

Slovenská technická univerzita v Bratislave

FAKULTA INFORMATIKY A INFORMAČNÝCH TECHNOLÓGIÍ

FIIT-5208-5622

Bc. Filip Bednárik

Extraktcia informácií z textu

Diplomová práca

Študijný program:

Informačné systémy

Študijný odbor:

9.2.6 Informačné systémy

Miesto vypracovania:

Ústav informatiky a softvérového inžinierstva,
FIIT STU, Bratislava

Vedúci práce:

Ing. Marián Šimko, PhD.

Máj 2016

Zadanie diplomovej práce

Meno študenta: **Bc. Filip Bednárik**

Študijný program: Informačné systémy

Študijný odbor: Informačné systémy

Názov práce: **Sémantická anotácia dát pomocou inteligentného zameraného st'ahovača**

Samostatnou výskumnou a vývojovou činnosťou v rámci predmetov Diplomový projekt I, II, III vypracujte diplomovú prácu na tému, vyjadrenú vyššie uvedeným názvom tak, aby ste dosiahli tieto ciele:

Všeobecný cieľ:

Vypracovaním diplomovej práce preukážte, ako ste si osvojili metódy a postupy riešenia relatívne rozsiahlych projektov, schopnosť samostatne a tvorivo riešiť zložité úlohy aj výskumného charakteru v súlade so súčasnými metódami a postupmi študovaného odboru využívanými v príslušnej oblasti a schopnosť samostatne, tvorivo a kriticky pristupovať k analýze možných riešení a k tvorbe modelov.

Špecifický cieľ:

Vytvorte riešenie zodpovedajúce návrhu textu zadania, ktorý je prílohou tohto zadania. Návrh bližšie opisuje tému vyjadrenú názvom. Tento opis je záväzný, má však rámcový charakter, aby vznikol dostatočný priestor pre Vašu tvorivosť.

Riadťe sa pokynmi Vášho vedúceho.

Pokial' v priebehu riešenia, opierajúc sa o hlbšie poznanie súčasného stavu v príslušnej oblasti, alebo o priebežné výsledky Vášho riešenia, alebo o iné závažné skutočnosti, dospejete spoločne s Vašim vedúcim k presvedčeniu, že niečo v texte zadania a/alebo v názve by sa malo zmeniť, navrhnite zmenu. Zmena je spravidla možná len pri dosiahnutí kontrolného bodu.

Miesto vypracovania: Ústav informatiky a softvérového inžinierstva FIIT STU v Bratislave v spolupráci s DATALAN, a.s., Bratislava

Vedúci práce: **Ing. Miroslav Liška, PhD.**

Termíny odovzdania:

podľa harmonogramu štúdia platného pre semester, v ktorom máte príslušný predmet (Diplomový projekt I, II, III) absolvovať podľa Vášho študijného plánu

Predmety odovzdania:

V každom predmete dokument podľa pokynov na www.fiit.stuba.sk v časti:
home > Informácie o > štúdiu > organizácia štúdia > diplomový projekt.

V Bratislave dňa 17. 2. 2014


prof. Ing. Pavol Návrat, PhD.
riaditeľ Ústavu informatiky a softvérového
inžinierstva



Návrh zadania diplomovej práce

Finálna verzia do diplomovej práce¹

Študent:

Meno, priezvisko, tituly: Filip Bednárik, Bc.
Študijný program: Informačné systémy
Kontakt: filip.bednarik@memes.sk

Výskumník:

Meno, priezvisko, tituly: Miroslav Líška, Ing. PhD.

Projekt:

Názov: Sémantická anotácia dát pomocou inteligentného zameraného štahovača
Názov v angličtine: Intelligent focused crawler with semantic data annotation module
Miesto vypracovania: Datalan, a.s.
Oblast problematiky: Vyhľadávanie informácií

Text návrhu zadania²

Internetové stránky sú jedným z najpopulárnejších zdrojov informácií. Väčšina stránok ale poskytuje dôležité informácie len vo forme dokumentov určených pre pochopenie ľuďmi. Pri potrebe automatizácie spracovania takýchto informácií strojom nastáva problém. Aby stroj mohol spracovať informácie s ohľadom na ich význam, potrebuje mať znalosti o danej doméne a zvolať správnu metódu ich extrakcie. Problémom je efektívne získavanie takýchto znalostí zo štrukturovaných dokumentov. Analyzujte spôsoby získavania dát a transformácie vybraných častí dokumentov na znalosti pomocou sémantickej anotácie. Zamerajte sa na oblasť prehliadania webových portálov, získavanie potrebných dokumentov a anotáciu dát nachádzajúcich sa v ňom. Navrhnite metódy anotácie dát, spôsobu vyhľadania dát na stránke a optimalizáciu preliezania štahovača. Vytvorte model spracovania dokumentov na znalosti s pomocou ontológií. Uvažujte použitie riešenia v praxi a prispôsobte podľa toho aj výkon a použiteľnosť riešenia. Skúmajte rôzne spôsoby anotácie dát, vyhľadávanie informácií v dokumente a optimalizáciu štahovača na zameraný štahovač. Experimentujte s použitím riešenia v praxi na konkrétnej doméne. Pri riešení využite technológie vyhľadávania informácií v texte a metódy inteligentného preliezania webových portálov. Overte rýchlosť, kvalitu a presnosť anotácie a prechádzania portálov použitím Vašej navrhnutej metódy v porovnaní s inými metódami.

