PROPOSITION DE SUJET DE THESE

SIGLE ET NOM DU LABORATOIRE : UR 7537 BIOSTM - Biostatistique, Traitement et Modélisation des données biologiques
NOM DE L'EQUIPE : NA
DIRECTEUR DE THESE : PROF. YVES ROZENHOLC, DIRECTEUR DE L'EQUIPE
ADRESSE : 4 AVENUE DE L'OBSERVATOIRE, 75006 PARIS

TITRE DE LA THESE : SUIVI DYNAMIQUE, DETECTION D'ASSOCIATIONS ET DE RUPTURES DANS LES MODELES DE LANGAGE

CO-ENCADRANT EVENTUEL : CHRISTINE KERIBIN
EQUIPE DU CO-ENCADRANT : EQUIPE PROBABILITES ET STATISTIQUES
LABORATOIRE : LABORATOIRE DE MATHEMATIQUES D'ORSAY (UMR 8628)

PRESENTATION DU SUJET

2 à 4 pages (références comprises), structurées comme suit :

1. *le contexte scientifique du projet ;*

L'actualité des derniers mois a mis un focus fort sur les mega-modèles génératifs de langage comme GPT. Voulant être généralistes, ces mega-modèles impliquent des centaines de milliards de paramètres et consomment une puissance de calcul et donc d'électricité prodigieuse tant dans leur phase d'apprentissage que dans leur phase d'utilisation. Ces mega-modèles souffrent de plus d'une rigidité liée à leur représentation nécessairement extrêmement riche et l'apprentissage de ces centaines de milliards de paramètres. Cette rigidité contraint leur mise à jour à être faite *off-line* en relançant l'estimation des paramètres à intervalle régulier pour prendre en compte les évolutions de l'information et du langage. D'un autre côté, il existe des modèles spécialisés de langage plus petits et donc plus économiques qui offrent de très bonnes performances mais souffrent de leur spécialisation.

D'un point de vue statistique, la représentation du langage, de ses formes stables et de ses évolutions peuvent être vues comme répondant à des problèmes de détection de rupture ou de clustering, d'une part pour différencier des éléments « stables » du langage, d'autre part pour détecter des changements intervenants au court du temps dans un ou plusieurs éléments « stables ». Il devient alors possible d'imaginer entrainer plusieurs petits modèles de langue sur chaque élément stable et développer un modèle dynamique de langage pour prendre en compte les variations détectées --éventuellement dans

des contextes très déséquilibrés. L'idée est alors, en lieu et place d'apprendre les liens entre les mots ou les concepts, d'apprendre les dérivées de ces liens.

2. *les questions posées ;*

Dans ce contexte, les questions posées sont :
1/ Est-il possible de construire des modèles génératifs de langage plus économiques pouvant suivre et prendre en compte les variations du langage ?
2/ Des méthodes classiques de détections de changements dans les moments d'ordre 1 ou 2 du langage (fréquences, co-occurrences) sont-elles à même de signaler qu'un modèle de langage doit être remis à jour en particulier dans des contextes très déséquilibrés ?
3/ Est-il possible de segmenter à l'aide de méthodes de statistiques classiques (clustering, co-clustering, clustering de graphes, clustering variationnel, …) des matrices d'occurrence de langage pour repérer des sous-ensembles de langage « stables » et des sous-ensembles du langage « variables » ?
4/ Enfin, comment se positionnent les méthodes proposées issues de l'apprentissage statistique avec celles provenant du monde du traitement automatique du langage telles que les réseau de neurones profonds

3. *les sources de données qui seront utilisées ;*

Depuis janvier 2020, l'UR 7537 BioSTM a construit une base de données Dataverse[1], protégées sur ses serveurs, des parties publiques des médias européens, des journaux et des réseaux sociaux. A l'aide du module Python, pyDataverse, qui permet d'effectuer et d'automatiser des tâches liées à l'archivage, il a été possible de construire des "aspirateurs" de données et métadonnées qui collectent les données publiques disponibles automatiquement. Pour ces collectes par des flux automatisés, une attention toute particulière a été portée sur l'extraction et le choix des métadonnées collectées afin de garantir les utilisations futures, faciliter l'utilisation future de moteurs de recherche et l'échange entre les applications. Le contenu de chaque article a été automatiquement indexé en métadonnées (par exemple pays, origine, langue, titre, date, … mais aussi vocabulaire contrôlé comme les institutions, les noms, …) dans un format compatible RDF[2] basé sur JSON[3] à travers une API sémantique spécifique de Dataverse

