Extrakcia právnych odkazov pomocou modelov transformérov

Vedúci práce: RNDr. Peter Gurský, PhD.

Konzultant: RNDr. Dávid Varga

Študent: Bc. Nicol Fedurcová

UNIVERZITA PAVLA JOZEFA ŠAFÁRIKA V KOŠICIACH

Prečo je extrakcia právnych odkazov dôležitá?

- Využiteľnosť v súdnej vede :
  - Nájdenie rozhodnutí, týkajúce sa podobných prípadov
  - Skúmanie vzťahov medzi rozhodnutiami
- Aplikácia v systéme na vyhľadávanie súdnych rozhodnutí
- Zjednodušenie automatizovaného spracovania rozhodnutí

Výzvy pri vyhľadávaní a anotovaní právnych odkazov

HLAVNÝ PROBLÉM: súdne rozhodnutia sú písané prirodzeným jazykom

- Odkazy nemajú jednotnú štruktúru
- Používanie skratiek, značiek, skratkových slov, ktoré nemusia byť vždy vysvetlené
- Každý sudca má vlastný štýl písania rozhodnutia
- Štruktúra rozhodnutí je naprieč rôznymi krajinami rôzna

Motivácia

- Aktuálna tématika modelov transformérov
- Práca na datasete Slovenských súdnych rozhodnutí
- Využiteľnosť výsledku práce v súdnej vede
- Aplikácia metódy v systéme na vyhľadávanie súdnych rozhodnutí
- Vylepšenie presnosti existujúcej metódy

Ciele práce

1. Vypracovať prehľad metód extrakcie z textov s využitím transformerov 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.

Postup práce

- Prehľad použiteľných jazykových modelov
- Hľadanie príčin zlyhania extrakcie odkazov pôvodnou metódou
- Vytvorenie metódy na nájdenie rozhodnutí s viacerými možnosťami výberu pri aliase
- Anotovanie rozhodnutí s viacerými možnosťami výberu pri aliase, Vytvorenie prvej verzie Choice checkera
- Modifikovanie metódy na nájdenie rozhodnutí s viacerými možnosťami výberu pri aliase

Hľadanie príčin zlyhania extrakcie odkazov pôvodnou metódou

Príklad A)

Súd podľa § 101 ods. 2 O.s.p. konal a rozhodol v neprítomnosti navrhovateľa a jeho právneho zástupcu, ktorí svoju neúčasť na pojednávaní ospravedlnili, súhlasili s pojednávaním v ich neprítomnosti.

```
_id: "O.s.p"
▾ law_ids: Array
    0: "99/1963"
```

Príklad B)

§ 62 ods. 1 zákona č. 36/2005 Z. z. o rodine a o zmene a doplnení niektorých zákonov - ďalejlen \"ZoR\"

Podľa § 78 ods. 2 ZoR ak dôjde k zrušeniu alebo zníženiu výživného pre maloleté dieťa

```
_id: "ZoR"
▾ law_ids: Array
    0: "94/1963"
    1: "530/2003"
    2: "36/2005"
```

Príklad C)

Uvedené ust. § 23 ZoR predpokladá objektívnuexistenciu rozvratu vzťahov bez ohľadu na subjektívne

Postup práce

1. Vytvorenie metódy na nájdenie a preformátovanie znenia konkrétneho odkazu
2. Vytvorenie test-makera, ktorého výstup bude vstupom pre nový choice checker
3. Vytvorenie novej (zjednodušenej) verzie choice checkera (kontrolóra viacerých možností pri aliase – bez kontroly existencie odkazu)
4. Zostavenie a manuálne anotovanie testovacej sady na porovnanie choice checkerov s rôznymi parametrami
5. Porovnanie výsledkov rôznych parametrov choice checkera

