Pistes pour l'optimisation de modèles de parsing syntaxique

Rayan Ziane, Natasha Romanova

PISTES POUR L'OPTIMISATION DE MODÈLES DE PARSING SYNTAXIQUE

Rayan ZIANE*°, Natasha ROMANOVA*

*CRISCO, Université Caen Normandie; °LLL, Université d'Orléans

INTRODUCTION

Avec l'avènement de technologies d'apprentissage profond, toute analyse syntaxique automatique d'un corpus textuel présuppose l'utilisation de "systèmes d'annotation" (Renwick & Kraif 2024) : 1) un analyseur syntaxique automatique, 2) un gros modèle de langue, ou LLM (issu d'un apprentissage non-supervisé), et 3) un corpus d'entraînement pour l'entraînement supervisé qui permet l'acquisition du schéma d'annotation (en l'occurrence, parsing et tagging dans le système Universal Dependencies (UD), de Marneffe et al., 2021) et des traits linguistiques spécifiques.

La distance en ce qui concerne la variété linguistique (dont les variétés diachronique et diatopique) ainsi que celle de genre ou de registre a un impact sur la performance. L'attention portée à la qualité du corpus d'entraînement pour l'apprentissage supervisé, néanmoins, peut être étendue aux autres étapes du processus d'annotation agile afin d'optimiser l'injection des caractéristiques du corpus cible dans le système de parsing. Nous avançons deux hypothèses : 1) que l'introduction d'une étape supplémentaire avant l'acquisition du schéma d'annotation peut apporter un gain de performance et 2) que la qualité et échantillonnage des "sous-corpus de finetuning" (post-apprentissage du schéma d'annotation) peut augmenter les performances du système d'annotation lors d'entraînements progressifs.

PRÉSENTATION DES CORPUS

Corpus "Guernesey" (GNS) : transcription du registre Crime I, conservé aux Archives du Tribunal de l'île anglo-normande de Guernesey. La source est un texte juridique (dépositions) du 16ème siècle, en français qui présente une forte variation dialectale (normande) au niveau des graphies et de la grammaire - Corpus cible/sous-corpus de finetuning.

