Institut National des Langues et Civilisations Orientales

Département Textes, Informatique, Multilinguisme

Nœuds, arêtes, clusters... À quoi pense BERT ?

Évaluation de modèles de langue par réseaux complexes : application à l'étiquetage de séquences en chinois classique

MASTER

TRAITEMENT AUTOMATIQUE DES LANGUES

Parcours :

Ingénierie Multilingue

par

Clément BUON

Encadrant :

Pierre Magistry

Année universitaire 2023/2024

TABLE DES MATIÈRES

Liste des figures **5**

Liste des tableaux **5**

Avant-propos **7**

Introduction et motivation **9**

I Contexte général **11**

1 Description du projet **13**
1.1 Cadre du stage 13
1.2 Tâches confiées 15

2 État de l'art **17**
2.1 Modèles BERT 17
2.2 Chinois classique et ressources TAL 21
2.3 Approche par réseaux complexes 25

II Méthode, expériences et résultats **31**

3 Méthodes et expériences **33**
3.1 Motivation et question de recherche 33
3.2 Corpus utilisés 35
3.3 Organisation des expériences 39
3.4 Pré-entraînement de modèles 42

4 Résultats **43**
4.1 Résultats en mesures d'évaluation 43
4.2 Mesures sur les graphes issus des réseaux d'attention 44
4.3 Corrélations entre mesures d'évaluation et propriétés des graphes . . . 45

IIIDiscussion et perspectives **47**

5 Discussion **49**
5.1 Discussion des mesures d'évaluation 49
5.2 Discussion des mesures sur les graphes issus des réseaux d'attention . . 50
5.3 Discussion des corrélations entre mesures d'évaluation et propriétés des graphes 50

6 Perspectives **53**

Conclusion générale **55**

Bibliographie **57**

A Annexes **59**

Liste des figures

1.1 Distribution des langues sinitiques dans le monde chinois. Source : *Wikipedia* . . . 14
2.1 Représentation d'un perceptron multicouche . . . 18
2.2 Décomposition d'un sinogramme . . . 22
2.3 De gauche à droite, formes traditionnelle et simplifiée d'un sinogramme . . . 23
2.4 Distribution de loi de puissance . . . 27
2.5 Réseaux de cooccurrence de la phrase d'exemple . . . 28
2.6 Extraction d'un graphe reliant les tokens les plus proches dans la matrice d'attention du modèle SIKU-BERT . . . 28
3.1 Proportion de chaque tag dans le corpus CHisIEC . . . 35
3.2 Pipeline globale pour la génération des réseaux . . . 39
5.1 Lois de puissance selon le paramètre γ . . . 51

Liste des tableaux

3.1 Ensemble des 22 étiquettes pour l'annotation en parties du discours du corpus ACC, avec exemples et traductions en anglais . . . 38
4.1 Classement des corpus par ordre décroissant de F-mesure moyenne . . . 43
4.2 Classement des modèles par rang cumulé décroissant . . . 44
4.3 Étendue des F-mesures obtenues par les modèles HuggingFace par corpus . 44
4.4 F-mesures moyennes obtenues par le modèle Chinese BERT pré-entraîné sur le corpus Wikisource . . . 44
4.5 Moyennes des valeurs calculées sur les graphes issus des réseaux d'attention de chaque modèle . . . 45
A.1 Mesures obtenues avec le modèle SIKU-BERT N : nombre de nœuds, E : nombre d'arêtes, $< k >$: degré moyen, r : coefficient de la loi de puissance estimée par le modèle, R2 : *fitting coefficient*, L : longueur moyenne des plus courts chemins, C : coefficient de clustering, D : diamètre du graphe . . 60
A.2 Mesures obtenues avec le modèle Chinese BERT N : nombre de nœuds, E : nombre d'arêtes, $< k >$: degré moyen, r : coefficient de la loi de puissance estimée par le modèle, R2 : *fitting coefficient*, L : longueur moyenne des plus courts chemins, C : coefficient de clustering, D : diamètre du graphe . . 61
A.3 Mesures obtenues avec le modèle Jihuai N : nombre de nœuds, E : nombre d'arêtes, $< k >$: degré moyen, r : coefficient de la loi de puissance estimée par le modèle, R2 : *fitting coefficient*, L : longueur moyenne des plus courts chemins, C : coefficient de clustering, D : diamètre du graphe . . . 62

A.4 Mesures obtenues avec le modèle Guwen BERT N : nombre de nœuds, E : nombre d'arêtes, $< k >$: degré moyen, r : coefficient de la loi de puissance estimée par le modèle, R2 : *fitting coefficient*, L : longueur moyenne des plus courts chemins, C : coefficient de clustering, D : diamètre du graphe . . 63
A.5 Mesures obtenues avec le modèle Guji BERT N : nombre de nœuds, E : nombre d'arêtes, $< k >$: degré moyen, r : coefficient de la loi de puissance estimée par le modèle, R2 : *fitting coefficient*, L : longueur moyenne des plus courts chemins, C : coefficient de clustering, D : diamètre du graphe . . 64
A.6 Mesures obtenues avec le modèle RoBERTa N : nombre de nœuds, E : nombre d'arêtes, $< k >$: degré moyen, r : coefficient de la loi de puissance estimée par le modèle, R2 : *fitting coefficient*, L : longueur moyenne des plus courts chemins, C : coefficient de clustering, D : diamètre du graphe 65
A.7 Mesures obtenues avec notre modèle Chinese BERT pré-entraîné sur le corpus Wikisource N : nombre de nœuds, E : nombre d'arêtes, $< k >$: degré moyen, r : coefficient de la loi de puissance estimée par le modèle, R2 : *fitting coefficient*, L : longueur moyenne des plus courts chemins, C : coefficient de clustering, D : diamètre du graphe 66
A.8 Tableau des modèles avec les métriques de précision, rappel et F-mesure pour chaque corpus 67

AVANT-PROPOS

Remerciements

Je tiens à remercier de tout cœur l'équipe ERTIM qui m'a accompagné pendant tout ce stage, qui a été à mon écoute, avec qui on a pu échanger, rire, se livrer, et bien entendu travailler. Je remercie tout particulièrement mon encadrant, Pierre Magistry, qui m'a guidé avec bienveillance tout au long de mon stage, et qui s'est montré très flexible compte tenu de mes engagements en dehors.

Merci à ma famille, qui m'a soutenu pendant le stage et la rédaction de ce mémoire.

Je remercie Fanny et Tifanny, avec qui nous nous sommes serré les coudes tout au long du master.

Merci à Bénédicte, Samuel, Mathieu, Damien, Ilaine, Johanna, Stéphane de constituer une équipe aussi accueillante et chaleureuse, j'ai eu beaucoup de plaisir à travailler au sein de cette équipe.

Je remercie Agathe et Laura, sans qui les moments de doute et d'inquiétude auraient été insurmontables, et avec qui malgré les médisances de certains et quelques parties de console (occasionnelles) où le ton a pu monter, je suis resté en excellent termes. Merci pour tous les moments que nous avons passés ensemble entre les murs de la Maison de la Recherche !

Résumé

Ce mémoire explore les relations entre la structure interne des têtes d'attention de modèles de langue dont l'architecture se base sur le Transformer, et les performances en fin de tâche pour du POS-tagging et de la Reconnaissance d'Entités Nommées sur des corpus en chinois classique. Pour explorer cette question, nous étudions six modèles déployés sur HuggingFace et un modèle pré-entraîné par nous-même, sur 3 corpus présentant chacun des variétés différentes de chinois classique. Nous aboutissons à une division en deux familles de modèles : ceux pour lesquels des corrélations significatives existent, et ceux pour lesquels rien ne corrèle.

Mots-clés : BERT, chinois classique, introspection de modèle, LLM, réseaux complexes

Introduction et motivation

Le chinois classique, malgré sa parenté avec le chinois mandarin moderne, reste une langue relativement peu dotée en traitement automatique des langues (TAL). Les défis à relever sont nombreux pour obtenir des données exploitables, puis produire des modèles pré-entraînés ou autres solutions permettant d'extraire les connaissances contenues dans les documents d'époque. Ce constat est d'ailleurs vrai pour les autres langues anciennes.

Tout d'abord, les sources historiques sont nombreuses, mais un important travail de numérisation et d'annotation des données est nécessaire pour les rendre exploitables. Les méthodes actuelles de Reconnaissance Optique de Caractères (ROC, OCR en anglais) ne sont pas infaillibles et ajoutent du bruit dans les données. Par ailleurs, l'annotation des corpus est particulièrement coûteuse en raison de la rareté des experts en langues anciennes. La première campagne internationale dédiée à l'étiquetage de séquences en chinois classique s'est déroulée en 2022, dans le cadre du *workshop* «Language Technologies for Historical and Ancient Languages »(LT4HALA). En comparaison, la première campagne similaire pour l'anglais a eu lieu en 2000 à l'occasion de la CoNLL-2000.

De plus, le terme "chinois classique" (en chinois, 古文 guwen ou 文言文 wenyanwen) amalgame une multitude de réalités linguistiques. Il s'agit globalement de la langue de l'administration et de la littérature en usage dans la Chine impériale, ce qui couvre près de 2500 ans d'histoire. Elle sert ainsi de langue véhiculaire, en contraste avec l'infinie diversité des langues vernaculaires spécifiques à chaque région. Par commodité, on accepte généralement d'appeler "chinois classique" la langue de l'époque de Confucius (551–479 Avant l'Ère Commune, AEC) à la fin de la dynastie Han (202 AEC –220 Ère Commune, EC). Ces enjeux de délimitation temporelle sont cruciaux, car ils sont un facteur important dans la construction de corpus : faut-il se concentrer sur une variété précise ? combiner plusieurs variétés dans un même corpus ?

Ces considérations nous ont conduit à questionner la représentation que les modèles de langue ont du chinois classique. En particulier, comment ces représentations diffèrent-elles par rapport au chinois moderne ? Comment se structure le réseau de neurones au contact d'un corpus classique ? Quels outils sont à notre disposition pour étudier ces questions introspectives ?

Plan de lecture

Le présent mémoire est organisé de la façon suivante :

- Description du projet dans lequel s'inscrit ce stage (chapitre 1) ;

- État de l'art (chapitre 2) : nous présentons l'achitecture Transformer à la base des modèles BERT, la situation du chinois classique dans le TAL aujourd'hui, et enfin l'approche en réseaux complexes appliquée au TAL ;
- Méthodes et expériences (chapitre 3) : la méthode employée, les données mobilisées et la chaîne de traitement que nous avons suivie ;
- Résultats (chapitre 4) : les résultats obtenus pour les expériences menées, ainsi que les tests statistiques opérés ;
- Discussion (chapitre 5) : nous commentons les résultats obtenus, proposons des explications pour les meilleurs et les moins bons scénarios et donnons les limites de l'approche envisagée ;
- Perspectives (chapitre 6) : nous proposons une ouverture pour poursuivre ce travail et approfondir les questions abordées en introduction.

partie I

Contexte général

CHAPITRE 1

DESCRIPTION DU PROJET

Sommaire

1.1 Cadre du stage 13

1.2 Tâches confiées 15

Introduction

Ce chapitre présente le projet dans lequel s'inscrit mon stage de fin d'études, ainsi que les tâches qui m'ont été confiées dans ce cadre.

1.1 Cadre du stage

Ce stage s'est déroulé au sein de l'Équipe de Recherche Textes, Informatique, Multilinguisme (ERTIM)[1] de l'Inalco, chargé de mener des projets de recherche en Traitement Automatique des Langues. Le stage s'inscrit dans le cadre du projet de recherche «Diversité des Langues Sinitiques et Humanités Numériques»[2] (DLSHN), coordonné par Pierre Magistry et financé l'Agence Nationale de Recherche.

L'objectif de ce projet de recherche est de développer les outils TAL pour les langues sinitiques (ou langues chinoises), abordées sous l'angle de la variation. Dans la langue courante, on désigne souvent par 'chinois' les langues les plus réprésentées comme le mandarin (langue officielle de plusieurs États) ou le cantonais. Toutefois, la famille est langues sinitiques est bien plus riche et diverse, comme illustré dans la figure 1.1. Alors que le mandarin figure parmi les langues les mieux dotées aux côtés de l'anglais, la majorité des autres langues chinoises reste peu outillée.

[1] https://ertim.inalco.fr/
[2] https://anr.fr/Projet-ANR-23-CE38-0004

FIG. 1.1 : Distribution des langues sinitiques dans le monde chinois. Source : *Wikipedia*

Trois axes de variation sont explorés dans le projet DLSHN : "temporelle (diachronique), géographique (dialectale/diatopique) et grapholinguistique (rapport langue-écriture)". En particulier pour l'axe géographique, une équipe se charge d'étudier le teochew, langue parlée par une importante diaspora chinoise installée en France. Pour l'axe temporel, un corpus tiré du Shun-Pao, premier journal quotidien imprimé en sinogramme entre 1872 et 1949, est à diaposition de l'équipe. Il permet d'explorer la reconnaissance d'entités nommées, avec un intérêt non seulement linguistique pour l'étude en diachronie du chinois, mais aussi un intérêt historique en collaboration avec l'équipe ENP-China[3]. L'axe grapholinguistique est quand à lui au cœur des préoccupations de la communauté du TAL pour les langues sinitiques. En effet, le système d'écriture chinois pose des problèmes spécifiques, qu'il est nécessaire de résoudre pour continuer à enrichir les outils.

