

Pour des IA Francophones

Open Lab n°3 - 26 novembre 2020



Piaf

etalab^{gouv.fr}

Programme

9h30 - Restitution : le modèle et l'application

- Entraînement d'un modèle open-source : à tester pour vos réutilisations
- Création de l'application: PiafAPI, PiafBot & PiafAgent
- Prochaines étapes, et ambition

10h15 - Échanges

11h - Ateliers

- Identifier les cas d'usage
- Vos documents dans notre application: let's do it !

12h - Conclusion



Rendre accessible aux administrations les IA de question-réponse

Car un jeu de données d'entraînement ne suffit pas

Qu'est ce qu'un modèle de Question-réponse ?

Wikipédia est une encyclopédie universelle et multilingue, créée par Jimmy Wales et Larry Sanger le 15 janvier 2001. Il s'agit d'une œuvre libre, c'est-à-dire que chacun est libre de la rediffuser.

Gérée en wiki dans le site web wikipedia.org grâce au logiciel MediaWiki, elle permet à tous les internautes d'écrire et de modifier des articles. Elle est devenue en quelques années l'encyclopédie la plus fournie et la plus consultée au monde.



Quand ?



Introduction – Situer Piaf dans Etalab



Accompagnement de projets d'IA publics via des appels à manifestation d'intérêts



Des outils mutualisés, type plateforme d'annotation, outils, de l'open data
- PIAF
- pseudonymisation



Une communauté : data scientists, chercheurs...
& des guides



Introduction – L'origine du projet

CÉDRIC VILLANI

Mathématicien et député de l'Essonne

DONNER UN SENS À L'INTELLIGENCE ARTIFICIELLE

POUR UNE STRATÉGIE
NATIONALE ET EUROPÉENNE

D'abord, une politique offensive visant à favoriser l'accès aux données, la circulation de celles-ci et leur partage. Les données sont la matière première de l'IA contemporaine et d'elles dépend l'émergence de nombreux usages et applications. Il est tout d'abord urgent d'accélérer et d'étoffer la politique d'ouverture des données publiques (*open data*), en particulier s'agissant des données critiques pour les applications en IA. La démarche d'*open data* fait l'objet d'une politique volontariste depuis plusieurs années, notamment sous l'impulsion de la loi pour une République numérique³ : cet effort, important, doit être soutenu. La puissance publique doit par ailleurs amorcer de nouveaux modes de production, de collaboration et de gouvernance sur les données, par la constitution de « *communs de la donnée* »⁴. Il lui revient ainsi d'inciter les acteurs économiques au partage et à la mutualisation de données voire



Introduction – Les premières réalisations

2019
Novembre
OpenLab 1

Lancement de la campagne d'annotation – *PiafAnno*



2020
Février
OpenLab 2

Publication des annotations sur data.gouv.fr



Mai
LREC

Publication d'un article de recherche dans la conference LREC

Natif > Multilingue

Introduction – PiafAnno

Question 1 / 5

Union européenne 2 / 5

Toutefois, au début des années 1990, la Commission européenne propose dans ses rapports « Europe 2000 » et « Europe 2000+ », une régionalisation relative aux dynamiques transnationales et rapprochements transfrontaliers au sein des États membres. Huit ensembles se détachent alors : l'aire des capitales, l'Arc atlantique, l'Arc méditerranéen, la diagonale continentale, la mer du Nord, les nouveaux Länder allemands et les régions ultrapériphériques. Cependant, compte tenu des élargissements de 1995 et 2004, cette régionalisation nécessite une actualisation en y ajoutant notamment l'espace Baltique et en considérant l'Europe centrale et orientale.

Cliquer sur la réponse dans le texte

Allez-y : posez ici une question en utilisant vos propres mots ! (La réponse doit être dans le texte)

Comment la commission de l'UE appelle-t-elle ses rapports à la fin du 20ème siècle ?

