- Marko Turpeinen (1001 Lakes, Finland)
- Walter Daelemans (University of Antwerp, NCC Belgium)
- Svetla Koeva (Bulgarian Academy of Sciences, NCC Bulgaria)
- Maciej Ogrodniczuk (Polish Academy of Sciences, NCC Poland)
- Marta Villegas (Barcelona Supercomputing Center, NCC Spain)
- **François Yvon (LIMSI/CNRS, NCC France)**

EUROPEAN LANGUAGE GRID

META NET
META FORUM 2020

Recent evolution of the French landscape

François Yvon (LIMSI, CNRS)
francois.yvon@limsi.fr

01/02/03-12-2020 META-FORUM 2020 – Piloting the European Language Grid (virtual conference)
http://www.european-language-grid.eu

Background: the French roadmap on AI

- Evolutions in the academic sector
- Etalab / DataHubs
- Opening government / administrative data
- Recent initiatives in the private sector

- National computing infrastructure (2020-)
- National Resarch programs
  - National excellence clusters, chair program in AI, joint AI calls, etc.
- Open science, open data policy

The French academic landscape: a welcome clari(n)fication

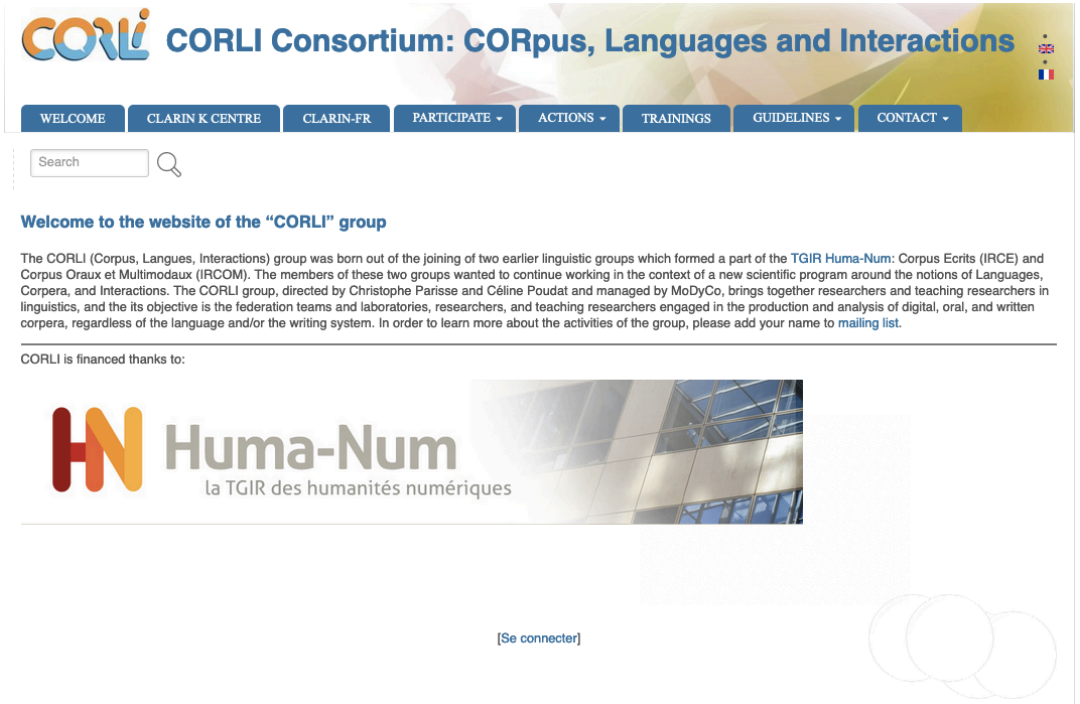
Funding, strategic vision: HumaNum

Scientific planning & coordination: CORLI

Resources & softwares :

