Offre n°2024-07719

Ingénieur en traitement des langues et développement de Modèles de reconnaissance de la parole

Type de contrat : CDD

Niveau de diplôme exigé : Bac + 5 ou équivalent

Fonction : Ingénieur scientifique contractuel

Contexte et atouts du poste

Motivation et Contexte

À travers le projet COLaF (Corpus et Outils pour les Langues de France), Inria a pour objectif de contribuer au développement de corpus et d'outils libres pour le français et les autres langues de France (alsacien, breton, corse, occitan, etc). La promotion et sauvegarde de ces langues dépend de la disponibilité des technologies linguistiques, mais ces langues sont largement ignorées par les industriels.

La principale difficulté au développement de technologies linguistiques variées est le manque de données. En particulier, les données audio ont besoin d'une transcription pour la plupart des applications. Mais transcrire manuellement des données audio est coûteux en temps, nécessite la participation d'un.e locuteur.trice de bon niveau, et peut résulter en des données inconsistentes en l'absence d'orthographe standard. Afin d'augmenter la quantité de données audio annotées pour diverses langues de France, et de développer la première brique de chaines de traitement variées pour ces langues, nous souhaitons développer une chaine de traitement pour l'entrainement de systèmes de reconnaissance de la parole (ASR, automatic speech recognition).

Les corpus et modèles réalisés seront mis à disposition autant que possible gratuitement, afin de soutenir le développement de la communauté de la technologie linguistique des langues de France.

Mission confiée

Missions :

La mission principale est de créer un système de reconnaissance de la parole (ASR) adapté au contexte des langues peu dotées. Le modèle devra être souple afin de s'adapter à des sources de données variant dans leur quantité et qualité. Il s'agit de types d'enregistrements variés : longues interviews, phrases isolées, émissions de TV, etc.

Le système devra automatiser un maximum d'étapes afin de pouvoir être facilement déployé sur plusieurs langues.

Le système s'appuierai sur le système ASR existant Whisper, qu'il faudra adapter et ré-entrainer sur des données dans les langues que nous souhaitons couvrir. La mission comprend aussi l'amélioration du système en y ajoutant d'autres outils, phases de prétraitement, systèmes pour améliorer les données d'entrainement, etc.

La mission inclue d'interagir avec des chercheurs experts dans les langues concernées pour évaluer et améliorer le système de façon itérative.

Nos langues de travail incluent les langues régionales (picard, occitan, provençal, etc) ainsi que les créoles à base française et les langues des territoires d'outre-mer (kanak, etc). Un intérêt particulier peut être accordé à une langue selon les connaissances et intérêts du ou de la candidat.e.

Principales activités

- Collecte et prétraitement de données en langues de France, via collaboration avec fournisseurs de données institutionnels et associatifs
- Développement d'une pipeline d'entrainement de système ASR, du nettoyage des données audio jusqu'à la publication du modèle
- Adaptation du système à différents scénarios et types de données
- Analyse des résultats et adaptation du modèle de façon itérative
- Préparation, documentation et publication de jeux de données et de modèles
- Promotion des modèles publiés afin de lancer une communauté autour du traitement des langues concernées

Compétences

- Master en machine learning, informatique, linguistique informatique ou équivalent
- Solides connaissances en machine learning
- Expérience en ASR ou traitement du signal appréciée
- Connaissance d'une langue de France appréciée mais pas nécessaire, tel que corse, occitan, picard, ou créole à base française
- Capacité à travailler en équipe
- Capacité à suggérer des solutions originales

Avantages

- Restauration subventionnée
- Transports publics remboursés partiellement
- Congés: 7 semaines de congés annuels + 10 jours de RTT (base temps plein) + possibilité d'autorisations d'absence exceptionnelle (ex : enfants malades, déménagement)
- Possibilité de télétravail (après 6 mois d'ancienneté) et aménagement du temps de travail
- Équipements professionnels à disposition (visioconférence, prêts de matériels informatiques, etc.)
- Prestations sociales, culturelles et sportives (Association de gestion des œuvres sociales d'Inria)
- Accès à la formation professionnelle
- Sécurité sociale

Rémunération

A partir de 2692 € brut/mois selon diplômes et expériences

Informations générales

- **Thème/Domaine :** Langue, parole et audio
  Ingénierie logicielle (BAP E)
- **Ville :** Villers lès Nancy
- **Centre Inria :** Centre Inria de l'Université de Lorraine
- **Date de prise de fonction souhaitée :** 2024-06-24
- **Durée de contrat :** 3 ans
- **Date limite pour postuler :** 2024-07-03

Contacts

- **Équipe Inria :** MULTISPEECH
- **Recruteur :**
  Ouni Slim / slim.ouni@loria.fr

A propos d'Inria

Inria est l'institut national de recherche dédié aux sciences et technologies du numérique. Il emploie 2600 personnes. Ses 215 équipes-projets agiles, en général communes avec des partenaires académiques, impliquent plus de 3900 scientifiques pour relever les défis du numérique, souvent à l'interface d'autres disciplines. L'institut fait appel à de nombreux talents dans plus d'une quarantaine de métiers différents. 900 personnels d'appui à la recherche et à l'innovation contribuent à faire émerger et grandir des projets scientifiques ou entrepreneuriaux qui impactent le monde. Inria travaille avec de nombreuses entreprises et a accompagné la création de plus de 200 start-up. L'institut s'efforce ainsi de répondre aux enjeux de la transformation numérique de la science, de la société et de l'économie.

Attention: Les candidatures doivent être déposées en ligne sur le site Inria. Le traitement des candidatures adressées par d'autres canaux n'est pas garanti.

Consignes pour postuler

Sécurité défense :
Ce poste est susceptible d'être affecté dans une zone à régime restrictif (ZRR), telle que définie dans le décret n°2011-1425 relatif à la protection du potentiel scientifique et technique de la nation (PPST). L'autorisation d'accès à une zone est délivrée par le chef d'établissement, après avis ministériel favorable, tel que défini dans l'arrêté du 03 juillet 2012, relatif à la PPST. Un avis ministériel défavorable pour un poste affecté dans une ZRR aurait pour conséquence l'annulation du recrutement.

Politique de recrutement :
Dans le cadre de sa politique diversité, tous les postes Inria sont accessibles aux personnes en situation de handicap.