Machine Learning and NLP on Diplomatic Documents

Practical applications of NLP in digital humanities

Diplomatic Documents of Switzerland (Dodis)

- Dodis is the centre of excellence for studies in the history of Swiss foreign policy and international relations
- Researchers at Dodis manually process ~1.5million documents from the federal administration per year, most of the documents come from the Swiss Federal Archives
- From those documents, Dodis publishes 1‘000-1’500 per year in the database Dodis
- In addition, Dodis publishes a selection of documents for each research year in a printed volume
- Additional documents are included in thematic publications

5.11.1968 (Tuesday)

Language: French

Circular (Circ)

Domaines dans lesquels une collaboration américano-suisse en recherche fondamentale aurait les plus grandes chances de se réaliser et d'avoir des effets positifs. Les sciences de l'environnement notamment offriraient de bonnes perspectives de collaboration.

File reference: o.320.USA

How to cite: dodis.ch/30266 Copy

Printed in

Sacha Zala et al. (ed.)

Diplomatic Documents of Switzerland, vol. 24, doc. 115

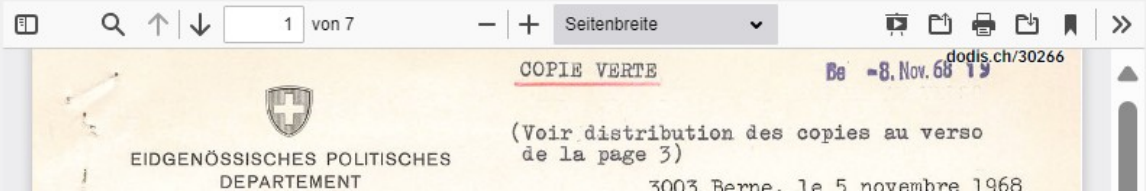
Zürich/Locarno/Genève 2012

more… | How to cite: Sacha Zala et al. (ed.), Copy

▸ **Repository**

Transcription DDS Facsimile

Zoom PDF



Tags

Science

United States of America (USA) (General)

Persons

Author

Tavel, Charles Hubert (1918–2010)

Signatory

Thalmann, Ernesto (1914–1993)

Mentioned

Barandun, Silvio

Beckler, David Berger, Fritz

Brunner, Kurt

Hornig, Donald Frederick (1920–2013)

Isliker, Henri Sträuli, Peter

Tavel, Charles Hubert (1918–2010)

Thürlimann, Bruno

Organizations

Author

FDFA/Directorate of International Organizations

Addressee

COPIE VERTE

Be -8. Nov. 68 19

EIDGENÖSSISCHES POLITISCHES DEPARTEMENT
DÉPARTEMENT POLITIQUE FÉDÉRAL

o.320.USA.1.- VE/bi.

Bitte dieses Zeichen in der Antwort wiederholen
Prière de rappeler cette référence dans la réponse

(Voir distribution des copies au verso de la page 3)