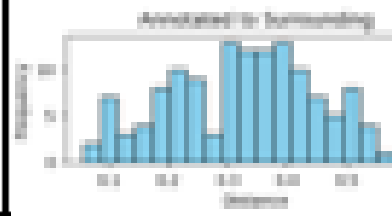
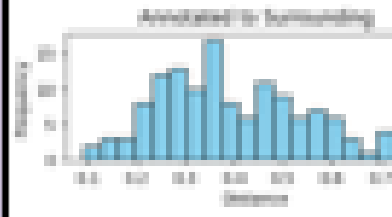
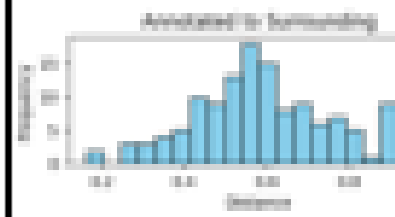
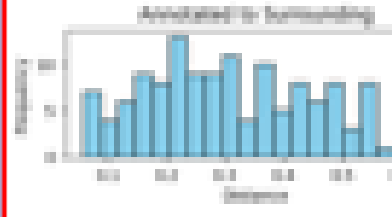
Testovací dataset


```
{
    "url": "/SK/ZZ/1977/61/19930822.html#paragraf-127.odsek-4",
    "law": "61/1977",
    "version": "19930822",
    "fragment": "paragraf-127.odsek-4",
    "section": 4,
    "letter": null,
    "start": 5937,
    "end": 5953,
    "paragraph": "127",
    "zakonnik_start": 5949,
    "zakonnik_end": 5953,
    "extracted_from": "global from db",
    "alias_text": "",
    "aliases": "ZoCP",
    "alias_options": [
        "61/1977",
        "526/1990",
        "566/2001",
        "326/2005",
        "8/2009"
    ],
    "annotated_url": "/SK/ZZ/2001/566/20130610.html#paragraf-127.odsek-4",
    "existing_options": [
        {
            "zakonnik": "566/2001",
            "provision_text": "566/2001, o cenných papieroch a investičných službách a o zmene a doplnení niektorých zákonov      (zákon o cenných papieroch), och
            "version": "20130610.html"
        },
        {
            "zakonnik": "8/2009",
            "provision_text": "8/2009, o cestnej premávke a o zmene a doplnení niektorých zákonov, evidencia vozidiel, evidovanie vozidiel a evidenčné čísla, evid
            "version": "20160101.html"
        }
    ],
    "surrounding_text": "neskorších predpisov /ZoCP/  za porušenie ustanovenia § 129 ods. 3 v spojení             sustanovením § 128 ods.1 písmeno b/ c/ h/ ZoCP
}
```

Výsledok choice checkera V2


```
"5C/265/2015__2016-06-02__ECLI:SK:OSNO:2016:5815205480.3": [
    {
        "start": 5085,
        "end": 5100,
        "checker_url": "/SK/ZZ/2005/36/20160101.html#paragraf-78.odsek-2",
        "surrounding_text": "nosti  rodičov  k  deťom je ich zákonná  povinnosť, ktorá  trvá dočasu, kým  deti nie sú sch
        "existing_options": {
            "94/1963": {
                "zakonnik": "94/1963",
                "distance_from_surrounding_text": "0.64754695",
                "version": "20040101.html",
                "provision_text": "94/1963, Funkcia opatrovníka je čestnou povinnosťou."
            },
            "36/2005": {
                "zakonnik": "36/2005",
                "distance_from_surrounding_text": "0.15579677",
                "version": "20160101.html",
                "provision_text": "36/2005, o rodine a o zmene a doplnení niektorých zákonov, výživné, spoločné ustanover
            }
        }
    }
],
```

Výsledky

Model	Total	Matched	Unmatched	Distance	Window size	Accuracy	Avg_Ann_to_Surr	Med_Ann_to_Surr	Avg_Chos_to_Surr	Med_Chos_to_Surr	Avg_Ann_to_Chos	Med_Ann_to_Chos	
all-MiniLM-L6-v2	132	80	52	cosine	800	60,60606061	0,327162382	0,34336418	0,298077145	0,308030395	0,073831754	0,06124678	
paraphrase-multilingual-MiniLM-L12-v2	132	97	35	cosine	800	73,48484848	0,399497231	0,361825885	0,373926797	0,3483851	0,096437067	0,0884207	
distiluse-base-multilingual-cased-v2	132	87	45	cosine	800	65,90909091	0,597916833	0,580486025	0,565329815	0,56241157	0,095588588	0,08853113	
kinit/slovakbert-sts-stsb	132	93	39	cosine	800	70,45454545	0,312877411	0,299257395	0,297329678	0,27970916	0,052623097	0,02573776	

Porovnanie s nemeckým prístupom – Darji et al. (2023)

A Dataset of German Legal Reference Annotations:

- Fine-tuned German BERT
- Klasifikácia do 20 najcitovanejších paragrafov
- F1 ≈ 0.9
- Obmedzenie na 20 tried
- Potreba rozsiahlej anotovanej množiny
- Iný jazykový kontext

