MUNI
LAW

Datasety využité pro extrakci citací soudních rozhodnutí

Tereza Novotná

Ústav práva a technologií, Právnická fakulta, Masarykova Univerzita

GAČR – Exaktní hodnocení aplikační relevance judikatury (GA17-20645S)

Motivace

- Proč jsou datasety a databáze soudních rozhodnutí důležité?
  - Empirická data, automatické zpracování, vyhledávací nástroje, právní informační systémy
- Metody zpracování přirozeného jazyka
- Specifika právního jazyka
- Inspirace ze zahraničí (eur-lex.eu, rechtspraak.nl)

Osnova

- 3 datasety soudních rozhodnutí, které byly využity k analýze:

1) Dataset textů soudních rozhodnutí

2) Anotovaný dataset využitý pro segmentaci textů rozhodnutí

3) Anotovaný dataset využitý pro extrakci citací z textů rozhodnutí

1) Dataset soudních rozhodnutí

- Autorský tým: Jakub HARAŠTA a Tereza NOVOTNÁ

- Dataset textů soudních rozhodnutí opatřených metadaty
- Ústavní soud, Nejvyšší a Nejvyšší správní soud

Motivace

- Rozhodnutí přístupná veřejnosti jsou:
  - Dostupné pouze na webových stránkách soudů nebo
  - Přes aplikace soukromých poskytovatelů (paywall, netransparentnost)
  - Nekomplexní
  - Neintuitivní vyhledávání
  - Nevhodné formáty
- Nepřístupnost nezpracovaných dat
- Nevhodné pro automatické zpracování textu

Získání dat

- Žádost o informaci dle zákona č. 106/1999 Sb.
- Různé odpovědi
  - NS neposkytl rozhodnutí vůbec, rozhodnutí byla stáhnuta z webových stránek nsoud.cz
  - NSS a ÚS ano, ale v nevhodných formátech
- Úroveň technologií

Data processing

- Zpracovávání
  - Formátování – Apache Tikka
  - Extrakce souvisejících metadat (soubor, spisová značka nebo číslo jednací, datum rozhodnutí, identifikace soudu)
  - Textové přílohy
  - CSV seznamy

- Statistiky
  - Od 1. 1. 1993 do 30. 9. 2018
  - 237 723 rozhodnutí všech soudů, 460 524 867 slov

Výsledek

- Všechna rozhodnutí v elektronické podobě z daného období (ÚS, NSS)
- Pouze část rozhodnutí, která jsou zveřejněna na webové stránce soudu (NS)
- Publikace datasetu – Czech Court Decisions Corpus https://lindat.mff.cuni.cz/repository/xmlui/handle/11372/LRT-3052
- Související publikace se připravuje

Future work

- Doplnění dalších metadat (typ rozhodnutí, publikace ve Sbírce)
- Aktualizace datasetů
- Není součástí tohoto projektu

2) Anotovaný dataset využitý pro segmentaci textů rozhodnutí

- Autorský tým: Jakub HARAŠTA, Jaromír ŠAVELKA, František KASL, Jakub MÍŠEK

- Definice daných částí soudního rozhodnutí

MUNI LAW

Motivace

- Nestrukturovaná data
- Stejné informace v různých částech rozhodnutí mají různý význam
- Přesnější výsledky při dalším automatickém zpracovávání soudních rozhodnutí

Segmentace

- Trénovací dataset 350 rozhodnutí ÚS, NS a NSS
- Definované části rozhodnutí: hlavička, procesní historie, podání a reakci protistrany, argumentace, zápatí, disent, poznámka pod čarou
- Segmentace prováděna anotacemi úseků rozhodnutí, jedním či dvěma anotátory (v 70 případech) a supervidována editorem

Výsledky

- Dataset publikován: https://lindat.mff.cuni.cz/repository/xmlui/handle/11372/LRT-2901
- Související publikace: HARAŠTA, Jakub, Jaromír ŠAVELKA, František KASL a Jakub MÍŠEK. Automatic Segmentation of Czech Court Decisions into Multi-Paragraph Parts. *Jusletter IT, Weblaw AG, 2019, roč. 4, 23*. Mai 2019, s. 1-10. ISSN 1664-848X.

MUNI LAW

Výsledky a future work

- Využití při automatické extrakci citací z rozhodnutí za předpokladu, že jsou definovány části, ve kterých se citace nachází
- Zpřesnění výsledků

3) Anotovaný dataset využitý k extrakci referencí z textů rozhodnutí

- Autorský tým: Jakub HARAŠTA, Jaromír ŠAVELKA, František KASL, Adéla KOTKOVÁ, Pavel LOUTOCKÝ, Jakub MÍŠEK, Daniela PROCHÁZKOVÁ, Helena PULLMANNOVÁ, Petr SEMENIŠÍN, Tamara ŠEJNOVÁ, Nikola ŠIMKOVÁ, Michal VOSINEK, Lucie ZAVADILOVÁ a Jan ZIBNER

- Anotace referencí v textech soudních rozhodnutí za účelem automatizace extrakce citací

MUNI LAW

Motivace

- Nedostupnost PRÁVNÍCH textových korpusů v českém jazyce
  - Kríž, Vincent and Hladká, Barbora, 2014, Czech Court Decisions Dataset, LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.
  - Kríž, Vincent and Hladká, Barbora, 2017, Czech Legal Text Treebank 2.0, LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Anotace

- Dataset 350 rozhodnutí ÚS, NS a NSS
- Trénovací dataset na rozpoznávání referencí v textu soudních rozhodnutí
- Dvoustupňová metodologie:
  - Manuální označení reference v textu vždy dvěma anotátory
  - Každé rozhodnutí poté zkontrolováno supervizorem z důvodu možné nekoherence označení referencí

Výsledky

- Automatické rozpoznání a extrakce citací v soudních rozhodnutích
- Dataset publikován:

  https://lindat.mff.cuni.cz/repository/xmlui/handle/11234/1-2647
- Související publikace:
  - HARAŠTA, Jakub, Jaromír ŠAVELKA, František KASL, Adéla KOTKOVÁ, Pavel LOUTOCKÝ, Jakub MÍŠEK, Daniela PROCHÁZKOVÁ, Helena PULLMANNOVÁ, Petr SEMENIŠÍN, Tamara ŠEJNOVÁ, Nikola ŠIMKOVÁ, Michal VOSINEK, Lucie ZAVADILOVÁ a Jan ZIBNER. Annotated Corpus of Czech Case Law for Reference Recognition Tasks. In *Sojka, P., Horák, A., Kopeček, I., Pala, K.. Text, Speech, and Dialogue: 21st International Conference*. Cham: Springer Nature Switzerland AG, 2018. s. 239-250, 12 s. ISBN 978-3-030-00793-5. doi:10.1007/978-3-030-00794-2_26.

Závěr

- Vytvoření 3 volně dostupných datasetů pro zpracování soudních rozhodnutí
- Odhadem 1600 člověkohodin ve zpracovávání
- Zveřejňování dat jako cesta k lepším nástrojům ke zpracovávání soudních rozhodnutí?

Děkuji za pozornost.

GAČR – Exaktní hodnocení aplikační relevance judikatury (GA17-20645S)

MUNI LAW