L'objectif du projet est également de questionner "les représentations formelles (normalisation et vectorisation des données) et les choix de corpus à la base de tout traitement de langues sinitiques". Il s'agit d'un travail plus bas-niveau, qui demande une introspection profonde dans les méthodes employées pour rendre exploitable un jeu de données en machine learning.

[3] https://enepchina.hypotheses.org/

Au sein de l'équipe ERTIM, nous avons accès à un GPU (NVIDIA L40S) d'une capacité de 48GB. C'est sur cette machine que nous avons lancé l'ensemble de nos expériences.

1.2 Tâches confiées

Dans le cadre de ce stage, nous avons d'abord exploré les questions de représentation des données à travers la tokenisation. L'état actuel des connaissances privilégie les tokenizers de type Byte-Pair Encoding (BPE), car ils permettent de ne perdre aucune information entre l'encodage et le décodage. Du point de vue de l'Intelligence Artificielle Explicable (IAE, en anglais XAI pour eXplainable AI), on peut se demander si le vocabulaire appris par ce tokenizer a du "sens", d'un point de vue humain et sémantique : les éléments du vocabulaire d'un modèle correspondent-ils à une représentation mentale que les sinophones peuvent avoir ?

La diversité des caractères chinois, estimés à près de 8 000 dans la langue moderne courante, conduit à une taille de vocabulaire souvent élevée. Il s'agit d'un facteur de complexification des calculs, en temps comme en espace. Nous avons exploité les ressources du CHaracter Information Service Environment (CHISE)[4], qui regroupe des informations de décomposition des sinogrammes en composants (clés et radicaux de sinogrammes, opérateurs binaires de composition...). Cette ressource a l'avantage de croiser différentes graphies pour un même caractère, et d'ainsi prendre en charge la variabilité de l'écriture dans une certaine mesure. Les tentatives n'ont pas permis de réduire significativement la taille du vocabulaire, car beaucoup de caractères n'avaient pas de décomposition et étaient représentés un à un dans le vocabulaire.

Par la suite, nous avons été chargés de benchmarker divers modèles de langue appliqués à l'extraction d'information en chinois classique. Nous avons notamment testé des modèles de type BERT, et des modèles beaucoup plus lourds de type Llama, sur des tâches de reconnaissance d'entités nommées.

Enfin, nous avons entrepris les recherches qui ont mené à ce mémoire de recherche, à savoir explorer la structure interne des modèles de langue au contact de tâches d'étiquetage de séquences. Le code et les données utilisés dans le cadre de ce stage sont accessibles au lien suivant : https://github.com/cbuontal/memoireM2.

[4] https://www.chise.org/

CHAPITRE 2

ÉTAT DE L'ART

Sommaire

2.1 Modèles BERT **17**
2.1.1 Réseaux neuronaux et perceptron multicouche 17
2.1.2 Mécanisme d'attention 18
2.1.3 Spécifités du modèle BERT 20
2.2 Chinois classique et ressources TAL **21**
2.2.1 Corpus disponibles 21
2.2.2 Éléments de description du chinois classique 22
2.2.3 Modèles existants 24
2.3 Approche par réseaux complexes **25**
2.3.1 Éléments de théorie des graphes 25
2.3.2 Réseaux complexes 26
2.3.3 Applications des réseaux complexes au TAL 27

Introduction

Dans ce chapitre, nous décrivons les modèles de langue à base de transformer qui se sont imposés depuis près de 10 ans. Puis nous donnons des éléments de chinois classique, ainsi que les corpus et modèles disponibles dans cette langue. Enfin, nous présentons une approche en réseaux complexes appliquée au chinois classique, qui s'inscrit dans le courant de l'IA explicable (eXplainable AI, XAI).

2.1 Modèles BERT

Cette section présente modèles de langue de type Transformer, tels qu'introduits en 2017 par l'article de Vaswani [Vaswani and al., 2017]. Cet article est fondateur car il passe de l'état de l'art précédent (réseaux récurrents) aux modèles transformers, dont l'innovation principale est l'ajout du mécanisme d'attention.

2.1.1 Réseaux neuronaux et perceptron multicouche

En apprentissage automatique, les modèles de langue sont des réseaux de neurones avec une visée applicative en TAL. Un réseau de neurones est schématiquement une succession de petites boîtes noires, les neurones, qui prennent de

l'information en entrée, la traitent, et produisent une sortie. Cette sortie peut constituer l'information d'entrée d'un autre neurone. Plusieurs neurones peuvent traiter une même information, et former ce qu'on l'on appelle une couche de neurones. Avec une telle organisation, on a donc une couche d'entrée, des couches intermédiaires appelées couches cachées, et une couche de sortie. Cette architecture est à la base du deep learning, et s'appelle le Perceptron multicouche (en anglais MLP, Multi-Layer Perceptron).

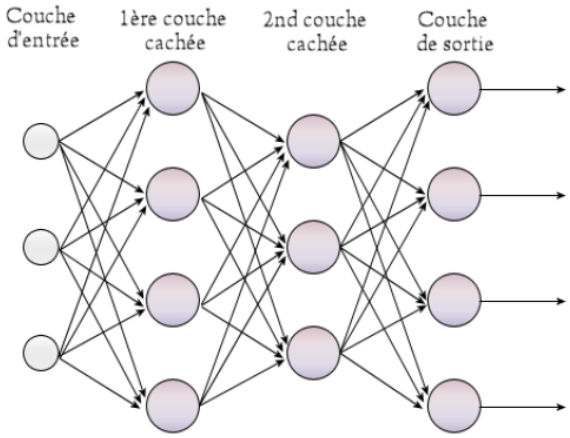

FIG. 2.1 : Représentation d'un perceptron multicouche

Le perceptron multicouche est un modèle dit "à action directe", car l'information ne se déplace que dans une direction : de la couche d'entrée vers la couche de sortie.

L'optimisation des paramètres du modèle se fait par l'algorithme de descente de gradient. Qualitativement, on cherche à ajuster un ensemble de poids qui représentent l'information que le modèle a sur les données. Pour ce faire, on minimise une fonction de coût, et on cherche un minimum (si possible global). La recherche de ce minimum revient à minimiser le gradient de la fonction de coût (équivalent de la fonction dérivée, pour une fonction à plusieurs variables). Se pose le problème de disparition du gradient (*vanishing gradient problem*) pour désigner le problème suivant : à mesure que l'on avance dans les couches du modèle récurrent, le modèle se souvient de moins en moins du contexte le plus lointain.

Pour palier ce problème, on a introduit les modèles récurrents comme LSTM (Long Short-Term Memory [Hochreiter and Schmidhuber, 1997], litt. *mémoire à court ou long terme*) ou GRUs (Gated Recurrent Units, litt. *unité récurrente fermée*). Dans cette architecture, chaque bloc du réseau se "souvient" du résultat de l'opération précédente, en plus de l'information en entrée de la couche courante. Cela permet un contrôle plus fin des informations précédentes, et une meilleure représentation du contexte (dans le cas de données textuelles, le contexte gauche). Cette architecture a porté ses fruits en TAL, par exemple en traduction automatique [Cho et al., 2014].

2.1.2 Mécanisme d'attention

Dans une tâche de machine learning appliquée à des données textuelles, on doit rendre les données manipulables par la machine. C'est l'étape de vectorisation du corpus : à chaque token i on associe un vecteur dit "embedding" $\vec{E_i}$. Pour chaque embedding, on construit deux vecteurs :

- *query vector* $\vec{Q}_i$ (vecteur requête), dont le but est d'interroger sur le contexte du token. On peut les concaténer pour former la matrice requêtes Q (query matrix),
- *key vector* $\vec{K}_i$ (vecteur clé), qui représente les informations sur le token i. De même on peut formé la matrice clés K (key matrix).

Muni de ces données, on utilise les méthodes de l'algèbre linéaire pour créer de l'information, comme expliqué dans les vidéos de vulgarisation de [Brown, 2024]. On construit la **matrice d'attention** qui est la matrice des produits scalaires $A = KQ^T = (\vec{K}_i \cdot \vec{Q}_j)_{i,j}$, pour tous les couples (i, j) de tokens. Cette matrice est carrée, et sa dimension est le carré de la taille du contexte choisie, notée d_k. Elle contient donc l'ensemble des informations contextuelles que le token i apporte sur le token j. Le produit scalaire mesure à quel point chaque paire de vecteurs $(\vec{K}_i, \vec{Q}_j)$ est "proche" dans l'espaces des embeddings. C'est une sorte de mesure de similarité. Lorsque ce score est élevé, cela signifie que l'information contextuelle de la paire d'embeddings $(\vec{E}_i, \vec{E}_j)$ est particulièrement significative.

Ensuite, on normalise des poids par la taille de la matrice, puis on applique une fonction *softmax* (pour obtenir une distribution de probabilité) :

$$\texttt{softmax}(\frac{KQ^T}{\sqrt{d_k}})$$

La matrice de ces scores après normalisation et application du softmax s'appelle *attention pattern*. À cette étape, il peut intervenir un procédé dit de "masking" : on peut choisir d'ignorer (masquer) des valeurs qui ne sont pas pertinentes pour la tâche en présence. Par exemple, pour de la génération de texte dans une langue s'écrivant de gauche à droite, seul le contexte gauche est important. On peut masquer tous les éléments de la matrice d'attention qui représentent l'influence qu'a un token i sur un token j situé avant lui (contexte droit), ce qui économise du temps de calcul.

On définit ensuite une matrice dite *value matrix* M_V qu'on multiplie par les embeddings E_i pour obtenir les *value vectors* de chaque embedding : $\vec{V}_i = M_V \cdot \vec{E}_i$. On pondère ensuite ces vecteurs avec les poids de la matrice *attention pattern* : $\texttt{softmax}(\frac{\vec{K}_j \cdot \vec{Q}_i}{\sqrt{d_k}}) * \vec{V}_i$, pour chaque ligne i. On somme tous ces vecteurs pour obtenir l'ensemble des informations contextuelles de tous les tokens j par rapport à un token i :

$$\Delta\vec{E}_i = \sum_j \texttt{softmax}(\frac{\vec{K}_i \cdot \vec{Q}_j}{\sqrt{d_k}}) * \vec{V}_i$$

Cette étape permet de trier les embeddings pertinents ou non pour représenter le contexte d'un autre. On ajoute le delta obtenu à l'embedding de départ, pour mettre à jour l'information du modèle sur ce token :

$$\vec{E}'_i = \vec{E}_i + \Delta\vec{E}_i$$

Ce procédé est ce que l'on appelle une tête d'attention. Dans la réalité, les couches d'attention des modèles Transformers contiennent plusieurs têtes d'attention par couche. L'article [Vaswani and al., 2017] donne une architecture avec 12 couches de

12 têtes d'attention, 6 pour l'encoder, 6 pour le decoder, largement réutilisée dans les travaux suivants. Les couches cachées des modèles Transformer sont une alternance de perceptrons multicouches et de couches d'attention.

2.1.3 Spécifités du modèle BERT

Le modèle BERT (pour *Bidirectional Encoder Representations from Transformers*) est un modèle de langue développé par Google, présenté dans l'article [Devlin et al., 2019] à l'occasion de la conférence de la *North American Chapter of the Association for Computational Linguistics* en 2019. Son architecture se base sur les Transformers, et conserve globalement la structure du papier de [Vaswani and al., 2017].

Ce modèle est dit bidirectionnel, ce qui signifie que pour un token donné, il prend en compte à la fois le contexte gauche et le contexte droit. Cela permet de mobiliser davantage d'information pour représenter le contexte d'un token donné. Cette caractéristique est notamment due au pré-entraînement des modèles BERT, composé de deux tâches non supervisées :

- Masked Language Modeling (MLM) : le modèle cache aléatoirement 15% des mots et les remplace avec '[MASK]' 80% du temps, un token aléatoire 10% du temps, et conserve le token lui-même 10% du temps. Le modèle choisit ensuite le mot masque à l'aide des contextes gauche et droit. On minimise l'erreur de prédiction en minimisant une fonction de coût.
- Next Sentence Prediction (NSP) : on constitue des paires de phrases A et B ou B est 50% du temps la véritable phrase qui suit A, et 50% une autre phrase choisie aléatoirement.

Les auteurs montrent que le modèle BERT surpasse les modèles existants sur diverses tâches : compréhension du langage naturel, question/réponse et génération de texte. Par ailleurs, ils présentent à la fois un modèle *BERT base* à 100 millions de paramètres et un plus gros modèle *BERT large* à 300 millions de paramètres. Ils observent que le plus gros modèle obtient systématiquement de meilleures performances.

2.2 Chinois classique et ressources TAL

Dans ce travail, j'utilise trois corpus pour le fine-tuning de modèles sur des tâches d'étiquetage de séquences. Après les avoir présentés, je donne des éléments de description du chinois tel qu'il existe dans ces corpus. Je présente enfin les modèles pré-entraînés à disposition.

