

Université de Montréal

Préentraînement d'un modèle ELECTRA

par

Simon Théorêt

Département d'informatique et de recherche opérationnelle
Faculté des arts et des sciences

Mémoire présenté en vue de l'obtention du grade de
Maître ès sciences (M.Sc.)
en informatique,

December 20, 2024

Université de Montréal

Faculté des arts et des sciences

Ce mémoire intitulé

Préentraînement d'un modèle ELECTRA

présenté par

Simon Théorêt

a été évalué par un jury composé des personnes suivantes :

Nom du président du jury

(président-rapporteur)

Nom du directeur de recherche

(directeur de recherche)

Nom du membre de jury

(membre du jury)

Résumé

Le logiciel Antidote permet de corriger des textes en français et en anglais. Il détecte plusieurs milliers de types d'erreurs orthographiques et grammaticales. Le logiciel dispose d'un modèle ELECTRA capable de détecter efficacement les erreurs en français. L'équipe de TAL de Druide désire mettre en place un système similaire pour la langue française . Dans le cadre de ce stage, le but est de créer un modèle ELECTRA capable de détecter les erreurs grammaticales en français. Pour ce faire, plusieurs approches ont été testées et les résultats des derniers modèles sont prometteurs. On remarque entre autre une hausse importante des performances en introduisant une tâche intermédiaire et en faisant une recherche d'hyperparamètres.

Mots clés: Apprentissage automatique, Apprentissage profond, Apprentissage machine, Traitement de texte, Détection de mot manquants, Encodeur, Transformers, BERT, Electra, Réseaux de neurones, Intelligence artificielle.

Table des matières

Résumé	5
Liste des tableaux	9
Liste des figures	11
Liste des sigles et des abréviations	13
Remerciements	15
Introduction	1
Chapitre 1. Druide et ELECTRA	3
1.1. Contraintes	3
1.2. Méthode de préentraînement ELECTRA	3
1.3. Architecture ELECTRA	4
1.4. Affinage pour la détection d’erreurs	5
1.5. Infrastructures en place	5
Chapitre 2. Entraînement de modèles initiaux	7
2.1. Normalisation des données et entraînement d’un jetoniseur	7
2.2. Préentraînement initial	8
2.3. Premiers modèles affinés	9
2.4. Entraînement d’un modèle avec casse	10
Chapitre 3. Modèle final	13
Chapitre 4. Autre techniques	15
Chapitre 5. Conclusions	17

Références	19
Annexe A. Le titre.....	21
A.1. Section un de l'Annexe A.....	21
Annexe B. Les différentes parties et leur ordre d'apparition	23

Liste des tableaux

2.1	Hyper-paramètres utilisés pour l'entraînement des modèles	8
2.2	Résultats d'évaluation des trois modèles sur la tâche ELECTRA	9
2.3	Résultats des modèles préliminaires sur corpus d'évaluation de la tâche de correction.....	10
2.4	Résultats d'évaluation des trois modèles et du modèle avec casse sur la tâche ELECTRA	10

Liste des figures

1.1	Exemple de la méthode ELECTRA. Figure provenant de [1].....	4
2.1	Mesure F-0.5 des trois premiers modèles selon le nombre d'itérations sur données d'évaluation pour la détection d'erreurs en français.	9
2.2	Mesure F-0.5 du modèle avec casse selon le nombre d'itérations sur données d'évaluation pour la détection d'erreurs en français.	11

Liste des sigles et des abréviations

MLM	Modélisation de langage avec masque, de l'anglais <i>Masked Language Modeling</i>
TAL	Traitement automatique du langage
NER	Reconnaissance d'entités, de l'anglais <i>Named-Entity Recognition</i>
DDP	Parallélisme distribué des données, de l'anglais <i>Distributed Data Parallel</i>

Remerciements

Je tiens remercier Joss, pour sa précieuse aide tout au long de mon stage. Je n'aurais pas pu demander un meilleur superviseur.

Je remercie aussi Momo pour son support moral constant.

