Pour des IA Francophones

Open Lab n°3 - 26 novembre 2020

Liberté • Égalité • Fraternité
RÉPUBLIQUE FRANÇAISE
Premier ministre
Direction interministérielle du numérique

Piaf

etalab gouv.fr

Programme

9h30 - Restitution : le modèle et l'application

- Entrainement 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 !

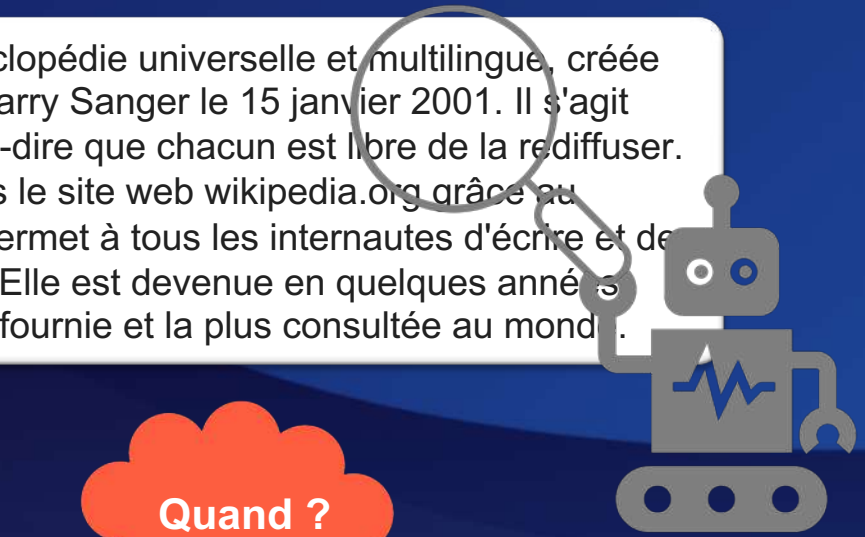
12h - Conclusion

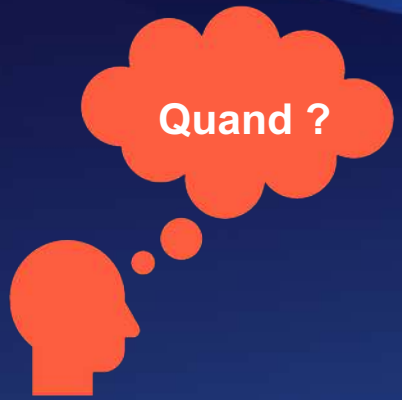
Rendre accessible aux administrations les **IA de question-réponse**

Car un jeu de données d'entrainement 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.





Introduction – Situer Piaf dans Etalab

Lab-IA

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

Piaf

https://www.etalab.gouv.fr/datasciences-et-intelligence-artificielle

Introduction – L’origine du projet

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 numérique[3] : 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

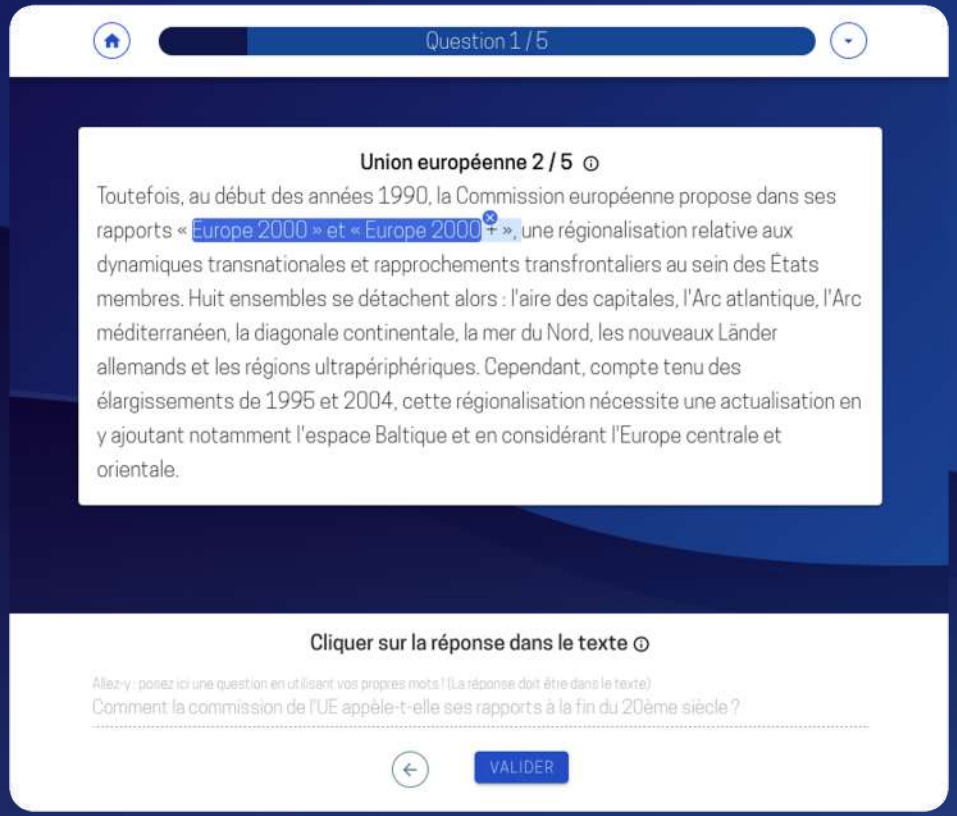
Publication des annotations sur data.gouv.fr

