Actes de la 6e conférence conjointe Journées d'Études sur la Parole (JEP, 33e édition),
Traitement Automatique des Langues Naturelles (TALN, 27e édition),
Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition)
Nancy, France, 08-19 juin 2020

Qu'apporte BERT à l'analyse syntaxique en constituants discontinus ? Une suite de tests pour évaluer les prédictions de structures syntaxiques discontinues en anglais.

Maximin Coavoux
Univ. Grenoble Alpes, CNRS, Grenoble INP, LIG,
38000 Grenoble, France
maximin.coavoux@univ-grenoble-alpes.fr

RÉSUMÉ

Cet article propose d'analyser les apports d'un modèle de langue pré-entraîné de type BERT (*bidirectional encoder representations from transformers*) à l'analyse syntaxique en constituants discontinus en anglais (PTB, Penn Treebank). Pour cela, nous réalisons une comparaison des erreurs d'un analyseur syntaxique dans deux configurations (i) avec un accès à BERT affiné lors de l'apprentissage (ii) sans accès à BERT (modèle n'utilisant que les données d'entraînement). Cette comparaison s'appuie sur la construction d'une suite de tests que nous rendons publique. Nous annotons les phrases de la section de validation du Penn Treebank avec des informations sur les phénomènes syntaxiques à l'origine des discontinuités. Ces annotations nous permettent de réaliser une évaluation fine des capacités syntaxiques de l'analyseur pour chaque phénomène cible. Nous montrons que malgré l'apport de BERT à la qualité des analyses (jusqu'à 95 en F_1), certains phénomènes complexes ne sont toujours pas analysés de manière satisfaisante.

ABSTRACT

What does BERT contribute to discontinuous constituency parsing ? A test suite to evaluate discontinuous constituency structure predictions in English.

We propose to analyse the contributions of a pretrained language model such as BERT to discontinuous constituency parsing of English (Penn Treebank). To do so, we perform a comparison of a parsing model in two experimental configuration (i) BERT fine-tuning (ii) without BERT. The comparison relies on the construction of a test suite that we release publicly. We manually annotate the sentences from the development section of the Penn Treebank with information about syntactic phenomena causing discontinuities. We use these annotations to evaluate the syntactic capabilities of a parser for each target phenomenon. Our experiments show that despite the contributions of BERT to very high scores (approaching 95 F_1), certain complex syntactic phenomena are still not identified reliably.

MOTS-CLÉS : Analyse syntaxique en constituants discontinus, analyse d'erreur, discontinuités syntaxiques, suite de tests.

KEYWORDS: Discontinuous constituency parsing, error analysis, syntactic discontinuities, test suite.

Volume 2 : Traitement Automatique des Langues Naturelles, pages 189–196.
HAL : hal-02784767.

