Extrakcia právnych odkazov pomocou modelov transformérov

Analýza a návrh riešenia

Autor: Bc. Nicol Fedurcová

Vedúci: RNDr. Peter Gurský, PhD.

Konzultant: RNDr. Dávid Varga

Pracovisko: Ústav informatiky

Úvod

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 extraktor odkazov môžu byť nápomocnými čiastkovými

krokmi k dosiahnutiu automatizovaného spracovania súdnych rozhodnutí s víziou umožnenia prehľadného vyhľadávania v súdnych rozhodnutiach.

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.

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.

Neúspešné extrakcie

Prvý krok našej práce spočíval v oboznámení sa s pravidlovým systémom a dôvodmi neúspešných extrakcií. Z predošlej činnosti tímu bolo zostavených 20 rozhodnutí, ktoré boli manuálne anotované a slúžili na testovanie doterajších metód. Nad každým z týchto rozhodnutí sme spustili pravidlový systém a analyzovali sme dôvody neúspešných extrakcií. Analyzovali sme prípady:

- **false negatives** - keď nejaký z odkazov mal byť nájdený, ale systém ho nenašiel
- **false positives** - keď nejaký z odkazov bol nájdený, ale nemal byť

Zozbierané poznatky sa nachádzajú v tabuľke 1 na nasledujúcej strane. Zahrnuli sme tam tie prípady, ktoré sa viackrát opakovali alebo boli niečím výrazné alebo charakteristické. Prvok je daný extrahovaný/neextrahovaný odkaz. Stav hovorí o tom, či išlo o false negative alebo false positive prípad. V stĺpci kontext uvádzame tiež to, ako sa odkazy vyskytovali v texte. V stĺpci dôvod sa nachádza pravdepodobný dôvod zlyhania extrakcie.

prvok	stav	kontext	dôvod
/SK/ZZ/1995/233/#paragraf-243k; 243j	false negative	v § 243i až 243k neustanovuje	rozsah bez zopakovania §
/SK/ZZ/1963/99/#paragraf-205.odsek-2	false negative	(§205ods.2\n O.s.p.).\n	znaky nových riadkov
/SK/ZZ/1964/40/#paragraf-18.odsek-1	false negative	\n \n Podľa §18 ods.1 Občianskeho zákonníka,	znaky nových riadkov
/SK/ZZ/2005/301/#paragraf-371.odsek-1.pismeno-c	false negative	za podmienky uvedenej v § 371 ods. 1 písm. c/ Tr. poriadku.	neidentifikovalo Tr. poriadku ako 2005/301
/SK/ZZ/2005/36/#paragraf-26	false negative	Ustanovenia § 24, § 25 a 26 sa použijú primerane.	rozsah bez zopakovania §
/SK/ZZ/1963/99/#paragraf-174.odsek-3	false negative	neoprávnenou osobou (§ 174 ods. 3 O.s.p.).	neidentifikovalo O.s.p
/SK/ZZ/2005/300/#paragraf-36.pismeno-a	false positive	v zmysle § 36 písmeno l), § 36 písmeno n) a § 37 písmeno h)	identifikovalo a ako súčasť rozsahu
/SK/ZZ/2015/160/#paragraf-430	false negative	na podanie dovolania. (§ 430 CSP).\n \n \n	znaky koncov riadkov + neidentifikovalo CSP
/SK/ZZ/1963/97/#paragraf-11	false negative	konania (§ 10 a 11 zák. č. 97/1963 Zb.	rozsah bez zopakovania §
/SK/ZZ/2015/160/#paragraf-364	false negative	odvolania (§ 364 CSP).\n \n \n	znaky koncov riadkov + neidentifikovalo CSP
/SK/ZZ/1963/99/#paragraf-75.odsek-5.veta-4	false negative	Podľa § 75 ods. 5 veta druhá, štvrtá Občianskeho súdneho poriadku	číslovka slovom
/SK/ZZ/1963/99/#paragraf-115.odsek-2	false positive	Podľa § 115 a) ods. 2 O.s.p. pojednávanie	neidentifikovalo 115a
/SK/ZZ/1963/99/#paragraf-200.odsek-1	false positive	Podľa § 200 ea) ods. 1 O.s.p.	neidentifikovalo 200ea
/SK/ZZ/2015/160/#paragraf-29	false positive	podľa § 29 ods. 1 C.m.p. zastavil.	neidentifikalo C.m.p. + označenie čísla rímskou číslicou
/SK/ZZ/1992/71/#paragraf-10.odsek-1	false negative	Podľa § 10 ods. 1 cit. zákona ak nebol	nedoplnilo citovaný zákon
/SK/ZZ/2015/160/#paragraf-244	false negative	Podľa § 243 až § 245 CSP,	neidentifikovalo rozsah
/SK/ZZ/1995/233/#paragraf-57.odsek-1.pismeno-a	false negative	podľa § 57 ods. 1 písm. c) a h) Exekučného	identifikovalo a ako súčasť rozsahu