2.2.1 Corpus disponibles

Cette section présente les principaux corpus en chinois classique utilisés dans ce travail, selon qu'ils sont annotés ou non.

Il est souvent difficile d'avoir accès à des jeux de données en chinois classique, comme commenté dans l'article de [Blouin and al., 2023]. En effet, les entreprises privées majeures en Chine, comme les acteurs publics (Beijing National Library, Shanghai Library), ont certes numérisé une quantité importante de documents, mais n'ont pas diffusé le contenu de manière facilement exploitable pour le TAL. Les auteurs mentionnent notamment des données consultables à travers une interface, mais pas téléchargeables pour de la fouille de données[1]. Ou encore, l'impasse dans les négociations avec les acteurs privés pour pouvoir télécharger les données textuelles, même avec des conditions les plus strictes de confidentialité des données.

Corpus non annotés

Les corpus non annotés suivants sont d'une taille importante à l'échelle d'une langue peu dotée. Ils permettent de pré-entraîner un modèle de langue de zéro.

- Le corpus Daizhige 殆知阁[2] comprend environ 3,4 millions de tokens de chinois classique. Il est issu de la littérature classique, et comprend des textes bouddhiques (comme le Tripitaka Koreana), confucéens (les Cinq classiques dont le Chunqiu, les Quatre Livres dont les Entretiens de Confucius), taoïstes... Il comprend également des textes ayant trait aux arts, à la poésie ou à la médecine. Il s'agit du plus gros corpus de chinois classique accessible librement en ligne.
- Wikisource[3] est également une source importante de texte en chinois classique. L'agrégation des ressources disponibles dans cette langue donne un corpus de texte brut de plus de 24 millions de caractères, dont 17 884 caractères uniques.

Corpus annotés

- CHisIEC (Chinese Historical Information Extraction Corpus) [Tang and al., 2024] est un corpus produit par l'Université de Pékin, qui prend comme base les Vingt-Quatre Livres (二十四史 Ershisi shi). Il s'agit d'une série d'annales retraçant l'histoire de la Chine de 3000 AEC jusqu'à la fin de la dynastie Ming au 17ème siècle. Les auteurs ont sélectionné 22 volumes de 12 livres à annoter, en entitées nommées et en extraction de relation, donnant lieu à deux sous-corpus. Le corpus annoté en entités nommées comprend 135 713 caractères et 14 194 entités.

[1] Par exemple, sur le site de la National Library of China : http://read.nlc.cn/user/category
[2] https://github.com/garychowcmu/daizhigev20
[3] https://zh.wikisource.org/

- Tagging Xing'an huilan (TXAHL)[4], de l'équipe Huma-Num DISTAM (DIgital STudies Africa, Asia, Middle East) a pour base une compilation des mémoires du ministère des Peines, compilées dans le Xing' an huilan 刑案汇览 (Une vue d'ensemble des affaires pénales). Celle-ci rassemble 5400 documents judiciaires du milieu du 19ème siècle. 389 documents sont annotés en entités nommées (lieu, criminel, victime, mention de textes de loi...).

- Ancient Chinese Corpus (ACC) [Zhang and al., 2022], rendu public à l'occasion de la *shared task* EvaHan2022, la première dédiée au chinois classique. Ce corpus est basé sur le Zuozhuan et comprend 180 000 sinogrammes, pour 195 000 unités (incluant les mots et la ponctuation). Il est segmenté et annoté en parties du discours par la Nanjing Normal University, selon 17 tags.

2.2.2 Éléments de description du chinois classique

Le chinois classique est une langue de la super-famille sino-tibétaine. Son système d'écriture emploie des caractères appelés sinogrammes. Chaque caractère désigne un lemme entier. Certains sont des idéogrammes, qui évoquent une idée ou un concept, d'autres des idéophonogrammes lorsqu'ils contiennent une partie phonétique. C'est le cas du caractère de la Figure 1.1, qui représente le caractère 媽 ma3, qui signifie "mère". La partie de gauche, en rouge, est le caractère 女 nü3, qui signifie "femme". La partie de droite est 馬 ma3, qui signifie "cheval". Ainsi, ce caractère est un idéophonogramme, avec la clé (partie de gauche) portant la sémantique, et la partie de droite donnant une information purement phonétique.

FIG. 2.2 : Décomposition d'un sinogramme

Il est important de noter que deux formes d'écriture coexistent pour écrire les langues chinoises : les caractères simplifiés et les caractères traditionnels. Officiellement, les caractères simplifiés sont en usage en Chine continentale et à Singapour, tandis que les caractères traditionnels le sont à Hong Kong, Macao et Taïwan. Par exemple, le caractère signifiant Han (ethnie majoritaire en Chine) s'écrit 汉 en forme simplifiée et 漢 en forme traditionnelle. La simplification des sinogrammes telle qu'on la connaît est le résultat de longues mutations qui ont commencé dès le début du 20ème siècle en Chine, avant d'être définitivement arrêtée le 24 juin 1986. Cette différence est importante dans les corpus, car un même caractère en forme simplifiée ou traditionnelle n'est pas encodé de la même façon par Unicode. Ainsi, un modèle peut inclure dans son vocabulaire la forme 汉 mais pas la forme 漢.

Ajoutons à ce sujet que le passage d'un caractère traditionnel à un caractère simplifié est surjectif. Cela signifie que deux caractères traditionnels différents peuvent avoir une même forme simplifiée. Par exemple les sinogrammes 刮 (gua1, raser) et 颳 (gua1, souffler (vent)) ont tous deux été simplifiés en 刮. Ainsi, pris isolément, ce caractère est ambigu dans le système simplifié. On peut affirmer notamment qu'un

[4] https://distam.hypotheses.org/12692

FIG. 2.3 : De gauche à droite, formes traditionnelle et simplifiée d'un sinogramme

corpus en caractères traditionnels subit une perte d'information lorsqu'il est converti en caractères simplifiés.

Du point de vue morphologique, le chinois classique ne présente aucune flexion ni verbale ni substantivale. Il existe quelques éléments de morphologie dérivationnelle, avec des mécanismes telles que la réduplication ou l'affixation. La conversion lexicale est aussi très courante, à savoir un même mot qui change de catégorie (partie du discours) sans changement formel [Jacques, 2022].

Le chinois classique est analytique et majoritairement SVO. Les démonstratifs, adjectifs, propositions relatives etc. précèdent la tête nominale [Barnes et al., 2009].

On distingue traditionnellement les "mots pleins" et les "mots vides" [Peyraube, 2008]. La première catégorie désigne les mots lexicaux (noms, adjectifs, verbes...) tandis que la deuxième désigne les mots grammaticaux (particules, adpositions...). Ces mots-vides structurent la phrase, et sont en nombre fini. Remarquons d'ailleurs que le lexique du chinois classique est principalement monosyllabique, à la différence du mandarin moderne qui est plutôt disyllabique ([Blouin and al., 2023]). Les caractères sont souvent très polysémiques, avec parfois une dizaine de sens possibles pour un même sinogramme.

Le chinois classique désigne largement la langue vernaculaire de la Chine impériale, ce qui recouvre près de 2500 ans d'histoire. Généralement, on restreint l'acception à la période du 5ème siècle AEC à 221 EC, respectivement de la fin de la période des Printemps et Automnes à l'émergence de la dynastie des Qin. Ce choix correspond à une production majeure dans l'histoire intellectuelle chinoise, celle des Annales des Printemps et Automne (春秋 Chunqiu). Ces annales décrivent les principaux événements politiques de la période des Zhou Orientaux, environ 771 à 481/453 AEC. Ce texte fait partie des Cinq Classiques, qui sont des piliers de la culture confucéenne [Nylan, 2001]. Ce canon de l'histoire intellectuelle a fait l'objet d'innombrables gloses et commentaires, comme le 左传 Zuozhuan (4ème siècle AEC). Cet état de la langue est représenté dans le corpus AAC.

Pour le corpus TXAHL, écrit dans une langue du 19ème siècle, il serait plus correct de parler de "chinois transitionnel" comme dans [Blouin and al., 2023]. Les auteurs désignent ainsi la langue chinoise "qui a émergé et s'est développée entre la fin du 19ème siècle et le début des années 1950". Ils décrivent également les transformations importantes que la langue a subies à cette époque : l'ouverture au monde crée le besoin de se représenter des concepts nouveaux, d'où de nombreux imports du japonais. Les mots importés étaient souvent disyllabiques, norme qui s'est imposée en chinois moderne aujourd'hui. La langue a également connu des transformations pour des mots très basiques : pronoms, démonstratifs, verbes... Autant de bouleversements qui éloignent cette langue de transition du chinois des Printemps et Automnes.

Le corpus CHisIEC est plus hétérogène, car il regroupe des documents historiques couvrant 1830 ans d'histoire de la Chine impériale. Le corpus contient des documents

de la dynastie Han (206 AEC - 220 EC) jusqu'aux Ming (1368 - 1644). Cette diversité est intéressante comparés aux autres corpus, plus homogènes.

2.2.3 Modèles existants

Les modèles présentés dans cette section ont tous pour base le modèle Chinese BERT[5]. Comme pour le modèle BERT original, ce modèle est entraîné sur une tâche de Masked Language Modeling.

Les modèles suivants poursuivent le pré-entraînement de Chinese BERT :

- SikuRoBERTa[6] a été diffusé par la Nanjing Normal Univerisity, à l'occasion de la shared task EvaHan2022. Le pré-entraînement a été poursuivi sur le Siku Quanshu 四库全书 (Les Livres complets des Quatre Magasins). Il s'agit d'une encyclopédie commissionnée par l'empereur Qianlong des Qing, en 1772. L'ouvrage entier comprend 36 381 volumes en 2,3 millions de pages, pour un total de 997 millions de mots. Le modèle a une taille de vocabulaire de 29 791 tokens.
- Jihuai/bert-ancient-chinese[7], prolonge le pré-entraînement de Chinese BERT avec un corpus de 6 fois la taille du Siku Quanshu. Le souci des auteurs était d'inclure un maximum de sinogrammes traditionnels rares dans le vocabulaire. La plus grande diversité de données donne une taille de vocabulaire plus grande, à 38 208 tokens.

Les modèles suivants sont pré-entraînés sur des corpus de chinois classique :

- Guwen-BERT[8] est produit par une équipe du Beijing Institute of Technology. Il utilise l'architecture du modèle RoBERTa, avec un pré-entraînement sur le corpus Daizhige. Le corpus est entièrement converti en caractères simplifiés. Le modèle a un vocabulaire de 23 292 tokens.
- RoBERTa-classical-chinese[9] se base sur Guwen-BERT, et prend en charge les caractères simplifiés et traditionnels. Il a une taille de vocabulaire de 26 318 tokens.
- Guji-BERT[10] est produit par une équipe de chercheurs de Nanjing, qui prend pour base l'architecture de Guwen BERT. Le modèle est entraîné sur le corpus Daizhige, cette fois en mélangeant les caractères simplifiés et traditionnels, pour un vocabulaire de 34 709 tokens [Wang and al., 2023].

Ainsi on a une famille de modèles qui a été entraîné à la fois sur du chinois moderne et sur du chinois classique, et une autre famille qui n'a été pré-entraînée que sur du chinois classique.

[5] https://huggingface.co/google-bert/bert-base-chinese
[6] https://huggingface.co/SIKU-BERT/sikuroberta
[7] https://huggingface.co/Jihuai/bert-ancient-chinese
[8] https://huggingface.co/ethanyt/guwenbert-base
[9] https://huggingface.co/KoichiYasuoka/roberta-classical-chinese-base-char
[10] https://huggingface.co/hsc748NLP/GujiBERT_jian_fan

2.3 Approche par réseaux complexes

Ce travail se veut être dans le prolongement de l'article de [Zheng and Xiao, 2024]. La méthode qu'ils emploient consiste à modéliser les données et le modèle de langue à l'aide de graphes. J'introduis des éléments de théorie des graphes, puis de réseaux complexes, avant de présenter plus en détail les travaux de [Zheng and Xiao, 2024] dans le cadre des réseaux complexes appliqués au TAL.

2.3.1 Éléments de théorie des graphes

La théorie des réseaux complexes descend d'une branche des mathématiques, la théorie des graphes [Bretto et al., 2012]. Celle-ci serait née grâce au fameux problème des sept ponts de Königsberg, résolu par le mathématicien Leonhard Euler (1707-1783) en 1735. De nos jours, cette discipline occupe également les informaticiens, et trouve des applications dans une grande variété de domaines, jusque dans les sciences sociales. En effet, outre des objets physiques comme dans le problème de Königsberg, la théorie des graphes permet également de modéliser des systèmes abstraits, éventuellement dynamiques : le réseau internet, un réseau de câbles électriques, les réseaux trophiques (chaîne alimentaire) en écologie...