[1] Le logiciel Dataverse, une application web open-source, développée par l'Université de Harvard, qui permet la construction et le partage de collections de données et leurs meta-données, qui a déjà été validée par l'UE et est déjà utilisées par de nombreux centres de recherches en France.
[2] RDF, *Resource Description Framework*, est une norme du World Wide Web Consortium (W3C) pour l'échange de données sur le Web.
[3] JSON, *JavaScript Object Notation* (JSON) est un format standard utilisé pour représenter des données structurées d'une manière similaire aux objets JavaScript. Il fournit des formats de fichiers et d'échange de données standard ouverts, utilisés dans l'échange électronique de données, comme les applications web avec les serveurs.

qui permet d'importer/exporter des métadonnées en JSON-LD, un format RDF basé sur JSON. Un moteur de recherche basé sur Elasticsearch[4] est disponible pour interroger cette base de données.

Cette base de données regroupe, en provenance de tout pays européen dans leur langue d'origine, aujourd'hui plus de 100 000 000 (cent millions) de documents avec leurs meta-données associées et continue à être alimentée.

4. *les méthodes ;*

Partant, à l'aide du moteur de recherche basé sur Elasticsearch, d'extractions de la forme « Covid IN Title », « Publication FROM day1 to day2 », « Langage IS French », « Country IS France »… et de leur composition booléennes avec les opérateurs AND et OR, il est possible d'extraire de notre Dataverse les articles en français contenant le terme « Covid » dans le titre, publiés semaine après semaine en Français, qui ont été publiés en France, en Belgique ou au Luxembourg et de faire des sous-ensembles séparés sur lesquels vont être estimés des moments d'ordre 1 et 2 c'est-à-dire les occurrences et les co-occurences. Il est alors possible de comparer ces estimations, soit entre pays pour une même semaine, soit au sein d'un même pays entre les semaines, etc.
Fort de la connaissance de l'histoire récente liée à l'apparition du SARS-Cov-2, il est possible de savoir à quelle période le langage en lien avec le terme « Covid » a évolué ou comment entre deux pays de langues françaises les discussions ou questionnements ont pu être communs ou différents.
Les termes « Autorisation », « Pangolin », « Masque », « Manque », « Balade », « Plage ou Montagne », « Vaccination », « ARN », « Anti-Vax » … sont alors autant de marqueurs dont l'historicité de l'apparition (ou son absence d'apparition) au sein du langage et des media est connue dans chacun des pays considérés. La comparaison du langage à travers les articles avant ou après l'apparition de ces termes permettent de produire des exemples de variations temporelles et internationales du langage.

La première partie de la thèse consistera, après la constitution de ces ensembles d'articles, à appliquer les méthodes classiques de détection de changements dans les moments d'ordre 1 ou 2 (fréquences, co-occurrences) et les méthodes de statistiques classiques de segmentation/classification (clustering, co-clustering, clustering de graphes, clustering variationnel, …) au regard de l'historicité de la pandémie pour repérer automatiquement des détections connues dans le langage et voir s'il est possible de segmenter le langage en sous-ensembles de langage « stables » et des sous-ensembles du langage « variables ». Ici, les méthodes de traitement automatique du langage seront utilisées comme des boites noires à appliquer sur les ensembles détectés différents pour déceler comment les modèles génératifs sont sensibles à ces changements.

[4] Elasticsearch est un moteur de recherche plein texte distribué et multi-tenant avec une interface web HTTP et des documents JSON sans schéma.

Enfin, la question se posera de considérer directement comment les modèles de langage sont construits en amont et aval d'un changement et d'étudier si la détection statistique peut ou non apporter une information de type : « le modèle doit être réinitialisé » , offrant ainsi une alarme précoce. Pour ce faire, le candidat ou la candidate étudiera les variations d'un modèle de langage comme BERT au court du temps pour voir à quelle « distance » un changement influe sur celui-ci et comparera cette « distance » à celle obtenue par les méthodes de détections statistiques.

5. *puissance de l'étude/nombre de sujets ;*

Non applicable ici. Toutefois, si l'échelle de la semaine n'est pas pertinent statistiquement à la détection, il sera possible d'adapter le suivi à un temps plus long comme le mois, le trimestre, …

6. *le calendrier prévisionnel (présenté sous forme d'un échéancier semestriel, doit être suffisamment précis pour constituer un document de référence sans pour autant traiter du détail. Le calendrier doit inclure la période de rédaction et celle d'examen par les rapporteurs) ;*

Semestre 1
Apprentissage de Dataverse, Elasticsearch, formulation d'une requête, constitution des ressources en français par semaine et pour les trois pays France, Luxembourg, Belgique.
Constitution d'une bibliographie sur la détection de ruptures et les modèles de langage.