¹ Vytlačiť obojstranne na jeden list papiera

² 150-200 slov (1200-1700 znakov), ktoré opisujú výskumný problém v kontexte súčasného stavu vrátane motivácie a smerov riešenia

Literatúra³

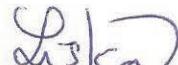
- Huang, Rui and Lin, Fen and Shi, Zhongzhi. Focused Crawling with Heterogeneous Semantic Information. Proceedings of the 2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology - Volume 01. Washington : IEEE Computer Society, 2008, Zv. WI-IAT '08, s. 525--531
- Yung-Chun Chang, Pei-Ching Yang, and Jung-Hsien Chiang. Ontology-Based Intelligent Web Mining Agent for Taiwan Travel. Proceedings of the 2009 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology - Volume 03. Washington : IEEE Computer Society, 2009, Zv. WI-IAT '09, s. 421--424.

Vyššie je uvedený návrh diplomového projektu, ktorý vypracoval(a) Bc. Filip Bednárik, konzultoval(a) a osvojil(a) si ho Ing. Miroslav Líška, PhD. a súhlasi, že bude takýto projekt viesť v prípade, že bude pridelený tomuto študentovi.

V Bratislave dňa 10.2.2014



Podpis študenta

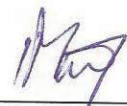


Podpis výskumníka

Vyjadrenie garanta predmetov Diplomový projekt I, II, III

Návrh zadania schválený: áno / nie⁴

Dňa: 14. 2. 2014



Podpis garanta predmetov

³ 2 vedecké zdroje, každý v samostatnej rubrike a s údajmi zodpovedajúcimi bibliografickým odkazom podľa normy STN ISO 690, ktoré sa viažu k téme zadania a preukazujú výskumnú povahu problému a jeho aktuálnosť (uvedte všetky potrebné údaje na identifikáciu zdroja, pričom uprednostnite vedecké príspevky v časopisoch a medzinárodných konferenciách)

⁴ Nehodiace sa prečiarknite

Návrh zadania diplomovej práce

Revízia č.: 1¹

Študent:

Meno, priezvisko, tituly: Filip Bednárik, Bc.
Študijný program: Informačné systémy
Kontakt: filip.bednarik@memes.sk

Výskumník:

Meno, priezvisko, tituly: Marián Šimko, Ing. PhD.

Projekt:

Názov: Extrakcia informácií z textu
Názov v angličtine: Information extraction from text
Miesto vypracovania: Ústav informatiky a softvérového inžinierstva, FIIT STU,
Bratislava
Oblast' problematiky: Vyhľadávanie informácií

Text návrhu zadania²

Aktuálne prebieha informatizácia a elektronizácia spoločnosti. Množstvo dokumentov sa elektronizuje a prechádza sa na elektronické formy dokumentov, aby sa posilnila možnosť rýchleho vyhľadávania, získali sa z textu cenné informácie, ale aj zjednodušila samotná práca s dokumentmi. Veľmi dôležitá je automatizácia spracovania textových dokumentov a automatické získavanie informácií z nich. Problémom je veľké množstvo dokumentov a prenos extrakcie informácií z nich.

Analyzujte metódy pre extrakciu informácií z textu, rozpoznanie pomenovaných entít, číselných hodnôt, dátumov, osôb a adres. Zamerajte sa na oblasť analýzy a spracovania textu v slovenskom jazyku a extrakciu informácií z neštrukturovaných alebo čiastočne štrukturovaných dokumentov. Preskúmajte možnosť využitia ontológií.

Navrhnite metódu pre vybrané úlohy extrakcie informácií. Experimentujte s použitím riešenia v praxi na konkrétnej doméne. Pri riešení využite metódy predspracovania textu a metódy spracovania prirodzeného jazyka. Overte úspešnosť riešenia v porovnaní so štandardnými metódami a ľudským spracovaním na množstve správne extrahovaných informácií.

¹ Vytlačiť obojstranne na jeden list papiera

² 150-200 slov (1200-1700 znakov), ktoré opisujú výskumný problém v kontexte súčasného stavu vrátane motivácie a smerov riešenia

Literatúra³

- Sanchez-Cisneros, D., & Aparicio Gali, F. (2013). UEM-UC3M: An Ontology-based named entity recognition system for biomedical texts. Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013), 622-627
- Cunningham, H. (2006). Information Extraction, Automatic. In Encyclopedia of Language & Linguistics (Vol. 5, pp. 665-677)

Vyššie je uvedený návrh diplomového projektu, ktorý vypracoval(a) Bc. Filip Bednárik, konzultoval(a) a osvojil(a) si ho Ing. Marián Šimko, PhD. a súhlasí, že bude takýto projekt viest.

V Bratislave dňa 8.12.2015



Podpis študenta



Podpis výskumníka

Vyjadrenie garanta predmetov Diplomový projekt I, II, III

Návrh zadania schválený: áno / nie⁴

Dňa: 8. 12. 2015



Podpis garanta predmetov

³ 2 vedecké zdroje, každý v samostatnej rubrike a s údajmi zodpovedajúcimi bibliografickým odkazom podľa normy STN ISO 690, ktoré sa viažu k téme zadania a preukazujú výskumnú povahu problému a jeho aktuálnosť (uvedte všetky potrebné údaje na identifikáciu zdroja, pričom uprednostnite vedecké príspevky v časopisoch a medzinárodných konferenciách)

⁴ Nehodiace sa prečiarknite

ANOTÁCIA

Extrakcia informácií z textu

Študijný program: Informačné systémy

Autor: Bc. Filip Bednárik

Vedúci diplomovej práce : Ing. Marián Šimko, PhD.