Un graphe est schématiquement un ensemble de points, qu'on appelle sommets (*nodes* en anglais). Chaque paire de sommets est éventuellement reliée par une arête (*edges*). Ces arêtes peuvent avoir une direction ou non, c'est-à-dire pointer d'un sommet sur un autre, ou sur les deux à la fois. Dans ce cas, on dira que le graphe obtenu est orienté, et ses arêtes sont appelées des arcs. Un graphe présentant à la fois des arêtes et des arcs est dit mixte.

Formellement, un graphe est donc un couple :

$$G = (V, E)$$

avec G l'ensemble des sommets, et V l'ensemble des arêtes, qui peuvent être des paires (u, v) si on a par exemple une arête partant de u et pointant sur v, ou des ensembles $\{u, v\}$ dans le cas non-orienté.

Pour décrire les graphes, on introduit plusieurs mesures au-delà du nombre de sommets et d'arêtes :

- On appelle **degré d'un sommet** le nombre de liens (arêtes ou arcs) qui relient ce sommet, que l'on peut noter $deg(S)$, avec S le sommet décrit,
- Pour une paire de sommets S_1, S_2, on peut définir la **distance** qui les sépare par le plus court chemin qui les relie, c'est-à-dire le chemin qui emprunte le moins d'arcs / arêtes possible,
- Le **diamètre** d'un graphe, noté D, est la distance maximale séparant deux sommets du graphe, autrement dit le plus court chemin le plus long qui existe,
- On peut introduire la **longueur moyenne des plus courts chemins** L, qui permet de mesurer à quel point les sommets sont "proches" les uns des autres,

- Le **coefficient de clustering**, noté C, permet de mesurer le regroupement des sommets dans le réseau. Formellement, on a :

$$C = \frac{3 * |\text{triangles}|}{|\text{paire de voisins distincts}|}$$

où $|x|$ désigne le nombre d'élements de la catégorie x. Par construction, on a $0 \leq C \leq 1$. Ce coefficient s'interprète comme la proportion de paires de voisins connectés dans le graphe.

Parmi les graphes les plus courants, on a par exemple la structure d'arbre (graphe orienté d'un seul tenant, comportant un unique sommet avec un arc sortant que l'on appelle racine), très courante en informatique.

2.3.2 Réseaux complexes

On appelle réseau complexe un graphe particulièrement dense (en nombre de sommets et d'arêtes), dont les propriétés topologiques sont difficiles à décrire. Les recherches dans ce domaine sont relativement récentes [Albert and Barabási, 2002] et prometteuses, dans la mesure où cet outil permet de décrire des systèmes complexes en sciences naturelles comme en sciences humaines [Strogatz, 2001].

Les travaux de [Watts and Strogatz, 1998] proposent d'utiliser le coefficient de clustering et la longueur moyenne des plus courts chemins pour caractériser les réseaux "réels", c'est-à-dire décrivant des structures existant dans la nature. Deux caractéristiques sont observées dans ces réseaux : l'effet "petit monde", et l'invariance d'échelle.

Un réseau est dit **petit monde** s'il présente des distances courtes et un coefficient de clustering élevé. Cela signifie que les mots du réseaux sont très reliés les uns aux autres (C élevé) et que les plus courts chemins sont relativement courts par rapport au nombre de sommets dans le réseau.

On dit qu'un réseau est **invariant d'échelle** lorsque la distribution des degrés de ses sommets a pour distribution une loi de puissance (*power law*). Formellement on a

$$P(k) \sim k^{-\gamma}$$

où $P(k)$ est la proportion de sommets reliés à k autres nœuds, et γ le paramètre de la loi de puissance. Qualitativement, cette distribution implique l'existence de sommets très connectés aux autres, que l'on appelle *hubs*.

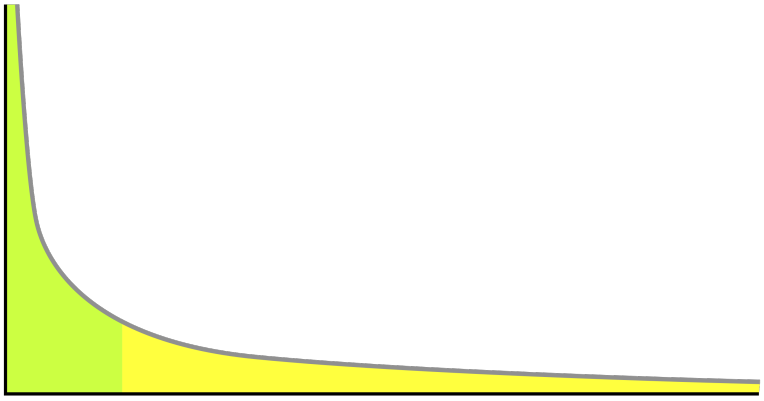

FIG. 2.4 : Distribution de loi de puissance

La partie verte représente un nombre important de tokens faiblement reliés ; et la queue de la partie jaune le faible nombre de tokens très fortement reliés : les hubs

2.3.3 Applications des réseaux complexes au TAL

En TAL, les réseaux complexes permettent de modéliser les relations entre tokens dans un corpus, qu'elles soient sémantiques [Mancio et al., 2012] comme syntaxiques [Salman et al., 2023]. Il est démontré que deux grandes propriétés caractérisent les réseaux lexicaux [Ferrer et al., 2004] : la propriété de petit monde (*small world*), et l'invariance d'échelle (*scale-free*).

Les auteurs de [Zheng and Xiao, 2024] utilisent une méthodologie à base de réseaux complexes pour mettre en regard les réseaux de co-occurrence observés dans divers corpus en chinois classique, et les réseaux d'attention issus de modèles BERT pré-entraînés. L'objectif est de comprendre comment le modèle se structure afin de représenter au mieux la langue et ses propriétés. Ils montrent notamment que les propriétés de petit monde et d'invariance d'échelle sont respectées. Les chercheurs ont testé Siku-BERT et Chinese BERT, et montrent que le premier a un réseau davantage *sparse*, c'est-à-dire avec moins d'arêtes, des degrés plus faibles, un plus petit coefficient de clustering et une longueur moyenne des plus courts chemins supérieure par rapport à Chinese BERT. Ces résultats suggèrent que les têtes d'attention de Siku-BERT organisent de manière plus cohérente la représentation de la langue, alors que Chinese BERT crée des liens superflus entre sommets.

Pour mener ce travail à bien, les auteurs ont utilisé le corpus ACC, et les modèles SIKU-BERT et Chinese BERT.

Les auteurs construisent des réseaux de cooccurrence en mots et en caractères. Dans chaque cas, ils relient les mots/caractères qui se suivent dans le texte par une arête. Par exemple, la phrase « 鄭伯使許大夫百里奉許叔以居許東偏 »donne les réseaux présentés à la figure 2.5.

En étudiant les propriétés des graphes ainsi générés, l'équipe de recherche montre que ces réseaux issus des données ont les propriétés de petit monde et d'invariance d'échelle introduits précédemment. Ainsi, dans le cas de ces travaux, l'invariance d'échelle suggère que la capacité d'un mot ou caractère à se composer en collocation est inversement proportionnelle à sa proportion dans le vocabulaire. Autrement dit, seul un petit nombre de mots ou caractères ont une forte propension à se composer

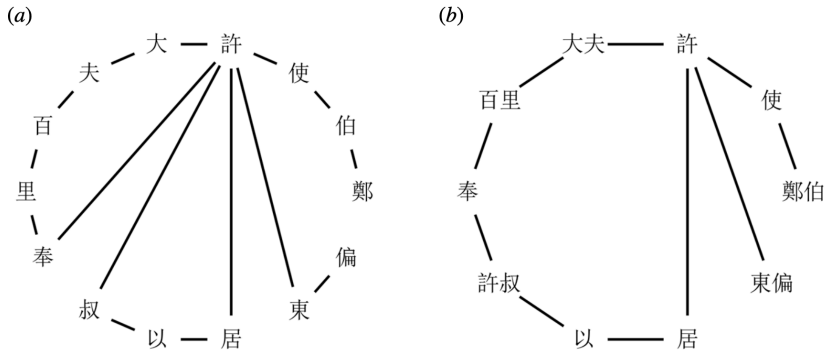

FIG. 2.5 : Réseaux de cooccurrence de la phrase d'exemple

(a) est le réseau de caractères, (b) le réseau de mots

avec d'autres.

En parallèle, les auteurs génèrent des graphes de cooccurrence à partir du modèle SIKU-BERT. Ils procèdent ainsi : pour chaque séquence donnée au modèle, chaque tête d'attention du modèle génère une matrice d'attention entre les caractères de la phrase. On relie alors chaque token à celui avec lequel il a le plus gros score dans la matrice d'attention (*the most attending character*) comme illustré dans la figure 2.6. Les chercheurs montrent que ces réseaux générés à partir des têtes d'attention du modèle ont majoritairement les propriétés de petit monde et d'invariance d'échelle.

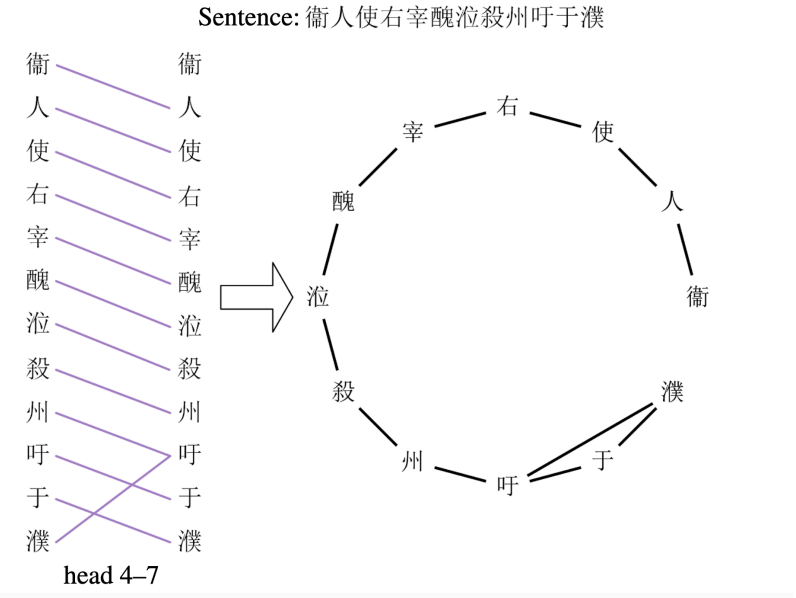

FIG. 2.6 : Extraction d'un graphe reliant les tokens les plus proches dans la matrice d'attention du modèle SIKU-BERT

Étant donnée la phrase d'exemple, on construit le réseau de cooccurrence avec les paires de mots les plus reliées dans la matrice d'attention. Muni de cette table, on crée le graphe correspondant. On répète cette opération pour chaque tête d'attention, que l'on nomme par exemple 'tête 4-7' pour la 7ème tête d'attention de la 4ème couche d'attention du modèle.

Les chercheurs calculent ensuite la similarité entre les graphes issus des réseaux d'attention et ceux issus du corpus. La similarité est significative, ce qui suggère que les têtes d'attention sont capables d'organiser leur connaissance du chinois classique en adéquation avec les observations des corpus.

Les auteurs poursuivent en analysant les têtes d'attention groupées par couche, en comparant avec Chinese BERT (qui n'est donc pas pré-entraîné sur des données en chinois classique) et enfin en mesurant l'influence du fine-tuning sur une tâche de pos-tagging sur l'organisation des têtes d'attention. Les diverses tâches mises en œuvre montrent comment chacune modifie la structure interne du modèle de langue, vue à travers la modélisation en réseau complexe.

partie II

Méthode, expériences et résultats

CHAPITRE 3

MÉTHODES ET EXPÉRIENCES

Sommaire

3.1 Motivation et question de recherche **33**
3.2 Corpus utilisés **35**
3.2.1 Chinese Historical Information Extraction Corpus (CHisIEC) 35
3.2.2 Tagging Xing'An HuiLan (TXAHL) 36
3.2.3 Ancient Chinese Corpus (ACC) 37
3.3 Organisation des expériences **39**
3.3.1 Génération des graphes 39
3.3.2 Étude statistique 40
3.4 Pré-entraînement de modèles **42**

Introduction

Cette section présente le cheminement que j'ai suivi pour aboutir à ma problématique, les corpus utilisés et les tâches spécifiques à chacun, avant d'expliquer la chaîne de traitement que j'ai mis en place pour mes expériences.

3.1 Motivation et question de recherche

L'article de [Zheng and Xiao, 2024] étudie notamment la relation entre la tâche de TAL mise en œuvre et la façon dont se structure le modèle. Je me suis questionné sur la façon dont un modèle se représente le chinois classique. Comme mentionné précédemment, le chinois classique comprend une liste finie de mots grammaticaux, que la tradition chinoise a appelés "mots vides". Puisque ces mots sont en nombre fini, et sont largement présents dans le corpus, j'ai formulé l'hypothèse que les nœuds des réseaux complexes associés à ces mots devaient être des 'hubs', au sens de la propriété d'invariance d'échelle. Plusieurs questions se succèdent alors : les observations de petit monde et d'invariance d'échelle à d'autres corpus et d'autres modèles ? Et dans le cas où l'on observe l'invariance d'échelle, peut-on identifier les hubs du réseau comme étant des mots vides de la langue ?

