COVID Information Commons (CIC) Research Lightning Talk

Transcript of a Presentation by Jane Pan (Princeton University), September 22, 2021

Title: Contradiction Detection of COVID-19 Randomized Controlled Trials via BERT Language Models

YouTube Recording with Slides

September 2021 CIC Webinar Information

Transcript Editor: Julie Meunier

Transcript

Lauren Close:

Slide 1:

Passons rapidement à notre prochain intervenant aujourd'hui, Jane Pan. Jane est en fait l'une des trois lauréates de notre tout premier défi étudiant de rédaction de documents de premier cycle CIC [COVID Information Commons]. Elle a remporté la première place. Le défi a eu lieu plus tôt ce printemps, et nous accueillons Jane avec beaucoup d'enthousiasme pour partager ses recherches avec la communauté élargie du CIC. Alors, Jane, c'est à toi.

Jane Pan:

Slide 2:

Génial, d'accord, laissez-moi partager mon écran rapidement. J'espère que cela fonctionne. Ça ne marche pas ? D'accord, super, d'accord. Alors, j'espère que vous avez tous passé un excellent début d'automne. Je m'appelle Jane. J'ai obtenu mon diplôme de Columbia ce printemps dernier et je suis maintenant à Princeton pour mes études supérieures. Je suis très heureuse de présenter le travail que j'ai réalisé pendant ma dernière année de premier cycle avec le professeur Chunhua Weng de l'Institut de science des données de l'Université de Columbia. Notre projet porte sur la détection des contradictions dans les études contrôlées randomisées sur la COVID-19 en utilisant des modèles linguistiques de masse tels que BERT.

Slide 3:

Un peu de contexte d'abord. Les résultats contradictoires dans les études cliniques sont un problème de longue date pour les universitaires, les chercheurs et les médecins, surtout dans un domaine caractérisé par des publications en grand volume. Une étude a révélé qu'un tiers des études cliniques originales sont soit remises en question, soit incapables d'être reproduites, et une autre a révélé qu'un quart des essais

contrôlés randomisés, en particulier, sont ouvertement contredits par des découvertes ultérieures. Et c'est un problème qui est devenu particulièrement tangible lors de l'épidémie de COVID-19. Nous avons tous entendu parler de l'hydroxychloroquine, dont les premières études cliniques étaient vraiment optimistes, et plus tard les résultats ont été contredits de manière assez décisive. Ainsi, analyser et interpréter les résultats d'un ensemble volumineux et en constante évolution est un défi très important dans des scénarios sensibles au temps comme la pandémie mondiale. Donc, pour nous, faciliter le processus d'identification des études contradictoires ou concordantes serait vraiment crucial pour les scientifiques qui pourraient vouloir, par exemple, mener des revues systématiques, identifier ce qui pourrait causer des résultats différents entre deux études, évaluer la véracité d'une revendication de recherche et caractériser l'état du consensus ou de la maturité sur une question de recherche particulière.

Slide 4:

Ainsi, nous formulons ce problème comme une tâche standard d'inférence de langage naturel, ou tâche NLI, et la revendication ou l'objectif est de classer une paire de phrases comme contradictoires, impliquantes, ou en accord et neutres, ce qui signifie que les revendications des phrases ne sont pas liées ou ne s'entraînent ni ne se contredisent mutuellement. Donc, l'objectif du modèle linguistique est plus formellement donné une paire de phrases x1 et x2 avec un jeton de classification en masse CLS et une matrice de paramètres. Nous choisissons une étiquette qui maximise la probabilité que l'état final de CLS soit cette étiquette pour ce x spécifique. Et nous choisissons des modèles de langage en masse ici, spécifiquement BERT, car ils ont historiquement eu de très bonnes performances dans les tâches NLI. Nous utilisons des modèles pré-entraînés existants comme modèle de base pour nos projets. L'objectif est d'utiliser l'apprentissage par transfert en adaptant ces modèles à notre tâche spécifique de NLI et nous considérons trois modèles de base. Le premier étant le modèle BERT générique qui est pré-entraîné sur BookCorpus et Wikipedia, puis deux modèles spécifiques au domaine. BioBERT qui est pré-entraîné sur les résumés et articles de PubMed et ClinicalBERT qui est pré-entraîné sur les notes cliniques de MIMIC 3.

Slide 5:

Donc, pour nous, la considération la plus cruciale était la rapidité avec laquelle le modèle s'adapterait à de nouvelles questions de recherche, car en pratique, vous voudriez que le modèle trouve des contradictions dans de nouvelles recherches sur lesquelles il n'aurait peut-être pas été pré-entraîné. Ainsi, à cette fin, nous savions que nous avions besoin d'un ensemble de données avec des domaines de recherche et des questions qui n'avaient jamais été vus auparavant par les modèles de base. Et donc, nous avons créé notre propre ensemble de données. Nous avons annoté manuellement un nouvel ensemble de données en utilisant LitCOVID, une base de données publiquement disponible d'articles COVID-19 sur PubMed. Cela s'explique par le fait que la COVID-19 est très récente et qu'elle n'aurait pas été présente dans les données utilisées pour pré-entraîner les modèles de base. Et conformément à d'autres méthodes d'annotation de rapports analytiques biomédicaux, nous avons identifié 15 questions de recherche distinctes et 103 études qui y répondaient. Nous extrayons manuellement une phrase de chaque résumé qui aborde directement la question de recherche, puis deux annotateurs indépendants ont étiqueté manuellement les paires comme contradictoires, entraînantes ou neutres par rapport à la

question de recherche. Ainsi, toutes les étiquettes qui avaient des conclusions divergentes ont été rejetées.