Introduction

Le domaine du traitement automatique des langues connaît une explosion fulgurante de techniques, de jeux de données et de modèles permettant de résoudre de nouveaux problèmes. Néanmoins, bon nombre de ces applications restent hors de portée des organisations désirant mettre en application des outils d'apprentissages automatique. En effet, la plupart des modèles de langues récents sont préentraînés sur des corpus majoritairement anglophones, avec des jetoniseurs spécialisés pour traiter le contenu anglophone. Ces deux facteurs limitent les modèles préentraînés disponibles ainsi que leur performance sur des tâches avec un corpus non anglophone.

Druide Inc. est une compagnie basée à Montréal dont le principal produit est Antidote, un logiciel de correction orthographique et grammaticale. Leur logiciel phare fait déjà usage de l'apprentissage profond pour leur moteur de correction en anglais, en plus d'utiliser un correcteur symbolique pour certains types d'erreurs. Le modèle utilisé en production pour la correction en anglais fait près de 2 corrections sur 3 et représente une part importante du moteur de correction. L'équipe de Druide désire mettre en place un modèle de correction similaire, mais adapté à la langue française. En particulier, ils désirent préentraîner un modèle ELECTRA avec un corpus et un jetoniseur francophones pour que le modèle puisse détecter les erreurs grammaticales présentes dans les textes des utilisateurs d'Antidote.

Pour la réalisation du projet, nous disposons d'un jeu de données d'environ 40 GB de données non structuré. De plus, l'entraînement du modèle se fait localement sur une machine ayant accès à 3 NVIDIA RTX A4000, disposant chacune de 16 GB de mémoire VRAM.

Chapitre 1

Druide et ELECTRA

L'équipe de Druides dispose de deux modèles déjà en place pour la correction des erreurs. Néanmoins, leur modèle en anglais corrige une plus grande gamme d'erreurs. Druides désire améliorer leur moteur de correction en français à l'aide de l'apprentissage profond.

1.1. Contraintes

Le modèle doit être intégré dans le logiciel principal de Druides, Antidote. Or, le logiciel Antidote est déployé sur les ordinateurs personnels des usagers. Cela implique d'importantes contraintes quant aux ressources disponibles pour l'exécution du modèle, notamment en ce qui a trait à la consommation de mémoire. De plus, le logiciel Antidote se doit d'être rapide, puisque attendre plusieurs minutes pour la correction d'un texte volumineux dégrade la qualité de l'expérience des utilisateurs. En d'autres mots, le modèle doit être rapide durant l'inférence. Finalement, le logiciel antidote cible deux systèmes d'exploitation: Windows et MacOS. Le déploiement du modèle sur les machines des usagers se fait à l'aide des bibliothèques ONNX[2] et CoreML. Il est donc nécessaire que le modèle soit supporté par les deux bibliothèques. En résumé, nous avons des limites quant aux ressources disponibles durant l'inférence ainsi que des contraintes quant aux couches et modèles utilisables.

Ces contraintes ont poussé l'équipe du TAL de Druides à sélectionner des petits modèles Transformers[5] avec encodeur. Ces derniers contiennent environ 14 millions de paramètres.

1.2. Méthode de préentraînement ELECTRA

La méthode ELECTRA[1] est une méthode inspirée de la modélisation de langage avec masque (Masked Language Modeling; MLM), mais qui se veut plus efficace et rapide que le MLM. La méthode ELECTRA consiste à entraîner deux modèles: un petit modèle, appelé le générateur, et le modèle final, appelé le discriminant. Le générateur reçoit des jetons masqués

