

Xiaoou Wang

6 Rue Colonna D'Istria, 06300 Nice

0652724252 | xiaoou.wang@inria.fr | aiprototypes.com/ | Google Scholar | linkedin.com/in/xiaoou-wang/

Résumé

Diplômé en linguistique et en traitement automatique des données textuelles, programmeur chevronné, je suis passionné par tout ce qui ressort à l'apprentissage automatique. Créateur du site [AI Prototypes](#), j'ai un état d'esprit orienté solutions mais j'ai aussi su garder une fibre académique dans ma façon d'aborder les problématiques. J'envisage toujours les problèmes et les défis avec une perspective positive et surtout, j'adore travailler avec différents acteurs passionnés par l'intelligence artificielle.

Formation

Université Paris Nanterre

Nanterre, France

M2 en traitement automatique des langues, mention Très Bien

2020 - 2021

- Formation en gestion de projet, apprentissage automatique, développement Web et base de données, etc.

Université Sorbonne-Nouvelle

Paris, France

M1 en traitement automatique des langues, mention Très Bien

2019 - 2020

- Formation en gestion informatique, programmation orientée objet, fouille de textes, algorithme et enrichissement de corpus, etc.

Université d'Aix-Marseille

Aix-en-Provence, France

Master en linguistique expérimentale

2012 - 2013

- Mémoire : Design d'une interface basée sur un clone orofacial pour l'acquisition phonétique des langues

Imperial College London

Coursera

Mathematics for Machine Learning

2020

- Formation en algèbre linéaire, analyse multivariée et analyse en composantes principales.

Expériences professionnelles

CNRS&Inria

Sophia Antipolis, France

Ingénieur dans le cadre du projet ANR PRCE ATTENTION pour la lutte contre la désinformation

2022 - 2023

- Mener une veille informatique dans le domaine de la génération automatique de textes pour des agents conversationnels en débat.
- Encadrer une stagiaire pour l'augmentation d'un corpus de fake news avec [Inception](#).
- Développer un module basé sur l'apprentissage joint pour l'identification des arguments en BIO et la prédiction des relations argumentatives dans un corpus de textes médicaux.
- **Résultats** : 1 livrable sur l'état de l'art de la génération d'arguments, 2 papiers de recherche.
- **Soft Skills** : Réunions avec un comité composé de juristes, de start-ups et d'ethnologues pour le suivi du projet.

APIL les French tech du NLP, projet soutenu par le ministère de la culture

A distance, France

Rédacteur technique de 6 études de cas, parfois sur la base d'interviews, EC choisies ci-dessous :

2022 - 2023

- [Exploiter les textes biomédicaux de manière semi-automatisée](#)
- [Au-delà de la simple dichotomie positif-négatif, découvrez l'analyse de sentiments à base d'aspects](#)
- [Extraire des réponses à des questions de prospects à partir des avis clients](#)
- [Réaliser des POC avec une quantité de données limitée ? Prioriser la qualité des données et connaître 3 solutions courantes](#)

Université Catholique de Louvain, France Education International

Louvain-la-Neuve, Belgique

Chercheur Junior

2021 - 2022

- Mettre en place la prédiction automatique des scores des examens internationaux de français.
- Co-rédiger la [documentation de FABRA](#), toolkit destiné à évaluer la lisibilité d'un texte en français.
- Entraîner des modèles LSA et des modèles n-gramme et intégrer ces derniers sous forme de services dans le toolkit [FABRA](#) avec FLASK.
- Piloter l'annotation d'un grand corpus des productions d'apprenants avec [Label Studio](#).
- Créer un module de correction automatique pour le français basé sur l'architecture Transformer.
- Développer un système de classification automatique d'erreurs à base de features morphosyntaxiques.
- **Résultats** : 3 papiers de recherche, création d'un module de l'analyse linguistique basée sur les erreurs, mise en ligne d'une démo pour le toolkit d'analyse de complexité.
- **Soft Skills** : Encadrement d'une stagiaire pour l'annotation d'un grand corpus, réunion hebdomadaire avec France Education International pour comprendre les données, leurs besoins et élaborer des timesheets.

Lattice, lettre de recommandation

Montrouge, France

Stage financé par le LABEX EFL (10 étudiants parmi 7 universités)

01/02/2021 - 31/07/2021

- Obtenir une nouvelle performance sur le grand défi d'Intelligence Artificielle Winograd.
- Créer un nouveau jeu de données en français.
- Mener des expériences de fine-tuning sur les nouvelles données.
- **Résultats** : Performance SOTA publiée dans un papier de recherche, création d'un nouveau jeu de données.
- **Soft Skills** : Animation des réunions hebdomadaires avec 3 chercheurs associés pour le suivi du stage.