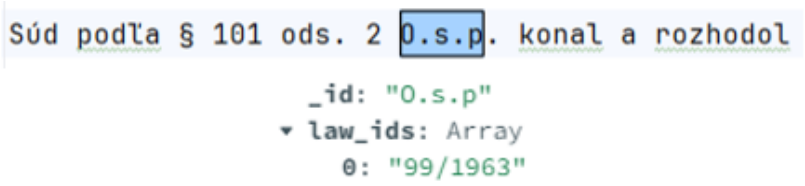
Tabuľka 1: Poznatky z neúspešných extrakcií

Rozhodovanie pri viacerých možnostiach aliasu

Jeden z prípadov kedy extrakcia odkazov zlyháva je prípad, že v rozhodnutí je použitá skratka, ktorá v ňom nie je vysvetlená, resp. jej vysvetlenie nebolo zachytené systémom. Deje sa tak často v prípade, že v súdnickom žargóne je nejaká skratka očividná a tak jej vysvetlenie nie je nutné.

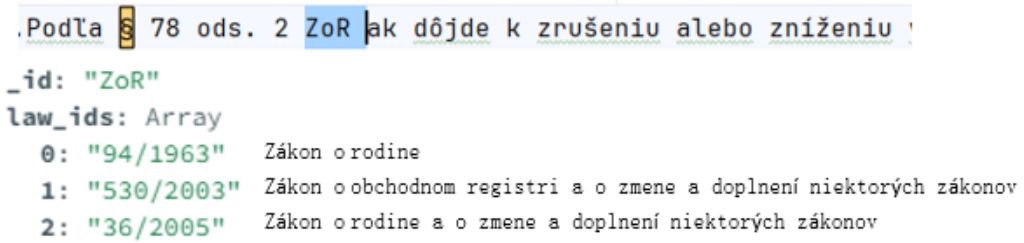
Prípady

Pre skratky ktoré majú jednoznačne priraditeľný zákon sa v systéme zasubstituuje vysvetlenie tejto skratky, ako je napríklad vidieť na obr. 1.



Obr. 1: Príklad použitia skratky s jednoznačne priraditeľným zákonom

Problémom sú však skratky, ku ktorým nie je možné jednoznačne priradiť odpovedajúci zákon. Môže sa tak stať napríklad vtedy, keď zákony majú rovnaké iniciálky. Príklad takéhoto prípadu sa nachádza na obr. 2



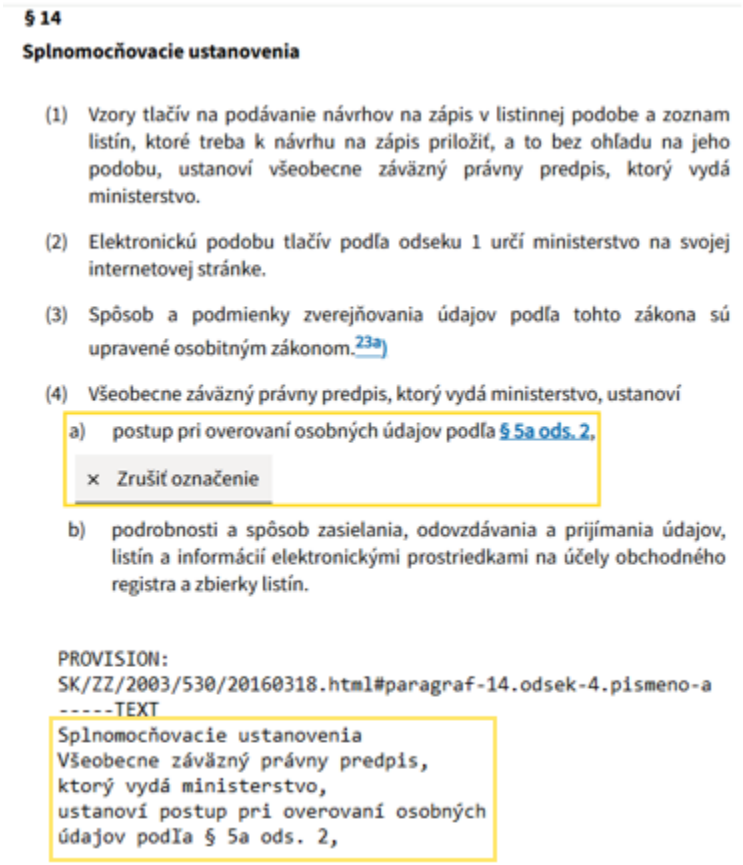
Obr. 2: Príklad použitia skratky s nejednoznačne priraditeľným zákonom

Vysvetlenie danej skratky záleží od kontextu v akom je použitá a teda otvára priestor na riešenie pomocou zahrnutia modelov transformérov. Pred vytvorením samotnej metódy sme však museli urobiť niekoľko krokov, ktoré nám jej použitie umožnia.