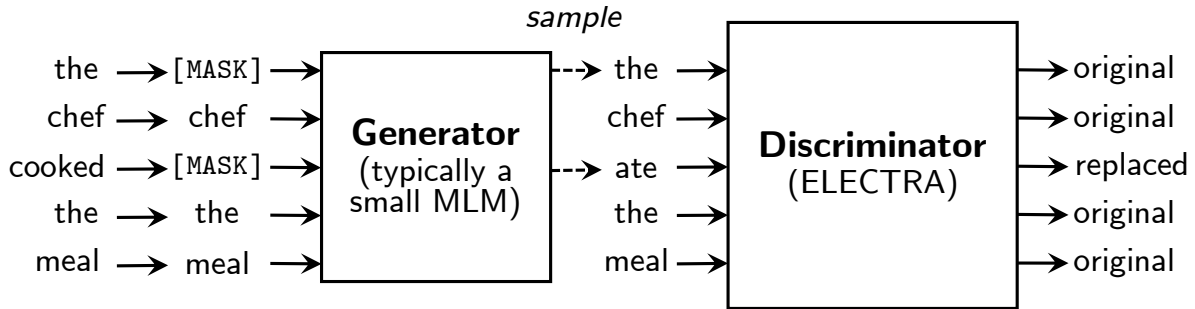


Fig. 1.1. Exemple de la méthode ELECTRA. Figure provenant de [1].

et doit prédire quel était le jeton original situé à la position du masque. Les prédictions du modèle sont échantillonnées, de façon à obtenir une nouvelle séquence, potentiellement différente de la séquence originale. Le discriminant reçoit la nouvelle séquence et a pour tâche de prédire quels jetons sont corrompus et lesquels n'ont pas été modifiés par le générateur. Seul le discriminant est réutilisé pour l'affinage. La méthode est visualisée dans la figure 1.1.

Trois éléments rendent l'entraînement du discriminant plus facile. Premièrement, le générateur dispose de significativement moins de capacité que le discriminant. En effet, ce dernier contient en général 3 à 4 fois plus de paramètres (en excluant les couches de projections *embeddings*) que le générateur. De plus, les entrées du discriminant sont échantillonnées depuis la distribution engendrée par le générateur, au lieu de sélectionner les entrées les plus probables selon la distribution du générateur. Finalement, les poids du générateur sont initialisés aléatoirement et ce dernier est entraîné en même temps que le discriminant, rendant la tâche de plus en plus difficile au fur et à mesure que le générateur s'entraîne. Ces trois facteurs rendent la tâche du discriminant plus facile et permettent de générer des erreurs similaire à ce que le modèle rencontrera en production.

La méthode ELECTRA a été choisie pour deux raisons: c'est une méthode de préentraînement similaire à la correction d'erreurs dans un texte et la méthode ELECTRA permet d'augmenter l'efficacité du préentraînement en atteignant des performances similaires aux performances du MLM en moins d'itérations.

1.3. Architecture ELECTRA

Le modèle ELECTRA utilise une architecture basée sur les modules d'encodeurs des Transformers [5]. L'usage d'une architecture basée sur les transformeur permet d'obtenir une représentation contextuelle pour tous les jetons d'une séquence. Cette avancée a été marquée par l'arrivée de [3], un modèle Transformers (Bidirectional Encoder Representation

from Transformers). Ce modèle a été développé par Google en 2018 et comprend de nombreuses versions de différentes tailles et entraînés sur différentes tâches. Les deux versions canoniques de BERT sont BERT-BASE et BERT-LARGE. Ces deux versions comprennent respectivement 12 couches et 24 couches, chacune étant composée de 768 unités de large, divisées en 12 têtes d'attention multiples. ELECTRA ajoute une nouvelle version plus petite de BERT, dénommée ELECTRA-small. Celle-ci consiste en 12 couches de 256 unités de large. Ces modèles à base d'encodeur sont composées de trois parties:

- Un **jetoniseur**, qui s'occupe de traiter le texte entrant et de le convertir en une séquence de d'entiers.
- Une **couche de projection**, qui permet d'associer à chacun des jetons d'entrées une représentation vectorielle qui dépend de la position du jeton dans la séquence ainsi que du jeton lui-même.
- Un module d'**encodeur**, qui permet d'obtenir une représentation contextualisée des entrées. Cette représentation est apprise et varie selon la tâche finale du modèle ainsi que le corpus utiliser pour entraîner le modèle.