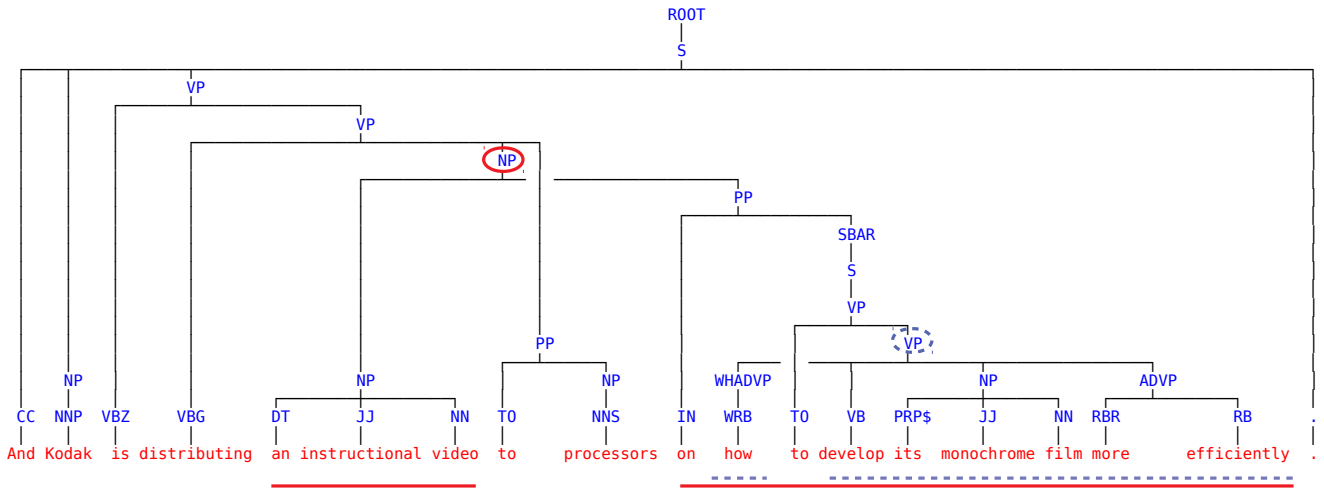

FIGURE 1 – Un arbre syntaxique discontinu issu du Discontinuous Penn Treebank (Evang, 2011). Les discontinuités sont dues à un syntagme prépositionnel (PP) extraposé et à une extraction-*Qu*. Les deux constituants discontinus correspondant ainsi que leurs paires d'empans respectives sont indiquées en rouge / trait plein (NP) et indigo / trait pointillé (VP).

1 Introduction

Dans le cas des langues pour lesquelles on dispose de beaucoup de ressources linguistiques, les analyseurs syntaxiques en constituants qui combinent des systèmes de scorage basés sur les réseaux de neurones et des modèles de langue pré-entraînés tels que BERT (Devlin *et al.*, 2019) obtiennent d'excellents résultats (Kitaev & Klein, 2018; Kitaev *et al.*, 2019; Zhou & Zhao, 2019), atteignant fréquemment plus de 95 F_1 sur le Penn Treebank. Cependant, l'évaluation exclusivement quantitative des outils de traitement automatique des langues (TAL) peut être trompeuse. En effet, un unique score F_1 est difficile à interpréter, et pénalise de la même façon des erreurs sans importance ou des erreurs qui ont un impact sur l'interprétation de la phrase. Ainsi, pour comprendre les capacités des analyseurs actuels, il est nécessaire d'utiliser des méthodes d'évaluation alternatives qui ciblent des phénomènes spécifiques.

Dans cet article, nous construisons une suite de tests pour automatiser l'analyse des erreurs commises sur les phrases présentant des discontinuités syntaxiques du Penn Treebank. Dans un arbre en constituants discontinus (figure 1), il est possible que les tokens dominés par un nœud représentent plusieurs empans (*spans*) non contigus de la phrase. Par exemple, le groupe nominal (NP) discontinu de la figure 1 matérialise l'extraposition en fin de phrase d'un syntagme prépositionel qui est un modifieur du NP. Ce nœud domine « *an instructional video* » et « *on how [...] efficiently* », mais pas le syntagme prépositionnel (PP) « *to processors* ». De telles structures relèvent de grammaires formelles légèrement sensibles au contexte, telles que les LCFRS (Vijay-Shanker *et al.*, 1987, *linear context-free rewriting system*), c'est-à-dire des grammaires d'une classe plus expressive que les grammaires hors-contexte (CFG). Ce type de structure est particulièrement adéquat pour représenter les phénomènes syntaxiques relatifs aux variations d'ordres des mots, et constitue donc des donneés intéressantes pour évaluer les analyseurs syntaxiques.