Provision retriever

Na to, aby sme mohli rozhodovať, ktorá z možností je na základe kontextu v ktorom sa odkaz na paragraf zákona používa potrebujeme vedieť sémantiku daného písmena/odseku/paragrafu/ zákona.

Na to sme vytvorili tzv. Provision retriever, ktorý pre konkrétny odkaz dokáže nájsť znenie písma/odseku/paragrafu a vrátiť vo forme prirodzeného textu, teda spojením elementu s jeho rodičovskými elementami až po nadpis paragrafu. Výsledkom je prirodzene znejúci odsek textu, ktorý poskytuje informáciu o tom, čoho sa daný odkaz týka. Príklady výstupu Provision retrievera pre konkrétny odkaz je možné vidieť na obr. 3.



Obr. 3: Príklad výstupu Provision retrievera pre konkrétny odkaz

Neskôr v ďalších moduloch sme k tomuto prirodzenému textu pridali aj nadpisy vyšších úrovní až po názov zákona v prípade, že sa dané nadpisy nachádzajú v databáze (v kolekcii `paragraf_texts`), ktorá je výsledkom práce, ktorá predchádzala práci vykonanej na tejto diplomovej práci.

Multichoice finder

Na preskúmanie tohto problému sme najprv potrebovali získať zbierku rozhodnutí, kde sa prípady s nejednoznačne priraditeľným zákonom nachádzajú. Vytvorili sme tzv. Multichoice finder, ktorý dokáže identifikovať takéto rozhodnutia.

Na jeho vytvorenie bolo potrebné čiastočne modifikovať pôvodný pravidlový systém tak, aby sme získali prístup k tomu na základe čoho bolo pri identifikácii odkazu vybrané číslo zákona. Modifikovali sme teda výstupný formát extrahovaných paragrafov pridaním parametrov: `extracted_from, alias_options` a naplnením parametra `aliases`.

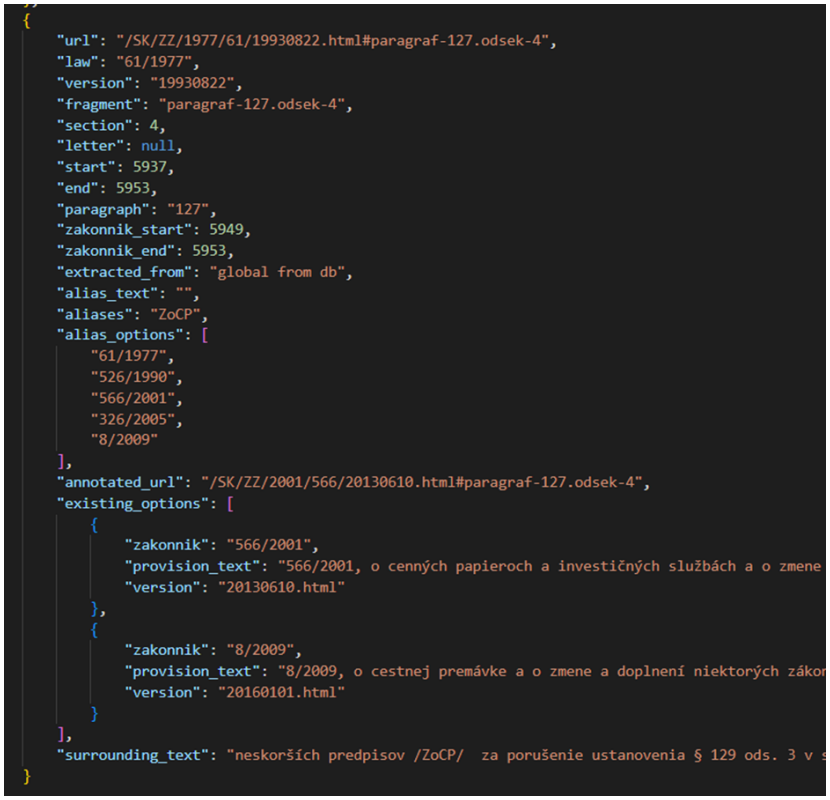
Na vstup tejto metódy sme dali 3500 súdnych rozhodnutí, z ktorých bolo na výstupe identifikovaných 132 prípadov, kedy došlo k výberu z viacerých možností pri aliase. Teda ide o situácie ako tá, ktorá je zobrazená na obr.2

Test set maker

Následne sme vybrané rozhodnutia spracovali pomocou tzv. Test set maker-a. Pre každé z vybraných rozhodnutí sme do testovacej sady zahrnuli tie extrahované odkazy, pre ktoré je potrebné vybrať medzi viacerými možnosťami pre interpretovanie skratky zákona. Každý z takýchto paragrafov má v testovacej sade päť pre nás kľúčových parametrov:

