Extrakcia právnych odkazov pomocou modelov transformérov

Rozšírené zadanie diplomovej práce

Autor: Bc. Nicol Fedurcová

Vedúci: RNDr. Peter Gurský, PhD.

Konzultant: RNDr. Dávid Varga

Pracovisko: Ústav informatiky

Táto diplomová práca sa zaoberá extrakciou odkazov zo súdnych rozhodnutí. Ministerstvo spravodlivosti Slovenskej Republiky zverejnilo štyri milióny súdnych rozhodnutí, ktoré slúžia ako základ tejto práce. V súdnych rozhodnutiach sa nachádzajú odkazy rôzneho charakteru smerujúce napríklad na:

- Zákony
- Články ústavy
- Iné súdne rozhodnutia
  - Slovenských súdov
  - Európskych súdov

Súdne rozhodnutia sú špecifickým typom textov, ktoré sú polo-štruktúrované a písané prirodzeným jazykom. Štýl značiek, skratiek a celkovej organizácie dokumentu je teda silne závislý na konkrétnom sudcovi, ktorý dokument píše. Práve preto je automatizované spracovanie takýchto textov pomerne náročné a podporné nástroje ako napríklad extraktov odkazov môžu byť nápomocnými čiastkovými krokmi k dosiahnutiu automatizovaného spracovania súdnych rozhodnutí s víziou umožnenia prehľadného vyhľadávania v súdnych rozhodnutiach.

Ciele práce:

1. Vypracovať prehľad metód extrakcie z textov s využitím transformérov a prehľad transformer-modelov vhodných pre slovenské právne texty.

2. Identifikovať a analyzovať dôvody neúspešných extrakcií v existujúcom pravidlovom systéme extrahujúcom odkazy na súdne rozhodnutia a právne predpisy.

3. Navrhnúť a implementovať rozšírenie existujúceho riešenia s využitím transformer modelu.

4. Porovnať úspešnosť novej metódy s existujúcim pravidlovým systémom na rozšírenom anotovanom datasete právnych textov.

Práca vychádza z článku (*1*) a už existujúceho pravidlového systému na extrakciu odkazov na zákony a iné súdne rozhodnutia. Tento systém dosahuje 92.05% *F1-skóre* pri extrakcii odkazov na zákony a 81.36% *F1-skóre* pri extrakcii odkazov na iné súdne rozhodnutia.

Nadväzujúc na výsledky tohto článku je hlavným cieľom tejto práce spresnenie procesu extrakcie odkazov s využitím modelov transformérov. Transforméry sú špecifickým typom neurónových sietí, ktoré sú schopné uchopiť kontext písaného textu prostredníctvom zachytenia vzťahov v sekvenčných dátach - v našom prípade súdnych rozhodnutiach písaných prirodzeným jazykom. Majú teda potenciál na vylepšenie aktuálne existujúceho pravidlového systému, ktorý sa napríklad pri nejednoznačnom značení nerozhoduje podľa kontextu súdneho rozhodnutia ale podľa výsledkov algoritmu zvolí prvú z možností.

Avšak extrakcia odkazov zo súdnych rozhodnutí len pomocou transformérov bez pomocných mechnizmov akými je napríklad aj tento pravidlový systém je mimoriadne náročná úloha. Článok (*2*) sa zaoberal čiastkovou úlohou - predikciu odkazov 20 najcitovanejších zákonov, s využitím predtrénovaného modelu German BERT. Na týchto 20 najcitovanejších zákonoch dosiahli 0.92% /textitF1-skóre pri predikcii základného zákonu (bez písmen a ďalších podrobností), pri predikcii celého odkazu 0.80% /textitF1-skóre. Tieto výsledky sú zaujímavé, avšak je potrebné si uvedomiť, že súdne rozhodnutia sa odkazujú na veľké množstvo zákonov, ktoré nielen že obsahujú viaceré (rôzne podrobné) členenia ale zároveň podliehajú zmene v čase, či vznikajú nové verzie rozhodnutí, pričom však v starších rozhodnutiach sa stále odkazovalo na ich staršie verzie.

Práve pre tieto skutočnosti našim hlavným cieľom nie je využívať transformér model samostatne ale v kombinácii z ďalšími metódami, či už novými, alebo existujúcimi.

Postup práce

Táto téma ponúka rôzne čiastkové smerovania a využitie rôznych typov metód a ich kombinácií, preto je mimoriadne dôležitá systematická práca.

- Zorientovanie sa v problematike právnych textova
- Analýza súčasného pravidlového systému
- Analýza dôvodov neúspešných extrakcií v pravidlovom systéme

- Vytvorenie prehľadu modelov transformérov
- Vytvorenie prehľadu metód extrakcie prvkov z textov s využitím modelov transformérov
- Rozšírenie a prispôsobenie anotovanej sady
- Zlúčenie pravidlového systému s modelom transformeru
- Vyhodnotenie úspešnosti novej metódy na rozšírenom anotovnaom datasete
- Porovnanie novej metódy s existujúcou metódou

Literatúra

1. Dávid Varga et al. (2023) *Extraction of Legal References from Court Decisions*, Vol. 3498, pp. 89–95.
2. Harshil Darji, Jelena Mitrović, and Michael Granitzer (2023) *A Dataset of German Legal Reference Annotations*, Proceedings of the Nineteenth International Conference on Artificial Intelligence and Law.
3. Daniel Jurafsky and James H. Martin (2024) *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition with Language Models*, 3rd ed.
4. Tommaso Agnoloni et al. (2017) *Linking European Case Law: BO-ECLI Parser, an Open Framework for the Automatic Extraction of Legal Links*, Legal Knowledge and Information Systems, IOS Press, pp. 113–118.

Podpis vedúceho práce: _____