Une suite de tests similaire à celle que nous proposons existe pour l'allemand (Maier *et al.*, 2014, Discosuite). Les auteur·es de Discosuite ont sélectionné les 1500 premières phrases présentant des discontuités syntaxiques issues du corpus arboré de l'allemand Tiger (Brants *et al.*, 2004), et ont

annoté pour chacune quel phénomène syntaxique est à l'origine de la discontinuité. Ce jeu de tests permet une évaluation fine des analyseurs syntaxiques, sur des phénomènes cibles réputés difficiles à analyser (extrapositions, *scrambling*, ...). À notre connaissance, un tel jeu de tests n'existe pas pour d'autres langues. Certains travaux présentent une analyse d'erreurs sur le Penn Treebank (Evang, 2011; Coavoux *et al.*, 2019), mais d'une part effectuent une analyse manuelle, arbre par arbre, et d'autres part ne rendent pas publiques leurs annotations. À l'inverse, nous rendons publique notre suite de tests ainsi que des scripts d'évaluation qui permettent d'obtenir de manière automatique une évaluation par phénomène d'un analyseur syntaxique.

En effet, nous utilisons la suite de tests pour analyser et comparer les erreurs commises par un analyseur à l'état de l'art dans deux configurations expérimentales distinctes : (i) entraînement utilisant uniquement les données d'entraînement du Penn Treebank (ii) entraînement utilisant (et affinant) les vecteurs de BERT comme source complémentaire d'information. Nous montrons que même si BERT permet d'atteindre des scores approchant 95% de F_1, il ne permet toujours pas d'analyser certains phénomènes cibles de manière satisfaisante.

2 Processus d'annotations

Nous extrayons tous les arbres présentant au moins un constituant discontinu dans le corpus de validation [1] de la version discontinue du Penn Treebank (Evang & Kallmeyer, 2011; Evang, 2011). Nous ignorons les phrases pour lesquelles la discontinuité provient uniquement du rattachement de la ponctuation. Cela correspond en tout à 266 phrases, c'est-à-dire environ 16% des phrases du corpus. Ensuite, pour chaque phrase, nous annotons manuellement quels phénomomènes syntaxiques sont à l'origine des discontinuités, en suivant la classification de Evang (2011) reprise par Coavoux *et al.* (2019) : (i) extractions à longue distance, (ii) citations avec proposition principale en incise, (iii) extraposition à gauche d'une citation (iv) extraposition-*it* (v) autres extrapositions (vi) inversion sujet-verbe. Nous exemplifions ces phénomènes ci-dessous, en matérialisant en gras le constituant discontinu principal :

i *[...] the worst thing* ***that*** *anyone* ***can do*** *[...]* (phrase 53 du corpus de développement) ;

ii ***The stock market has lost some precursory power ,*** *analysts at the Columbia center claim* ***, because of the growing impact of international developments .*** (1190) ;

iii ***Currently , average pay for machinists is $ 13.39 an hour ,*** *Boeing* ***said .*** (238) ;

iv *But* ***it*** *remains to be seen* ***whether their ads will be any more effective .*** (64) ;

v *[...] as* ***the news*** *spread* ***that Wall Street was moving up*** *[...]* (300) ;

vi ***Says*** *James Norman , the mayor of Ava , Mo.* ***: " I do n't invest in stocks .*** (1575).

Cette classification couvre tous les cas présents dans le corpus [2].

Pour certaines phrases, les discontinuités sont dues à plusieurs occurrences de phénomènes distincts dans la même phrase, par exemple la phrase de la figure 1 comporte une extraposition et une extraction. Il s'agit de 21 phrases avec 2 occurrences et d'une seule phrase avec 3 occurrences. Nous rendons disponibles ces annotations [3] en ligne.

1. Nous avons choisi de travailler sur le corpus de validation pour que les phrases annotées soient distinctes du corpus d'entraînement. Cela permettra d'évaluer n'importe quel analyseur syntaxique entraîné sur le split standard de ce corpus.

2. En pratique, il serait possible d'utiliser une granularité encore plus fine, par exemple distinguer parmi les extractions, celles qui sont dues à une question directe et celles qui sont dues à des propositions relatives.