En termes de performance sur les tâches TAL, j'ai souhaité explorer la relation entre les indicateurs des réseaux complexes issus des matrices d'attention des modèles, et leurs performances en fin de tâche (en termes de précision, rappel et de f-mesure).

Dans le cadre de ce mémoire, j'ai choisi de me poser la question suivante : dans quelle mesure peut-on établir un lien entre la structure interne des modèles après fine-tuning et leurs performances sur les tâches TAL de POS-tagging et de reconnaissance d'entités nommées ?

3.2 Corpus utilisés

Ce travail utilise trois corpus disponibles librement en ligne : CHisIEC, TXAHL et ACC (voir section 1.2.1). Ces corpus ont été constitués pour deux tâches différentes : la reconnaissance d'entités nommées (*Named Entity Recognition*) et l'étiquetage en parties du discours (*POS tagging*).

Corpus	Tâche	Nb de caractères	Nb de caractères uniques	Époque
CHisIEC	NER	135 713	3 668	200 AEC - 1630
TXAHL	NER	128 691	2 563	1644-1911
ACC	POSt	228 290	3 342	475-221 AEC

3.2.1 Chinese Historical Information Extraction Corpus (CHisIEC)

Ce corpus a été établi pour des tâches de reconnaissance d'entités nommées et d'annotation sémantique (*entity linking*). Dans le cadre de mon travail, je me concentre exclusivement sur la portion du corpus étiquetée en entitées nommées.

Les annotations portent sur quatre catégories définies de la manière suivante par les auteurs de [Tang and al., 2024] :

- Personne (*Person*) : Acteurs centraux de l'histoire à travers les époques et les dynasties, dont les empereurs, politiciens, acteurs du monde culturel, des généraux, et d'autres membres influents de la société ;
- Lieu (*Location*) : régions, capitales, comtés, toponyme et palaces ;
- Titre officiel (*Official*) : retrace l'histoire des différents titres utilisés au fil des dynasties (roi, empereur...), ce qui permet de retracer les caractéristiques socio-politiques de chaque époque ;
- Livre (*Book*)

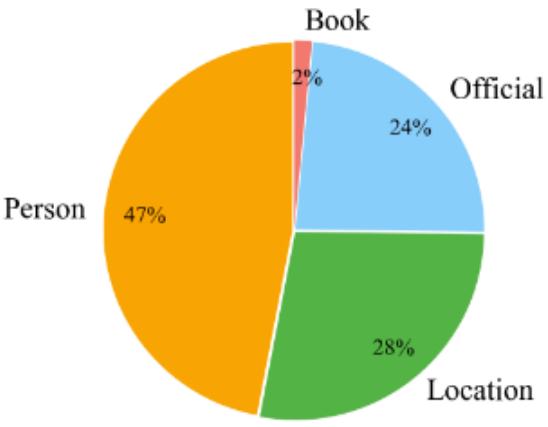

FIG. 3.1 : Proportion de chaque tag dans le corpus CHisIEC

Les tags sont inégalement répartis dans le corpus : la catégorie "Personne" regroupe 47% des entités annotées, puis viennent les catégories "Lieu" et "Titre officiel" avec respectivement 28% et 24% des annotations. La catégorie "Livre" est très minoritaire, avec seulement 2% des annotations. Cette disparité s'explique par la nature des Vingt-Quatre Livres, dont les narratifs citent davantage de personnages historiques que d'ouvrages.

3.2.2 Tagging Xing'An HuiLan (TXAHL)

Ce corpus[1] regroupe des sources juridiques de l'époque des Qing (1644-1911), ce qui permettrait la constitution de bases de données sur l'histoire du droit et les pratiques judiciaires en Chine. Les annotations ont été réalisées avec la plateforme Kairntech[2]. Les informations annotées peuvent être réparties en deux catégories :

- celles identifiables à l'aide listes (noms de province, titres des fonctionnaires...) ;
- celles qui nécessitent un modèle de reconnaissance (noms de criminels, textes de loi cités...).

Le corpus comprend 389 des 5 400 documents que comprend le Xing'An HuiLan, divisés en

- sous-corpus d'entraînement (*train*) de 309 documents ;
- sous-corpus de test (*test*) de 40 documents ;
- sous-corpus de validation (*validation*) de 40 documents ;

Les étiquettes pour la reconnaissance d'entités sont les suivantes :

- Fonction 職 : titre officiel de la personne qui a pris en charge le cas ;
- Province 省 : nom de province ;
- Toutes les autres lois : Toutes les lois citées dans une affaire, hormis "jugement" (cas suivant). Le but de l'étiquette 'toutes les autres lois' est de fournir une liste de toutes les dispositions légales citées dans une affaire spécifique ;
- Jugement 判 : toutes les lois citées dans un procès et utilisées comme base pour le verdict, ce qui inclut la peine prononcée. En général, on a une étiquette "jugement" par procès ;
- Criminel 首 : principal criminel dans un procès ;
- Complice 從 : noms des complices dans le procès ;
- Autres parties : nom des autres personnes mentionnées dans le procès, qui ne sont ni le criminel ni un·e complice ;
- Date : date du document (située à la fin du procès)

[1] https://distam.hypotheses.org/12692
[2] https://kairntech.com/

- Nom du cas : toutes les informations mentionnée dans un précédent judiciaire (date, province, nom de criminel etc.), qui en général apparaissent entre les caractères 查 et 案
- Peine prononcée : la peine réelle issue du verdict final (y compris l'éventuelle atténuation de la peine, ou la modification de la peine prononcée dans un jugement antérieur. Cette étiquette marque la peine finale réellement prononcée dans une affaire, qui peut inclure soit la peine légale, soit la peine atténuée ;
- Lieu : tous les lieux mentionnés dans un procès, à l'exception des provinces

Dans l'échantillon utilisé par les chercheurs, trois étiquettes sont représentées : criminel, complice et autre partie.

3.2.3 Ancient Chinese Corpus (ACC)

Ce corpus a été publié le 18 Octobre 2017 par la Nanjing Normal University. Il contient le ZuoZhuan segmenté en mots et annoté en parties du discours. Les données rendues disponibles pour la *shared task* EvaHan2022[3] sont divisées en trois datasets : un corpus d'entraînement et un corpus de test basés sur le ZuoZhuan, et un deuxième corpus de test qui utilise des sources littéraires chinoises de la même époque. Dans le cadre de mon travail, j'utilise le corpus d'entraînement qui est le plus gros des trois.

L'annotation en parties du discours se fait selon 17 catégories d'après le guide [4] :

Le format des données est du type "TOKEN/TAG", avec une séparation par des espaces.

Notons que certaines de ces catégories relèvent plutôt de la reconnaissance d'entités nommées, comme "nr / person", ou "ns / location".

[3] https://circse.github.io/LT4HALA/2022/EvaHan.html

[4] https://github.com/CIRCSE/LT4HALA/blob/master/2022/data_and_doc/EvaHan_guidelines_v1_training.pdf

ID	Tag	Part-of-speech	Example (Chinese_English Trans)
1	a	adjective	大 _big
2	c	conjunction	則 _then
3	d	adverb	不 _not
4	f	locative	前 _front
5	j	combined	焉 _at there
6	m	number	一 _one
7	n	noun	人 _human
8	nr	person	孔子 _Confucius
9	ns	location	齊 _Qi(state name)
10	p	prepositional	於 _at
11	q	classifier	匹 _classifier for horse and wolf
12	r	pronoun	吾 _me
13	s	onomatopoeia	譆譆 _LOL
14	t	time	五月 _the fifth month
15	u	aux	之 _of
16	v	verb	如 _go
17	vs	causative usage of predicate	驚 _make someone shocked
18	vw	benefactive usage of predicate	泣 _cry for
19	vy	conative usage of predicate	寶 _take something as treasure
20	y	modal	乎 _interrogative
21	w	punctuation	。_full stop
22	x	others	abc

TAB. 3.1 : Ensemble des 22 étiquettes pour l'annotation en parties du discours du corpus ACC, avec exemples et traductions en anglais

3.3 Organisation des expériences

Cette section présente l'organisation des expériences : le génération des graphes issus des matrices d'attention du modèle, puis les outils mobilisés pour l'analyse statistique des données collectées.

3.3.1 Génération des graphes

Dans le cadre de ce travail, j'ai étudié 6 modèles, entraînés chacun sur 3 corpus, comme illustré à la figure 3.2. Pour chaque paire de corpus et de modèle, j'ai généré les graphes des têtes d'attention :

- avant le fine-tuning, pour voir comment le modèle est formé sans avoir vu les données ;
- après fine-tuning, pour voir comment le modèle se structure au contact du modèle et de la tâche associée.

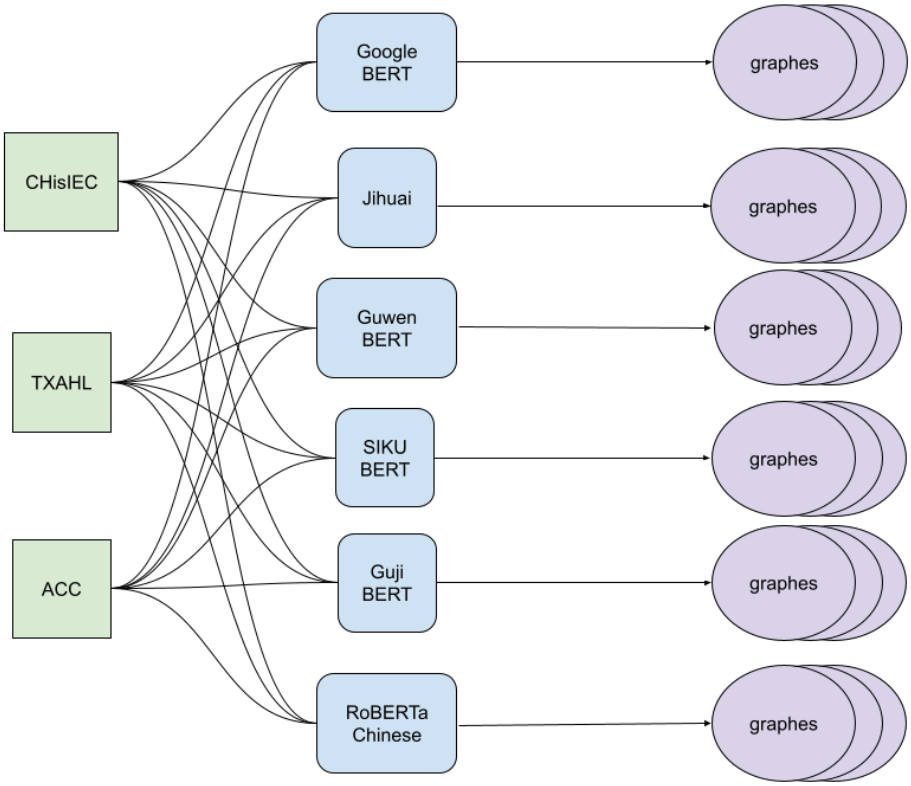

FIG. 3.2 : Pipeline globale pour la génération des réseaux

Chaque paire de modèle et de corpus donne lieu à 144 (12 couches d'attention avec chacune 12 têtes d'attention) fichiers textuels de co-occurrences. À l'aide de la bibliothèque NetworkX[5], j'ai généré des graphes à partir de ces fichiers de co-occurrences,

[5] https://networkx.org/

selon le même procédé que les auteurs de [Zheng and Xiao, 2024]. J'obtiens un total de 3*6*144 = 2 592 graphes, pour les modèles non fine-tunés.

Pour le fine-tuning, je réalise 30 runs pour chaque paire de modèle et de corpus. Ceci afin de gommer la variation aléatoire entre chaque entraînement, et constituer un échantillon de taille significative pour les analyses statistiques. J'obtiens un total de 6*3*30*144 = 7 776 fichiers de co-occurrences, que je convertis en graphes également.

Pour chaque graphe obtenu, je calcule toute sortes d'indicateurs : nombre de nœuds, nombre d'arêtes, coefficient de clustering, longueur moyenne des plus courts chemins... Comme mentionné à la section 1.3.1. Ces données sont agrégées et font l'objet d'analyses statistiques présentées plus bas.

3.3.2 Étude statistique

J'utilise la librairie Python Polars[6] pour mes analyses statistiques.

Je rassemble dans un DataFrame les informations de mesures d'évaluation suivantes (accompagnées de leur définition) :

- Précision P : mesure le nombre d'éléments correctement étiquetés (vrais positifs) rapporté au nombre d'éléments présents dans l'hypothèse (vrais positifs + faux positifs).

$$\mathrm{P} = \frac{Vrais_positifs}{Vrais_positifs + Faux_positifs}$$

- Rappel R : mesure le nombre d'éléments correctement étiquetés (vrais positifs) rapporté au nombre d'éléments présents dans la référence (vrais positifs + faux négatifs).

$$\mathrm{R} = \frac{Vrais_positifs}{Vrais_positifs + Faux_negatifs}$$

- F-mesure $F1$: moyenne harmonique du rappel et de la précision.

$$F = \frac{2PR}{P + R}$$

Ainsi que les indicateurs sur les réseaux complexes présentées précédemment :

- Nombre de nœuds N,
- Nombre d'arêtes E,
- Degré moyen $< k >$,
- Paramètre de la loi de puissance r,

[6] https://pola.rs/

- Coefficient de détermination R^2, qui mesure à quel point la distribution observée des degrés correspond à la loi de puissance théorique
- Longueur moyenne des plus courts chemins L,
- Coefficient de clustering C,
- Diamètre du réseau D

Muni de ces données agrégées, je calcule notamment un coefficient de corrélation classique de Pearson entre les métriques d'évaluation et les mesures sur les graphes :

$$r = \frac{\mathrm{Cov}(X, Y)}{\sigma_X \sigma_Y}$$

où $\mathrm{Cov}(X, Y)$ est la covariance de deux séries X et Y, et σ_X l'écart-type de la série X.