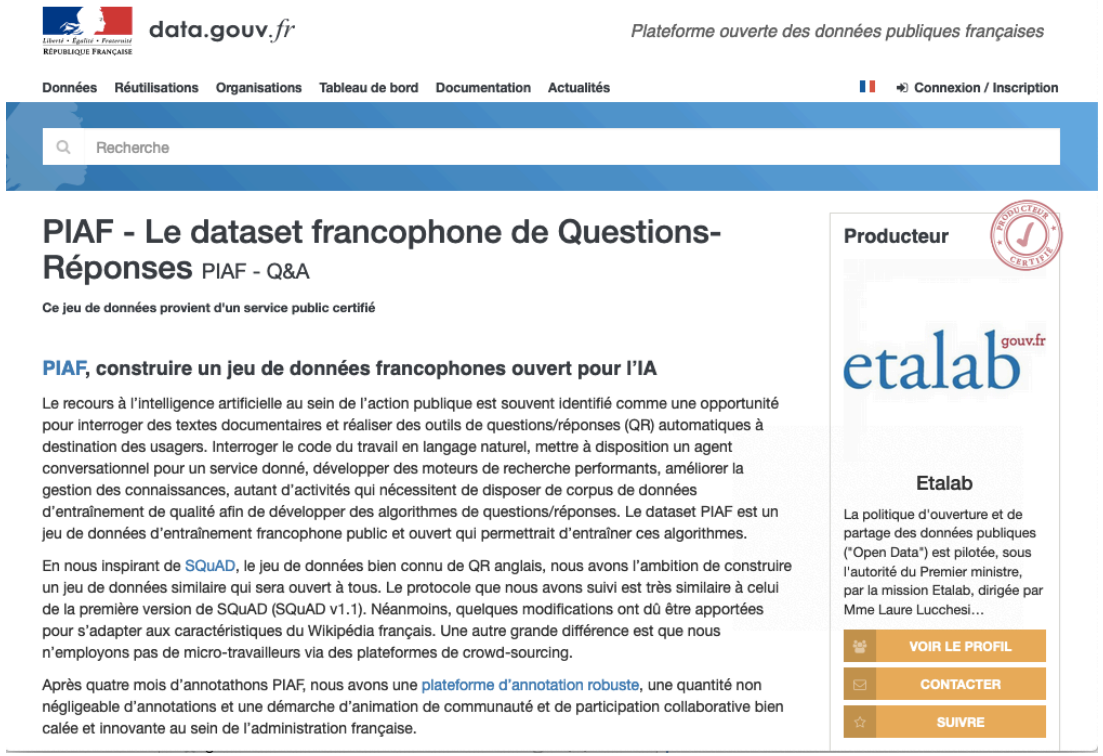
CoCoon, Ortholang, Pangloss, etc

Valorisation : ELDA/ELRA + LREC + LRE



Emerging initiatives from the French state

« Pour permettre que soient exploitées au mieux les données, nous allons méthodiquement mais résolument procéder à une ouverture proactive de nos données. Cela passe d'abord et avant tout par l'ouverture des données publiques. »



data.gouv.fr

Plateforme ouverte des données publiques françaises

Données Réutilisations Organisations Tableau de bord Documentation Actualités

Connexion / Inscription

Recherche

PIAF - Le dataset francophone de Questions-Réponses PIAF - Q&A

Ce jeu de données provient d'un service public certifié

PIAF, construire un jeu de données francophones ouvert pour l'IA

Le recours à l'intelligence artificielle au sein de l'action publique est souvent identifié comme une opportunité pour interroger des textes documentaires et réaliser des outils de questions/réponses (QR) automatiques à destination des usagers. Interroger le code du travail en langage naturel, mettre à disposition un agent conversationnel pour un service donné, développer des moteurs de recherche performants, améliorer la gestion des connaissances, autant d'activités qui nécessitent de disposer de corpus de données d'entraînement de qualité afin de développer des algorithmes de questions/réponses. Le dataset PIAF est un jeu de données d'entraînement francophone public et ouvert qui permettrait d'entraîner ces algorithmes.

En nous inspirant de SQuAD, le jeu de données bien connu de QR anglais, nous avons l'ambition de construire un jeu de données similaire qui sera ouvert à tous. Le protocole que nous avons suivi est très similaire à celui de la première version de SQuAD (SQuAD v1.1). Néanmoins, quelques modifications ont dû être apportées pour s'adapter aux caractéristiques du Wikipédia français. Une autre grande différence est que nous n'employons pas de micro-travailleurs via des plateformes de crowd-sourcing.

Après quatre mois d'annotathons PIAF, nous avons une plateforme d'annotation robuste, une quantité non négligeable d'annotations et une démarche d'animation de communauté et de participation collaborative bien calée et innovante au sein de l'administration française.

Producteur

etalab gouv.fr

Etalab

La politique d'ouverture et de partage des données publiques ("Open Data") est pilotée, sous l'autorité du Premier ministre, par la mission Etalab, dirigée par Mme Laure Lucchesi…

VOIR LE PROFIL

CONTACTER

SUIVRE

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 applications; in contrast, having to deal with languages, NLP applications are naturally bound to language specificities.

Over the years, the NLP community has produced several resources to tackle tasks we call, for simplicity, *upstream* (such as Part-Of-Speech tagging, Dependency Parsing, etc.), targeting multiple languages and enabling the construction of effective automated systems. Still, for tasks we refer to as *downstream*, i.e. those which enable the development of value-added end products such as Question Answering (QA) or Conversational Agents, the current 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;
2. we release a native French QA dataset;
3. we provide baselines using state-of-the-art methodologies.