3003 Berne, le 5 novembre 1968

Au Secrétariat général du
Département fédéral de l'intérieur
3003 Berne

Au Délégué aux questions
d'énergie atomique
3003 Berne

Au Conseil suisse de la science
3003 Berne

Au Fonds national suisse
de la recherche scientifique
3001 Berne

A l'Ecole polytechnique fédérale
8006 Zurich

A l'Ecole polytechnique de
l'Université de Lausanne
1000 Lausanne

A la Faculté des sciences
de l'Université de Bâle
4000 Bâle

A la Faculté des sciences
de l'Université de Berne
3000 Berne

A la Faculté des sciences
de l'Université de Fribourg
1700 Fribourg

A la Faculté des sciences
de l'Université de Genève
1200 Genève

A la Faculté des sciences
de l'Université de Lausanne
1000 Lausanne

A la Faculté des sciences
de l'Université de Neuchâtel
2000 Neuchâtel

A la Faculté des sciences de
l'Université de Zurich (Phil. II)
8006 Zurich

A la Faculté de médecine
de l'Université de Bâle
4000 Bâle

A la Faculté de médecine de
l'Université de Berne
3000 Berne

A la Faculté de médecine de
l'Université de Genève
1200 Genève

A la Faculté de médecine
de l'Université de Lausanne
1000 Lausanne

A la Faculté de médecine
de l'Université de Zurich
8006 Zurich

Be -8. Nov. 68 19

Messieurs,

Les possibilités d'une coopération bilatérale plus étroite entre la Suisse et les Etats-Unis, dans des domaines de recherche fondamentale, ont fait l'objet ces derniers mois de divers entretiens, à Washington et à Berne. Le plus concret de ces entretiens a eu lieu à la fin de juin entre notre Conseiller scientifique à Washington,

./.

12784/6

Dodis workflow

- Each researcher receives ~50 archive boxes from the federal archive per week
- The researcher scans trough the content of the boxes and tries to find the most relevant documents:
  - Relevance is defined as: “A key document to showcase foreign policy decisions“
- Relevant documents are selected during the weekly meeting of the research group
  - Research assistants scan the document
  - Research assistants manually transcribe the document
  - Research assistants manually annotate the document
  - Research assistants manually index the document in our database

Future challenges

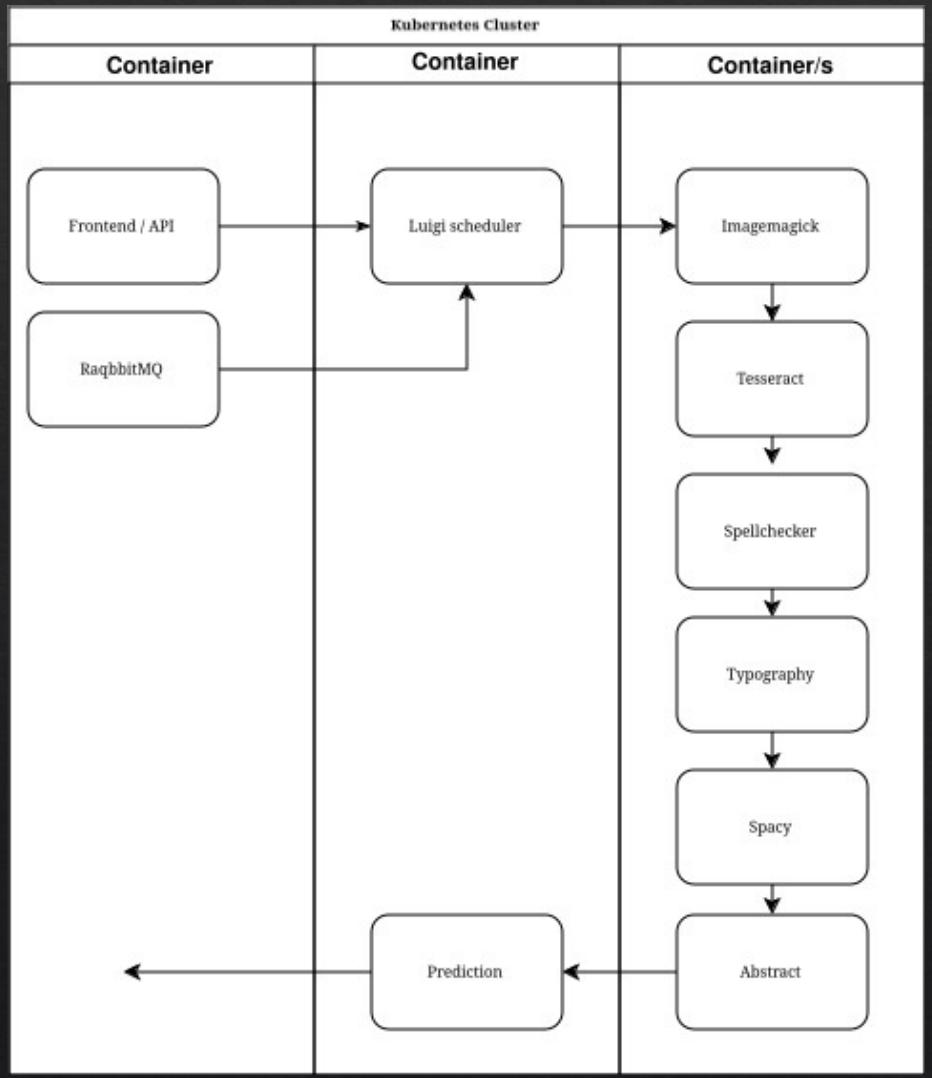
- Based on the “Bundesgesetz über die Archivierung” (Federal Act on Archiving) archival records become available after the expiry of a retention period of 30 years (for some documents extended retention periods apply)
- Most of those documents are analogue and must be digitized and transcribed
- In the future, we will have more scanned documents
- In the future, we will have more and more digital born documents
- In the future, we will have much more documents in general