3. https://gitlab.com/mcoavoux/disco-eval-ptb/-/releases/v1.0_taln2020

	Corpus de validation						Corpus de test					
	P	R	F_1	Disc. P	Disc. R	Disc. F_1	P	R	F_1	Disc. P	Disc. R	Disc. F_1
-BERT (Coavoux & Cohen, 2019)	91.5	91.3	91.4	76.1	66.4	70.9	91.3	90.6	90.9	73.3	62.1	67.3
+BERT	94.9	94.9	94.9	80.5	77.6	79.0	95.0	94.5	**94.8**	76.5	70.9	**73.6**
Δ	+3.4	+3.6	+3.5	+4.4	+11.2	+8.1	+3.7	+3.9	+3.9	+3.2	+8.8	+6.3
Corro (2020)									**94.8**	90.8	49.7	64.2

TABLE 1 – Le modèle sans BERT est le modèle pré-entraîné fourni avec le code de l'analyseur et décrit par Coavoux & Cohen (2019).

3 Expériences

Analyseur syntaxique Nous utilisons l'analyseur discontinu décrit par Coavoux & Cohen (2019) et librement disponible [4]. Il s'agit d'un analyseur statistique par transitions, paramétré par un réseau de neurones modulaire :

— un mot-forme est représenté par la concaténation d'un plongement lexical standard ($\mathbf{w}$) et de la sortie ($\mathbf{c}$) d'un bi-LSTM basé sur la séquence de ses caractères ;

— la séquence de vecteurs pour une phrase $([\mathbf{w}_1;\mathbf{c}_1],[\mathbf{w}_2;\mathbf{c}_2],\ldots,[\mathbf{w}_n;\mathbf{c}_n])$ est ensuite donnée en entrée à un bi-LSTM qui calcule des représentations contextualisées de chaque token ;

— enfin, à chaque étape de l'analyse, un réseau à propagation avant (*feedforward*) prédit une action de parsing à partir des vecteurs contextualisés de tokens extraits d'une configuration de l'analyseur (se reporter à Coavoux & Cohen, 2019, pour plus de détails).

Nous avons augmenté cet analyseur d'une option qui permet d'utiliser les vecteurs contextualisés issus de BERT comme représentation de chaque token, concaténés aux représentations lexicales originelles de l'analyseur. En d'autres termes, nous donnons la séquence $([\mathbf{w}_1;\mathbf{c}_1,\mathbf{b}_1],[\mathbf{w}_2;\mathbf{c}_2,\mathbf{b}_2],\ldots,[\mathbf{w}_n;\mathbf{c}_n,\mathbf{b}_n])$ au bi-LSTM qui cacule les représentations contextualisées, où $\mathbf{b}_i$ est la sortie de BERT pour le token i.

Nous affinons les paramètres de BERT lors de l'entraînement. Nous utilisons le modèle de base de BERT qui prend en compte la casse (Devlin *et al.*, 2019, `bert-base-cased`) via l'interface de la bibliothèque `transformers` [5] (Wolf *et al.*, 2019).

Évaluation générale Nous présentons les résultats des deux modèles dans la table 1. Nous utilisons l'évaluateur `discodop` [6] (van Cranenburgh *et al.*, 2016) avec ses paramètres standards. Ce paramétrage ignore la ponctuation et les racines des arbres. Il neutralise également la distinction entre ADVP (syntagmes adverbiaux) et PRT (particule). Nous présentons les précisions (P), rappels (R) et F mesures (F_1) calculés soit sur l'ensemble des constituants, soit uniquement sur les constituants discontinus (Disc. P/R/F_1). Par exemple, pour l'arbre de la figure 1, ces dernières mesures ne considèreraient que les deux constituants discontinus. Toutes ces mesures prennent en compte les étiquettes des constituants, c'est-à-dire que pour qu'un constituant prédit soit considéré juste, il faut qu'il ait la bonne étiquette et qu'il couvre le(s) bon(s) empan(s).