- **aliases**: o akom aliase sa rozhoduje
- **alias_options**: aké zákony prislúchajú danému aliasu
- **annotated_url**: aký je reálny odpovedajúci zákon/paragraf/odsek/písmeno spolu s verziou vzhľadom na čas vydania rozhodnutia, ktorý sme ručne doplnili po anotácii daného rozhodnutia
- **existing_options**: spomedzi možností pre interpretáciu aliasu, ktoré sú také možnosti, že v nich existuje daný paragraf/odsek/písmeno a zároveň informácie o danej možnosti:
  - **zakonik**: číslo zákona
  - **provision_text**: text zákona spolu s nadpismi až po názov zákona (ak sú k dispozícii)
  - **version**: verzia zákona, v ktorej existuje daný paragraf/odsek/písmeno a je najnovšia vzhľadom na dátum vydania rozhodnutia
- **surrouding_text**: okolie odkazu v súdnom rozhodnutí

Na obr. 4 sa nachádza príklad `.json` objektu so všetkými parametrami, ktoré sa preň v testovacej sade nachádzajú.



Obr. 4: Príklad jedného objektu odkazu zo súdneho rozhodnutia v testovacej sade

Vhodné modely

Vektorová reprezentácia viet

Vektorová reprezentácia viet predstavuje neštruktúrovaný text transformovaný na numerické vektory v mnohorozmernom priestore. Tieto vektory sú navrhnuté tak, aby efektívne kódovali sémantický a syntaktický význam celých viet. Vety, ktoré sú si významovo podobné, by mali byť v tomto mnohorozmernom vektorovom priestore umiestnené blízko seba, čo umožňuje kvantifikovať a porovnávať ich sémantickú podobnosť. Vektorová reprezentácia slov mapuje jednotlivé slová na vektory, ktoré kódujú ich význam. Vektorová reprezentácia viet sa snaží zachytiť význam celej vety ako jednu koherentnú a komplexnú jednotku.

Proces tvorby vektorovej reprezentácie viet v Transformér modeloch zahŕňa niekoľko kľúčových krokov: počiatočné kódovanie tokenov a pozičných informácií, iteratívnu kontextualizáciu pomocou viacnásobného samooznačovania, ktorá dynamicky váži vzťahy medzi slovami, a následnú agregáciu týchto kontextuálnych

informácií do jedného hustého vektora, často prostredníctvom špeciálneho tokenu alebo priemerovania.

Pri porovnávaní textu zákona s kontextom v súdnom rozhodnutí je kľúčové vybrať modely, ktoré dokážu efektívne zachytiť sémantickú podobnosť medzi vetami. Nasledujúce modely generujú pomerne kvalitné vetné vektory, ktoré sú následne využívané našou metódou Choice checker na výpočet toho, ako veľmi sú si dva texty podobné. Choice checker potom určí, ktorý odkaz na zákon najlepšie zodpovedá danému textu.

All-MiniLM-L6-v2

Model All-MiniLM-L6-v2 (alebo aj mini BERT) pochádza z renomovanej rodiny modelov SentenceTransformers a je špeciálne navrhnutý pre úlohy zisťovania podobnosti viet. Hoci bol trénovaný primárne na anglických dátach, jeho architektúra s iba šiestimi transformer vrstvami zaisťuje vynikajúcu rovnováhu medzi kompaktnosťou, rýchlosťou a presnosťou. Práve preto pre nás slúžil ako veľmi vhodný kandidát na štartovací model, s ktorým sme pracovali na začiatku pre jeho rýchlosť.

Aj napriek tomu, že bol trénovaný primárne na dátach v anglickom jazyku, dokázal prekvapivo robustne reprezentovať texty v slovenčine. Pri porovnávaní slovenských právnych textov dokázal efektívne generovať kvalitné vetné vektory, ktoré sú kľúčové pre metódu Choice checker.

Paraphrase-multilingual-MiniLM-L12-v2

Paraphrase-multilingual-MiniLM-L12-v2 je multilingválny model schopný generovať vetné vektory pre viac ako 50 jazykov vrátane slovenčiny. Tento model bol špeciálne trénovaný na úlohách parafrázovania a identifikácie podobnosti viet, preto sme ho vybrali ako jeden z kandidátnych modelov pre porovnávanie znenia zákona so súvisiacim kontextom v súdnom rozhodnutí. Jeho architektúra MiniLM s dvanástimi vrstvami priniesla vyššiu reprezentatívnu silu v porovnaní s modelmi s menšou hĺbkou.