Máj 2016

Extrakcia informácií z textov v prirodzenom jazyku, je aktuálnou téμou v sfére informačných technológií. Získavanie informácií z textu, je kľúčové pre vyhľadávanie, štatistiky, ale aj pri automatizovaní úloh, ktoré sa ešte dnes vykonávajú ručne. V tejto práci sa zaoberáme analýzou a spracovaním textu v prirodzenom jazyku, rozpoznávaním pomenovaných entít, extrakciou informácií a ich využitím na anonymizáciu citlivých osobných údajov. Navrhujeme metódu založenú na moderných technológiách, ktorá môže mať signifikantný prínos pre akademickú aj komerčnú sféru. Metóda kladie dôraz na pokrytie a presnosť vďaka zabudovaným pravidlám, ale berie do úvahy aj možné zmeny vďaka využitiu metód strojového učenia.

Kľúčové slová : extrakcia informácií, extrakcia pomenovaných entít, anonymizácia textu, stemovanie, lematizácia

ANNOTATION

Information extraction from text

Degree Course: Information systems

Author: Bc. Filip Bednárik

Supervisor : Ing. Marián Šimko, PhD.

May 2016

Information extraction form text and natural language processing is current problem in information technology. Information gathering from text is key component of search engines, statistics engines and also takes part in solving automatization tasks, which are even today still being manually processed. We introduce method for de-identification of documents using text analysis, natural language processing, named entity recognition and information extraction in this thesis. Furthermore, we propose method based on modern technologies, which could have significant benefits for both academic and business sphere. Our method emphasizes on recall and precision thanks to predefined rules, but also takes changes into account while using machine learning concepts.

Keywords : information extraction, named entity recognition, text de-identification, stemming, lemmatization

Pod'akovanie

Chcel by som sa pod'akovať vedúcemu projektu Ing. Mariánovi Šimkovi, PhD. za odbornú pomoc, informácie a postrehy k mojej práci.

Čestné prehlásenie

Podpísaný Bc. Filip Bednárik týmto vyhlasujem, že som diplomovú prácu s názvom „**Extraktcia informácií z textu**“ vypracoval samostatne s použitím uvedenej literatúry.

Som si vedomý zákonných dôsledkov v prípade, ak hore uvedené údaje nie sú pravdivé.

V Bratislave, máj 2016

..... podpis

Obsah

1	Úvod.....	1
2	Spracovanie prirodzeného jazyka	3
2.1	Tokenizácia	4
2.1.1	Segmentácia slov	4
2.1.2	Segmentácia viet.....	5
2.2	Lematizácia a stemovanie	6
2.3	Určovanie slovných druhov a pádov (POS značkovanie).....	7
2.3.1	Dagger.....	7
2.3.2	Naivný pravidlový korpusový POS tagger	8
2.3.3	TreeTagger.....	8
2.3.4	CRFPOSTagger	8
3	Extrakcia informácií	9
3.1	Koncept	9
3.2	Špecifická slovenského jazyka	10
3.3	Existujúce nástroje pre extrakciu informácií	11
3.3.1	GATE Developer	11
3.3.2	CoreNLP	11
3.3.3	OpenNLP	13
3.3.4	NLTK.....	13
4	Rozpoznávanie pomenovaných entít (NER).....	14
4.1	Metódy podľa prístupu.....	14
4.1.1	Lingvistické metódy	14
4.1.2	Metóda založená na štatistických modeloch.....	15
4.2	Metódy podľa typu extrahejanej entity	18
4.2.1	Všeobecne použiteľné dostupné riešenia	18
4.2.2	Extrakcia dátumov a iných časových údajov	22
4.2.3	Číselné hodnoty	24
4.2.4	Osoby (mená a priezviská)	24
4.2.5	Firmy.....	25
4.2.6	Adresy a geolokačné entity	26
4.2.7	Telefónne čísla a emaily	27
4.2.8	Doménovo špecifické entity	27
4.3	Otvorené problémy a nedostatky technológií	29
5	Extrakcia informácií pre anonymizáciu dát	30
5.1	Anonymizácia - charakteristika problému	30
5.2	Špecifická súdnych rozhodnutí.....	30