Le modèle entraîné avec BERT (+BERT) obtient une mesure F_1 de 94.8 sur le corpus de test. Ce résultat est identique à celui publié récemment par Corro (2020), qui est le seul autre résultat utilisant

4. https://gitlab.com/mcoavoux/discoparset
5. https://github.com/huggingface/transformers
6. https://github.com/andreasvc/disco-dop

Phénomène	Effectif	Reconnus		Partiellement reconnus		Précision		Rappel		F_1	
		-BERT	+BERT	-BERT	+BERT	-BERT	+BERT	-BERT	+BERT	-BERT	+BERT
Évaluation avec étiquettes											
Extraction	93	65.6	83.9	81.7	91.4	81.6	87.3	77.2	86.2	79.4	86.7
Citation extraposée	73	89.0	93.2	91.8	94.5	94.4	93.4	90.7	94.7	92.5	94.0
Autre extraposition	39	23.1	59.0	25.6	64.1	90.9	93.1	20.0	54.0	32.8	68.4
Citation avec incise	16	0.0	0.0	93.8	100.0	48.9	53.7	46.0	58.0	47.4	55.8
Extraposition-*it*	12	41.7	75.0	50.0	83.3	85.7	83.3	50.0	83.3	63.2	83.3
Inversion sujet-verbe	6	50.0	66.7	66.7	83.3	100.0	71.4	50.0	62.5	66.7	66.7
Extraction et citation extraposée	6	83.3	100.0	100.0	100.0	100.0	100.0	94.1	100.0	97.0	100.0
Évaluation sans étiquettes											
Extraction	93	65.6	84.9	81.7	91.4	81.3	89	78.7	89	80	89
Citation extraposée	73	89	93.2	91.8	94.5	94.4	93.4	90.7	94.7	92.5	94
Autre extraposition	39	23.1	61.5	25.6	66.7	90.9	96.4	20.4	55.1	33.3	70.1
Citation avec incise	16	81.2	87.5	93.8	100	94.4	97.4	89.5	97.4	91.9	97.4
Extraposition-*it*	12	41.7	75	50	83.3	85.7	83.3	50	83.3	63.2	83.3
Inversion sujet-verbe	6	50	66.7	66.7	83.3	100	85.7	50	75	66.7	80
Extraction et citation extraposée	6	83.3	100	100	100	100	100	93.8	100	96.8	100

TABLE 2 – Évaluation par phénomène pour les deux modèles d'analyse (avec ou sans BERT), sur le corpus de validation. La précision, le rappel et le score F_1 sont calculés uniquement sur les constituants discontinus et correspondent aux mesures Disc. P/R/F_1 de la table 1, décomposées par phénomène.

BERT sur ce jeu de données. Le modèle +BERT obtient une amélioration de 4 points par rapport au modèle sans BERT (-BERT). L'apport de BERT est particulièrement fort sur les constituants discontinus (+8.1 en Disc. F_1 sur le corpus de développement). En particulier, nous observons un très fort effet sur le rappel : le modèle avec BERT est bien plus compétent pour détecter les phénomènes syntaxiques qui produisent des discontinuités. Cet effet est d'autant plus important que les travaux en analyse syntaxique discontinue obtiennent constamment une précision bien plus elevée que le rappel (Maier, 2015; Stanojević & G. Alhama, 2017; Coavoux *et al.*, 2019), ce qui est lié à la rareté des constituants discontinus dans les données d'entraînement. Malgré ces améliorations, le modèle +BERT obtient une mesure F_1 de seulement 73.6 sur le corpus de test, signe que les phénomènes syntaxiques à l'origine des discontinuités ne sont toujours pas identifiés de manière satisfaisante. Pour cette raison, nous proposons une analyse plus fine des résultats de ces analyseurs à l'aide de la suite de tests que nous avons construite.