← VALIDER

Introduction – Contribution & Communauté

+9500 contributions



+700 contributeurs



Introduction – LREC

Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 5481–5490

Marseille, 11–16 May 2020

© European Language Resources Association (ELRA), licensed under CC-BY-NC

Project PIAF: Building a Native French Question-Answering Dataset

Rachel Keraron^{≡,*}, Guillaume Lancrenon^{≡,†}, Mathilde Bras^{≡,†},
Frédéric Allary^{*}, Gilles Moyses^{*}, Thomas Scialom^{*◊},
Edmundo-Pavel Soriano-Morales[‡], Jacopo Staiano^{*}

≡ equal contribution

** reciTAL, Paris (France)*

†Etalab, DINUM, Prime Minister's Office, Paris (France)

◊ Sorbonne Université, CNRS, LIP6, F-75005 Paris, France

{rachel, frederic, gilles, thomas, jacopo}@recital.ai

{guillaume.lancrenon, mathilde.bras, pavel.soriano}@data.gouv.fr

Abstract

Motivated by the lack of data for non-English languages, in particular for the evaluation of downstream tasks such as Question Answering, we present a participatory effort to collect a native French Question Answering Dataset. Furthermore, we describe and publicly release the annotation tool developed for our collection effort, along with the data obtained and preliminary baselines.

Keywords: Question Answering, Annotation, Crowdsourcing

1. Introduction

Along with the availability of massive amounts of data, the increase in computational power has in recent years allowed the development of Deep Learning techniques, leading to significant advancements in the fields of Computer Vision (CV), and Natural Language Processing (NLP), among others. Visual information can, to some extent, be

considered to generalize across cultures in many real-world baselines, and provide details on the implementation of the open source annotation tool we developed. Such tool allows volunteers to participate in crowdsourced QA dataset collection campaigns.

In summary, we make the following contributions:

1. we develop and release a novel annotation tool to collect large-scale QA data in a participatory scenario;

Le pivot : du jeu de données aux usages



Un constat

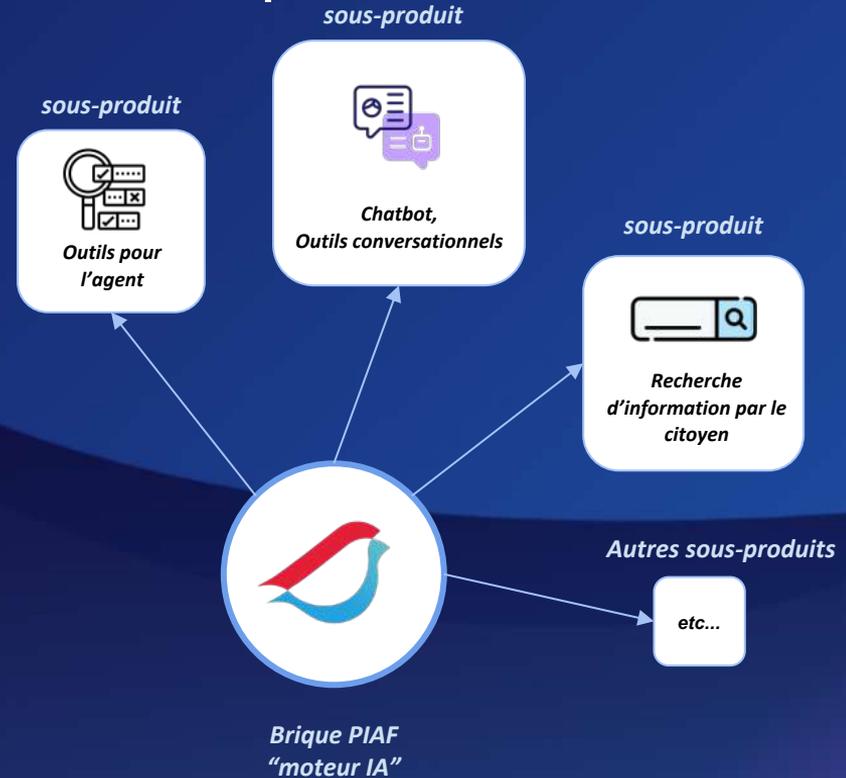
- Service-public.fr
- Le code du travail
- La cours de cassation
- sante.fr
- CNIL
- Data.gouv.fr
- ...

Une brique technique, des produits adaptés aux besoins

PIAF est la brique technique centrale, que des **sous produits** viennent interroger.