Laboratoire de Sciences Cognitives et Psycholinguistique, ENS

Paris, France

Vacation de recherche, contrat accessible publiquement

01/11/2014 - 30/04/2015

- Ecrire des scripts Praat en vue de l'annotation phonétique et l'alignement automatique du corpus Buckeye (307,000 mots, 40 locuteurs)

GIPSA-Lab, Département Parole&Cognition, lettre de recommandation

Grenoble, France

Stage

01/04/2013 - 30/09/2013

- Mise en place d'un entraînement phonétique à l'aide d'un clone oro-facial

iFLYTEK

Anhui, Chine

Responsable de l'équipe d'annotation

30/06/2015 - 30/04/2018

- Coordination des stratégies d'annotation pour de grands corpus.
- Documentation de projet.
- **Résultats** : Création de plusieurs corpus propriétaires pour l'entraînement de modèles de reconnaissance vocale.
- **Soft Skills** : Médiateur entre annotateurs, diverses équipes MOE et MOA, calibrage des objectifs et des étapes, analyse du besoin métier.

Compétition

Responsable d'une équipe obtenant la deuxième place au Défi Fouille de Textes 2021

Lille, France

28e Conférence sur le Traitement Automatique des Langues Naturelles

Juin 2021

- Comparer des mesures de similarité textuelle pour l'évaluation automatique de copies d'étudiants.
- **Résultats** : 2ème place en termes de précision et publication d'un papier de recherche.
- **Open source** : Publication de French NLP, State of the art toolkit for Natural Language Processing in French sur PyPI.
- **Soft Skills** : Animation des réunions sur l'identification des approches possibles et les performances de chaque approche, coordination des tâches et management des deadlines.

Enseignements

Traduction automatique par règles et par réseaux de neurones, intervention à l'Université Sorbonne Nouvelle

Correction automatique neuronale s'inspirant de la traduction automatique, intervention à l'UCLouvain

Compétences

Programmation Python, R, HTML/CSS, Php, JavaScript, Praat, Matlab, SQL.

Frameworks Pandas, Numpy, Streamlit, Scikit-Learn, Pytorch, FairSeq, Transformers, Beautiful Soup, Flask.

Esprit critique Reviewers pour de nombreuses conférences scientifiques (LREC, ACL, etc.)

Divers Shell (Bash/Zsh), Git, LaTeX (Overleaf/R Markdown), Documentation avec Sphinx.

Soft Skills Gestion de projet, Teamwork, Problem-solving, Documentation.

Langues Très haut niveau en français et anglais, natif en chinois.

Projets choisis

Web Scraping, [Complete tutorial on scraping French news from Le Monde](#)

Programmation orientée objet,

[Understand objected-oriented programming \(OOP\) by building a minimal Web Scraping framework](#)

Analyse multivariée en R, [Analyse en composantes principales \(PCA\) : prépositions d'inclusion en français](#)

Annotation de Tweets en registres en Php,

[Déploiement d'une interface d'annotation \(pour tester utilisez les valeurs par défaut\)](#)

Parsing, [Comparer Spacy, StanfordNLP, TreeTagger sur un corpus formel et un corpus oral](#)

Classification de textes,

[Classification de textes : du TF-IDF aux word embeddings en passant par des features expertes](#)

Normalisation de textes, [How to build a spell checker with deep learning](#)

Gestion de documentation, [Documentation en Sphinx du projet Winograd au Lattice](#)

Gestion de documentation, [Fabra, un toolkit en complexité lexicale](#)

Gestion de projet, [Rédaction de cahier des charges fonctionnel](#)

Gestion de projet, [Gestion de projet avec la méthode agile](#)

Déploiement, [Déployer une app Django sur Digitalocean](#)

Bash, [Représentation de la guerre commerciale sur Internet, projet réalisé avec Rmarkdown et Bash](#)

Centres d'intérêt

Cuisine Expert en cuisine chinoise de diverses écoles, j'adore les moments de convivialité culinaire.

Sport et nature Grand amateur de natation et de randonnées, j'ai également un bon niveau en ski, snowboard et badminton, mes 3 sports favoris.

Pour les papiers scientifiques, voir mon profil [Google Scholar](#)