Muni de ces données, je peux organiser les données par ordre décroissant de performances du modèle sur la tâche finale. Ainsi j'étudie si les mesures sur les graphes sont corrélées (ou non) aux indicateurs de précision, rappel et F-mesure.

3.4 Pré-entraînement de modèles

Afin d'étudier l'impact des données d'entraînement pour le pré-entraînement sur la structuration du modèle. En particulier, on peut se demander quel impact le pré-entraînement sur des données en chinois moderne a sur la compréhension du classique. En effet, puisque les deux langues diffèrent sur une multitude de plans, on peut supposer que les architectures permettant de comprendre chaque langue sont différentes.

Pour étudier cette question, j'ai pré-entraîné le modèle Chinese BERT sur le corpus non-annoté Wikisource, comme présenté précédemment. Je l'ai entraîné sur 200 époques, pour un temps de calcul de près de 64 heures. Les moyens matériels de l'équipe ERTIM sont bien entendu incomparables avec ceux d'une firme comme Google.

On ne peut pas espérer les mêmes performances qu'avec un pré-entraînement comme Google en termes de mesures d'évaluation. Toutefois, il serait intéressant d'inspecter l'architecture du modèle avec l'approche en réseaux complexes. Par exemple, le réseau est-il plus *sparse*, clusterisé avec un pré-entraînement sur le moderne ou le classique ?

CHAPITRE 4

RÉSULTATS

Sommaire

4.1 Résultats en mesures d'évaluation **43**
4.2 Mesures sur les graphes issus des réseaux d'attention **44**
4.3 Corrélations entre mesures d'évaluation et propriétés des graphes . **45**

4.1 Résultats en mesures d'évaluation

Les résultats sur les tâches TAL sont inégaux. La table 4.1 présente les moyennes des mesures d'évaluation agrégées selon tous les modèles, sur chaque corpus. Elle illustre que la tâche sur TXAHL est moins bien réussie par les modèles, alors que les tâches sur ACC et CHisIEC ont des scores moyens acceptables.

Corpus	Mean Precision	Mean Recall	Mean F1
ACC	0.807581	0.824682	0.815673
CHisIEC	0.790286	0.828028	0.806558
TXAHL	0.479880	0.527447	0.502037

TAB. 4.1 : Classement des corpus par ordre décroissant de F-mesure moyenne

Les résultats des métriques d'évaluation obtenus pour chaque modèle, pour chaque tâche/corpus, sont présentés en annexe A.8. Le classement est sur la base de la F-mesure moyenne obtenue sur l'ensemble des runs. À partir des rangs obtenus, on peut faire un classement des modèles qui arrivent le plus souvent en tête. Nous nous proposons de sommer les rangs de chaque modèle, puis de les classer par ordre décroissant de rang cumulé. On obtient un classement des modèles dans le tableau 4.2 ci-après.

Le modèle le plus performant sur toutes les tâches est RoBERTa. Il n'est dépassé que sur le corpus CHisIEC par le modèle Guji BERT, comme illustré dans l'annexe A.8.

Globalement, les modèles pré-entraînés issus de HuggingFace réalisent des scores corrects et relativement proches les uns des autres sur les tâches ACC et CHisIEC. Il y a davantage de disparité pour le corpus TXAHL, comme illustré dans le tableau 4.3.

Modèle	Score
RoBERTa	4
Guji BERT	7
Jihuai	9
SIKU-BERT	11
Chinese BERT	16
Guwen BERT	16
Chinese BERT pré-entraîné sur Wikisource	21

TAB. 4.2 : Classement des modèles par rang cumulé décroissant

Corpus	Étendue des F-mesures obtenues
ACC	77,61% - 85,64% = 8,03%
CHisIEC	86,39% - 89,3% = 2,91%
TXAHL	46,08% - 68,61% = 22,53%

TAB. 4.3 : Étendue des F-mesures obtenues par les modèles HuggingFace par corpus

Mon modèle pré-entraîné sur le corpus Wikisource est classé dernier dans toutes les tâches. Il réalise un score honorable sur le corpus ACC, mais médiocres sur les autres corpus, comme illustré dans le tableau 4.4 ci-après.

Corpus	F-mesure moyenne
ACC	71,72%
CHisIEC	34,69%
TXAHL	4,68%

TAB. 4.4 : F-mesures moyennes obtenues par le modèle Chinese BERT pré-entraîné sur le corpus Wikisource

4.2 Mesures sur les graphes issus des réseaux d'attention

Dans cette section, je m'intéresse aux valeurs calculées sur les graphes issus des réseaux d'attention de chaque modèle, pour caractériser à quel point ces réseaux sont clusterisés. Le tableau 4.5 présente les moyennes sur les nombres de nœuds, nombre d'arêtes, longueur moyenne des plus courts chemins, coefficients de clustering et diamètres.

On peut observer que de manière générale, les modèles deviennent plus clusterisés au contact du corpus ACC, avec un coefficient de clustering proche de 50%. Par ailleurs, pour ce corpus, le degré moyen de chaque noeud est supérieur aux autres. Cette grandeur est nettement inférieure pour le corpus TXAHL. Cela suggère que les réseaux se comportent globalement de la même façon pour les corpus CHisIEC et TXAHL du point de vue de l'organisation des têtes d'attention.

Corpus	Modèle	Mean N	Mean E	Mean k	Mean L	Mean C	Mean D
ACC	Guwen BERT	3272	65070	39	2.39	0.5114	5.35
	Guji BERT	3272	65592	40	2.40	0.4963	5.43
	SIKU-BERT	3272	65389	40	2.40	0.4982	5.40
	Chinese BERT	3272	65411	40	2.40	0.4978	5.38
	Jihuai	3272	65600	40	2.40	0.4969	5.40
	RoBERTa	3272	65973	40	2.40	0.5014	5.36
	C.BERT (Wiki)	3272	66033	40	2.40	0.4953	5.36
CHisIEC	SIKU-BERT	3665	67981	37	2.53	0.3410	5.71
	Guji BERT	3665	68218	37	2.54	0.3382	5.73
	RoBERTa	3665	68863	38	2.53	0.3419	5.69
	C.BERT (Wiki)	3665	68412	37	2.53	0.3366	5.67
	Chinese BERT	3665	68053	37	2.53	0.3393	5.72
	Guwen BERT	3665	68975	38	2.53	0.3413	5.68
	Jihuai	3664	68285	37	2.53	0.3391	5.72
TXAHL	SIKU-BERT	2115	28281	27	2.59	0.3238	5.49
	Guji BERT	2115	28358	27	2.59	0.3222	5.49
	Chinese BERT	2115	28340	27	2.59	0.3239	5.49
	RoBERTa	2115	28552	27	2.58	0.3282	5.41
	Guwen BERT	2115	28263	27	2.56	0.3500	5.33
	Jihuai	2115	28497	27	2.59	0.3211	5.49
	C.BERT (Wiki)	2115	28332	27	2.58	0.3213	5.44

TAB. 4.5 : Moyennes des valeurs calculées sur les graphes issus des réseaux d'attention de chaque modèle

4.3 Corrélations entre mesures d'évaluation et propriétés des graphes

Les tableaux A.1 à A.7 montrent les coefficients de corrélation calculés entre chaque grandeur calculée à partir des graphes issus des têtes d'attention de chaque modèle. Les coefficients supérieurs à 50% en valeur absolue sont mis en gras.

Il se dégage deux groupes de modèles :

- 1) ceux pour lesquels les coefficients de corrélation sont souvent supérieurs à 50% : SIKU-BERT, Chinese BERT, Jihuai et Guji BERT ;
- 2) ceux pour lesquels les coefficients sont systématiquement très proches de 0 : Guwen BERT, RoBERTa et notre modèle Chinese BERT pré-entraîné sur Wikisource.

Pour SIKU-BERT par exemple, le tableau A.1 montre que pour toutes les tâches, le nombre d'arêtes, le degré moyen et le coefficient de la loi de puissance r présentent un coefficient de corrélation positif et important par rapport aux mesures d'évaluation. Cela suggère que pour ce modèle, plus les nœuds des graphes sont fortement reliés, meilleures sont les performances en fin de tâche. Pour les corpus ACC et TXAHL, le coefficient de clustering corrèle également plutôt bien avec les mesures d'évaluation (respectivement de l'ordre de 49% et 52%). Enfin pour le corpus

TXAHL uniquement, les longueurs moyennes des plus courts chemins et le diamètre des graphes corrèlent négativement avec les performances : plus les chemins entre deux cœuds du graphe sont courts, meilleurs sont les résultats.

Ce constat est partagé pour les modèles du groupe 1), notamment Jihuai et Guji BERT. Le cas de Chinese BERT est limite, avec des corrélations intéressantes pour ACC et CHisIEC, mais aucune pour TXAHL comme illustré dans l'annexe A.2.

Pour les modèles du groupe 2), aucune mesure d'évaluation ne corrèle avec les grandeurs des graphes, comme pour Guwen BERT en annexe A.4. Cela suggère que pour ces modèles, la structure interne du réseaux d'attention, ou du moins la vue que nous en avons par l'approche en réseaux complexes, est décorrélée des performances en fin de tâche.

partie III

Discussion et perspectives

CHAPITRE 5

DISCUSSION

Sommaire

5.1 Discussion des mesures d'évaluation **49**
5.2 Discussion des mesures sur les graphes issus des réseaux d'attention **50**
5.3 Discussion des corrélations entre mesures d'évaluation et propriétés des graphes **50**

5.1 Discussion des mesures d'évaluation

La tâche de détection d'entités nommées sur TXAHL est la moins bien réussie par les modèles. Ces résultats peuvent s'expliquer pour deux raisons :

- la nature des données : le corpus TXAHL contient des données écrites dans un chinois "transitionnel", qui n'est ni le chinois moderne sur lequel Chinese BERT (et donc les modèles basés sur celui-ci) est pré-entraîné, ni le chinois classique de 200 AEC à 500 EC.
- la quantité de données : ce corpus est le plus petit des trois utilisés dans notre travail. L'échantillon de 389 documents sur les 5 400 pourrait être étendu.

Le corpus Guwen BERT a des performances décevantes sur le corpus TXAHL et sur le corpus ACC, par rapport à la tâche sur le corpus CHisIEC. Cela est sans doute dû au pré-entraînement de ce modèle, qui a appris sur un corpus de caractères simplifiés, tandis que les modèles utilisés ici sont en sinogrammes traditionnels. Ce modèle doit rencontrer un certain nombre de caractères qui ne font pas partie de son vocabulaire, dits OOV(*Out Of Vocabulary*).

Notre modèle pré-entraîné sur le corpus Wikisource a des performances correctes pour le corpus ACC, mais mauvaises pour le corpus CHisIEC, et médiocres pour TXAHL. Cela s'explique pour moi par le pré-entraînement de ce modèle. Le corpus Wikisource est "classique", dans le sens où la langue est plus proche de celle employée dans ACC. Ainsi, le vocabulaire du modèle et sa compréhension du fonctionnement de la langue sont plus adaptés pour le ZuoZhuan que pour le chinois transitionnel de TXAHL, ou la variété des états de la langue représentée dans CHisIEC. De plus, la tâche de reconnaissance d'entités nommées est particulièrement difficile car il est impossible que ce modèle ait vu des noms de personne ou des titres officiels qui n'existaient pas à l'époque du corpus d'entraînement. Par ailleurs, un entraînement de 200 époques sur un corpus de 24 millions de tokens ne permettent pas

d'atteindre les performances de l'état de l'art, où l'on emploie de plus grandes quantités de données et on fait tourner plus longtemps le modèle. Pour se donner un ordre de grandeur, ce pré-entraînement a duré 64 heures, tandis que les entraînements par les géants du domaine se comptent en mois.

5.2 Discussion des mesures sur les graphes issus des réseaux d'attention

Les résultats ont montré que les coefficients de clustering de tous les modèles étaient globalement plus élevés pour le corpus ACC. Cela suggère qu'au contact de ce corpus, les réseaux se densifient. Cette hypothèse se confirme en examinant les degrés moyens, qui sont les plus élevés pour ce corpus, ce qui signifie que les nœuds sont en moyenne plus reliés aux autres. Le diamètre est aussi en moyenne plus petit que pour les autres corpus, même si cette différence est de l'ordre de quelques dixièmes.