5.3 Existujúce riešenia pre anonymizáciu textu.....	32
5.3.1 Deid.....	32
5.3.2 NLM-Scrubber.....	33
5.3.3 MITRE Identification Scrubber Toolkit (MIST)	33
5.3.4 Riešenia pre anonymizáciu v slovenčine	34
5.4 Zhrnutie existujúcich riešení a ich problémy	36
6 Otvorené problémy	37
7 Ciele	38
8 Metóda	40
8.1 Predspracovanie	40
8.1.1 Sanácia dokumentu	40
8.1.2 Tokenizácia	41
8.1.3 Segmentácia viet.....	41
8.1.4 Lematizácia.....	41
8.1.5 Stemovanie.....	41
8.1.6 Označovanie slovných druhov	42
8.2 Rozpoznávanie entít a extrakcia dodatočných informácií	42
8.2.1 Hybridné NER	43
8.3 Automatická anonymizácia.....	45
8.4 Charakteristika dokumentov a doménové obmedzenia	46
9 Implementácia a technológie	48
9.1 Spoločná architektúra pri práci s prirodzeným jazykom	48
9.2 Decrete crawler	49
9.3 Brat.....	49
9.3.1 Konfigurácia nástrojov	50
9.3.2 Konfigurácia anotácie	50
9.3.3 Konfigurácia vizuálnych prvkov	51
9.3.4 Konfigurácia klávesových skratiek.....	52
9.4 ASuR tools	52
9.4.1 Nástroje pre prácu so slovníkmi a anotovanými dokumentmi	52
9.4.2 Nástroje pre trénovanie štatistických modelov	53
9.4.3 Nástroje pre vyhodnocovanie úspešnosti riešení	56
9.5 NLP tools	57
9.5.1 SpaceHighlightingSanitizer	57
9.5.2 PTBTokenizerUtils	57
9.5.3 SSplitUtils	57
9.5.4 LemmatizerUtils	58

9.5.5	SlovakStemmer.....	58
9.5.6	TaggerUtils	58
9.5.7	NERUtils.....	58
9.5.8	AnonymizerUtils.....	58
9.6	NLP web	59
9.6.1	Tokenizácia	59
9.6.2	Segmentácia viet	60
9.6.3	Lematizácia	60
9.6.4	Stemovanie.....	60
9.6.5	POS značkovanie	60
9.6.6	Rozpoznávanie pomenovaných entít	60
9.6.7	Anonymizácia	60
10	Experiments	62
10.1	Experiment č.1 – Predspracovanie textu v slovenskom jazyku	62
10.2	Experiment č.2 – Určovanie slovných druhov (POS Tag).....	62
10.3	Experiment č.3 – NER súdnych rozhodnutí.....	63
10.4	Experiment č.4 Anonymizácia súdnych rozhodnutí	65
11	Zhodnotenie	66
12	Použité zdroje	68
A.	Príloha - Obsah DVD nosiča	70
B.	Príloha - Zdroje údajov	71
C.	Príloha - Optimalizácia stemera.....	72
D.	Príloha - Inštrukcia č. 24/2011 Ministerstva spravodlivosti Slovenskej republiky	73
E.	Príloha - Funkčná špecifikácia aplikácie ASuR	75
F.	Príloha - Technická špecifikácia aplikácie ASuR.....	81
G.	Inštalačný manuál pre ASuR web.....	83
H.	Príloha - Analýza možnosti využitia ontológií	84
I.	Príloha - Experimenty s metódami predspracovania textu a extrakcie informácií .	87

Slovník skratiek

Gazetteer – súbor obsahujúci zoznam entít, slovník

NLP – Natural Language Processing – Spracovanie prirodzeného jazyka

IE – Information Extraction – Extrakcia informácií

IR – Information Retrieval – Vyhľadávanie informácií

HMM – Hidden Markov Model – Skryté Markovovské modely sú generatívou metódou strojového učenia

MEMM – Maximum Entropy Markov Model - Markovovské modely maximálnej entropie sú diskriminačnou metódou strojového učenia

CRF – Conditional Random Fields - Podmienené náhodné polia sú diskriminačnou metódou strojového učenia

POS – Parts of Speech – Slovné druhy

NER – Named Entity Recognition – Rozpoznávanie pomenovaných entít

RDF – Resource Description Framework – Štandard pre výmenu dát na internete

URI – Uniform Resource Identifier - Spojitý reťazec znakov, ktorý identifikuje abstrakt alebo fyzický zdroj¹

OWL - Web Ontology Language – Sémantický webový jazyk navrhnutý pre reprezentáciu komplexných znalostí o veciach, skupinách a vzťahoch medzi vecami.²

¹ <http://www.ietf.org/rfc/rfc3986.txt>

² http://semanticweb.com/semantic-web-impact-on-enterprise-software-part-1_b703

Zoznam použitých obrázkov

Obrázok 1. Porovnanie chybovosti oddelovačov viet na korpusoch OntoNotes a MASC v anglickom jazyku	6
Obrázok 2. Architektúra CoreNLP ^[18]	12
Obrázok 3. Componenty ANNIE	19
Obrázok 4. Používateľské rozhranie aplikácie Ontea 1.0	21
Obrázok 5. Anonymizovaný text pomocou nástroja NLM-Scrubber	33
Obrázok 6. Postup krovov anonymizácie pomocou nástroja MIST	34
Obrázok 7. Snímka časti obrazovky aplikácie Súdny Manažment, určená pre písanie súdneho rozhodnutia	35
Obrázok 8. Snímka časti obrazovky aplikácie Súdny Manažment, určená pre anonymizáciu súdneho rozhodnutia	35
Obrázok 9. Diagram znázorňujúci kroky pri anonymizácii dokumentu	40
Obrázok 10. Diagram znázorňujúci kroky predspracovania	40
Obrázok 11. Diagram metódy rozpoznávania entít	43
Obrázok 12. Diagram časti anonymizácie	45
Obrázok 13. Príklad anonymizovaného súdneho rozhodnutia, v ktorom sme ručne farebne anotovali entity a vzťahy medzi nimi	47
Obrázok 14. Obrazovka pridania novej anotácie v aplikácii Brat	51
Obrázok 15. Používateľské rozhranie NLP web so zobrazením vizualizácie výsledkov NER	59
Obrázok 16. Grafové zobrazenie výsledkov experimentu	64
Obrázok 17. Počet nesprávne pozitívnych nálezov	72
Obrázok 18. Počet nesprávne negatívnych nálezov	72
Obrázok 19. Diagram funkčný požiadaviek na aplikáciu ASuR	75
Obrázok 20. Integračné požiadavky na aplikáciu ASuR	76
Obrázok 21. Diagram toku údajov pri trénovaní štatistického modelu	77
Obrázok 22. Diagram toku údajov pri anonymizácii	78
Obrázok 23. Diagram prípadov použitia	79
Obrázok 24. Diagram komponentov aplikácie ASuR	81
Obrázok 25. Prehľad jazyka OWL 2	86