Mai
LREC

Publication d'un article de recherche dans la conference LREC

Natif > Multilingue

Introduction – PiafAnno

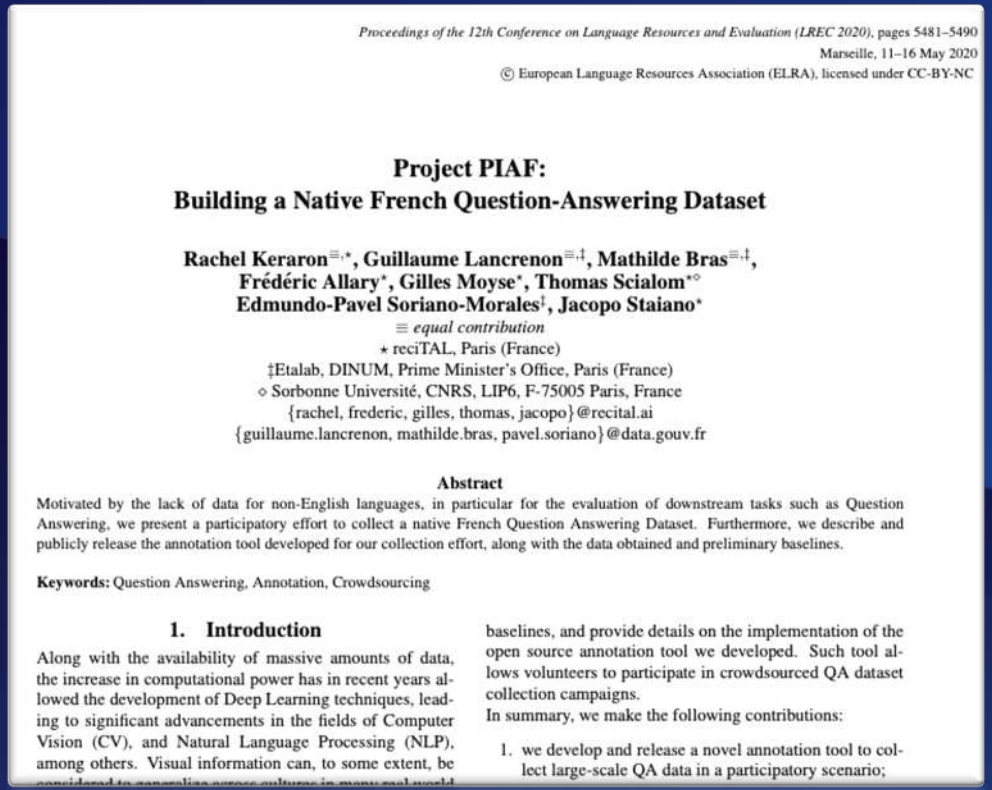


Introduction – Contribution & Communauté

+700 contributeurs

+9500 contributions

Introduction – LREC



Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 5481–5490
Marseille, 11–16 May 2020