Distiluse-base-multilingual-cased-v2

Model Distiluse-base-multilingual-cased-v2 predstavuje distilovanú verziu viacjazyčného modelu USE (Universal Sentence Encoder). Táto distilácia znamená, že model je menší a rýchlejší pri spracovaní, pričom si zachováva vysokú úroveň porozumenia medzi-jazykovým podobnostiam. Podporuje viac ako 15 jazykov, vrátane slovenčiny čo poskytuje značnú výhodu.

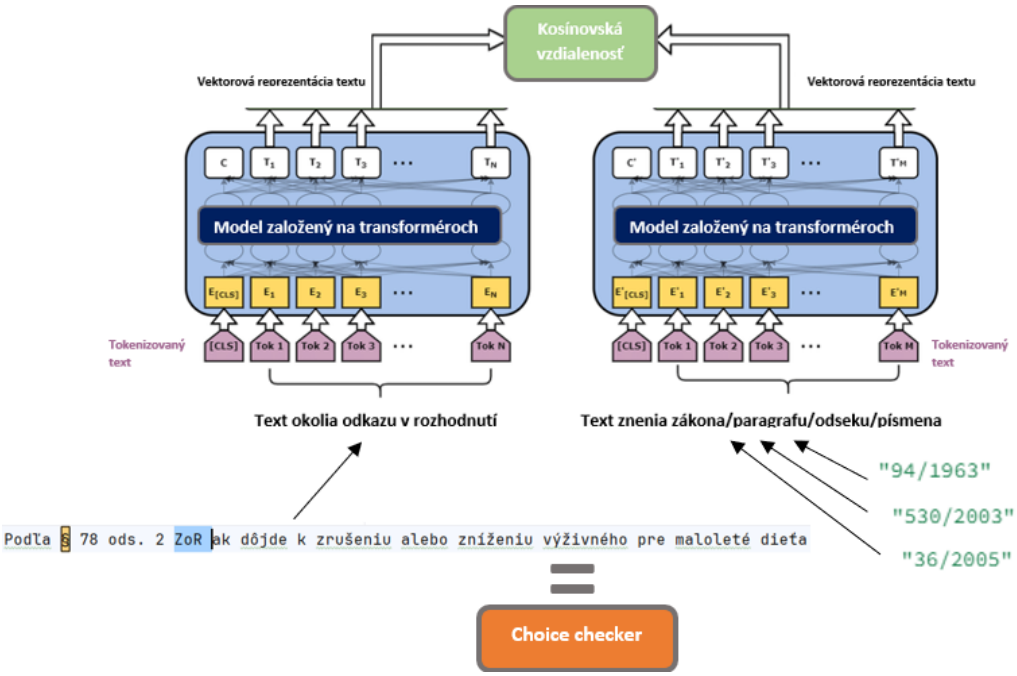
Kinit/slovakbert-sts-stsb

Kinit/slovakbert-sts-stsb je slovenský transformer model založený na architektúre BERT, špecificky doladený na úlohu zisťovania sémantickej podobnosti viet (STS – Semantic Textual Similarity). Bol trénovaný na dátach z úlohy STS-Benchmark prispôsobených pre slovenský jazyk. Model je optimalizovaný práve na porovnávanie významu viet v slovenčine, a preto je mimoriadne vhodný pre našu prácu. Dokáže presnejšie rozlíšiť jemné významové odtiene v prirodzenom jazyku vďaka tomu, že jeho embeddingy sú vysoko špecifické pre sémantickú podobnosť slovenských textov.

Choice checker

S využitím vyššie uvedených modelov sme následne vytvorili tzv. Choice checker, ktorý na základe kosínovskej vzdialenosti medzi vektorovou reprezentáciou textu okolia odkazu v rozhodnutí a vektorovou reprezentáciou znenia zákona/paragrafu/odseku/písmena určí vzdialenosti týchto dvoch elementov.

Tento krok zopakuje pre každú možnú interpretáciu použitej skratky, čím získame vzdialenosti, z ktorých možnosť s najmenšou hodnotou predstavuje text, ktorý by mal byť sémanticky najbližší k zneniu zákona. Na obr. 5 je zobrazený proces fungovania Choice checkera.



Obr. 5: Vizualizácia činnosti Choice checkera

Výstup, ktorý generuje Choice checker je spolu s testovacou sadou (výstup Test set maker-a) vstupom pre tzv. Comparer. Ten následne na základe anotovanej správnej možnosti porovná v koľkých prípadoch Choice checker správne určil ktorý zákon je správnou možnosťou. Výstup Choice checkera je možné vidieť na obr.6.

Obr. 6: Príklad objektu z výstupu Choice checkera

Výsledky

Po vytvorení všetkých podporných modulov sme otestovali Choice checker naprieč štyrmi modelmi vymenovanými v časti Vhodné modely. Testovali sme aj to, ako veľkosť okolia textu vplýva na presnosť (accuracy). Výsledky z testovania je možné vidieť v tabuľke 2.