Ce comportement différencié au contact du corpus ACC peut s'expliquer par la tâche de POS-tagging. La langue employée dans le corpus ACC est plus compacte : les phrases sont plus courtes, les mots sont plus souvent monosyllabiques, ce qui contribue à tisser des réseaux plus denses qu'avec des phrases plus longues (dépendances à plus longue distance). Par ailleurs, le corpus étant déjà segmenté en mots, les nœuds du réseau sont davantage identifiés, et leurs relations aux autres mots s'aligne mieux avec la réalité de la syntaxe.

5.3 Discussion des corrélations entre mesures d'évaluation et propriétés des graphes

Les résultats obtenus scindent les modèles sélectionnés en trois groupes : ceux pour lesquels on a des corrélations significatives entre les mesures d'évaluation et les propriétés des graphes issus des têtes d'attention (1), et ceux pour lesquels on n'a aucune corrélation (2).

Parmi les modèles du groupe (1), le nombre d'arêtes dans le graphe E corrèle positivement avec les mesures d'évaluation. Cela suggère que plus les nœuds sont reliés les uns aux autres, plus le modèle est performant. Ceci est à mettre en lien avec le degré moyen, qui est également positivement corrélé aux mesures de précision, rappel et F-mesure. En prenant ces deux indicateurs, on peut affirmer que plus le réseau est densément connecté, meilleures sont les performances. Cette suggestion est soutenue par le fait que le coefficient de clustering corrèle également positivement avec les performances.

Le coefficient de la loi de puissance aussi corrèle positivement avec les mesures d'évaluation. La figure 5.1 illustre l'allure des courbes de la fonction de puissance selon différentes valeurs du paramètre γ dans $f(x) = x^{-\gamma}$, appelé r dans les données. On voit que plus r est grand, plus on a d'éléments peu connectés dans la distribution, sans remettre en question les *hubs*. De ce point de vue, on pourrait supposer qu'un paramètre r élevé suggère un graphe avec davantage de nœuds, donc une taille de vocabulaire plus grande, d'où de meilleures performances.

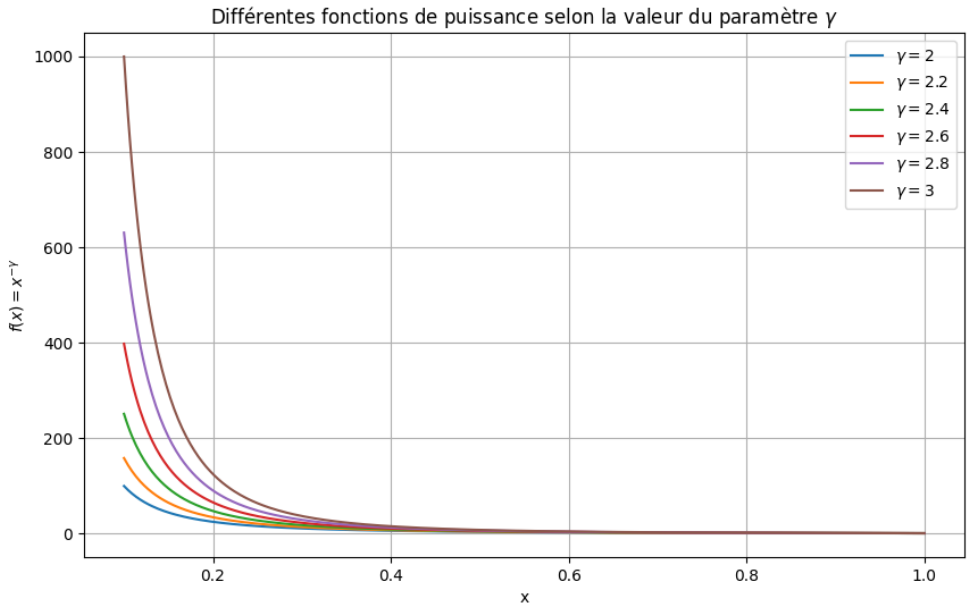

FIG. 5.1 : Lois de puissance selon le paramètre γ

Notons que les longueurs moyennes des plus courts chemins L et le diamètre D des graphes corrèlent négativement avec les performances. Ainsi plus les chemins entre deux nœuds sont longs en moyenne, moins les performances sont bonnes. Cela va dans le sens d'une conception "économique" du fonctionnement de ces réseaux : pouvoir relier deux nœuds rapidement pourrait être une condition de la performance, comme si chaque arc traversé dans le réseau était associé à un coût, qu'il fallait minimiser. Le diamètre est un indicateur plus globale de la taille du réseau, et sa corrélation négative avec les performances va dans le sens de notre hypothèse.

CHAPITRE 6

PERSPECTIVES

Ce travail invite à davantage d'expériences pour approfondir la question soulevée. En particulier, pour expliquer la division en deux groupes, les modèles qui corrèlent, et ceux qui ne corrèlent pas. On pourrait notamment étudier les relations entre les représentations en réseau complexe des corpus du modèle sans entraînement, et les métriques en fin de tâche. Ceci pourrait permettre d'évaluer l'impact du pré-entraînement du modèle et de son architecture, indépendamment du fine-tuning. On pourrait reformuler la question et se demander si l'état initial des têtes d'attention, avant leur réorganisation au cours du fine-tuning ne serait pas un déterminant des performances en fin de tâche.

Il serait intéressant d'étudier le même problème avec une autre langue, avec davantage de variabilité morphologique par exemple. Nos résultats se limitent au chinois classique, et même, aux variétés représentées dans les corpus. On pourrait aussi envisager de construire de manière différente les réseaux lexicaux issus des corpus : plutôt que de relier deux tokens consécutifs dans le texte, formuler une analyse syntaxique et relier chaque token à la tête du/des syntagme(s) au(x)quel il appartient. Cette approche permettrait peut-être de faire un parallèle plus évident entre les graphes obtenus et la réalité de la langue.

Le corpus ACC diffère des corpus CHisIEC et TXAHL sur deux points :

- la tâche TAL : POS-tagging, au lieu de la reconnaissance d'entités nommées (REN) ;
- la tokenisation : sur les mots, au lieu des caractères.

On pourrait notamment étudier dans quelle mesure la segmentation impacte les résultats. On peut remarquer que dans les données fournies, comme expliqué dans la section 3.2.3, l'annotation du corpus ACC en "parties du discours" contient deux étiquettes qui relèvent de la reconnaissance d'entités nommées : nom de personne et nom de lieu. Ainsi, il est possible de modifier le corpus pour entraîner les modèles sur une tâche de REN. De cette manière, on pourrait isoler la variable de la tokenisation et voir si les modèles se comportent de la même façon au contact des mêmes données, sur une tâche différente.

Les modèles testés ici sont tous basés sur Chinese BERT. On pourrait compléter ces expériences en ajoutant d'autres familles de modèle. Par exemple, les LLMs de type Llama ; mais ces expériences seraient assez énergivores compte tenu de l'ordre de grandeur du nombre de paramètres de ces modèles. Une autre piste serait

d'explorer les Tiny Language Models. Ces modèles de taille réduite peuvent rivaliser avec des modèles bien plus lourds, avec une consommation en ressources moindres. C'est une autre approche de l'IA explicable : créer des modèles-jouets *ad hoc* pour tester des hypothèses. Dans ce travail nous avons favorisé l'autre démarche, à savoir ouvrir la boîte noire des gros modèles les plus utilisés.

Conclusion générale

Dans ce mémoire, nous avons traité de la problématique suivante : dans quelle mesure peut-on établir un lien entre la structure interne des modèles après fine-tuning et leurs performances sur les tâches TAL de POS-tagging et de reconnaissance d'entités nommées ? Pour répondre à cette question, nous avons étudié six modèles librement disponibles, et un modèle pré-entraîné sur nos données. Nous entraînons ces modèles sur trois corpus, chacun représentant une ou plusieurs variété(s) de chinois classique ; nous représentons la structure des têtes d'attention à l'aide de réseaux complexes, avant de procéder à des mesures statistiques. En conclusion, il n'y a pas de corrélation évidente entre les mesures d'évaluation en fin de tâche (rappel, précision et F-mesure) et les grandeurs calculées sur les graphes issus des réseaux d'attention. Les modèles testés se divisent en deux groupes : ceux qui présentent des corrélation importantes, supérieures à 50% en valeur absolue, et ceux qui ne présentent aucune corrélation. Des travaux supplémentaires sont à envisager pour explorer ces deux familles de modèles, et mieux comprendre l'origine de ces deux comportements très caractérisés.

BIBLIOGRAPHIE

[Albert and Barabási, 2002] Albert, R. and Barabási, A.-L. (2002). Statistical mechanics of complex networks. *Reviews of modern physics*. – Cité page 26.

[Barnes et al., 2009] Barnes, A., Starr, D., and Ormerod, G. (2009). *Du's Handbook of Classical Chinese Grammar*. Great Britain : Alcuin Academics. – Cité page 23.

[Blouin and al., 2023] Blouin, B. and al. (2023). Unlocking transitional chinese : Word segmentation in modern historical texts. In *M. Hämäläinen et al. (eds) Proceedings of the Joint 3rd International Conference on Natural Language Processing for Digital Humanities and 8th International Workshop on Computational Linguistics for Uralic Languages. NLP4DH-IWCLUL 2023, Tokyo, Japan : Association for Computational Linguistics, pp. 92–101. Available at : https :/ /aclanthology.org/2023.nlp4dh-1.11 (Accessed : 17 November 2024).* – Cité pages 21 et 23.

[Bretto et al., 2012] Bretto, A., Faisant, A., and Hennecart, F. (2012). *Eléments de théorie des graphes*. Springer Paris. – Cité page 25.

[Brown, 2024] Brown, . B. . (2024). Neural networks. https://www.youtube.com/watch?v=aircAruvnKk&list=PLZHQObOWTQDNU6R1_67000Dx_ZCJB-3pi&ab_channel=3Blue1Brown. – Cité page 19.

[Cho et al., 2014] Cho, K., van Merrienboer, B., Gulcehre, C., Bougares, F., Schwenk, H., and Bengio, Y. (2014). Learning phrase representations using rnn encoder-decoder for statistical machine translation. *CoRR*. – Cité page 18.

[Devlin et al., 2019] Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT : Pre-training of deep bidirectional transformers for language understanding. In Burstein, J., Doran, C., and Solorio, T., editors, *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. – Cité page 20.

[Ferrer et al., 2004] Ferrer, I. C. R., RV, S., and R, K. (2004). Patterns in syntactic dependency networks. *Phys Rev E Stat Nonlin Soft Matter Phys.* – Cité page 27.

[Hochreiter and Schmidhuber, 1997] Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*. – Cité page 18.

[Jacques, 2022] Jacques, G. (2022). On the nature of morphological alternations in archaic chinese and their relevance to morphosyntax. *Bulletin of the School of Oriental and African Studies*. – Cité page 23.

[Mancio et al., 2012] Mancio, D. R., Oliveira, J. O. N., and COSTA, L. d. F. (2012). Structure–semantics interplay in complex networks and its effects on the predictability of similarity in texts. *Physica A : Statistical Mechanics and its Applications*. – Cité page 27.

[Nylan, 2001] Nylan, M. (2001). *The Five "Confucian" Classics*. Yale University Press. – Cité page 23.

[Peyraube, 2008] Peyraube, A. (2008). *Ancient Chinese*. Cambridge University Press. – Cité page 23.

[Salman et al., 2023] Salman, M., Haller, A., and et Méndez, S. J. R. (2023). Syntactic complexity identification, measurement, and reduction through controlled syntactic simplification. *International workshop on Learning with Knowledge Graphs (IWLKG)*. – Cité page 27.

[Strogatz, 2001] Strogatz, S. (2001). Exploring complex networks. *Nature*. – Cité page 26.

[Tang and al., 2024] Tang, X. and al. (2024). Chisiec : An information extraction corpus for ancient chinese history. *roceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*. – Cité pages 21 et 35.

[Vaswani and al., 2017] Vaswani, A. and al. (2017). Attention is all you need. *Google Research*. – Cité pages 17, 19 et 20.

[Wang and al., 2023] Wang, D. and al. (2023). Gujibert and gujigpt : Construction of intelligent information processing foundation language models for ancient texts. *arXiv.2307.05354*. – Cité page 24.

[Watts and Strogatz, 1998] Watts, D. and Strogatz, S. (1998). Collective dynamics of small-world networks. *Nature*. – Cité page 26.

[Zhang and al., 2022] Zhang, H. and al. (2022). Bert 4ever@evahan 2022 : Ancient chinese word segmentation and part-of-speech tagging based on adversarial learning and continual pre-training. *Proceedings of the Second Workshop on Language Technologies for Historical and Ancient Languages. LT4HALA 2022*. – Cité page 22.

[Zheng and Xiao, 2024] Zheng, J. and Xiao, X. (2024). A complex network approach to analyse pre-trained language models for ancient chinese. *Royal Society Open Science*. – Cité pages 25, 27, 33 et 40.