Semestre 2
Bibliographie sur le clustering, le bi-clustering et la segmentation Modèles de clustering et bi-clustering pour la segmentation du langage dans son évolution
Etude des occurrences et des co-occurrences par semaine. Recherche de détections dans les moments de premier et second ordre.

Semestre 3
Génération des modèles de langage à l'aide de BERT et comparaison des textes générés.
Ecriture et soumission du premier article

Semestre 4
Représentations graphiques des évolutions du langage
Ecriture et soumission du second article

Semestre 5
Rédaction de la thèse

Semestre 6
Soumission
Ecriture et soumission du troisième article
Préparation de la soutenance

7. *le thème de chacun des articles prévus. Une proposition de sujet de thèse doit comporter au moins deux articles originaux.*

Cette thèse doit donner lieu à trois articles :
- Détection de changements dans le langage et application aux modèles génératifs
- Représentation de l'évolution du langage – un exemple à partir du suivi des articles contenant « Covid » dans leur titre
- Segmentation/clusterisation dynamique du langage dans un contexte évolutif

8. *Références bibliographiques*

Rozenholc Y., *Nonparametric tests of change-points with tapered data*, J. Time Ser. Anal. 22 (2001), no. 1, 13–43.

Liu F., Cuenod C.-A., Thomassin-Naggara I., Chemouny S., and **Rozenholc Y.**, *Hierarchical segmentation using equivalence test (HiSET) : Application to DCE image sequences*, Medical Image Analysis 51 (2019), 125–143.

Clarté G., Ryder R. J., *A Phylogenetic Model of the Evolution of Discrete Matrices for the Joint Inference of Lexical and Phonological Language Histories*, arXiv preprint arXiv:2206.12473, (2022)

Antonazzo F., Biernacki C., **Keribin C.**, *Frugal Gaussian clustering of huge imbalanced datasets through a bin-marginal approach*, Statistics and Computing (2023)

Biernacki C., Jacques J., **Keribin C**., *A Survey on Model-Based Co-Clustering: High Dimension and Estimation Challenges*, Journal of Classification (https://inria.hal.science/hal-03769727/) (2023)

PREREQUIS, FORMATION :

LE CANDIDAT AURA UNE FORTE APPETENCE EN STATISTIQUE ET EN INFORMATIQUE AVEC UNE BONNE CONNAISSANCE DE PYTHON ET DE LIBRAIRIES COMME PANDA, SCIKIT-LEARN, PYTORCH. UNE CONNAISSANCE D'UN MODELE DE LANGAGE COMME BERT OU CANINE SERA UN PLUS DE MEME QUE LA CONNAISSANCE DE TYPE D'UN ENVIRONNEMENT COMME ANACONDA, JUPYTER, VISUAL STUDIO.

CONTACT POUR CE SUJET : YVES ROZENHOLC
EMAIL : YVES.ROZENHOLC@U-PARIS.FR
TELEPHONE : 06 63 68 82 65

ECOLE DOCTORALE PIERRE LOUIS DE SANTE PUBLIQUE A PARIS
EPIDEMIOLOGIE ET SCIENCES DE L'INFORMATION BIOMEDICALE

Directeur : Pierre-Yves Boëlle
Responsable pour l'Université Paris Cité : Isabelle Boutron

SPECIALITE DE LA THESE

Santé publique - Epidémiologie	☐
Santé publique - Epidémiologie clinique	☐
Santé publique - Epidémiologie sociale	☐
Santé publique - Epidémiologie génétique	☐
Santé publique - Biostatistique	☐
Santé publique - Biomathématiques	☐
Santé publique - Biostatistique et Biomathématiques	☐
Santé publique - Informatique médicale	☐
Santé publique - Imagerie biomédicale	☐
Santé publique - Bioinformatique	☐
Santé publique - Recherches sur les services de santé	☐
Santé publique - Economie de la santé	☐
Santé publique - Science des données	☒
Santé publique – Prévention et promotion de la santé	☐

SIGNATURE DU DIRECTEUR DE THESE

[redacted]

VISA DU DIRECTEUR DU LABORATOIRE
(DEROGATION DE SIGNATURE NON ACCEPTEE)

AVIS FAVORABLE ☒

[redacted]

Ecole Doctorale 393
Centre Biomédical des Cordeliers
15, rue de l'Ecole de Médecine 75006 Paris
www.ed393.upmc.fr
Contact : magali.moulie@sorbonne-universite.fr/Téléphone : 01.44.27.24.35