Tabuľka obsahujetieto stĺpce:

1. Model: názov modelu
2. Total: počet všetkých testovaných prípadov
3. Matched: koľkokrát Choice checker s využitím daného modelu správne vybral spomedzi ponúkaných možností
4. Unmatched: koľkokrát Choice checker s využitím daného modelu nesprávne vybral spomedzi ponúkaných možností
5. Distance: metrika použitá na meranie vzdialenosti (podobnosti) dvoch textov

6. Window size: veľkosť okolia odkazu v počtoch znakov (spolu sčítané pred aj za odkazom)

7. Accuracy: dosiahnutá presnosť

8. Avg_Ann_to_Surr: priemerná vzdialenosť medzi vektorom textu zákona, ktorý bol anotovaný ako správna možnosť a vektorom textu okolia odkazu z rozhodnutia

9. Med_Ann_to_Surr: stredová vzdialenosť medzi vektorom textu zákona, ktorý bol anotovaný ako správna možnosť a vektorom textu okolia odkazu z rozhodnutia

10. Avg_Chos_to_Surr: priemerná vzdialenosť medzi vektorom textu zákona, ktorý bol Choice checkerom vybraný ako správna možnosť a vektorom textu okolia odkazu z rozhodnutia

11. Med_Chos_to_Surr: stredová vzdialenosť medzi vektorom textu zákona, ktorý bol Choice checkerom vybraný ako správna možnosť a vektorom textu okolia odkazu z rozhodnutia

12. Avg_Ann_to_Chos: priemerná vzdialenosť medzi vektorom textu zákona, ktorý bol anotovaný ako správna možnosť a vektorom textu zákona, ktorý bol Choice checkerom vybraný ako správna možnosť

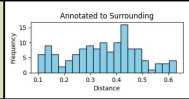
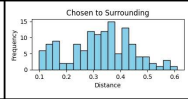
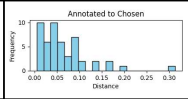
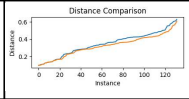
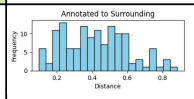
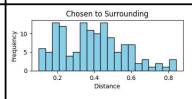
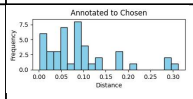
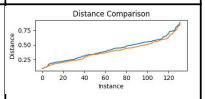
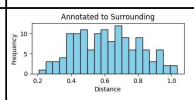
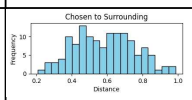
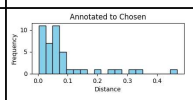
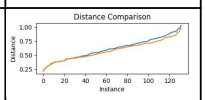
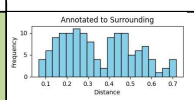
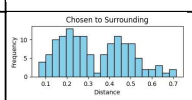
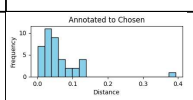
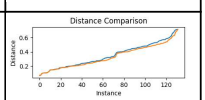
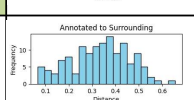
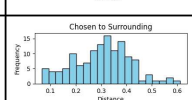
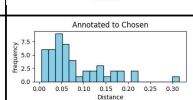
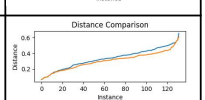
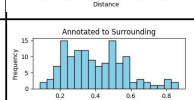
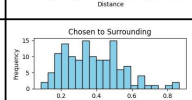
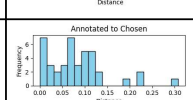
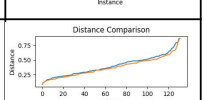
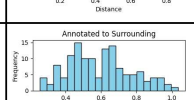
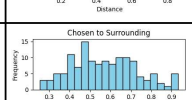
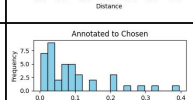
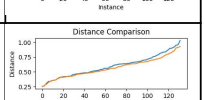
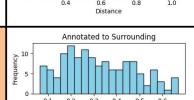
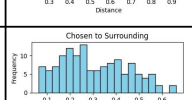
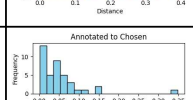
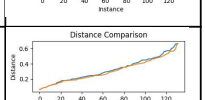
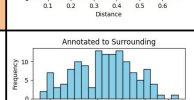
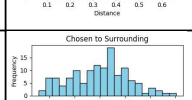
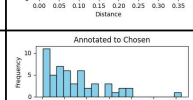
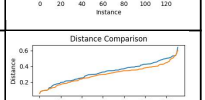
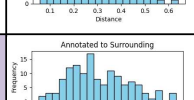
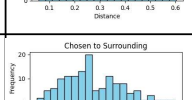
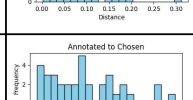
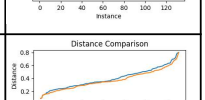
13. Med_Ann_to_Chos: stredová vzdialenosť medzi vektorom textu zákona, ktorý bol anotovaný ako správna možnosť a vektorom textu zákona, ktorý bol Choice checkerom vybraný ako správna možnosť