ANNEXES

Corpus	Grandeur	Précision	Rappel	F1
	N	**-0.999681**	**-0.999771**	**-0.999984**
	E	**0.517468**	**0.517305**	**0.517517**
	$< k >$	**0.533244**	**0.533083**	**0.533298**
	r	**0.679709**	**0.679696**	**0.679883**
ACC	R2	0.309836	-0.309427	-0.309715
	L	-0.047794	-0.047840	-0.047836
	C	0.489240	0.489449	0.489483
	D	-0.338006	-0.338256	-0.338249
	N	NaN	NaN	NaN
	E	**0.699022**	**0.700166**	**0.699814**
	$< k >$	**0.699022**	**0.700166**	**0.699814**
	r	**0.760056**	**0.760193**	**0.760366**
CHisIEC	R2	-0.351057	-0.351834	-0.351561
	L	0.292601	0.293901	0.293332
	C	0.037883	0.036673	0.037294
	D	-0.215579	-0.215230	-0.215448
	N	NaN	NaN	NaN
	E	**0.636292**	**0.638867**	**0.637782**
	$< k >$	**0.636292**	**0.638867**	**0.637782**
	r	**0.699097**	**0.701585**	**0.700557**
TXAHL	R2	-0.153421	-0.153678	-0.153568
	L	**-0.519302**	**-0.521187**	**-0.520429**
	C	**0.520195**	**0.521818**	**0.521174**
	D	**-0.515132**	**-0.516939**	**-0.516209**

TAB. A.1 : Mesures obtenues avec le modèle SIKU-BERT
N : nombre de nœuds, E : nombre d'arêtes, $< k >$: degré moyen, r : coefficient de la loi de puissance estimée par le modèle, R2 : *fitting coefficient*, L : longueur moyenne des plus courts chemins, C : coefficient de clustering, D : diamètre du graphe

Corpus	Grandeur	Précision	Rappel	F1
	N	**-0.999968**	**-0.999925**	**-0.999990**
	E	0.475146	0.475473	0.475319
	$< k >$	0.492966	0.493291	0.493139
	r	**0.621942**	**0.622235**	**0.622106**
ACC	R2	-0.247186	-0.247554	-0.247369
	L	-0.368461	-0.368101	-0.368307
	C	**0.597697**	**0.597691**	**0.597718**
	D	-0.374764	-0.374718	-0.374756
	N	NaN	NaN	NaN
	E	**0.540936**	**0.539671**	**0.540588**
	$< k >$	**0.540936**	**0.539671**	**0.540588**
	r	**0.624769**	**0.624666**	**0.624988**
CHisIEC	R2	-0.217269	-0.216231	-0.216956
	L	-0.323915	-0.323640	-0.323921
	C	0.443262	0.443496	0.443511
	D	-0.374654	-0.375084	-0.375039
	N	NaN	NaN	NaN
	E	0.197263	0.192554	0.197036
	$< k >$	0.197263	0.192554	0.197036
	r	0.214178	0.211467	0.214723
TXAHL	R2	-0.024137	-0.020826	-0.023093
	L	-0.161218	-0.160307	-0.162096
	C	0.225423	0.225383	0.227012
	D	-0.173879	-0.172683	-0.174830

TAB. A.2 : Mesures obtenues avec le modèle Chinese BERT
N : nombre de nœuds, E : nombre d'arêtes, $< k >$: degré moyen, r : coefficient de la loi de puissance estimée par le modèle, R2 : *fitting coefficient*, L : longueur moyenne des plus courts chemins, C : coefficient de clustering, D : diamètre du graphe

Corpus	Grandeur	Précision	Rappel	F1
	N	**-0.999964**	**-0.999527**	**-0.999902**
	E	0.263480	0.263501	0.263558
	$< k >$	0.289432	0.289441	0.289507
	r	**0.545693**	**0.545376**	**0.545628**
ACC	R2	-0.085324	-0.085434	-0.085412
	L	-0.416264	-0.416136	-0.416248
	C	**0.646134**	**0.645743**	**0.646037**
	D	-0.368419	-0.368228	-0.368421
	N	0.081918	0.081945	0.081960
	E	0.393181	0.393442	0.393474
	$< k >$	0.393002	0.393262	0.393294
	r	**0.687638**	**0.688118**	**0.688106**
CHisIEC	R2	-0.091683	-0.092323	-0.092086
	L	-0.370458	-0.370590	-0.370636
	C	0.473567	0.473149	0.473482
	D	-0.332091	-0.331791	-0.332057
	N	NaN	NaN	NaN
	E	**0.547012**	**0.548509**	**0.548115**
	$< k >$	**0.547012**	**0.548509**	**0.548115**
	r	**0.596705**	**0.601003**	**0.599227**
TXAHL	R2	-0.071795	-0.069079	-0.070520
	L	**-0.502302**	**-0.505583**	**-0.504266**
	C	**0.536148**	**0.542261**	**0.539523**
	D	-0.493467	-0.498201	-0.496089

TAB. A.3 : Mesures obtenues avec le modèle Jihuai
N : nombre de nœuds, E : nombre d'arêtes, $< k >$: degré moyen, r : coefficient de la loi de puissance estimée par le modèle, R2 : *fitting coefficient*, L : longueur moyenne des plus courts chemins, C : coefficient de clustering, D : diamètre du graphe

Corpus	Grandeur	Précision	Rappel	F1
ACC	N	**-0.998169**	**-0.998845**	**-0.999799**
	E	-0.074111	-0.075380	-0.074855
	$< k >$	-0.044539	-0.045792	-0.045237
	r	-0.063021	-0.061480	-0.062396
	R2	0.029196	0.030808	0.030011
	L	-0.008638	-0.008197	-0.008425
	C	0.002777	0.004424	0.003574
	D	0.027883	0.028871	0.028414
CHisIEC	N	NaN	NaN	NaN
	E	0.004850	0.006267	0.005588
	$< k >$	0.004850	0.006267	0.005588
	r	-0.062713	-0.062934	-0.062793
	R2	-0.033305	-0.034710	-0.034022
	L	-0.013091	-0.012261	-0.012611
	C	-0.052939	-0.054620	-0.053816
	D	0.014706	0.016016	0.015255
TXAHL	N	NaN	NaN	NaN
	E	0.049831	0.050834	0.050386
	$< k >$	0.049831	0.050834	0.050386
	r	0.004846	0.004451	0.004589
	R2	-0.042740	-0.043553	-0.043217
	L	0.032171	0.032898	0.032547
	C	-0.040346	-0.041265	-0.040898
	D	0.045861	0.046451	0.046188

TAB. A.4 : Mesures obtenues avec le modèle Guwen BERT
N : nombre de nœuds, E : nombre d'arêtes, $< k >$: degré moyen, r : coefficient de la loi de puissance estimée par le modèle, R2 : *fitting coefficient*, L : longueur moyenne des plus courts chemins, C : coefficient de clustering, D : diamètre du graphe

Corpus	Grandeur	Précision	Rappel	F1
	N	**-0.998601**	**-0.998835**	**-0.999907**
	E	0.411543	0.411945	0.412234
	$< k >$	0.431240	0.431645	0.431956
	r	0.365404	0.365008	0.365646
ACC	R2	-0.191139	-0.190327	-0.190962
	L	-0.417433	-0.418062	-0.418231
	C	**0.657711**	**0.658058**	**0.658655**
	D	-0.425365	-0.426562	-0.426440
	N	NaN	NaN	NaN
	E	**0.504274**	**0.503989**	**0.504258**
	$< k >$	**0.504274**	**0.503989**	**0.504258**
	r	0.428210	0.428181	0.428309
CHisIEC	R2	-0.169846	-0.169547	-0.169690
	L	-0.362978	-0.363574	-0.363361
	C	**0.523123**	**0.523773**	**0.523597**
	D	-0.409079	-0.409777	-0.409511
	N	NaN	NaN	NaN
	E	**0.617128**	**0.616455**	**0.617646**
	$< k >$	**0.617128**	**0.616455**	**0.617646**
	r	0.470951	0.470803	0.471502
TXAHL	R2	-0.138229	-0.139736	-0.139091
	L	-0.496391	-0.496398	-0.497071
	C	**0.558182**	**0.557113**	**0.558440**
	D	**-0.527971**	**-0.529130**	**-0.529189**

TAB. A.5 : Mesures obtenues avec le modèle Guji BERT
N : nombre de nœuds, E : nombre d'arêtes, $< k >$: degré moyen, r : coefficient de la loi de puissance estimée par le modèle, R2 : *fitting coefficient*, L : longueur moyenne des plus courts chemins, C : coefficient de clustering, D : diamètre du graphe

Corpus	Grandeur	Précision	Rappel	F1
ACC	N	NaN	NaN	NaN
	E	-0.003927	0.011663	0.009094
	$< k >$	-0.003927	0.011663	0.009094
	r	0.013438	-0.009306	0.001404
	R2	0.015276	-0.016907	-0.004827
	L	0.004893	0.001027	0.005800
	C	0.033682	-0.026937	-0.000067
	D	-0.005351	0.008543	0.006231
CHisIEC	N	NaN	NaN	NaN
	E	-0.013128	-0.025815	-0.023624
	$< k >$	-0.013128	-0.025815	-0.023624
	r	0.010634	-0.018877	-0.008303
	R2	0.007374	0.015284	0.013771
	L	-0.015170	-0.035595	-0.031576
	C	0.021734	0.018754	0.024120
	D	0.000186	0.008079	0.005034
TXAHL	N	NaN	NaN	NaN
	E	-0.004589	0.000072	-0.003050
	$< k >$	-0.004589	0.000072	-0.003050
	r	-0.001764	0.007753	0.002531
	R2	0.007982	0.005311	0.007024
	L	-0.005091	0.000607	-0.002972
	C	0.015728	0.015323	0.016142
	D	-0.014347	-0.018927	-0.017146

TAB. A.6 : Mesures obtenues avec le modèle RoBERTa
N : nombre de nœuds, E : nombre d'arêtes, $< k >$: degré moyen, r : coefficient de la loi de puissance estimée par le modèle, R2 : *fitting coefficient*, L : longueur moyenne des plus courts chemins, C : coefficient de clustering, D : diamètre du graphe

Corpus	Grandeur	Précision	Rappel	F1
	N	NaN	NaN	NaN
	E	0.017974	-0.030883	-0.005002
	$< k >$	0.017974	-0.030883	-0.005002
	r	0.018304	-0.017354	0.008137
ACC	R2	-0.013909	0.019186	-0.001525
	L	0.020773	-0.019226	0.009909
	C	-0.010178	0.000388	0.009909
	D	0.018875	-0.011149	-0.014468
	N	NaN	NaN	NaN
	E	-0.028550	-0.006439	-0.017357
	$< k >$	-0.028550	-0.006439	-0.017357
	r	-0.006271	0.012622	0.007013
CHisIEC	R2	0.032062	0.010175	0.021499
	L	-0.016866	0.006776	-0.002734
	C	0.029500	0.007358	0.019164
	D	-0.022002	-0.005805	-0.013436
	N	NaN	NaN	NaN
	E	0.033773	0.033773	0.033773
	$< k >$	0.033773	0.033773	0.033773
	r	0.019585	0.019585	0.019585
TXAHL	R2	-0.022789	-0.022789	-0.022789
	L	0.046899	0.046899	0.046899
	C	-0.044643	-0.044643	-0.044643
	D	-0.011086	-0.011086	-0.011086

TAB. A.7 : Mesures obtenues avec notre modèle Chinese BERT pré-entraîné sur le corpus Wikisource
N : nombre de nœuds, E : nombre d'arêtes, $< k >$: degré moyen, r : coefficient de la loi de puissance estimée par le modèle, R2 : *fitting coefficient*, L : longueur moyenne des plus courts chemins, C : coefficient de clustering, D : diamètre du graphe

Corpus	Modèle	Mean Precision	Mean Recall	Mean F1
ACC	Guwen BERT	0.7715	0.7809	0.7761
	Guji BERT	0.8314	0.8591	0.8449
	SIKU-BERT	0.8303	0.8523	0.8408
	Chinese BERT	0.8077	0.8434	0.8249
	Jihuai	0.8406	0.8596	0.8495
	RoBERTa	0.8448	0.8684	0.8564
	C. BERT (Wikisource)	0.7268	0.7092	0.7172
CHisIEC	SIKU-BERT	0.8701	0.8894	0.8792
	Guji BERT	0.8840	0.9040	0.8930
	RoBERTa	0.8772	0.9032	0.8896
	C. BERT (Wikisource)	0.2894	0.4330	0.3469
	Chinese BERT	0.8577	0.8710	0.8639
	Guwen BERT	0.8746	0.8994	0.8863
	Jihuai	0.8789	0.8964	0.8870
TXAHL	SIKU-BERT	0.6066	0.6539	0.6291
	Guji BERT	0.5925	0.6529	0.6214
	Chinese BERT	0.4608	0.5076	0.4828
	RoBERTa	0.6681	0.7053	0.6861
	Guwen BERT	0.4350	0.4903	0.4608
	Jihuai	0.5592	0.6194	0.5874
	C. BERT (Wikisource)	0.0371	0.0627	0.0468

TAB. A.8 : Tableau des modèles avec les métriques de précision, rappel et F-mesure pour chaque corpus