MASAPI

Project overview

Pavel Pecina
UFAL meeting, Jun 24, 2024
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
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 list of all such routers that are available and their price.
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
### MASAPI Home

**Užitečné odkazy**
- Webové stránky MASAPI
- Technická dokumentace
- Uživatelská příručka
- Sada testovacích vstupů dat

**Sponzoři**

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


<table>
<thead>
<tr>
<th>Služba</th>
<th>Adresa</th>
<th>Status</th>
</tr>
</thead>
<tbody>
<tr>
<td>MASAPI Server</td>
<td><a href="https://server-masapi.lingea.cz">https://server-masapi.lingea.cz</a></td>
<td>Type Error: Failed to fetch</td>
</tr>
<tr>
<td>Indexing Service</td>
<td><a href="https://indexingservice-masapi.lingea.cz/471d259-240a-4a63-9ac5-9f9d0b85523c/extract/keywords">https://indexingservice-masapi.lingea.cz/471d259-240a-4a63-9ac5-9f9d0b85523c/extract/keywords</a></td>
<td>Type Error: Failed to fetch</td>
</tr>
<tr>
<td>Faiss Index</td>
<td><a href="https://faissindex-masapi.lingea.cz/">https://faissindex-masapi.lingea.cz/</a></td>
<td>Type Error: Failed to fetch</td>
</tr>
<tr>
<td>MasapiQA</td>
<td><a href="https://qa-masapi.lingea.cz/">https://qa-masapi.lingea.cz/</a></td>
<td>Type Error: Failed to fetch</td>
</tr>
<tr>
<td>DocFusion</td>
<td><a href="https://docfusion-fastapi-masapi.lingea.cz">https://docfusion-fastapi-masapi.lingea.cz</a></td>
<td>Type Error: Failed to fetch</td>
</tr>
<tr>
<td>VUT Summarizer</td>
<td><a href="https://vutsum-masapi.lingea.cz/">https://vutsum-masapi.lingea.cz/</a></td>
<td>Type Error: Failed to fetch</td>
</tr>
<tr>
<td>UFAL Summarizer</td>
<td><a href="https://ufal-sum-masapi.lingea.cz/">https://ufal-sum-masapi.lingea.cz/</a></td>
<td>Type Error: Failed to fetch</td>
</tr>
<tr>
<td>Pdf2Text</td>
<td><a href="https://pdf2text-masapi.lingea.cz/upload">https://pdf2text-masapi.lingea.cz/upload</a></td>
<td>Type Error: Failed to fetch</td>
</tr>
<tr>
<td>Image Captioning</td>
<td><a href="https://imgcaption-masapi.lingea.cz/api">https://imgcaption-masapi.lingea.cz/api</a></td>
<td>Type Error: Failed to fetch</td>
</tr>
<tr>
<td>Opinion Mining</td>
<td><a href="https://opinion-masapi.lingea.cz/">https://opinion-masapi.lingea.cz/</a></td>
<td>Type Error: Failed to fetch</td>
</tr>
<tr>
<td>Diacritic Marker</td>
<td><a href="https://diacritic-marker-masapi.lingea.cz/">https://diacritic-marker-masapi.lingea.cz/</a></td>
<td>Type Error: Failed to fetch</td>
</tr>
<tr>
<td>Reranker</td>
<td><a href="https://reranker-masapi.lingea.cz">https://reranker-masapi.lingea.cz</a></td>
<td>Type Error: Failed to fetch</td>
</tr>
<tr>
<td>Data Toolkit</td>
<td><a href="https://datatoolkit-masapi.lingea.cz">https://datatoolkit-masapi.lingea.cz</a></td>
<td>Type Error: Failed to fetch</td>
</tr>
<tr>
<td>Sentence Splitter</td>
<td><a href="https://sentencesplit-masapi.lingea.cz/">https://sentencesplit-masapi.lingea.cz/</a></td>
<td>Type Error: Failed to fetch</td>
</tr>
</tbody>
</table>
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