Notre modèle ELECTRA utilise l'architecture ELECTRA-small.

1.4. Affinage pour la détection d'erreurs

Une fois le modèle ELECTRA préentraîné, il est nécessaire d'adapter le modèle pour que celui-ci soit en mesure de détecter efficacement les erreurs dans les textes des utilisateurs. *Druide* a développé une liste des différents types d'erreurs, permettant de classer les différents types d'erreurs en de grandes catégories, telles que les erreurs de virgules, les erreurs de mots manquants, les erreurs d'accord du nom, etc. Cette liste contient 750 différents types d'erreurs. Chaque erreur fait partie d'une de ces grandes catégories, et bon nombre de ces erreurs ont une sous-catégorie, précisant encore plus le contexte associée à l'erreur. La détection d'erreur est modélisée comme une tâche de détection d'entité nommée (DEN/NER), dans laquelle chaque jeton dispose d'une classe. Les classes d'erreurs sont représentées avec un identifiant, tandis que la classe représentant l'absence d'erreurs est représentée par l'identifiant *O*. Le modèle a comme objectif de spécifier la classe de chaque jeton de la séquence. Le schéma *IOB2* [4] est utilisé pour représenter sans ambiguïté les jetons contigus contenus dans la même erreur.

1.5. Infrastructures en place

Notre tâche principale consistait à préentraîner un modèle ELECTRA. Or, un modèle ELECTRA est déjà utilisé pour la tâche de correction en anglais. Ce dernier n'a pas été préentraîné par *Druide*. En effet, la librairie Transformers[6] permet un usage libre de

différents modèles ELECTRA préentraînés. De plus, il existe quelques modèles ELECTRA préentraînés sur des corpus francophone. Cependant, aucun d'entre eux ne respectent nos contraintes de tailles et de vitesse. Il est donc nécessaire d'entraîner un modèle à partir d'une initialisation aléatoire.

Nous disposons de deux corpus déjà préparés pour préentraîner et affiner un modèle Electra. Le corpus de préentraînement est une collection de textes non structuré provenant de nombreuses sources, notamment des manuels, des articles de blogues, des livres. Ce corpus de préentraînement est appelé corpus des Combis et représente 40 gigaoctes (Go) de données et 7 milliard de jetons. C'est un corpus deux fois plus grand que le corpus de préentraînement utilisé pour le préentraînement par Google du modèle ELECTRA de même taille. Pour l'affinage, Druide dispose d'un corpus contenant près de 100000 annotations sur des textes francophones. Ces annotations sont fournies par Druide et proviennent d'équipes de linguistes et d'annotateurs corrigeant des textes et classifiant les erreurs qu'ils y trouvent en fonction des types d'erreurs proposés par Druide.

Chapitre 2

Entraînement de modèles initiaux

Le préentraînement d'un modèle de langue se fait en trois étapes. Il est nécessaire de pré-traiter les données, de sélectionner un jetoniseur adapté à la tâche ainsi que d'entraîner le modèle sur la tâche de préentraînement.

2.1. Normalisation des données et entraînement d'un jetoniseur

La normalisation consiste à réduire le nombre de caractères différents contenus dans le corpus. C'est une étape importante puisqu'elle permet de réduire la taille du vocabulaire du jetoniseur sans pour autant perdre des éléments syntaxiques. Notre normalisation consistait à transformer tous les espaces en le même caractère d'espace (espaces insécables, tabulations), de transformer tous les guillemets (guillemets français, guillemets informatiques, etc.) en guillemets anglais, de retirer les espaces en trop et modifier les types d'apostrophes pour que ceux-ci soient uniformes. La normalisation modifie aussi certains les espacements entre certains mots, de façon à ce que par exemple le texte "11 ème étage" devienne "11ème étage".