14. plot Annotated to Surrounding: zobrazenie rozloženia frekvencie výskytov vzdialeností medzi vektorom textu zákona, ktorý bol anotovaný ako správna možnosť a vektorom textu okolia odkazu z rozhodnutia

15. plot Chosen to Surrounding: zobrazenie rozloženia frekvencie výskytov vzdialeností medzi vektorom textu zákona, ktorý bol choice checkerom vybraný ako správna možnosť a vektorom textu okolia odkazu z rozhodnutia

16. plot Annotated to Chosen: zobrazenie rozloženia frekvencie výskytov vzdialeností medzi vektorom textu zákona, ktorý bol anotovaný ako správna možnosť a vektorom textu zákona, ktorý bol choice checkerom vybraný ako správna možnosť

17. plot Distance Comparison (orage=chosen:surr, blue=annot:surr): porovnanie usporiadaných vzdialeností medzi vektorom textu zákona, ktorý bol choice checkerom vybraný ako správna možnosť a vektorom textu okolia odkazu z rozhodnutia (oranžovou) a vzdialeností medzi vektorom textu zákona, ktorý bol anotovaný ako správna (modrou)

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	plot: Annotated to Surrounding	plot: Chosen to Surrounding	plot: Annotated to Chosen	plot: Distance Comparison (orage=chosen:surr, blue=annot:surr)
all-MiniLM-L6-v2	132	82	50	cosine	400	62,121	0,348	0,360	0,323	0,332	0,065	0,048				
paraphrase-multilingual-MiniLM-L12-v2	132	90	42	cosine	400	68,182	0,422	0,419	0,392	0,389	0,092	0,082				
distiluse-base-multilingual-cased-v2	132	89	43	cosine	400	67,424	0,613	0,612	0,585	0,581	0,086	0,053				
kinit/slovakbert-sts-stsb	132	92	40	cosine	400	69,697	0,348	0,319	0,330	0,292	0,059	0,042				
all-MiniLM-L6-v2	132	84	48	cosine	600	63,636	0,334	0,339	0,304	0,313	0,081	0,064				
paraphrase-multilingual-MiniLM-L12-v2	132	91	41	cosine	600	68,939	0,403	0,381	0,378	0,359	0,082	0,074				
distiluse-base-multilingual-cased-v2	132	92	40	cosine	600	69,697	0,597	0,588	0,568	0,558	0,095	0,064				
kinit/slovakbert-sts-stsb	132	92	40	cosine	600	69,697	0,322	0,299	0,306	0,281	0,052	0,039				
all-MiniLM-L6-v2	132	80	52	cosine	800	60,606	0,327	0,343	0,298	0,308	0,074	0,061				
paraphrase-multilingual-MiniLM-L12-v2	132	97	35	cosine	800	73,485	0,399	0,362	0,374	0,348	0,096	0,088				
distiluse-base-multilingual-cased-v2	132	87	45	cosine	800	65,909	0,598	0,580	0,565	0,562	0,096	0,089				
kinit/slovakbert-sts-stsb	132	93	39	cosine	800	70,455	0,313	0,299	0,297	0,280	0,053	0,026				
all-MiniLM-L6-v2	132	82	50	cosine	1000	62,121	0,326	0,331	0,299	0,303	0,072	0,053				
paraphrase-multilingual-MiniLM-L12-v2	132	85	47	cosine	1000	64,394	0,400	0,361	0,371	0,354	0,082	0,066				
distiluse-base-multilingual-cased-v2	132	81	51	cosine	1000	61,364	0,606	0,583	0,571	0,564	0,091	0,079				
kinit/slovakbert-sts-stsb	132	89	43	cosine	1000	67,424	0,310	0,295	0,293	0,285	0,051	0,038				

Tabuľka 2: priebeh testovania modelov s rôznymi veľkosťami okolia textu odkazu z rozhodnutia

Najvyššiu presnosť 73,485% dosiahol model Paraphrase-multilingual-MiniLM-L12-v2 pri veľkosti okolia odkazu +-400 znakov, teda s okolím o 800 znakoch. Správne určil 97 odkazov z celkových 132.

Druhú najvyššiu presnosť 70,455% dosiahol model Kinit/slovakbert-sts-stsb pri veľkosti okolia odkazu +-400 znakov, teda s okolím o 800 znakoch. Správne určil 93 odkazov z celkových 132.

Ako je vidieť v tabuľke 2, ukázalo sa, že najvyššiu priemernú presnosť dosahoval model Kinit/slovakbert-sts-stsb naprieč všetkými veľkosťami okien a to 69,3%. Model Paraphrase-multilingual-MiniLM-L12-v2 taktiež dosiahol dobrú priemernú presnosť, a to 68,7%. Spomedzi všetkých modelov najviac zaostával model All-MiniLM-L6-v2.