Goal of our current developments

- We want to support the research group with a pipeline that:
  - Speeds up the transcription process
  - Supports the annotation process
  - Automates the integration into our DB
- Why now?
  - Recent developments in ML and LLMs have demonstrated the potential of applications in text-based domains
  - Dodis has always been open to new technologies
  - LLMs challenge core skills of historians.

Demo time: 🤖 Fancy-doc

- Born during an internal hackathon at Dodis
- Developed into a tool, we currently use in production. The tool is WIP, has limitations and bugs.
- https://fancy-ml.dodis.ch
- https://luigi.dodis.ch
- https://dodis.ch/62121
- https://fancy-ml.dodis.ch/predict/index.html?version=v0.0.3&doc=dodis-62121



🤖 How it works

- Convert PDF to images
- Use OCR to extract text from images
- Apply rules and a LLM to improve OCR output
- Do named entity recognition (NER)
- Do named entity linking (NEL)
- Summarize content
- Produce HTML and TEI-XML output

🤖 Detail: Spacy task

1) We started with Spacy and a pre-trained model at our hackathon
2) We added an «EntityRuler» to get better results and basic linking with our database.
3) We evaluated our XML-data and build a dataset to train a custom NER model
4) We trained a custom NER model to mimic a Dodis researcher
5) We built up a knowledge base with known entities and their distribution in our XML-data
6) We extracted training data for named entity disambiguation from our XML
7) We trained the «EntityLinker» to learn different entities

🤖 NER model evaluation

NER with de_core_news_lg

	P	F	R
LOC	41.18	72.85	52.61
PER	36.72	45.13	40.49
ORG	13.56	43.96	20.73

Model trained on Dodis data

	P	F	R
LOC	89.51	93.11	91.27
PER	88.63	87.00	87.81
ORG	74.19	68.24	71.09

🤖 NEL model evaluation

1) The evaluation from the NEL model was okish
   - Good for full names
   - Good for fuzzy matches
   - Not so good for functions
2) In the reality we often encounter functions and different abbreviations of names
   - Example: https://fancy-ml.dodis.ch/predict/index.html?version=v0.0.3&doc=dodis-62121
   - Staatsekretär → wrong, mixed up Secretary of State from Switzerland and Greece
   - Departementschef → wrong, mixed «Federal Department of Home Affairs» and not «Federal Department for Foreign Affairs»

🤖Challenges

- Limited ground truth. 8‘000 out of 50’000 documents are transcribed
  - We miss a lot of train data
  - We miss a lot of entities in our knowledge base
  - We miss a lot of words in our vocabulary
- Multilingual: most documents are multilingual. This is a challenge for all tasks
- Compute power: Until now, projects in the humanities have required little or no resources for computing power. This is changing and resources need to be adjusted accordingly.
- ML-Knowledge: Available resources also play an important role in the development of ML skills.
- Legal limitations