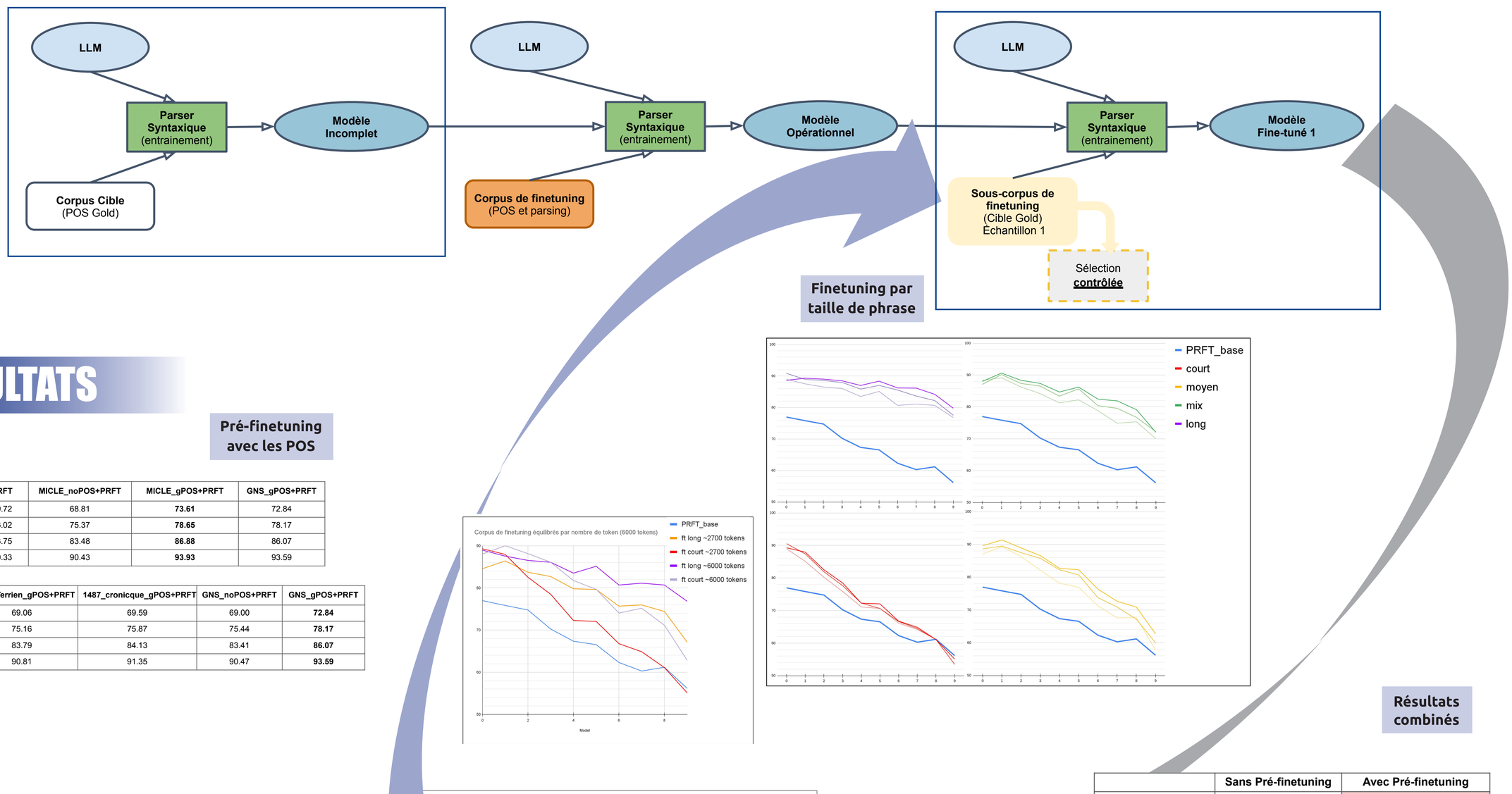
Corpus "Profiterole" (PRFT) : corpus de textes en ancien et moyen français (majoritairement littéraires) du 9ème au 15ème siècle, annotés en Universal Dependencies (Profiterole Old 2@14 et Profiterole Middle 2@14, Prévost et al., 2024) - Corpus de finetuning.

	GNS	PRFT
Tokens	40996	239262
Phrases	1271	20277
Taille moyenne de la phrase	32.25	11.79

CONFIGURATION D'ENTRAINEMENT

- Python
  - LLM: google-bert/bert-base-multilingual-cased
  - Parser: kirianguiller/BertForDeprel
- **Interface Web**
  - Arborator Grew

MÉTHODES



RÉSULTATS

Pré-finetuning avec les POS

	PRFT	MICLE_noPOS+PRFT	MICLE_gPOS+PRFT	GNS_gPOS+PRFT
LAS	69.72	68.81	**73.61**	72.84
UAS	76.02	75.37	**78.65**	78.17
DEPREL	83.75	83.48	**86.88**	86.07
UPOS	90.33	90.43	**93.93**	93.59

	PRFT	1578_Terrien_gPOS+PRFT	1487_cronicque_gPOS+PRFT	GNS_noPOS+PRFT	GNS_gPOS+PRFT
LAS	69.72	69.06	69.59	69.00	**72.84**
UAS	76.02	75.16	75.87	75.44	**78.17**
DEPREL	83.75	83.79	84.13	83.41	**86.07**
UPOS	90.33	90.81	91.35	90.47	**93.59**

Résultats combinés

	Sans Pré-finetuning	Avec Pré-finetuning
PRFT	68,07	**72,84**
ft100_court	72,44	**75,09**
ft100_moyen	78	**79,82**
ft100_mix	81,86	**83,96**
ft100_long	84,06	**85,77**

CONCLUSION

Les résultats des expériences menées montrent l'impact de la réutilisation des PoS en pré-entraînement pour la performance de modèles de parsing syntaxique ainsi que celui du recours au critère de la longueur de la phrase pour la constitution de corpus d'entraînement dans le cadre d'un processus itératif. Les expériences nous amènent à formuler deux recommandations pour une approche à l'annotation agile de treebanks : 1) Prendre en considération la distance et la proximité du corpus d'entraînement et du corpus cible en ce qui concerne les critères linguistiques (langue et variation diachronique et diatopique, genre etc); 2) tenir compte du statut du corpus existant et du corpus cible en ce qui concerne les traitements et enrichissements effectués (segmentation, annotation etc). Combinant les différents paramètres à de différentes étapes de réentraînement progressif on peut réduire l'intervention humaine et augmenter la qualité et la cohérence des annotations.

