Interlex – A Search Engine to Explore the Interconnectedness of Swiss Legal Texts

Selena Calleri and **Michelle Wastl** and **Bojan Peric** and **Andreas Abegg**
selena.calleri@gmail.com

Abstract

The Swiss legal landscape is tied to working with text by using non-digital approaches due to its multilingual and federalist nature. Digital approaches must overcome the plurality of formats, contents, and languages in order to provide useful solutions to legal practitioners. We present Interlex, a tool which tries to overcome these chasms by leveraging NLP technologies and knowledge from linguistics and law to make Swiss state and canton level court decisions accessible through their interconnectedness. The tool is built on a collection of 600'000+ web-crawled court decisions. The texts have been preprocessed to form a standardized corpus of Swiss court decisions that is regularly updated. The preprocessing of the texts included manual layout analysis of every court, scraping the text from different data formats, regex-based cleaning and paragraph splitting. In a next step, a fine-tuned BERT model has been used for legal sentence-boundary-detection before the sentences were embedded using a multilingual BERT-based sentence encoder to facilitate semantic similarity analysis in multiple languages. The data is then stored in a dynamically expandable database, which is regularly updated with new court decisions. This database serves as a backend for the Interlex web application. Interlex focuses on identifying so-called textual building blocks (TBBs) which are "prefabricated argument blocks." They are often repeated throughout court decisions and usually require years of practitioner experience to be identified. Apart from yearlong experience the only other approach to find these TBBs is using books or articles with precedents (dt. "Präjudizien") which indicate potentially relevant passages yet imply a) the continuous maintenance of the material, b) tedious manual labor to create it by multiple authors, and c) access to the collection. Our tool recognizes relevant and potential TBBs on sentence level by using four different levels of similarity: i) exact copy, ii) strongly similar wording, iii) similar meaning with edits, and iv) similar meaning but different words. These surface form and semantic level similarities are automatically detected by leveraging multiple edit differences and similarity scores. This approach differs from manual selection as every phrase is considered a potential TBB and then defined as such if it meets the interconnectedness criteria. Interlex allows for 2 modalities: either searching in full text for relevant keywords and exploring the found passages in the full text, or directly exploring TBBs based on a search term and finding the most relevant sentences, defined by their interconnectedness. This required a parametrization of TBBs, as for the first time a data-based approach was applied to identify them systematically in a comprehensive corpus of court decisions. The exploratory nature of the tool and the ambiguous definition of TBBs make the tool's evaluation complex. We employed sample-based qualitative analyses to evaluate data cleanliness, embedding model selection, and the interconnectedness score.