerne le succès des versions `cased` de BERT par rapport aux versions `uncased`. Pourquoi les modèles qui conservent les majuscules sont-ils plus performants alors que nos jeux de données sont entièrement en minuscules ? Rappelons que les modèles où les mots conservent leurs majuscules sont essentiellement utilisés pour la reconnaissance d'entités nommées pour laquelle ces informations supplémentaires sont importantes. En outre, la version `large-cased` de BERT donne les meilleurs résultats dans la plupart des cas, ou est à égalité avec les versions de `base`. Or nous savons que les gros modèles *transformers* sont plus difficiles à ajuster, surtout lorsque les jeux de données sont petits. Les bons résultats de `bert-large-cased` s'expliquent probablement par l'homogénéité relative de nos jeux de données composés de phrases de genres similaires (littérature, articles de presse).

6 Conclusion

Nous avons mené dans cette étude plusieurs expériences qui testent la capacité des modèles *transformers* à capter les catégories aspectuelles comme la télicité et la durée. Nous avons testé cette capacité en réalisant un *fine-tuning* de modèles de classification binaire au moyen notamment du jeu de données annotées pour la télicité et la durée de (Friedrich & Gateva, 2017). Le *fine-tuning* a été réalisé sur des modèles *transformers* de plusieurs architectures (BERT, RoBERTa, XLNet, ALBERT). Nous avons ainsi observé que malgré la taille réduite de nos jeux de données, certains modèles sont très efficaces pour la classification aspectuelle et que les performances sont considérablement améliorées lorsque la position du verbe dans la phrase est fournie au classifieur lors de l'entraînement. L'examen des erreurs des classifieurs nous a permis de caractériser les limites des modèles, notamment pour les phrases où l'information temporelle exprimée dans le contexte est à rebours de l'aspect verbal. Enfin, nous avons mis en évidence l'importance du contexte dans la prédiction de l'aspect en utilisant des masques d'attention.

Références

CELIK I. (2021). Text classification logistic regression from scratch. github.com/iremcelik/Text-Classification-Logistic-Regression-From-Scratch.

CHAMBERS N., CASSIDY T., MCDOWELL B. & BETHARD S. (2014). Dense Event Ordering with a Multi-Pass Architecture. In *Transactions of the Association for Computational Linguistics*, volume 2, p. 273–284. DOI : 10.1162/tacl_a_00182.

COSTA F. & BRANCO A. (2012). Aspectual Type and Temporal Relation Classification. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, p. 266–275, Avignon, France : Association for Computational Linguistics.

DEVLIN J., CHANG M.-W., LEE K. & TOUTANOVA K. (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)*, p. 4171–4186, Minneapolis, Minnesota : Association for Computational Linguistics. DOI : 10.18653/v1/N19-1423.

DODGE J., ILHARCO G., SCHWARTZ R., FARHADI A., HAJISHIRZI H. & SMITH N. (2020). Fine-tuning pretrained language models : Weight initializations, data orders, and early stopping. *arXiv preprint arXiv :2002.06305*.

DOWTY D. R. (1979). *Word Meaning and Montague Grammar : The Semantics of Verbs and Times in Generative Semantics and in Montague's Ptq*, volume 7. Springer.

FALK I. & MARTIN F. (2016). Automatic identification of aspectual classes across verbal readings. In * *Sem 2016 THE FIFTH JOINT CONFERENCE ON LEXICAL AND COMPUTATIONAL SEMANTICS*.

FRIEDRICH A. & GATEVA D. (2017). Classification of telicity using cross-linguistic annotation projection. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, p. 2559–2565.

FRIEDRICH A. & PALMER A. (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)*, p. 517–523.

FRIEDRICH A., PALMER A. & PINKAL M. (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)*, p. 1757–1768.

FRIEDRICH A. & PINKAL M. (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*, p. 2471–2481.

HOSSEINI M. J., CHAMBERS N., REDDY S., HOLT X. R., COHEN S. B., JOHNSON M. & STEEDMAN M. (2018). Learning Typed Entailment Graphs with Global Soft Constraints. In *Transactions of the Association for Computational Linguistics*, volume 6, p. 703–717. DOI : 10.1162/tacl_a_00250.

IDE N., BAKER C., FELLBAUM C., FILLMORE C. & PASSONNEAU R. (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).

KIM Y. (2014). Convolutional neural networks for sentence classification. *CoRR*, **abs/1408.5882**.