Project PIAF:
Building a Native French Question-Answering Dataset

Rachel Keraron[≡,⋆], Guillaume Lancrenon[≡,‡], Mathilde Bras[≡,‡],
Frédéric Allary[⋆], Gilles Moyse[⋆], 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
- …

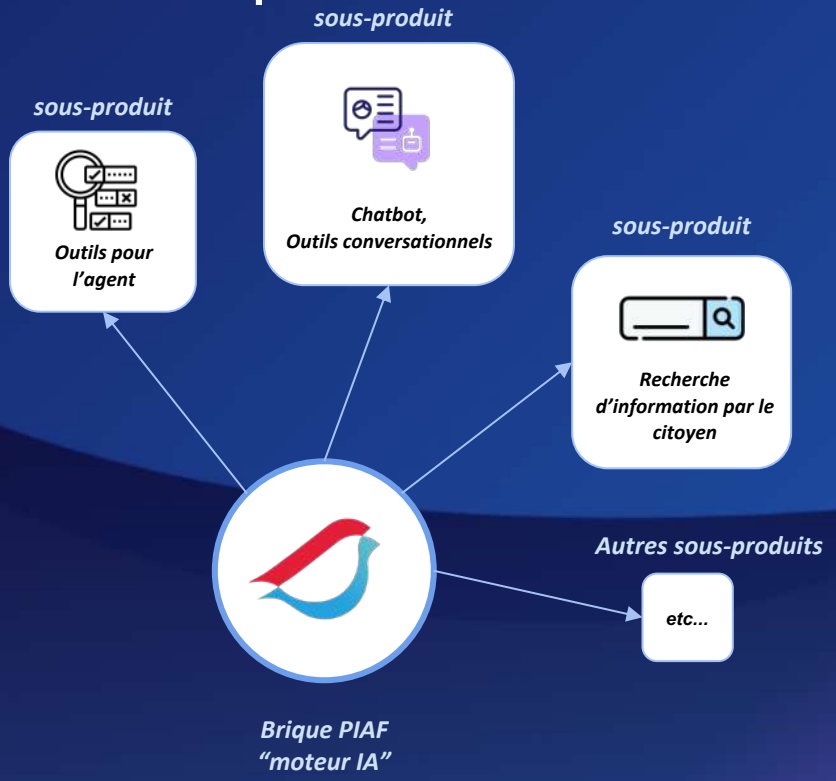
Une brique technique, des produits adaptés aux besoins

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

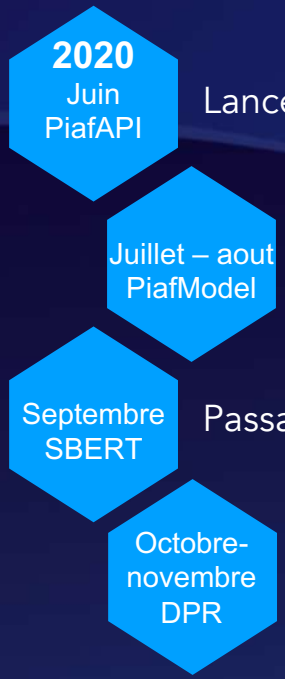
Aider à répondre aux mails reçus ?