Les sous-produits issus de PIAF vont répondre à des besoins concrets :

- **Pour l'agent** : améliorer l'outil de l'agent par une recherche d'information plus intelligente, aide à la réponse lors de question du citoyen ou de l'utilisateur.
- **Pour le citoyen** : accéder plus rapidement et plus efficacement à l'information et à une réponse ciblée.



Cas d'usage n°1 : service-public.fr

Comprendre les utilisateurs et leurs besoins



Agent
service-public.fr



Citoyens

Quels problèmes rencontrent les agents ?

Types de demandes traitées ?

Niveau de complexité ?

Sujets récurrents ?

Besoins des agents ?

Outils utilisés

Liste des réponses types

Prestataires qui sous-traitent

Quels problèmes rencontrent les utilisateurs du site ?

Quels retours ?

Comment arrivent-ils sur le site ?

Délai de réponses ?

etc ...

Cas d'usage n°1 : service-public.fr

Comprendre les utilisateurs et leurs besoins



Agent
service-public.fr



Citoyens



*Aider à répondre aux
mails reçus ?*

Un outil pour aider à chercher
dans les textes d'une **base de
connaissance**.



Chatbot ?

Un outil directement accessible sur le
site.

Travaux actuels : modèle et application



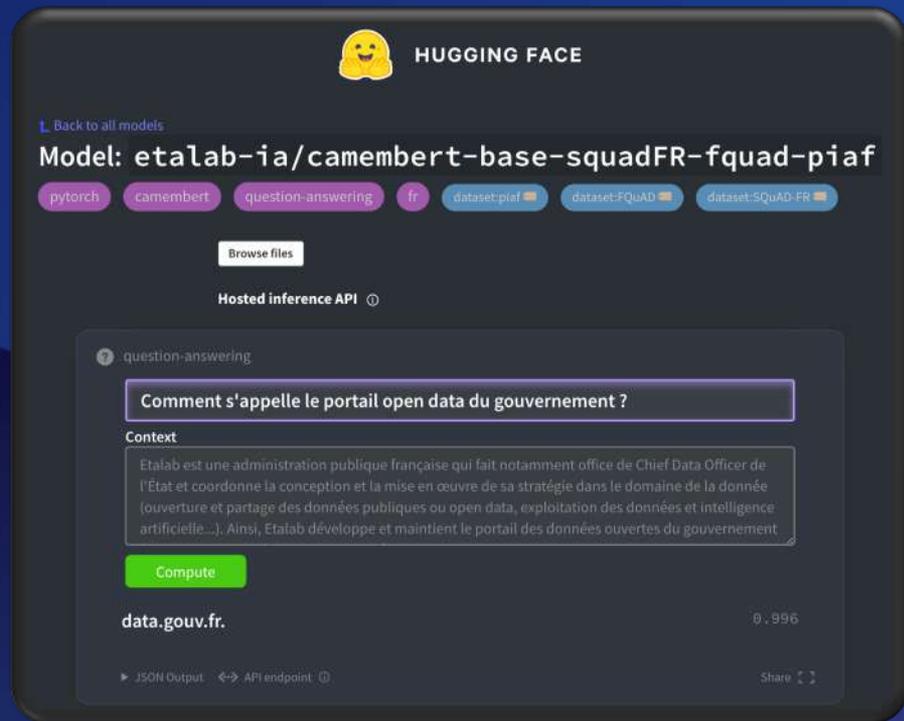
Modèle de Question-Réponse



+

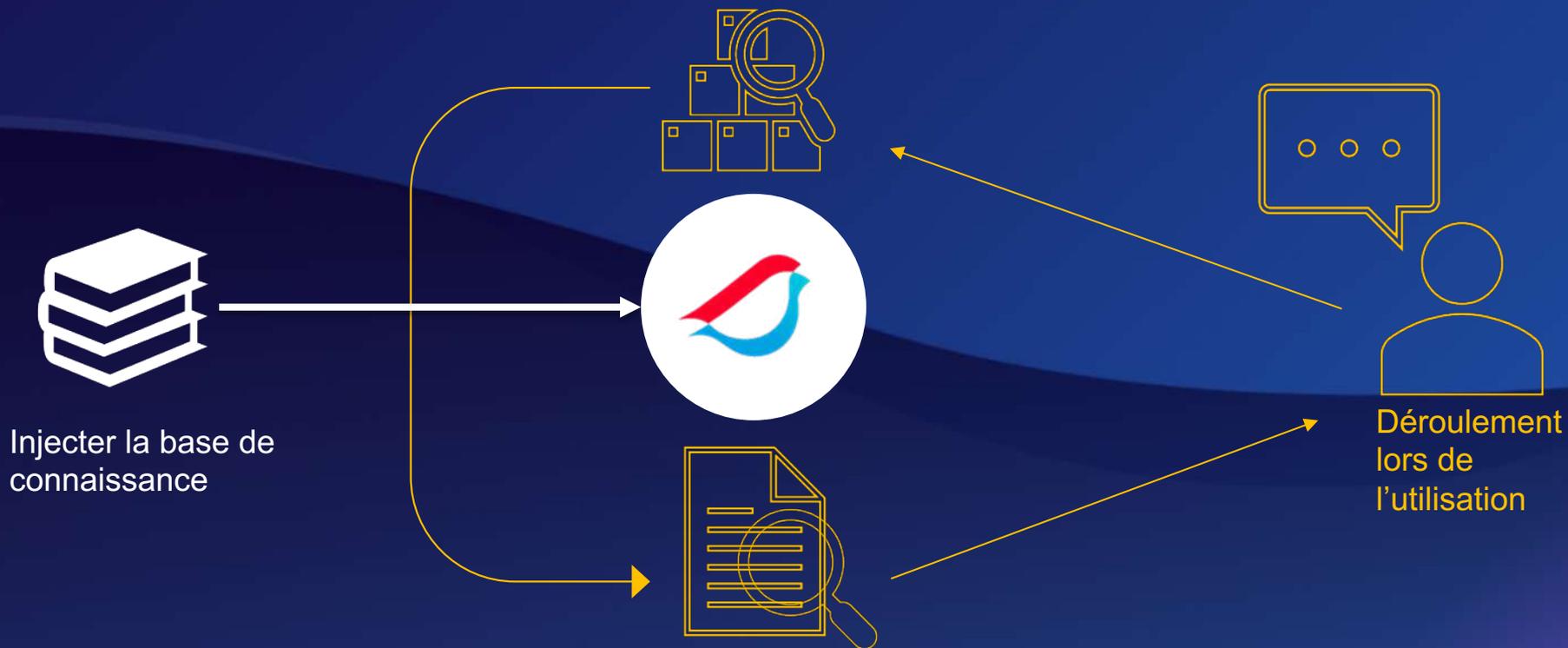


=