Zoznam použitých tabuliek

Tabuľka 1. Prehľad najpoužívanejších tokenizérov ^[7]	4
Tabuľka 2. Porovnanie POS značkovačov	8
Tabuľka 3. Porovnanie NER nástrojov	28
Tabuľka 4. Porovnanie nástrojov MIST, Deid a NLM-S ^[15]	36
Tabuľka 5. Vyhodnotenie experimentu č.1	62
Tabuľka 6. Vyhodnotenie experimentu č. 2	63
Tabuľka 7. Vyhodnotenie experimentu č. 3	63
Tabuľka 8. Vyhodnotenie experimentu č.4	65
Tabuľka 9. Špecifikácia testovacej zostavy	87
Tabuľka 10. Vyhodnotenie 1. iterácie Monte Carlo krížovej validácie	89
Tabuľka 11. Vyhodnotenie 2. iterácie Monte Carlo krížovej validácie	89
Tabuľka 12. Vyhodnotenie 3. iterácie Monte Carlo krížovej validácie	90
Tabuľka 13. Vyhodnotenie 4. iterácie Monte Carlo krížovej validácie	90
Tabuľka 14. Vyhodnotenie 5. iterácie Monte Carlo krížovej validácie	90

I. Problém

1 Úvod

Aktuálne prebieha informatizácia a digitalizácia spoločnosti vo svete aj na Slovensku. Počas tohto procesu nahradzame manuálne činnosti za automatické a papierové dokumenty za elektronické. Výhodou elektronických údajov a dokumentov je, že aj bežní ľudia k nim môžu mať prístup. Problémom sú však osobné údaje, ktoré tieto dokumenty obsahujú. Tieto osobné údaje treba pred zverejnením anonymizovať. Proces ručnej anonymizácie si vyžaduje veľké množstvo času, ale aktuálnosť zverejnenia dokumentov je často kľúčová a musí byť vykonávaná rýchlo. Preto sa pri úlohách anonymizácie dokumentov využívajú metódy extrakcie informácií z textu, ktoré sú rýchle a dostatočne spoľahlivé.

Našim cieľom je navrhnuť a implementovať riešenie, ktoré za pomoci analýzy a extrakcie informácií z textu dokáže získať pomenované entity z textu a následne rozhodnúť, či sú údaje osobného charakteru, podliehajú ochrane osobných údajov, a teda ich treba anonymizovať. Keďže metódy pre prácu s textom v prirodzenom jazyku, ktoré sa využívajú na extrakciu entít, sú doménovo závislé, budeme prácu overovať na konkrétnej doméne súdnych rozhodnutí.

Práca je rozdelená na tri časti. V prvej časti sa nachádza analýza, v ktorej približujeme problematiku práce s textom v prirodzenom jazyku a extrakcie entít. Vysvetľujeme metódy predspracovania textu, a zároveň porovnávame existujúce riešenia, najmä v slovenskom jazyku, ktorý je špecifický svojou gramatikou a bohatou morfológiou. Opisujeme lingvistické a štatistické prístupy k extrakcii pomenovaných entít a zároveň podrobne rozoberáme typy entít, ktoré sa najčastejšie rozpoznávajú. V rámci popisu entít porovnávame jednotlivé prístupy existujúcich riešení, k riešeniu problémov s extrakciou danej entity. Ďalej opisujeme problém anonymizácie dokumentov, doménovo špecifické vlastnosti a problémy a zároveň porovnávame existujúce riešenia na anonymizáciu voľných textov. Na konci uvádzame experimenty s existujúcimi riešeniami, vyhodnotené podľa štandardných metrík.

V druhej časti práce uvádzame hybridnú metódu rozpoznávania entít v texte, ktorá je založená na využití extrahovaných informácií pri predspracovaní a lingvistickým rozpoznávaním entít ako črt pre štatistický model. V rámci metódy uvádzame spôsob riešenia jednotlivých analyzovaných problémov a poskytujeme pohľad na aplikáciu metódy. V sekcií implementácie vysvetľujeme postupy a uvádzame konkrétnie riešenia a implementácie v súlade s prezentovanou metódou.

V tretej časti vyhodnocujeme doterajšiu prácu a prezentujeme výsledky, ktoré sme dosiahli počas vývoja riešenia.

