MASAPI

Project overview

Pavel Pecina
UFAL meeting, Jun 24, 2024

Charles University
Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics

From the proposal ...

- **Goal:** To develop a “multilingual personal assistant for decision support and preparation of high-quality informative texts”.

- **Implementation**: Locally running tool, indexing everything that the user looks at, including personal information (docs, web pages, emails, ...) with numerous functionalities

Functionalities:

- Semantic indexing of documents
- Semantic search
- Web search
- Question answering
- Document summarization
- Document fusion
- Opinion mining
- Machine translation
- Image captioning

Project details

Partners

- Lingea (Michal Kašpar, Michal Hala)
  - System integration and testing
  - Machine translation
  - Semantic Indexing
  - Image captioning
  - Diacritics restoration
- BUT FIT (Pavel Smrž, Martin Dočekal)
  - Named Entities
  - Extractive Summarization
  - Question Answering
- UFAL
  - Document Fusion
  - Abstractive Summarization
  - CorePipe, UDPipe, NameTag

Program: TAČR Trend

Timeline: 1/2021 – 4/2024

Budget:

- 31M CZK Total Expenditures
- 22M CZK provided by TAČR
- 4M CZK for CUNI

Example use-cases

Ex. 1: A Czech engineer specializing in electronics found an integrated circuit on the website aliexpress.com, which, according to the shop's description, solves a large part of the problem he is currently working on. The brief marketing description is in English, but the technical documentation, spanning about 200 pages, is available only in Chinese, across five documents. The user wants to get a general overview of the component and its application.

Ex. 2: A person working on a grant application. In the grant application proposal, the planned project outcomes have been reorganized, and it is necessary to check the consistency of all mentions of the outcomes in the project proposal.

Ex. 3: An IT person wants to buy a router with some given (minimal) parameters and need a lost of all such routers that are available and their price.

UFAL team members

Michal Auersperger

- Document fusion
- Extractive tunable summarization of multiple documents
- Based on TextRank algorithm and sentence clustering
- Very much appreciated by Lingea