Une fois le texte normalisé, il est possible d'entraîner un jetoniseur adapté à la tâche. En l'occurrence, nous avons sélectionné le jetoniseur Wordpiece. C'est le jetoniseur choisi par les auteurs de l'article de ELECTRA et est actuellement utilisé en production chez Druides. Il répond donc à nos contraintes. Pour l'entraînement du jetoniseur, nous utilisons le corpus des combis normalisé, comprenant environ 40 GO de données. Les hyper-paramètres sélectionnés pour le jetoniseur wordpiece sont donnés dans le tableau 2.1.

2.2. Préentraînement initial

Nous préentraînons le modèle à l’aide de l’implémentation originale de ELECTRA. Celle-ci est disponible au dépôt git suivant: <https://github.com/google-research/electra>. Notre machine dédiée à l’entraînement dispose de 3 GPUs de 16GB de VRAM chacun. Or, l’implémentation originale utilise moins de 16GB de VRAM et ne permet pas l’usage de DDP, limitant la vitesse d’entraînement ainsi que la taille de lot (*batch size*). Pour tirer parti au maximum des 3 GPUs disponibles, nous avons initialisé aléatoirement trois modèles ELECTRA et avons changé l’ordre des données. En entraînant 3 modèles sur le même jeu de données modifié, nous espérons être en mesure d’appliquer la méthode de la soupe [7] une fois les modèles affinés et faire un meilleur usage des ressources computationnelles à notre disposition. Les hyper-paramètres utilisés durant l’entraînement des trois modèles sont décrits dans le tableau 2.1. Nous avons divisé par 4 la largeur du générateur, tel que recommandé dans l’article original de ELECTRA [1].

Hyper-paramètres	Discriminant	Générateur
Nombre de couches	12	12
Taille des couches cachées	256	64
Taille du MLP cachés	15	15
Taille des couches de projection	128	128
Nombre de têtes d’attention	4	1
Taille du vocabulaire	30522	30522
Ignore la casse	oui	oui
% des jetons masqués	-	15
Taux d’apprentissage	5e-4	5e-4
Poids de la perte du générateur	-	1
Poids de la perte du discriminant	50	-
Nombre d’itération de <i>warmup</i>	10000	10000

Tableau 2.1. Hyper-paramètres utilisés pour l’entraînement des modèles

Une fois le préentraînement complété, nous avons obtenu les résultats suivants durant l’évaluation des modèles sur la tâche d’ELECTRA:

Malgré le changement de l’ordre des données, aucun des trois modèles n’a performé différemment des autres modèles. Cette remarque nous a poussé à utiliser la méthode de la soupe [7] dans le but d’améliorer la performance du modèle. Cette méthode sera discutée plus en détails dans la section 4

Métrique	Modèle 1	Modèle 2	Modèle 3
Exactitude du discriminant	0.952	0.949	0.950
AUC du discriminant	0.934	0.929	0.934
Perte du discriminant	0.135	0.142	0.138
Précision du discriminant	0.801	0.794	0.793
Rappel du discriminant	0.469	0.449	0.483
Perte totale	8.967	9.411	9.276
Exactitude du générateur	0.570	0.562	0.550
Perte du générateur	2.216	2.308	2.342
Exactitude du générateur	0.469	0.458	0.448

