

Citation Recognition in Legal Documents

Named Entity Recognition with Transformer-Based Deep Learning Methods

Graduate



Mirio Eggmann



Dario Glasl



Jasmin Fitz

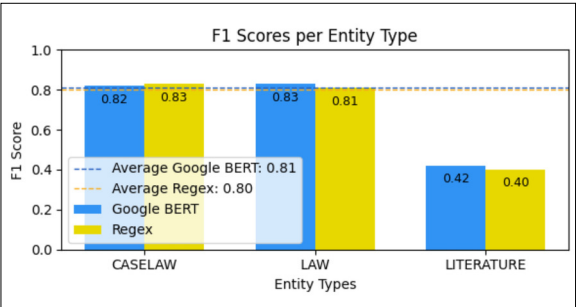
Definition of Task: This bachelor thesis develops a system for recognizing citations in legal documents. It is carried out in collaboration with Swisslex, an online research platform for legal documents. In order to link those documents in their web application, citations and their specific parts need to be identified in the text. The current citation recognition system is based on regular expression matching. This approach is not robust and is hard to maintain.

Approach: The task is divided into two stages: recognizing citations, as well as their parts. First, text chunks are processed using a fine-tuned Google BERT multilingual base model (uncased). This identifies citations and classifies them into the categories CASELAW, LAW and LITERATURE. The recognized citations are then passed to a second fine-tuned model, based on DistilBERT (uncased). This identifies the specific parts of a citation, such as COURT, DATE, ARTICLE, etc. Sparse data poses a challenge for recognizing parts of a citation. Few-shot prompting with an LLM yields good results, but experiments show that it is prohibitively slow in practice. Computational requirements of such large models hampers their adoption. Therefore, DistilBERT is fine-tuned via the supervision of Llama 3.3 70B through knowledge distillation. The 1'044 times smaller DistilBERT achieves similar performance compared to Llama. The final .NET based solution offers bulk processing as well as a web interface for user interaction. FastAPI is used to serve the fine-tuned models. Results are stored in an MS SQL database.

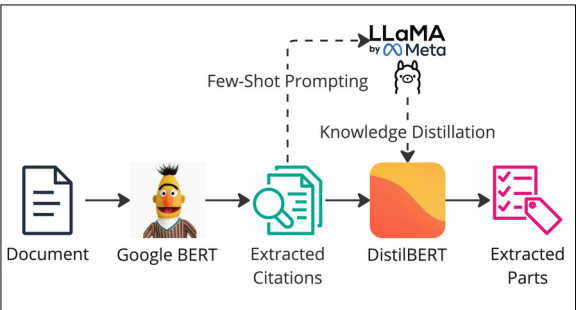
Result: The fine-tuned Google BERT achieves an F1 test score of 95.7% for recognizing citations. The proposed solution offers many benefits apart from the at par performance with the regex-based system. The

solution is more robust and can handle deviations such as typing errors and new citation formats more easily. Moreover, the cherry on top is the improved performance in recognizing parts of a citation with more granular labels. The fine-tuned DistilBERT recognizes parts with an F1 test score of 98.9%. Further improvement is possible via increasing diversity in the training samples.

Comparison of the regex- vs. Google BERT-based citation recognition, evaluated on the manually annotated test set
Own presentation



Processing pipeline
Own presentation



Web-application to test citation recognition
Own presentation

Citation Checker Web

Document Upload

Browse files

or

4a_216/2020

Citation Recognition

googlebert-multilingual-10M

Recognize Citations

Recognize Citations and Parts (Few-Shot)

Process time: 186 ms

Parts Recognition

distilbert-0.3M

Recognize Parts (BERT-based)

Recognize Parts (Few-Shot)

Process time: 347 ms

37 citations recognized with model and 28 citations linked in Swisslex. Chunks used in inference are marked with green start and red end lines.

In Erwägung,

dass selbständig eröffnete Zwischenentscheide über den Ausstand eines Gutachters nach Art. 92 BGG in der Regel zwar anfechtbar sind und nicht erst zusammen mit dem Endentscheid angefochten werden können, womit verhindert werden soll, dass ein Experte mit einem Gutachten betraut wird, dessen Ergebnisse wegen Befangenheit ohnehin nicht verwertet werden könnten (Urteil 4A_255/2011 vom 4. Juli 2011 E. 1.2);

dass Art. 92 BGG indessen nicht zum Tragen kommt, wenn das Gutachten bereits erstellt ist und es mithin lediglich darum geht, ob dieses im Recht liegende Beweismittel verwertbar ist, da es mit Bezug auf die Verwertbarkeit eines Beweismittels nicht darauf ankommen kann, ob die behauptete Unverwertbarkeit sich aus der Befangenheit des Gutachters oder aus anderen Gründen ergibt (Urteile 4A_255/2011 vom 4. Juli 2011 E. 1.2; 4A_269/2013 vom 7. Oktober 2013 E. 1.1; vgl. für den Ausstand einer Richterin bzw. eines Staatsanwalts: Urteile 4A_221/2016 vom 20.

Citation	Confidence	Type
Art. 92 Article BGG Law	99.94%	Law
4A_255/2011 Dossiernumber vom 4. Juli 2011 Date E. 1.2 Consideration	99.71%	Caselaw
Art. 92 Article BGG Law	99.91%	Law
4A_255/2011 Dossiernumber vom 4. Juli 2011 Date E. 1.2 Consideration	99.64%	Caselaw
4A_269/2013 Dossiernumber vom 7. Oktober 2013 Date E. 1.1 Consideration	99.65%	Caselaw
4A_221/2016 Dossiernumber vom 20. September 2016 Date E. 2.2 Consideration	99.64%	Caselaw
1B_77/2017 Dossiernumber vom 9. Mai 2017 Date E. 1.3 Consideration	99.65%	Caselaw
Art. 92 Article BGG Law	99.95%	Law
Art. 93 Article BGG Law	99.95%	Law
Art. 93 Article Abs. 1 Paragraph lit. a Letter BGG Law	99.96%	Law
Art. 93 Article Abs. 1 Paragraph lit. b Letter BGG Law	99.96%	Law
BGE Court 144 Volume III Subvolume 475 Part E. 1.2 Consideration	94.29%	Caselaw

Advisors

Prof. Dr. Mitra
Purandare, Benjamin
Plattner

Co-Examiner

Marcel Amsler, Sanitas
Management AG,
Zürich 1, ZH

Subject Area

Artificial Intelligence,
Software

Project Partner

Swisslex AG, Zürich

Eastern Switzerland University of Applied Sciences | Bachelor Theses 2025 | Bachelor of Science OST | Informatik