Mateusz Krubinski

- Abstractive summarization

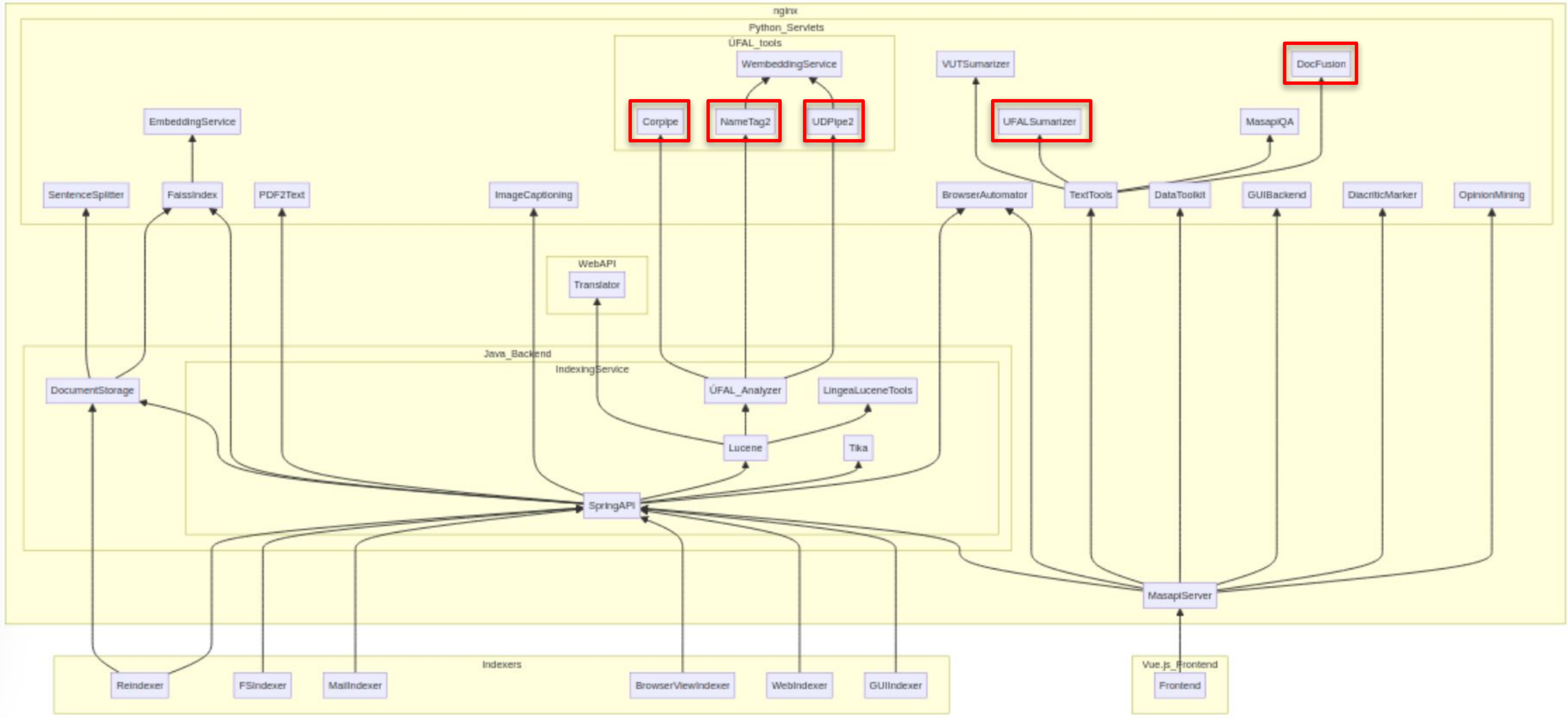
Michal Novák

- Coreference resolution
- Training data preparation
- CorePipe

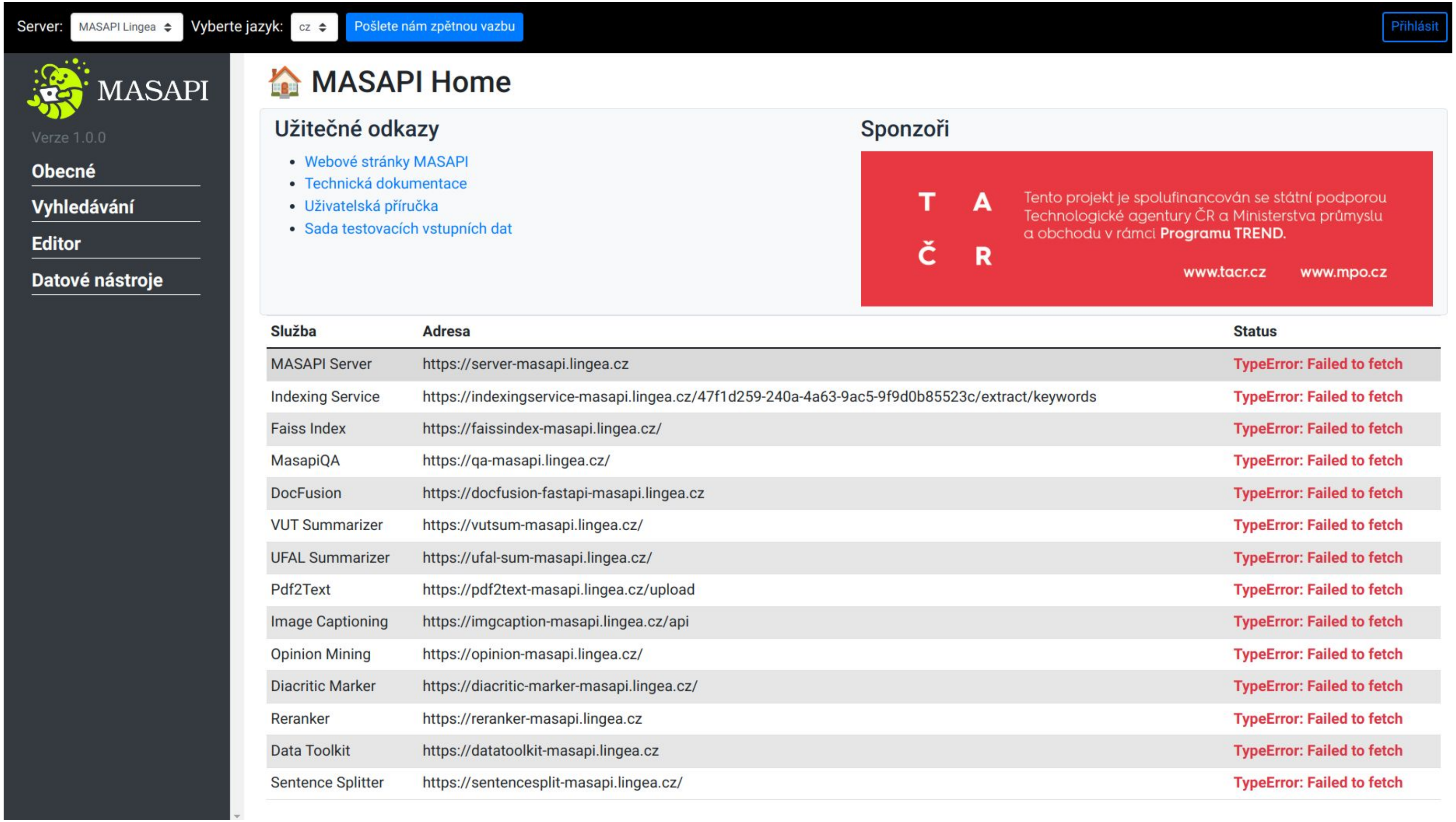
Dušan Variš

- Optimization of NameTag and UDPipe

System overview



Frontend



MASAPI Home

Užitečné odkazy

- Webové stránky MASAPI
- Technická dokumentace
- Uživatelská příručka
- Sada testovacích vstupních dat

Sponzoři

T A Č R

Tento projekt je spolufinancován se státní podporou Technologické agentury ČR a Ministerstva průmyslu a obchodu v rámci **Programu TREND.**

www.tacr.cz www.mpo.cz

Služba	Adresa	Status
MASAPI Server	https://server-masapi.lingea.cz	TypeError: Failed to fetch
Indexing Service	https://indexingservice-masapi.lingea.cz/47f1d259-240a-4a63-9ac5-9f9d0b85523c/extract/keywords	TypeError: Failed to fetch
Faiss Index	https://faissindex-masapi.lingea.cz/	TypeError: Failed to fetch
MasapiQA	https://qa-masapi.lingea.cz/	TypeError: Failed to fetch
DocFusion	https://docfusion-fastapi-masapi.lingea.cz	TypeError: Failed to fetch
VUT Summarizer	https://vutsum-masapi.lingea.cz/	TypeError: Failed to fetch
UFAL Summarizer	https://ufal-sum-masapi.lingea.cz/	TypeError: Failed to fetch
Pdf2Text	https://pdf2text-masapi.lingea.cz/upload	TypeError: Failed to fetch
Image Captioning	https://imgcaption-masapi.lingea.cz/api	TypeError: Failed to fetch
Opinion Mining	https://opinion-masapi.lingea.cz/	TypeError: Failed to fetch
Diacritic Marker	https://diacritic-marker-masapi.lingea.cz/	TypeError: Failed to fetch
Reranker	https://reranker-masapi.lingea.cz	TypeError: Failed to fetch
Data Toolkit	https://datatoolkit-masapi.lingea.cz	TypeError: Failed to fetch
Sentence Splitter	https://sentencesplit-masapi.lingea.cz/	TypeError: Failed to fetch

Conclusions

- Experienced PI
- Long history of cooperation with UFAL

- Highly ambitious proposal
- Joined writing very late
- Almost no chance to change the content

- Well-structured organization of the work
- BUT and CUNI as technology providers “only”, no complex interactions

- No training/test data provided by PI
- No project-specific evaluation, publications, …

- PI not very active in project organization, not very interested in some deeper collaboration

- Contribution to two papers, one dataset.

Thanks

- Lenka Fišerová
- Michal Auersperger
- Dušan Variš
- Michal Novák
- Mateusz Krubinksi