Tableau 2.2. Résultats d'évaluation des trois modèles sur la tâche ELECTRA

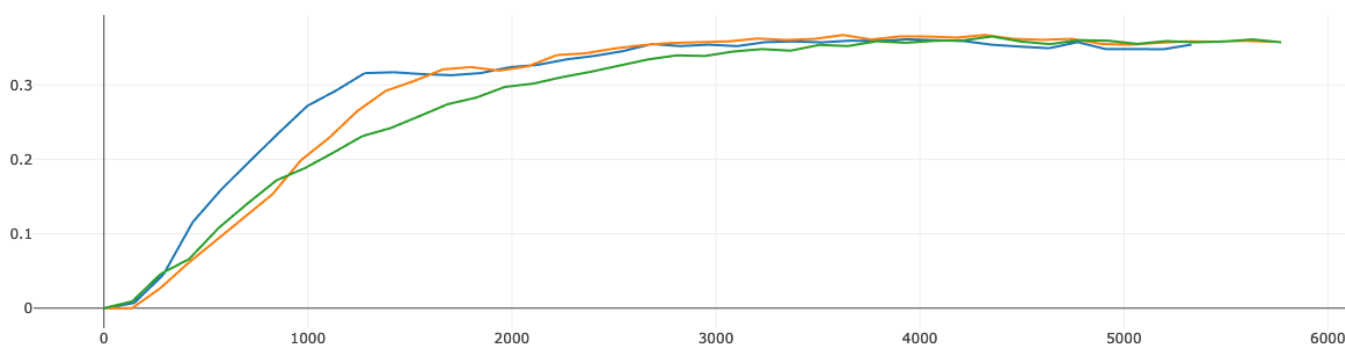


Fig. 2.1. Mesure F-0.5 des trois premiers modèles selon le nombre d'itérations sur données d'évaluation pour la détection d'erreurs en français.

2.3. Premiers modèles affinés

Une fois les trois premiers modèles initiaux préentraînés, nous avons entraîné ces trois modèles sur la tâche de détection des erreurs en français. Les meilleures métriques enregistrées sont détaillées dans le tableau 2.3

Les mesures F-0.5 obtenues sur l'ensemble de test durant l'affinage sont contenues dans la figure 2.1.

Modèle	Précision	Rappel	F0.5
Modèle initial 1	39.59	23.84	34.97
Modèle initial 2	40.22	24.36	35.63
Modèle initial 3	39.85	24.88	35.57

Tableau 2.3. Résultats des modèles préliminaires sur corpus d’évaluation de la tâche de correction

2.4. Entraînement d’un modèle avec casse

Nos premiers modèles preentraînés sont limités en terme de performance par leur jetoniseur. En effet, le jetoniseur WordPiece entraîné sur les données ignore la casse, et certains types d’erreurs, tel que les majuscules pour les noms propres, nécessite de connaître la casse des mots. Ce choix n’affecte que peu la performance du modèle durant le préentraînement, mais pose problème durant l’affinage sur la tâche de détection des erreurs. Nous avons donc entraîné un second jetoniseur WordPiece, cette fois-ci sensible à la casse et avec tous les autres hyper-paramètres identiques. Nous avons par la suite entraîné le modèle sur l’ensemble des données, encore une fois avec l’ordre des données mélangée et avec les hyper-paramètres du tableau 2.1. Les résultats du modèle avec casse ainsi que les modèles préliminaires sont contenus dans le tableau 2.4.

Métrique	Modèle 1	Modèle 2	Modèle 3	Modèle avec casse
Exactitude du discriminant	0.952	0.949	0.950	0.946
AUC du discriminant	0.934	0.929	0.934	0.931
Perte du discriminant	0.135	0.142	0.138	0.149
Précision du discriminant	0.801	0.794	0.793	0.795
Rappel du discriminant	0.469	0.449	0.483	0.471
Perte totale	8.967	9.411	9.276	10.31
Exactitude du générateur	0.570	0.562	0.550	0.499
Perte du générateur	2.216	2.308	2.342	2.848
Exactitude du générateur	0.469	0.458	0.448	0.403

Tableau 2.4. Résultats d’évaluation des trois modèles et du modèle avec casse sur la tâche ELECTRA

La performance du discriminant sensible à la casse est similaire à celles des autres discriminant. Cependant, le générateur sensible à la casse performe moins bien que les autres. Une explication possible est que la tâche de prédire le bon jeton en place en plus