Processus d'évaluation par phénomène Nous procédons comme suit pour évaluer automatiquement les prédictions des analyseurs sur le corpus de développement. À l'aide de l'évaluateur `discodoop`, nous récupérons une évaluation individuelle pour chaque phrase contenant une discontinuité. En particulier, nous récupérons son score Disc. F_1. Nous considérons que le phénomène annoté pour une phrase donnée a été (i) reconnu si la phrase obtient 100% en Disc. F_1 (ii) partiellement reconnu si son Disc. F_1 est strictement supérieur à 0 (c'est-à-dire si au moins un constituant discontinu a été bien prédit). Dans la table 2 nous rapportons ces résultats par phénomène pour chaque modèle, ainsi que la précision, le rappel et le F_1 (micro-moyenne sur l'ensemble des constituants discontinus) pour ce phénomène. Nous rapportons ces valeurs dans deux cas : le cas standard où on prend en compte les étiquettes des constituants (partie supérieure de la table), et le cas non étiqueté, où il suffit de prédire les bons empans pour qu'on considère que la prédiction d'un constituant est correcte (partie inférieure de la table). Cela nous permet d'isoler les cas où le système fait simplement des erreurs d'étiquetage des constituants.

Lorsqu'il y a plusieurs occurrences de phénomènes cibles pour une phrase, nous ne pouvons pas

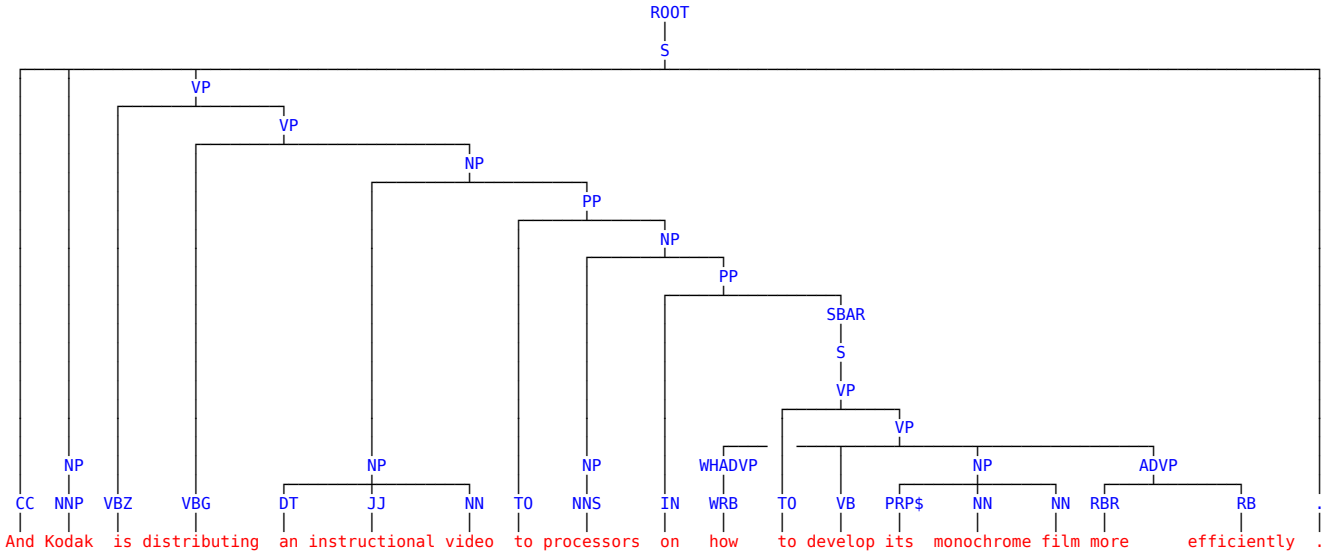

FIGURE 2 – Arbre prédit par le modèle -BERT pour la phrase de la figure 1, présentant des erreurs de rattachement prépositionnel. Le modèle +BERT ne fait aucune erreur sur cette phrase.

conclure de manière automatique lequel des phénomènes a été reconnu ou non (hormis quand le F_1 est à 0 ou 100). Par ailleurs, comme il y a très peu d'occurrences de phrases ayant une combinaison spécifique de 2 phénomènes cibles, nous rapportons des statistiques seulement pour la combinaison la plus fréquente (c'est-à-dire les phrases qui contiennent à la fois une extraction à longue distance et une citation extraposée), et nous ignorons les autres (moins de 6 occurrences chacune).