The screenshot shows the Hugging Face interface for the model `etalab-ia/camembert-base-squadFR-fquad-piaf`. The model is categorized as `pytorch`, `camembert`, `question-answering`, `fr`, `dataset:piaf`, `dataset:FQuAD`, and `dataset:SQuAD-FR`. It has a "Browse files" button and a "Hosted inference API" link. The interface shows a question: "Comment s'appelle le portail open data du gouvernement ?". The context provided is: "Etalab est une administration publique française qui fait notamment office de Chief Data Officer de l'État et coordonne la conception et la mise en œuvre de sa stratégie dans le domaine de la donnée (ouverture et partage des données publiques ou open data, exploitation des données et intelligence artificielle...). Ainsi, Etalab développe et maintient le portail des données ouvertes du gouvernement". A green "Compute" button is visible. The answer is `data.gouv.fr.` with a score of `0.996`. There are also links for "JSON Output", "API endpoint", and "Share".

L'application - PiafAPI



L'application - PiafBot

The screenshot shows the Service-Public.fr website interface. At the top, there are navigation tabs for 'PARTICULIERS', 'PROFESSIONNELS', 'ASSOCIATIONS', and 'ANNUAIRE DE L'ADMINISTRATION'. The main header includes the French Republic logo and the text 'Service-Public.fr Le site officiel de l'administration française'. A search bar contains the text 'Exemple : Passeport, maire de Montreuil, acte de naissance...'. Below this is a grid of service categories: 'PAPIERS - CITOYENNETÉ', 'FAMILLE', 'SOCIAL - SANTÉ', and 'TRAVAIL'. A chatbot window titled 'Service Public' is overlaid on the right side, displaying a welcome message and instructions for using the chatbot.