Slide 6:

Pour construire notre modèle, nous ajoutons des couches de classification non initialisées aux modèles de base et les affinons. Nous conservons les paramètres de la couche de base gelés pour le moment, car nous avons un ensemble d'entraînement relativement petit et nous n'affinons que les couches de classification. Pour notre ensemble d'entraînement, nous utilisons ManConCorpus, un corpus NLI médical annoté manuellement et disponible publiquement, très similaire à ce que nous avons fait, mais plus large, pas seulement lié à la COVID-19.

Slide 7:

Et nous réservons également une petite partie, environ 20 % de l'ensemble de données LitCOVID, pour l'entraînement. Et nous avons été très prudents pour éviter toute contamination entre les données de test et d'entraînement, car le modèle doit généraliser à des questions qu'il n'a jamais vues auparavant. Ainsi, à cette fin, nous avons supprimé toutes les paires qui mélangeaient des phrases de test et d'entraînement. Donc, cela signifie que si une question de recherche apparaissait dans l'ensemble d'entraînement, elle n'apparaîtrait pas dans l'ensemble de test et vice versa. Nous avons tokenisé les paires de phrases, et pour chaque modèle de base, nous avons entraîné deux modèles : un qui ajoutait cette petite partie de données LitCOVID à son ensemble d'entraînement et un qui n'utilisait que ManConCorpus. Et nous faisons cela parce que nous voulons voir dans quelle mesure le modèle s'améliore avec juste une toute petite partie de données spécifiques à la COVID-19 ajoutée.

Slide 8:

Voici donc les résultats de nos mesures de classification de tous les modèles. BioBERT et Clinical BERT avec les données LitCOVID sont les plus performants, ce qui a du sens car les modèles de base sont formés sur un domaine similaire à LitCOVID, et sans surprise, si vous ajoutez des données d'entraînement LitCOVID, cela fonctionne mieux qu'un modèle qui n'en utilise pas. Mais je voudrais noter l'amélioration, une amélioration très drastique avec une toute petite proportion de données d'entraînement COVID. Ainsi, les scores F s'améliorent dans une large mesure. Presque tous doublent, à l'exception de la colonne de précision, qui est assez forte pour tous les modèles.

Slide 9:

Ici, nous montrons le rappel de manière classe par classe et nous observons quelques motifs assez intéressants. De loin, la contradiction est une classe qui se comporte le mieux, même avec des modèles qui n'ont ajouté aucune donnée d'entraînement LitCOVID, et nous émettons l'hypothèse que cela pourrait être dû au fait que des termes de négation comme "no" ou "not" sont universels à travers les sujets et les domaines, donc le modèle peut probablement identifier les négations assez rapidement. Nous avons constaté que les données d'entraînement LitCOVID améliorent principalement les prédictions neutres. Vous pouvez voir que cela double presque le nombre de prédictions neutres correctes, ce qui est probablement dû au fait que maintenant qu'il a des données d'entraînement LitCOVID, il sait quels mots liés à la COVID ne suggèrent pas nécessairement de contradiction ou

d'entraînement. Et l'entraînement est relativement faible dans l'ensemble, sauf pour BioBERT avec LitCOVID, et nous pensons que cela pourrait être dû au fait que les corpus d'entraînement préalable de BioBERT proviennent en réalité de PubMed, il a donc pu apprendre des caractéristiques qui ont contribué à mieux identifier l'affirmation ou la négation textuelle dans un domaine biomédical.

Slide 10:

En résumé, nous avons des preuves solides montrant que les modèles BERT sont une approche valide pour la détection de la conjugaison dans le domaine biomédical. Nous avons trois modèles pré-entraînés qui n'ont besoin que d'une petite quantité de données d'entraînement pour une amélioration drastique des performances. Et juste quelques analyses d'erreurs très brièvement. Quelques motifs communs que nous avons trouvés étaient, comme vous l'avez vu plus tôt, des difficultés à identifier des termes mutuels, puis nous avons constaté une certaine confusion avec des abréviations ou des termes médicaux. Par exemple, HCQ et hydroxychloroquine, le modèle ne sait pas immédiatement qu'ils sont la même chose, il dira donc que c'est neutre ou sans rapport, et ainsi de suite.

Slide 11:

Ainsi, pour l'avenir, certaines questions intéressantes que nous pensons pourraient être abordées sont, par exemple, comment pouvons-nous sélectionner automatiquement la meilleure phrase sans avoir besoin de l'extraire manuellement ? Et il existe déjà certains outils de nomination de texte pour les études cliniques déjà disponibles, tels que Trialstreamer ou le picoparser des laboratoires Weng, donc je serais intéressée de voir comment ils pourraient intégrer cela avec un outil de détection de la conjugaison. Et aussi, nous sommes curieux de savoir si nous pouvons améliorer les performances du modèle en fournissant, vous savez, une liste d'acronymes ou de synonymes fournie par l'utilisateur pour ce domaine que le modèle est susceptible de rencontrer.

C’est tout pour ma présentation d’aujourd’hui. J’aimerai conclure par remercier Pr. Weng et Dr Hao Lui pour leur mentorat et l’aide dans mes recherches. Un grand merci aussi à Stan et Marguerite pour leur aide sur le projet. Merci pour votre attention et bonne semaine à tous.