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



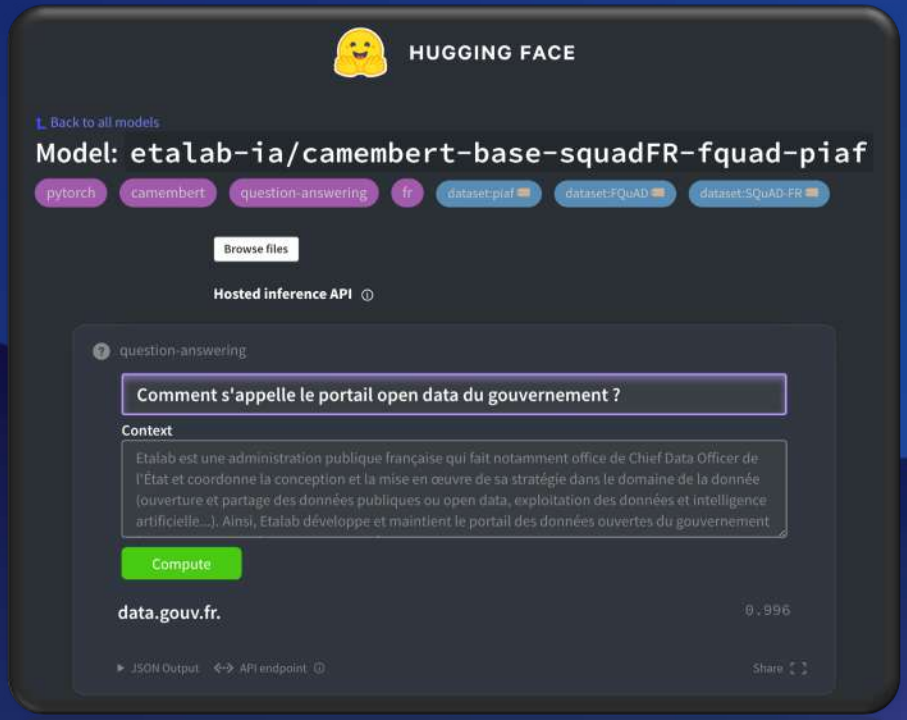
Lancement de cette seconde phase – ***PiafAPI***

Entrainement d'un **modèle de Question-réponse**

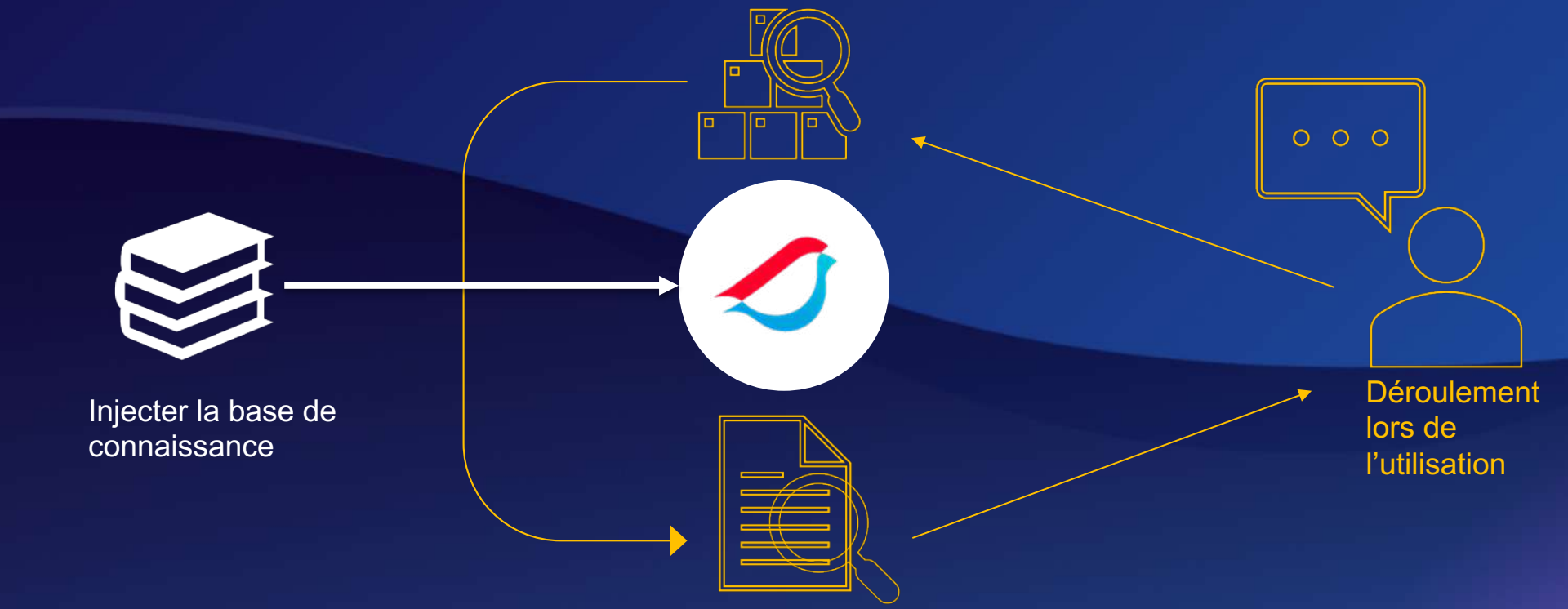
Passage à SBERT & sortie de ***PiafAgent***