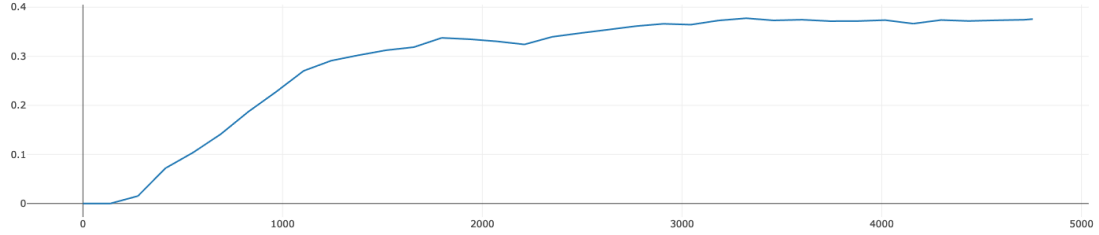


Fig. 2.2. Mesure F-0.5 du modèle avec casse selon le nombre d'itérations sur données d'évaluation pour la détection d'erreurs en français.

difficile avec la casse, puisque les jetons contenant la casse sont souvent syntaxiquement identiques, mais sont traités comme étant différents durant l'évaluation.

Nous avons complété l'affinage du modèle en l'entraînant sur la tâche de détection des erreurs en français. Nous avons réutilisé les hyper-paramètres des trois modèles précédents, contenus dans le tableau

Les mesures F-0.5 obtenues sur l'ensemble de test durant l'affinage pour le modèle avec casse sont contenues dans la figure 2.2.

Chapitre 3

Modèle final

Chapitre 4

Autre techniques

Chapitre 5

Conclusions

Références

- [1] Kevin Clark, Minh-Thang Luong, Quoc V. Le, et Christopher D. Manning, *Electra: Pre-training text encoders as discriminators rather than generators*, 2020.
- [2] ONNX Runtime developers, *Onnx runtime*, <https://onnxruntime.ai/>, 2021, Version: x.y.z.
- [3] Jacob Devlin, Ming-Wei Chang, Kenton Lee, et Kristina Toutanova, *Bert: Pre-training of deep bidirectional transformers for language understanding*, 2019.
- [4] Lance A. Ramshaw et Mitchell P. Marcus, *Text chunking using transformation-based learning*, 1995.
- [5] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, et Illia Polosukhin, *Attention is all you need*, 2023.
- [6] 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, et Alexander M. Rush, *Transformers: State-of-the-art natural language processing*, Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations (Online), Association for Computational Linguistics, October 2020, pp. 38–45.
- [7] Mitchell Wortsman, Gabriel Ilharco, Samir Yitzhak Gadre, Rebecca Roelofs, Raphael Gontijo-Lopes, Ari S. Morcos, Hongseok Namkoong, Ali Farhadi, Yair Carmon, Simon Kornblith, et Ludwig Schmidt, *Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time*, 2022.

Annexe A

Le titre

A.1. Section un de l'Annexe A

...texte...

Annexe B

Les différentes parties et leur ordre d'apparition

J'ajoute ici les différentes parties d'un mémoire ou d'une thèse ainsi que leur ordre d'apparition tel que décrit dans le guide de présentation des mémoires et des thèses de la Faculté des études supérieures. Pour plus d'information, consultez le guide sur le site web de la faculté (www.fes.umontreal.ca).

Ordre des éléments constitutifs du mémoire ou de la thèse		
1.	La page de titre	obligatoire
2.	La page d'identification des membres du jury	obligatoire
3.	Le résumé en français et les mots clés français	obligatoires
4.	Le résumé en anglais et les mots clés anglais	obligatoires
5.	Le résumé dans une autre langue que l'anglais ou le français (si le document est écrit dans une autre langue que l'anglais ou le français)	obligatoire
6.	Le résumé de vulgarisation	facultatif
7.	La table des matières, la liste des tableaux, la liste des figures ou autre	obligatoires
8.	La liste des sigles et des abréviations	obligatoire
9.	La dédicace	facultative
10.	Les remerciements	facultatifs
11.	L'avant-propos	facultatif
12.	Le corps de l'ouvrage	obligatoire
13.	Les index	facultatif
14.	Les références bibliographiques	obligatoires
15.	Les annexes	facultatifs
16.	Les documents spéciaux	facultatifs