Résultats par phénomène Nous rapportons l'ensemble des résultats dans la table 2. Les phénomènes produisant des discontinuités les mieux identifiés sont les extractions, les citations extraposées, ainsi que, dans le cas +BERT, les extrapositions-*it*. Notons qu'il s'agit de phénomènes pour lesquels l'analyseur peut s'appuyer sur des indices lexicaux forts : les mots-*Qu* (tels que *what*, *that*, *when*, ...) pour les extractions, les verbes de discours (*say*) pour les citations extraposées, et la présence de *it*. Ce sont également des phénomènes plutôt fréquents.

En revanche, les citations avec incise sont plutôt mal prédites (0% reconnues par les deux modèles). Leur structure est plutôt bien identifiée ($> 80\%$ de reconnaissance non étiquetée), mais les analyseurs font systèmatiquement au moins une erreur sur les étiquettes des constituants discontinus. Ces erreurs d'étiquetage sont sans doute liées à un manque de traits pour classifier les constituants discontinus (Coavoux & Cohen, 2019), ce que nous prévoyons de tester.

Enfin, les extrapositions et les inversions sujet-verbe sont plutôt mal identifiées ($\approx 50\%$ de reconnaissance pour le modèle -BERT). Les extrapositions sont particulièrement difficiles à prédire dans la mesure où elles nécessitent souvent des connaissances syntaxiques et sémantiques fines pour être identifiées. Par exemple, pour désambiguïser un des deux rattachements de syntagmes prépositionnels dans l'arbre de la figure 1, il faut savoir que le sujet de la video (*on how to [...]*) est un modifieur possible pour *video* mais pas pour *processors*. Le modèle +BERT ne fait aucune erreur sur cette phrase, alors que le modèle -BERT attachent les deux prépositions localement (figure 2) [7].

7. Reviewer 2 nous fait également remarquer que *distributing* et *video* ont plusieurs occurrences dans le corpus d'entraînement mais jamais dans des structures syntaxiques similaires à celles que l'on trouve dans cet exemple, suggérant que BERT apporte ici avant tout des informations sur la valence et les cadres de sous-catégorisation de ces lexèmes.

De manière générale, l'entraînement avec BERT améliore tous les résultats en terme de reconnaissance et de score F_1. En particulier, son effet sur la reconnaissance des extrapositions (de 23.1 à 59%) et sur les extrapositions-*it* (41.7 à 75%) est remarquable.

4 Conclusions

Cet article présente une suite de tests permettant d'évaluer automatiquement les prédictions d'un analyseur en constituants discontinus sur un ensemble de phénomènes cibles réputés difficiles à analyser. Nous rendons publique cette suite de tests. Enfin, nous présentons une analyse des erreurs d'un analyseur à l'état de l'art dans deux configurations expérimentales, selon qu'il est entraîné par affinage des paramètres de BERT ou sans BERT. L'analyse met en lumière un résultat prometteur : BERT permet d'obtenir le meilleur résultat publié à ce jour en analyse syntaxique discontinue sur le Penn Treebank, et améliore significativement la prédiction des discontuités syntaxiques présentes dans l'anglais journalistique de l'époque du PTB. Cependant, elle montre également qu'il reste une marge d'amélioration pour ces phénomènes. À l'avenir, nous prévoyons d'expérimenter avec d'autres méthodes d'apprentissage semi-supervisées pour traiter cette limitation des analyseurs actuels.