Naše riešenie:

- Neobmedzený počet identifikátorov
- Nevyžaduje rozsiahly tréningový dataset
- Využíva LLM + pravidlá
- Accuracy ≈ 73%

Postup práce

Vytvorenie metódy na extrakciu odkazov z Európskeho súdneho dvora jednou iteráciou

Iteratívne vylepšovanie metódy pridávaním rozširujúcich parametrov a vylepšením promptu

Doplnenie sady anotovaných rozhodnutí a porovnanie úspešnosti pre rôzne sady parametrov

Upravenie pôvodnej metódy pre extrakciu odkazov z Európskeho súdu pre ľudské práva

Porovnanie výsledkov

Extrakcia odkazov na Európsky súdny dvor

```
Annotated:
rozsudok Súdneho dvora EÚ zo dňa 6. októbra 1982 vo veci CILFIT, C-283/81
uznesenie Súdneho dvora EÚ zo dňa 16. novembra 2010, C-76/10
```

ChatGPT-5

```
{
  "odkazy": [
    {"odkaz":"rozsudok Súdneho dvora EÚ zo dňa 6. októbra 1982, CILFIT, C-283/81, body 9 a 10"},
    {"odkaz":"uznesenie Súdneho dvora EÚ zo dňa 16. novembra 2010, C-76/10, bod 79"}
  ]
}
```

Llama 3.3

```
{
  "odkazy": [
    {
      "text_odkazu": "§ 3 ods. 9 zák. č. 757/2004 Z.z. o súdoch, v znení účinnom od 1.5.2011"
    }
  ]
}
```

Llama 4

```
{
    "odkazy": [
        {
            "text_odkazu": "Nariadenie (ES) č. 805/2004 - Exekučné konanie v Európskej únii"
        }
    ]
}
```

Annotated:
rozsudok SD EÚ zo dňa 5.11.2019, Komisia/Poľsko (Nezávislosť všeobecných súdov), C-192/18, EU:C:2019:924
rozsudky SD EÚ zo dňa 25.7.2018, Minister for Justice and Equality (Nedostatky súdneho systému), C-216/18 PPU
zo dňa 24.6.2019, Komisia/Poľsko (Nezávislosť Najvyššieho súdu), C-619/18
rozsudok SD EÚ z 15. januára 2013,C-416/10, Križan a iní proti Slovenskej inšpekcii životného prostredia
rozsudok SD EÚ zo 6. októbra 2021, C-561/19, Consorzio Italian Management e Catania Multiservizi a Catania Multiservizi
Súdny dvor zaoberal práve v rozsudku C-585/18, C-624/18 a C-625/18
Súdny dvor aj v rozsudku C-824/18
rozsudok C-585/18, C-624/18 a C-625/18
rozsudku C-824/18
rozsudok C-561/19

ChatGPT-5

```
{
  "odkazy": [
    {
      "odkaz": "rozsudok SD EÚ zo dňa 5.11.2019, Komisia/Poľsko (Nezávislosť všeobecných súdov), C-192/18, EU:C:2019:924, bod 101 a citovaná judikatúra"
    },
    {
      "odkaz": "rozsudky SD EÚ zo dňa 25.7.2018, Minister for Justice and Equality (Nedostatky súdneho systému), C-216/18 PPU, EU:C:2018:586, bod 63 a citovaná judikatúra, ako aj zo dňa
                24.6.2019, Komisia/Poľsko (Nezávislosť Najvyššieho súdu), C-619/18, EU:C:2019:531, bod 121 rozsudku"
    },
    {
      "odkaz": "rozsudok SD EÚ z 15. januára 2013,C-416/10, Križan a iní proti Slovenskej inšpekcii životného prostredia, ods. 72"
    },
    {
      "odkaz": "rozsudok SD EÚ zo 6. októbra 2021, C-561/19, Consorzio Italian Management e Catania Multiservizi a Catania Multiservizi, bod 33"
    }
  ]
}
```

Llama 3.3

```
{
  "odkazy": [
    {
      "text_odkazu": "rozsudku C-585/18, C-624/18 a C-625/18"
    },
    {
      "text_odkazu": "rozsudku C-824/18"
    }
  ]
}
```

Llama 4

```
{
    "odkazy": [
        {
            "text_odkazu": "rozhodnutie Najvyššieho správneho súdu SR"
        }
    ]
}
```