V prílohách uvádzame zdroje údajov, obsah digitálneho média, graf procesu optimalizácie stemera, inštrukciu Ministerstva spravodlivosti Slovenskej republiky, technickú a funkčnú špecifikáciu ako aj inštalačný manuál a ďalšie časti analýzy a experimentov, ktoré obsahujú dodatočné informácie.

2 Spracovanie prirodzeného jazyka

Spracovanie prirodzeného jazyka (Natural Language Processing NLP), je časť počítačovej vedy, ktorá sa zaobrá interakciami medzi ľudským prirodzeným jazykom a počítačom. Ide hlavne o počítačové spracovanie a porozumenie informácií, ktoré nám často prirodzene vyplývajú z kontextu. Pravidlá, ktoré platia pre ľudský jazyk, sú komplikované a je ich veľké množstvo. Preto sa pri riešení problémov s prirodzeným jazykom využíva umelá inteligencia, strojové učenie a rôzne štatistické modely, ktoré dokážu na základe trénovania a učenia riešiť úlohy, v ktorých je veľké množstvo pravidiel. Množstvo metód, ktoré dnešné implementácie spracovania prirodzeného jazyka (OpenNLP³, CoreNLP⁴, GATE (General Architecture for Text Engineering)⁵, OnTea⁶) poskytujú, umožňujú priamo riešiť bežné problémy so spracovaním textu napísaného v prirodzenom jazyku. Riešenia niektorých úloh, ktoré tieto nástroje poskytujú (lematizácia, stemovanie, určovanie slovných druhov a pod.) sú však často len stavebnými prvkami, ktoré po správnom poskladaní dokážu riešiť väčšie problémy. Pri riešení komplikovanejších problémov, sa využíva predspracovanie ako základný stavebný prvok, ktorý pripraví text na ďalšie spracovanie a doplní informácie, ktoré môžu využiť ďalšie metódy. NLP rieši veľké množstvo úloh a problémov, pomocou rôznych metód. Vyberieme niektoré problémy a metódy, ktoré nás zaujímajú z pohľadu našej práce.

Medzi metódy predspracovania textu patria:

- Tokenizácia
- Lematizácia
- Stemovanie
- Segmentácia
- Rekonštrukcia morfológických značiek

Medzi úlohy, ktoré sa snaží NLP riešiť patria najmä:

- Extraktia informácií
- Vyhľadávanie informácií

Pri práci s prirodzeným jazykom sa vo všeobecnosti používajú tri typy metód: lingvistická, štatistická a kombinovaná Každá má svoje výhody a nevýhody a použitie závisí od konkrétneho prípadu problému a aplikácie.

³ <https://opennlp.apache.org/>

⁴ <http://nlp.stanford.edu/software/corenlp.shtml>

⁵ <https://gate.ac.uk/>

⁶ <http://ontea.sourceforge.net/>

2.1 Tokenizácia

Pred tým než môžeme vôbec spracovávať prirodzený jazyk, musíme jasne oddelit' hranice jednotlivých viet, slov, skratiek, prípadne iných prvkov. Tiež je vhodné rozdeliť text na tzv. tokeny, ktoré predstavujú stavebné časti textu. Na oddelovanie a nájdenie tokenov sa využíva tokenizácia. Na základe určitých znakov a pravidiel tokenizér rozdeľuje text na menšie časti. V oblasti spracovania prirodzeného jazyka existuje niekoľko výkladov pojmu token. Výklad závisí od rôznych cieľov pri spracovaní a často aj s rôznymi vlastnosťami jazyka [27]. Vo väčšine prípadov sa pri samostatnom použití slova tokenizér myslí nástroj pre segmentáciu na slová.

2.1.1 Segmentácia slov

Základným stavebným prvkom viet sú slová, a teda prvuú úlohou pri spracovaní prirodzeného jazyka je zvyčajne segmentácia textu na slová. Podľa aplikácie získaných tokenov je vhodné zvoliť rôzne prístupy pri spracovaní interpunkcie v texte [2]. Najjednoduchšia implementácia (tzv. jednoduchý oddelovač) oddeluje slová podľa medzier a interpunkcie. Komplikovanejšie implementácie zahŕňajú pravidlá pre detekciu čísel, skratiek, emailov, rôznych typov jednotiek a dokonca aj textových smajlíkov. Pri komplikovanejších implementáciách prevažujú dva základné prístupy. Lingvistický (pravidlový) prístup pozostáva z regulárnych výrazov a fixných pravidiel pre určovanie hraníc tokenov na základe medzier, znakov, kontextu a polohy vo vete. Výhodou takéhoto prístupu je rýchlosť, nízka závislosť na doméne a nie sú potrebné trénovacie dátá pre funkčnosť. Štatistický prístup na druhej strane využíva metódy strojového učenia pre vytvorenie štatistického modelu, na základe ručne tokenizovaného trénovacieho korpusu. Takýto prístup má výhodu v tom, že nie je potrebné poznáť gramatiku a lingvistický model jazyka. Tento model sa vytvorí pri trénovaní. Nevýhodou je potreba pomerne veľkej rôznorodej trénovacej množiny ručne tokenizovaných dokumentov.