Remerciements

Je remercie Jibril Frej, Caio Corro, ainsi que 3 relecteurices anonymes pour leurs remarques et suggestions sur cet article.

Références

BRANTS S., DIPPER S., EISENBERG P., HANSEN-SCHIRRA S., KÖNIG E., LEZIUS W., ROHRER C., SMITH G. & USZKOREIT H. (2004). Tiger : Linguistic interpretation of a german corpus. *Research on language and computation*, **2**(4), 597–620.

COAVOUX M. & COHEN S. B. (2019). Discontinuous constituency parsing with a stack-free transition system and a dynamic oracle. 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. 204–217, Minneapolis, Minnesota : Association for Computational Linguistics. DOI : 10.18653/v1/N19-1018.

COAVOUX M., CRABBÉ B. & COHEN S. B. (2019). Unlexicalized transition-based discontinuous constituency parsing. *Transactions of the Association for Computational Linguistics*, **7**, 73–89. DOI : 10.1162/tacl_a_00255.

CORRO C. (2020). Span-based discontinuous constituency parsing : a family of exact chart-based algorithms with time complexities from $\mathcal{O}(n^6)$ down to $\mathcal{O}(n^3)$. arXiv preprint : 2003.13785.

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.

EVANG K. (2011). Parsing discontinuous constituents in English. Mémoire de master, University of Tübingen.

EVANG K. & KALLMEYER L. (2011). PLCFRS parsing of English discontinuous constituents. In *Proceedings of the 12th International Conference on Parsing Technologies*, p. 104–116, Dublin, Ireland : Association for Computational Linguistics.

KITAEV N., CAO S. & KLEIN D. (2019). Multilingual constituency parsing with self-attention and pre-training. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, p. 3499–3505, Florence, Italy : Association for Computational Linguistics. DOI : 10.18653/v1/P19-1340.

KITAEV N. & KLEIN D. (2018). Constituency parsing with a self-attentive encoder. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, p. 2676–2686, Melbourne, Australia : Association for Computational Linguistics. DOI : 10.18653/v1/P18-1249.

MAIER W. (2015). Discontinuous incremental shift-reduce parsing. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1 : Long Papers)*, p. 1202–1212, Beijing, China : Association for Computational Linguistics. DOI : 10.3115/v1/P15-1116.

MAIER W., KAESHAMMER M., BAUMANN P. & KÜBLER S. (2014). Discosuite - a parser test suite for German discontinuous structures. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, p. 2905–2912, Reykjavik, Iceland : European Language Resources Association (ELRA).

STANOJEVIĆ M. & G. ALHAMA R. (2017). Neural discontinuous constituency parsing. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, p. 1666–1676, Copenhagen, Denmark : Association for Computational Linguistics. DOI : 10.18653/v1/D17-1174.

VAN CRANENBURGH A., SCHA R. & BOD R. (2016). Data-oriented parsing with discontinuous constituents and function tags. *Journal of Language Modelling*, **4**(1), 57–111.

VIJAY-SHANKER K., WEIR D. J. & JOSHI A. K. (1987). Characterizing structural descriptions produced by various grammatical formalisms. In *25th Annual Meeting of the Association for Computational Linguistics*, p. 104–111, Stanford, California, USA : Association for Computational Linguistics. DOI : 10.3115/981175.981190.

WOLF T., DEBUT L., SANH V., CHAUMOND J., DELANGUE C., MOI A., CISTAC P., RAULT T., LOUF R., FUNTOWICZ M. & BREW J. (2019). Huggingface's transformers : State-of-the-art natural language processing. arXiv preprint : 1910.03771.

ZHOU J. & ZHAO H. (2019). Head-driven phrase structure grammar parsing on Penn treebank. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, p. 2396–2408, Florence, Italy : Association for Computational Linguistics. DOI : 10.18653/v1/P19-1230.