Passage au DPR & ***PiafBot***

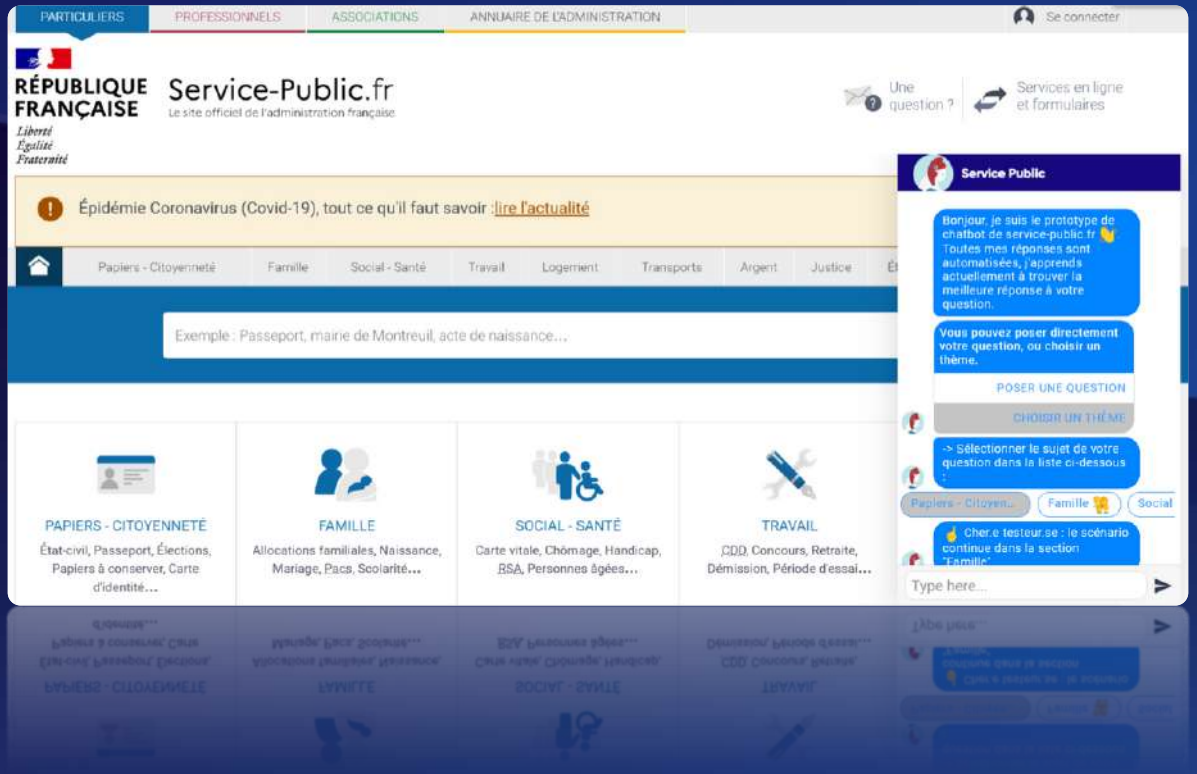
Modèle de Question-Réponse



L'application - PiafAPI



L'application - PiafBot



L'application - PiafAgent

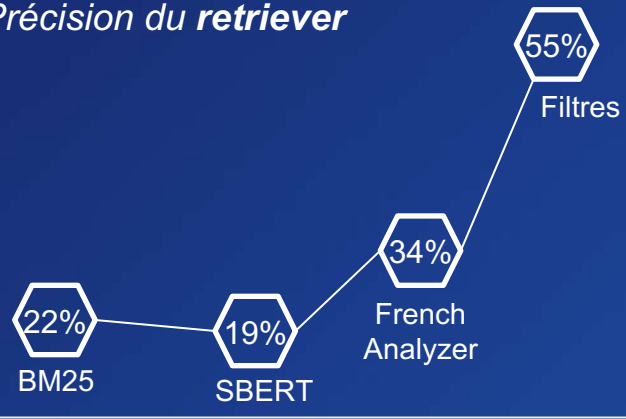
Précision et limites

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