V tabuľke 3 je vidieť, že pri veľkosti okna 600 bola dosahovaná najvyššia priemerná presnosť naprieč všetkými modelmi, hoci dve najvyššie presnosti boli dosiahnuté s veľkosťou okna 800.

Model	**Priemerný počet správne určených**	**Priemerná presnosť (%)**
all-MiniLM-L6-v2	82	62,1
paraphrase-multilingual-MiniLM-L12-v2	90,75	68,7
distiluse-base-multilingual-cased-v2	87,25	66,0
kinit/slovakbert-sts-stsb	91,75	69,3

Tabuľka 2: Výkonnosť jednotlivých modelov

Veľkosť okna	**Priemerná presnosť (%)**
100	84,25
400	88,25
600	89,75
800	89,25

Tabuľka 3: Priemerná presnosť všetkých modelov pri rôznych veľkostiach okna

Možné vylepšenie choice checkera

Je vidieť, že správna možnosť pre výber zákona nie vždy zodpovedá tej možnosti, ktorej vektor sa nachádza najbližšie k vektoru okolitého textu. V nie-

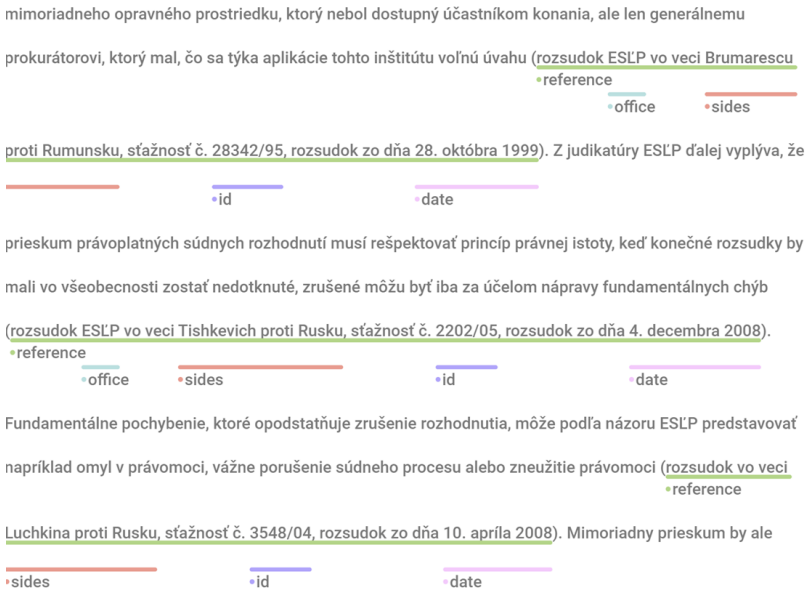
ktorých prípadoch je rozdiel medzi vzdialenosťou textu anotovanej možnosti k okoliu odkazu a vzdialenosťou textu vybranej možnosti veľmi malý, no jeho výsledkom je nesprávny výber. Preto je potrebné pridať do rozhodovacieho procesu dodatočnú informáciu.

Dobrým kandidátom na takúto informáciu je tzv. vektor zvyčajného použitia. Ten by predstavoval vektor vytvorený na základe okolí odkazu v rozhodnutiach spomedzi všetkých štyroch miliónov rozhodnutí. Získali by sme tým informáciu o tom, ako zvyčajne vyzerá kontext, v ktorom sa v rozhodnutiach odkazuje na daný zákon/paragraf/odsek/písmeno. Tento nápad teda predstavuje ďalšiu možnosť budúcej práce.

Okrem toho je možnosťou pre vylepšenie presnosti Choice checkera aj použitie modelu, ktorý je dotrénovaný na slovenských súdnych rozhodnutiach.

Extrakcia odkazov na ECHR

Po práci s vylepšením pravidlového systému je nasledujúcim krokom extrakcia odkazov na Európsky súd pre ľudské práva (ECHR). Na obr. 7 sa nachádza ukážka toho, akú štruktúru majú odkazy na ECHR. S ohľadom na špecifickú štruktúru bude metodika nášho riešenia zahŕňať dotrénovanie slovenského Named-Entity-Recognition modelu. Na dotrénovanie využijeme dataset oanotovaný modelom Llama na základe promptingu.



Obr. 7: Príklady odkazov na Európsky súd pre ľudské práva

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.

5. Hettiarachchi, Hansi et al. (2021) *DAAI at CASE 2021 Task 1: Transformer-based Multilingual Socio-political and Crisis Event Detection*, Proceedings of the 4th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE 2021), Association for Computational Linguistics, pp. 120—130.

6. Rajamanickam, Duraimurugan (2024) *Improving Legal Entity Recognition Using a Hybrid Transformer Model and Semantic Filtering Approach*