Travaux à venir

À la suite de l'étude menée, deux pistes de recherche se dégagent: 1) L'application de la méthode pour un processus de bootstrapping avec des langues différentes ; 2) Analyse qualitative des résultats afin d'optimiser l'annotation de certains types de phrases à différentes étapes du traitement.

BIBLIOGRAPHIE

DE MARNEFFE, M.-C., MANNING, C. D., NIVRE, J., & ZEMAN, D. (2021). Universal Dependencies. *Computational Linguistics*, 47(2), 255-308. https://doi.org/10.1162/coli_a_00402

DEVLIN, J., CHANG, M.-W., LEE, K., & TOUTANOVA, K. (2019). BERT: *Pre-training of Deep Bidirectional Transformers for Language Understanding* (No. arXiv:1810.04805). arXiv. https://doi.org/10.48550/arXiv.1810.04805

GROBOL, L., PRÉVOST, S., & CRABBÉ, B. (2021). Is Old French tougher to parse? *Proceedings of the 20th International Workshop on Treebanks and Linguistic Theories* (TLT, SyntaxFest 2021), 27-34. https://aclanthology.org/2021.tlt-1.3 CRIME I, 1563-1567. Guernsey Greffe (manuscrit).

GUIBON, G., TELLIER, I., PRÉVOST, S., CONSTANT, M., & GERDES, K. (2015). Analyse syntaxique de l'ancien français : quelles propriétés de la langue influent le plus sur la qualité de l'apprentissage ? *TALN 22*, http://www.atala.org/taln_archives/TALN/TALN-2015/taln-2015-long-017.pdf. https://hal.science/hal-01251006

GUILLER, K. (2020). Analyse syntaxique automatique du pidgin-créole du Nigeria à l'aide d'un transformer (BERT): Méthodes et Résultats. *Mémoire de Master, Sorbonne Nouvelle*.

MILETIC, A. (2018). Un treebank pour le serbe : *constitution et exploitations* [Phdthesis, Université Toulouse le Mirail - Toulouse II]. https://theses.hal.science/tel-02639473

PENG, Z., GERDES, K., & GUILLER, K. (2022). Pull your treebank up by its own bootstraps. In L. Becerra, B. Favre, C. Gardent, & Y. Parmentier (Éds.), *Journées Jointes des Groupements de Recherche Linguistique Informatique, Formelle et de Terrain (LIFT) et Traitement Automatique des Langues (TAL)* (p. 139-153). CNRS. https://hal.science/hal-03846834

PRÉVOST, S., GROBOL, L., DEHOUCK, M., LAVRENTIEV, A., & HEIDEN, S. (2024). Profiterole : un corpus morpho-syntaxique et syntaxique de français médiéval. *Corpus, 25*. https://doi.org/10.4000/corpus.8538

RENWICK, A., & KRAIF, O. (2024). Annotation de textes d'états de langue anciens : pour le redéploiement de l'existant. *Corpus, 25*. https://doi.org/10.4000/corpus.8286

REMERCIEMENTS

Nous remercions les Archives du Tribunal de Guernesey (Guernsey Greffe) et l'ancien archiviste de l'île Daryl Ogier pour nous avoir accordé l'accès au manuscrit de Crime I. Nous remercions aussi la Direction du Système d'Information de l'Université de Caen pour avoir soutenu ce travail en fournissant l'accès aux ressources informatiques. Cette recherche a été menée dans le cadre des projet ANR Franco-allemand MICLE (2021-2024) et le projet AUTOMATED (2023-2024) financé par la région Normandie, sous la direction de Professeur Pierre Larrivée (CRISCO, Université de Caen).