KOBER T., ALIKHANI M., STONE M. & STEEDMAN M. (2020). Aspectuality Across Genre : A Distributional Semantics Approach. In *Proceedings of the 28th International Conference on Computational Linguistics*, p. 4546–4562, Barcelona, Spain (Online) : International Committee on Computational Linguistics. DOI : 10.18653/v1/2020.coling-main.401.

KOBER T., BIJL DE VROE S. & STEEDMAN M. (2019). Temporal and Aspectual Entailment. In *Proceedings of the 13th International Conference on Computational Semantics - Long Papers*, p. 103–119, Gothenburg, Sweden : Association for Computational Linguistics. DOI : 10.18653/v1/W19-0409.

KRIFKA M. (1998). The origins of telicity. In *Events and grammar*, p. 197–235 : Springer.

LOÁICIGA S. & GRISOT C. (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.

MCCORMICK C. & RYAN N. (2019). BERT Fine-Tuning Tutorial with PyTorch. Retrieved January 24, 2021.

METHENITI E., VAN DE CRUYS T. & HATHOUT N. (2020). How Relevant Are Selectional Preferences for Transformer-based Language Models ? In *Proceedings of the 28th International Conference on Computational Linguistics*, p. 1266–1278, Barcelona, Spain (Online) : International Committee on Computational Linguistics. DOI : 10.18653/v1/2020.coling-main.109.

MOSBACH M., ANDRIUSHCHENKO M. & KLAKOW D. (2020). On the stability of fine-tuning BERT : Misconceptions, explanations, and strong baselines. *arXiv preprint arXiv :2006.04884*.

PEDREGOSA F., VAROQUAUX G., GRAMFORT A., MICHEL V., THIRION B., GRISEL O., BLONDEL M., PRETTENHOFER P., WEISS R., DUBOURG V., VANDERPLAS J., PASSOS A., COURNAPEAU D., BRUCHER M., PERROT M. & DUCHESNAY E. (2011). Scikit-learn : Nachine Learning in Python. *Journal of Machine Learning Research*, **12**, 2825–2830.

PENG Q. (2018). *Towards aspectual classification of clauses in a large single-domain corpus*. Edingburgh, UK : School of Informatics, University of Edinburgh.

ROGERS A., KOVALEVA O. & RUMSHISKY A. (2021). A Primer in BERTology : What we know about how BERT works. In *Transactions of the Association for Computational Linguistics*, volume 8, p. 842–866 : MIT Press.

SCHAPIRA D. (2019). diegoschapira/cnn-text-classifier-using-keras.

SIEGEL E. V. (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.

SIEGEL E. V. & MCKEOWN K. R. (2000). Learning Methods to Combine Linguistic Indicators :Improving Aspectual Classification and Revealing Linguistic Insights. In *Computational Linguistics*, volume 26, p. 595–627.

VERKUYL H. J. (1972). *On the compositional nature of the aspects*, volume 15. Springer.

WOLF T., DEBUT L., SANH V., CHAUMOND J., DELANGUE C., MOI A., CISTAC P., RAULT T., LOUF R., FUNTOWICZ M., DAVISON J., SHLEIFER S., VON PLATEN P., MA C., JERNITE Y., PLU J., XU C., SCAO T. L., GUGGER S., DRAME M., LHOEST Q. & RUSH A. M. (2020). Transformers : State-of-the-Art Natural Language Processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing : System Demonstrations*, p. 38–45, Online.

WRIGHT L. (2019). Meet ALBERT : a new 'Lite BERT' from Google & Toyota with State of the Art NLP performance and 18x fewer parameters. Retrieved January 24, 2021.

YANG Z., DAI Z., YANG Y., CARBONELL J., SALAKHUTDINOV R. R. & LE Q. V. (2019). XLNet : Generalized autoregressive pretraining for language understanding. *Advances in Neural Information Processing Systems*, **32**, 5753–5763.

ZHANG T., WU F., KATIYAR A., WEINBERGER K. Q. & ARTZI Y. (2020). Revisiting few-sample BERT fine-tuning. *arXiv preprint arXiv :2006.05987*.

ČERMÁK F. & ROSEN A. (2012). The Case of InterCorp, a multilingual parallel corpus. In *International Journal of Corpus Linguistics*, volume 13, p. 411–427. DOI : 10.1075/ijcl.17.3.05cer.