Vytvorenie metódy na extrakciu odkazov z Európskeho súdneho dvora

Vstup: text slovenského súdneho rozhodnutia
Výstup: štruktúrovaný JSON obsahujúci zoznam odkazov

Kľúčové prvky:

- Presne definovaný výstupný formát (text odkazu bez úprav)
- Nastavenie deterministického režimu (temperature = 0)
- Extrakcia všetkých výskytov vrátane opakovaných

Systematické vylepšovanie

- Veľkosti úsekov (100 – 3000 znakov, whole, by paragraph)
- Veľkosti kontextového okna
- Typu modelov
- Typy príkladov (pozitívne/negatívne)
- Teploty
- Reťazového myslenia
- Typy promptov
  - System
  - Kontext
  - Rola

Metódy využité pri extrakcii odkazov na Európsky súdny dvor

Chunking: Rozdelenie dlhých rozhodnutí na menšie úseky (100 – 3000 znakov) pre elimináciu straty pozornosti modelu

- Vlastný algoritmus na delenie viet (ošetrenie skratiek ako čl., ods.)

LLM Extrakcia: Využitie Llama 3.3

- Few-shot
- Definovaná JSON schéma
- Teplota, veľkosť kontextového okna,

Post-processing a validácia:

Fuzzy matching: Spätné priradenie extrahovaného textu k originálu

Regex validácia: Kontrola ID rozhodnutia (veci) a ECLI kódov a prekryvu + odstránenie duplikátov

Sekundárne delenie: Opätovné spracovanie "zlepených" odkazov ďalšou iteráciou modelu.

Vyhodnotenie

- True Positives
- False Positives
- False Negatives
- Precision, Recall, F1

	llama3.3	oss-gpt120b	llama 4	oss-gpt
size	42GB	65GB	67GB	13GB
params	70.6B	116.8B	108.6B	20.9B
context length	131 072	131 072	10 485 760	131 072
embedding length	8192	2880	5120	2880

Model	Chunk size	Acc	Recall	F1	Precision	Time	False neg	False pos	True Positives	Total amount of annotated
Llama 3.3	500	1	1	1	1	5:21	0	0	20	20
gpt-oss:120b	3000	1	1	1	1	2:24	0	0	20	20
Llama 4	1000	0.9	0.9	0.947	1	3:20	2	0	18	20
gpt-oss	100	0.85	0.85	0.919	1	12:37	3	0	17	20

Analýza chýb

Najčastejšie problémy:

- Spojenie viacerých odkazov do jedného
- Skrátenie textu odkazu
- False positives
- Pokles presnosti pri spracovaní celého dokumentu

Riešenia:

- Zavedenie split mechanizmu
- Fuzzy matching + kontrola ID
- Deterministické nastavenie modelu

Porovnanie s existujúcim pravidlovým systémom

Pravidlový systém:

- Presnosť ~ 81 %
- Zlyhávanie pri variáciách zápisu
- Citlivosť na formát

Kombinovaný prístup:

- Pravidlá + LLM + validácia ID
- Najvyššia stabilita
- Presnosť: až 1.0
- Výrazne vyšší recall pri optimalizovanom chunkovaní
- Robustnosť voči štylistickým variáciám