Tabuľka 1 Prehľad najpoužívanejších tokenizérov [7]

Č. Názov	Algoritmus	Rýchlosť	Otvorený Jazyk	
		slov/sek.	zdroj	
1	NLTK tokenizer	Jednoduchý oddelovač	>6000	Áno
2	OpenNLP tokenizer	Štatistický model maximálnej entropie	~400	Áno
3	Mallet tokenizer	Jednoduchý oddelovač	>6000	Áno
				Java

4	SPECIALIST NLP tokenizer	Jednoduchý oddelovač	~600	Áno	Java
5	Gump tokenizer	Lingvistické pravidlá	>6000	Áno	Gump
6	Dan Melamed's tokenizer	Lingvistické pravidlá	>6000	Áno	Perl
7	Qtoken	Jednoduchý oddelovač	~1500	Nie	Java
8	UIUC word splitter	Lingvistické pravidlá	> 6000	Áno	Perl
9	LT TTT tokenizer	Lingvistické pravidlá	~1000	Čiastočne	Perl/...
10	MedPost tokenizer	Lingvistické pravidlá	~750	Áno	Perl/C++
11	Brill's POS tagger	Učenie riadené chybami a založené na transformácii	~300	Áno	Java/C
12	Stanford POS Tagger	Štatistický model maximálnej entropie	~50	Áno	Java
13	MXPOST tagger	Štatistický model maximálnej entropie	~200	Nie	Java
14	Stanford PTBTokenizer	Lingvistické pravidlá	> 6000	Áno	Java

2.1.2 Segmentácia viet

Pre potreby ďalšieho spracovania, najmä pri určovaní vetnej syntaxe, morfologických značiek a zisťovaní významu, potrebujeme pri tokenizácii identifikovať hranice viet. Všetky tieto informácie môžeme ďalej využiť pri identifikácii a extrakcii pomenovaných entít. Veta alebo súvetie je ukončené interpunkčným znamienkom a začína veľkým písmenom. Tento základný poznatok sa využíva pri oddelovaní jednotlivých viet. Treba však bráť ohľad na výnimky, ktoré vznikajú použitím zátvoriek, úvodzoviek, skratiek, dátumov, poradovými číslami a podobne. V týchto prípadoch sa využíva interpunkcia aj v strede vety.

Základné pravidlá pri oddelovaní viet:

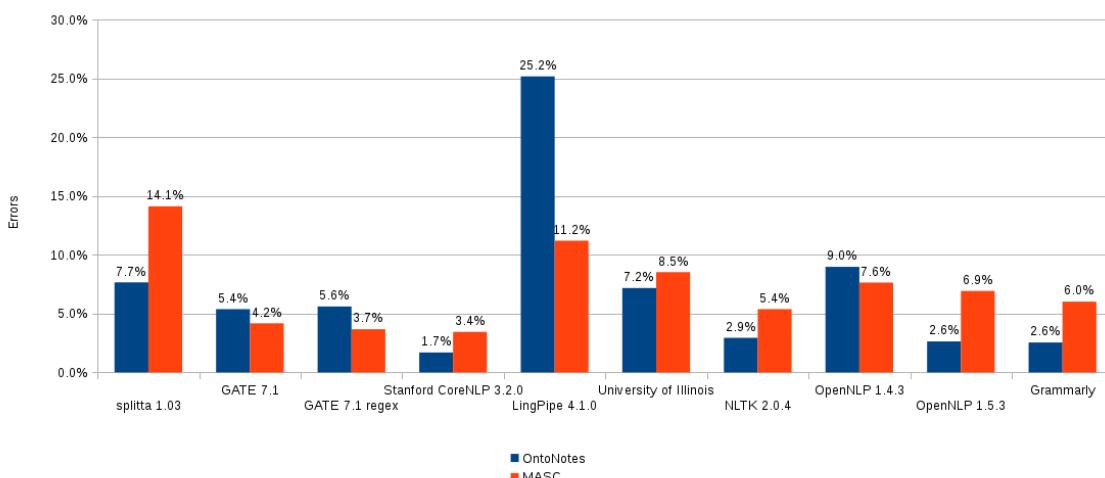
- Ak je na konci vety skratka, ďalšia bodka sa nepíše.
Spoločnosť s ručením obmedzeným je s.r.o.
- Aj keď po titule ide veľké písmeno nejde o novú vetu.
Jeho meno je Bc. Filip Bednárik.
- V zátvorkách sú najmä doplňujúce informácie

Pravidlá priamej reči:

- Uvádzacia veta sa končí dvojbodkou, priama reč sa začína veľkým písmenom.
- Priama reč sa končí čiarkou, otáznikom alebo výkričníkom, uvádzacia veta sa začína malým písmenom.
- Uvádzacia veta sa začína malým písmenom, končí čiarkou, druhá časť priamej reči sa začína malým písmenom.

Existuje niekoľko spôsobov riešenia problému segmentácie viet. Jednoduché oddelovače viet (tzv. sspliter) využívajú jednoduchú heuristiku, pre zisťovanie hraníc viet. Komplikovanejšie riešenia využívajú lingvistickej alebo štatistický model. Niektoré oddelovače sa spoliehajú na informácie z kroku segmentácie slov, ako napr. Stanford CoreNLP WordsToSentencesAnnotator.

V slovenskom jazyku existuje natréovaný štatistický model maximálnej entropie pre OpenNLP, ktorý natrénoval autor Bc. Ivan Šimko⁷ na predmete Vyhladávanie informácií na fakulte FIIT.