Navigation tabs: PARTICULIERS, PROFESSIONNELS, ASSOCIATIONS, ANNUAIRE DE L'ADMINISTRATION

Se connecter

RÉPUBLIQUE FRANÇAISE
Liberté
Égalité
Fraternité

Service-Public.fr
Le site officiel de l'administration française

Une question ? Services en ligne et formulaires

Épidémie Coronavirus (Covid-19), tout ce qu'il faut savoir : [lire l'actualité](#)

Papiers - Citoyenneté Famille Social - Santé Travail Logement Transports Argent Justice

Exemple : Passeport, maire de Montreuil, acte de naissance...

PAPIERS - CITOYENNETÉ
État-civil, Passeport, Élections, Papiers à conserver, Carte d'identité...

FAMILLE
Allocations familiales, Naissance, Mariage, Pacs, Scolarité...

SOCIAL - SANTÉ
Carte vitale, Chômage, Handicap, BSA, Personnes âgées...

TRAVAIL
CDD, Concours, Retraite, Démission, Période d'essai...

Service Public

Bonjour, je suis le prototype de chatbot de service-public.fr. Toutes mes réponses sont automatisées, j'apprends actuellement à trouver la meilleure réponse à votre question.

Vous pouvez poser directement votre question, ou choisir un thème.

POSER UNE QUESTION

CHOISIR UN THÈME

-> Sélectionner le sujet de votre question dans la liste ci-dessous

Papiers - Citoyen... Famille Social

👉 Cher à tester se : le scénario continue dans la section "Famille"

Type here...

L'application - PiafAgent



PIAFAgent & service-public.fr

Indice de confiance : 84 %

Congé maternité **d'une salariée du secteur privé** : Vous bénéficiez d'un congé de maternité durant la période qui se situe autour de la date présumée de votre accouchement. Sa durée est variable, en fonction du nombre d'enfants à naître ou déjà à charge. Il comporte une période de congé prénatal et un congé postnatal. Vous bénéficiez d'une indemnisation versée par la Sécurité sociale. Qui est concerné ? : Vous bénéficiez automatiquement d'un congé de maternité, en partie avant votre accouchement et en partie après. Le congé de maternit ...

[✓ Lien vers la fiche](#)

Indice de confiance : 64 %

Congé maternité **d'une salariée du secteur privé** : Vous bénéficiez d'un congé de maternité durant la période qui se situe autour de la date présumée de votre accouchement. Sa durée est variable, en fonction du nombre d'enfants à naître ou déjà à charge. Il comporte une période de congé prénatal et un congé postnatal. Vous bénéficiez d'une indemnisation versée par la Sécurité sociale. Qui est concerné ? : Vous bénéficiez automatiquement d'un congé de maternité, en partie avant votre accouchement et en partie après. Le congé de maternit ...

[✓ Lien vers la fiche](#)

[✓ Lien vers la fiche](#)

Précision et limites

Avec l'apparition des filtres sur les thèmes de service-public, nous atteignons un score d'environ 60% de paragraphes bien retrouvés (parmi environ 23.000) pour 400 Questions posés par les citoyens.

Il faut ajouter à cela les 50 à 70% de performances du modèle de QA sur ce cas pratique.

Précision du reader

50% à 70%

Précision du retriever



Piaf API peut :
retrouver des paragraphes avec beaucoup de justesse, et même avec des synonymes



Piaf API ne peut pas trouver les réponses précises

- si la question est trop longue (une phrase ou deux)
- trop compliqué (pas de raisonnement)
- si elle n'existe pas dans la base de connaissance

Prochaines étapes



Amélioration des performances



Mise en production



Lancement d'une offre SAAS

Piaf en 2021: donnez-nous vos données, nous faisons le reste

Questions-Réponses

ATELIERS

(proposez le votre)

Quels cas d'usage pour les algorithmes de questions-réponses ?

Ophélie & Robin

Quelles pistes d'améliorations techniques ?

Pavel et Guillaume

Comment mieux encourager l'IA francophone ouverte ?

Paul-Antoine & Rose

Merci !

piaf@data.gouv.fr