Úprava pôvodnej metódy pre extrakciu odkazov z Európskeho súdu pre ľudské práva

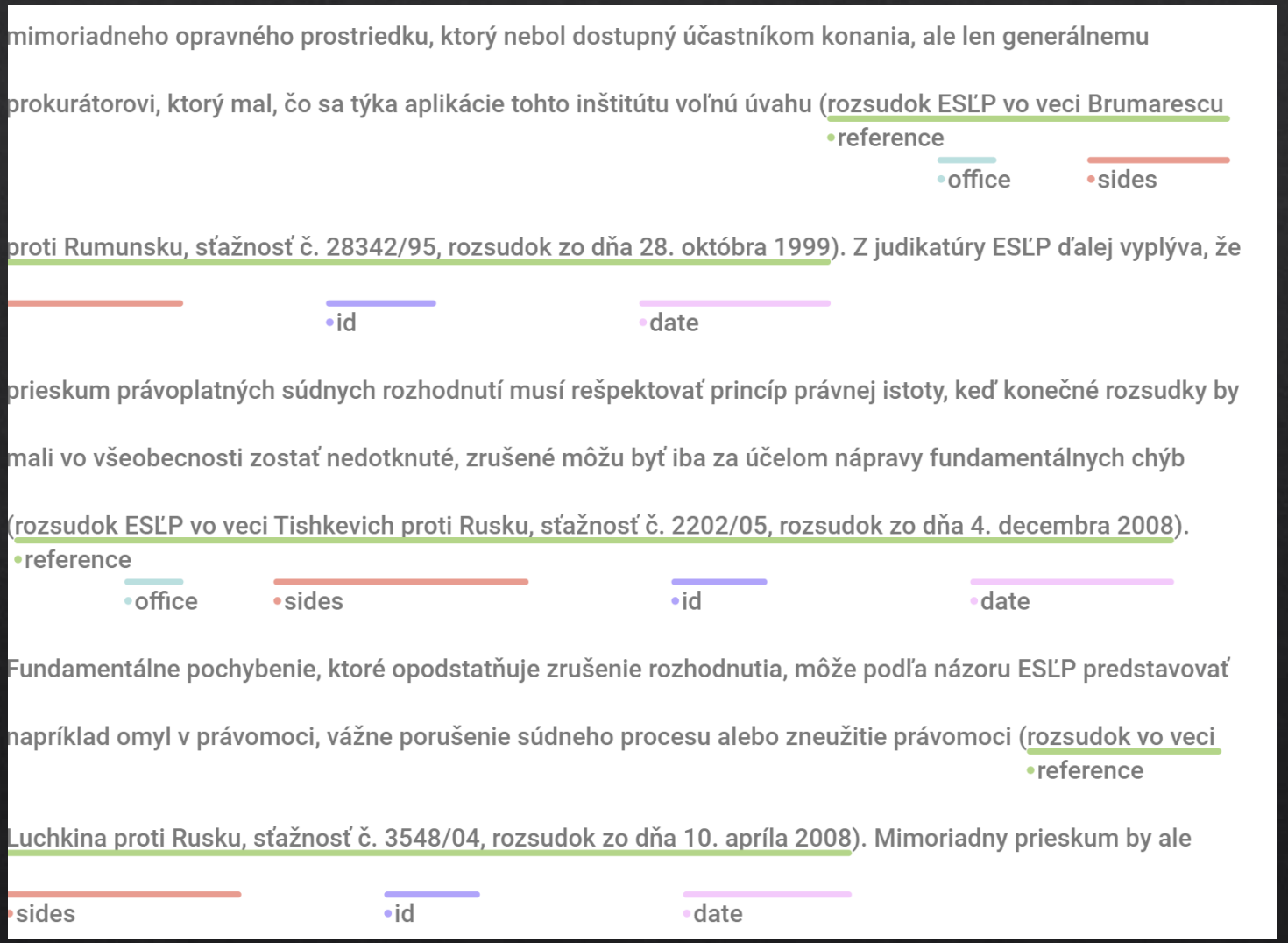
- Analýza rozdielov v štruktúre citácií
- Úprava kľúčových slov a identifikátorov
- Rozšírenie regex výrazov – nie vždy uvádzané ID
- Úprava promptu podľa typických vzorov ECHR

Rozdiely:

- Odlišné formátovanie identifikátorov
- Iné štylistické vzory citovania
- Častejšie implicitné odkazy (Borsa vs. Slovensko)

```
False negatives:  27
True positives:  78
False positives:  22
Total amount of annotated references:  77
Amount of decisions:  6
-------------
Accuracy 0.6141732283464567
Recall 0.7428571428571429
F1 0.7609756097560977
Precision 0.78
```

Anotovanie súdnych rozhodnutí



Literatúra

1. Dávid Varga, et al. „Extraction of Legal References from Court Decisions“. Vol. 3498, 2023, pp. 89–95,

2. Darji Harshil, Jelena Mitrović, and Michael Granitzer. "A Dataset of German Legal Reference Annotations." Proceedings of the Nineteenth International Conference on Artificial Intelligence and Law. 2023.

3. Elena Leitner, Georg Rehm and Julian Moreno-Schneider. “A Dataset of German Legal Documents for Named Entity Recognition“. Proceedings of the 12th Language Resources and Evaluation Conference (LREC 2020)

4.Malte Ostendorff, Till Blume, and Saskia Ostendorff. 2020. „Towards an open platform for legal information“. In Proceedings of the ACM/IEEE Joint Conference on Digital Libraries in 2020. 385–388.