2. Related Work

Several datasets for QA have been recently produced. The Stanford Question Answering Dataset (SQuAD) (Rajpurkar et al., 2016) consists of 100,000+ questions posed by crowdworkers on a set of Wikipedia articles. The answer to each question is a segment of text from the correspond-

Voice technologies : the “Voice Lab” initiative

« Nous faciliterons la création de plates-formes de partage de données entre acteurs publics et privés, avec une logique sectorielle. Il faut en effet que les acteurs économiques eux-mêmes aillent plus loin dans leur pratique de partage et de valorisation de leurs données »

A private sector public sector collaboration

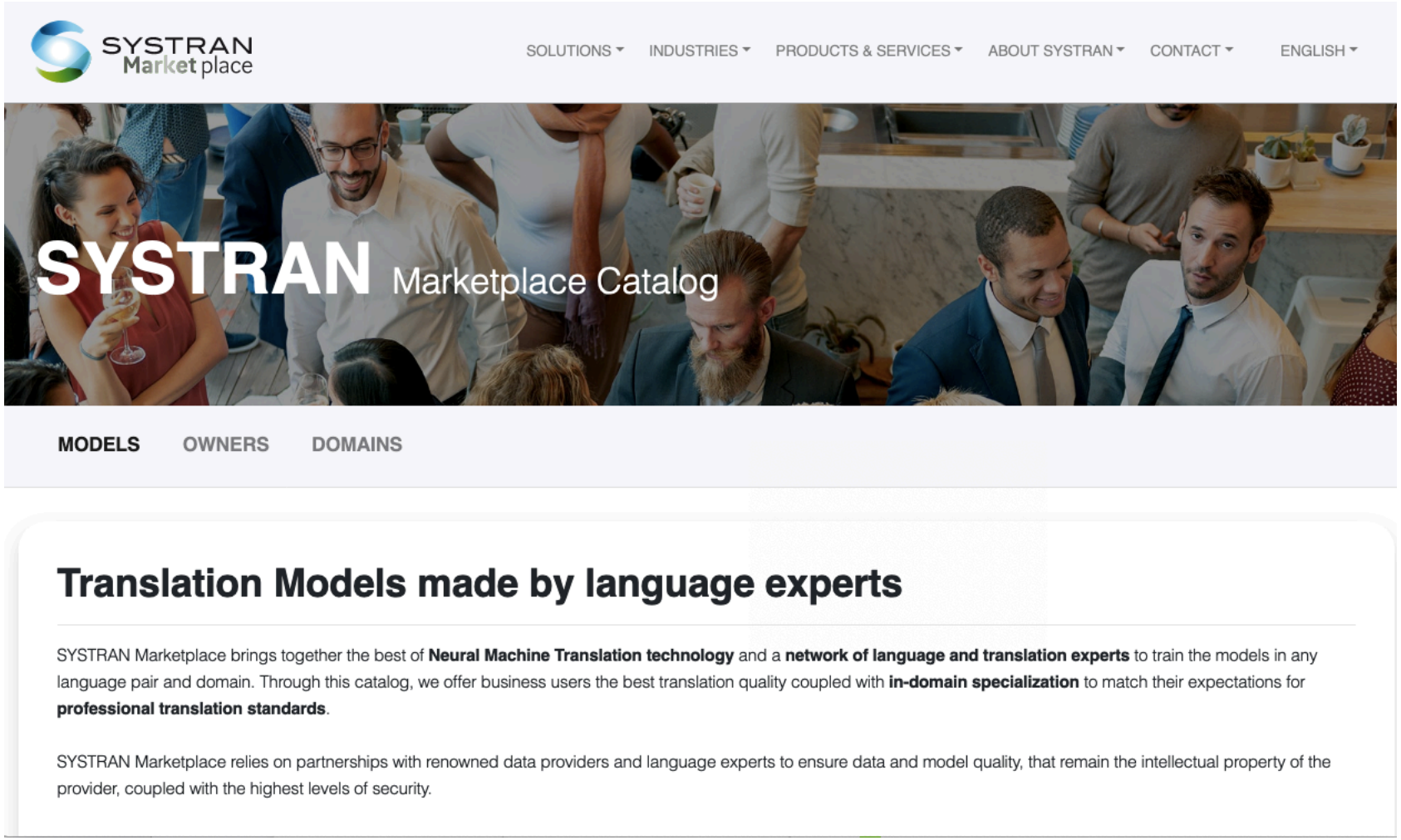
- Academic Labs, startups, SMEs, large groups

Ambitions:

- data collection, annotation & sharing (-> 100Kh transcript speech)
- interoperability of tools
- services & market place for voice technologies

Co-Funding: French government + private investments (2021-)

An open marketplace for translation systems



European Language Grid

Thank you!

The European Language Grid has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement № 825627 (ELG).

François Yvon (LIMSI, CNRS)
francois.yvon@limsi.fr

01/02/03-12-2020 META-FORUM 2020 – Piloting the European Language Grid (virtual conference)
http://www.european-language-grid.eu