Obrázok 1. Porovnanie chybovosti oddelovačov viet na korpusoch OntoNotes a MASC v anglickom jazyku⁸

2.2 Lematizácia a stemovanie

Lematizácia je proces, pri ktorom prevádzame morfológicky upravené slovo, na základný tvar pomocou slovníku. Problémom však môžu byť slová, ktoré sa ešte nedostali do slovníka. Pri takýchto slovách využijeme stemovanie.

⁷ <http://vi.ikt.ui.sav.sk/@api/deki/pages/1542/pdf>

⁸ <http://tech.grammarly.com/blog/posts/How-to-Split-Sentences.html>

„Stemovanie je proces redukcie slov na ich koreň, základný tvar. Koreňom budeme nazývať časť slova, ktorá je rovnaká pre všetky morfologické varianty slova. Koreň získaný stemovaním nemusí byť nutne zhodný s morfologickým koreňom slova. Zvyčajne stačí ak sú morfologické varianty projektované na rovnaký koreň. Cieľom je, aby slová s rovnakým významom a rovnakým základom slova mali rovnaký stem, a aby nebolo viac slov s rôznym významom s rovnakým stemom.“ [17]

Aj keď ide o proces predspracovania textu, ide o základnú stavebnú jednotku systémov pre spracovanie prirodzeného jazyka a extrakcie informácií. Úspešnosť lematizéra a stemera sa odzrkadľuje aj na úspešnosti ďalších metód, ktoré sa na toto predspracovanie spoliehajú. V niektorých prípadoch je dokonca žiaduce vykonávať lematizáciu a stemovanie súčasne s určovaním slovných druhov, keďže slovný druh môže byť rozhodujúci pri pokuse o rozlíšenie tvarových homoným.

2.3 Určovanie slovných druhov a pádov (POS značkovanie)

Pri určovaní slovných druhov a pádov, využívame značkovač tzv. POS značkovač, ktorý automaticky na základe pravidiel, slovníka, okolia slova a pozície vo vete, označuje slovné druhy a pády slov. Takáto úloha je pomerne komplikovaná aj pre ľudského anotátora, ktorý musí mať dobré znalosti gramatiky. Komplikovanosť značenia závisí práve na komplikovanosti gramatiky konkrétneho jazyka. V našej práci sa venujeme anotácii v slovenskom jazyku a práve slovenský jazyk má veľmi bohatú morfológiu a komplexné pravidlá pre skloňovanie, časovanie a odvodzovanie slov. Na túto úlohu sa najčastejšie využíva tvorba štatistického modelu na základe trénovacej množiny, ktorú je potrebné ručne anotovať človekom.

POS značkovanie má zmysel ako podporná technológia pri extrakcii informácií a zvyčajne sa vykonáva vo fáze predspracovania. Doplňa informácie a pridáva novú vlastnosť, ktorú je možné využiť pri klasifikácii pomenovaných entít.

V slovenčine na túto úlohu existuje niekoľko riešení.

2.3.1 Dagger

Jedným z riešení, ktoré využíva aj Slovenský Národný Korpus sa nazýva Dagger⁹ [8]. Toto riešenie využíva metódu skrytých markovovských modelov (z angl. hidden Markov model, HMM) pre klasifikáciu. Klasifikáciu HMM dopĺňajú o slabší klasifikátor na základe extrakcie sufíxov, vďaka ktorému tvary slov, ktoré sa

⁹ <https://github.com/hladek/dagger>

nevyskytujú v ručne anotovanom korpuse, dostávajú nenulovú percentuálnu šancu pri výbere vhodného kandidáta.

2.3.2 Naivný pravidlový korpusový POS tagger

Riešenie pre rozlíšenie medzi viacerými možnými slovnými druhmi, využíva jednoduchý spôsob redukcie viacznačnosti, (napr. slová viažuce sa s predložkou ‘v’ sa môžu vyskytovať len v páde lokál) založený na manuálne vytvorených pravidlach.¹⁰

Toto riešenie je veľmi rýchle a nepotrebuje trénovaciu množinu, avšak je pomerne nepresné.

2.3.3 TreeTagger

Riešenie rozširuje základný značkovač, založený na HMM, čím zvyšuje presnosť pri trénovaní na malom korpuse. Využíva rozhodovacie stromy pre získavanie spoľahlivejších odhadov pre kontextové parametre^[24]. Na stránke je dostupný natrénovaný model na Slovenskom Národnom Korpuse.

2.3.4 CRFPOSTagger

Toto riešenie využíva grafový pravdepodobnostný model podmienených náhodných polí. Architektúra tohto modelu je upravená pre vyššiu úspešnosť v slovenčine. Natrénované na Slovenskom Národnom Korpuse.¹¹

Tabuľka 2. Porovnanie POS značkovačov

Č.	Názov	Algoritmus	Otvorený zdroj	Jazyk
1	Dagger	HMM + klasifikátor suffíxov	áno	C
2	Naivný pravidlový korpusový POS tagger	Lingvistický	áno	Java
3	TreeTagger	HMM + rozhodovacie stromy	nie	Perl
4	CRFPOSTagger	CRF	áno	C++

¹⁰ http://text.fiit.stuba.sk/naivny_prawidlovy_pos_tagger.php

¹¹ http://text.fiit.stuba.sk/statisticky_pos_tagger.php

3 Extraktia informácií

Extraktia informácií v dnešnej dobe nachádza množstvo uplatnení pri rôznych úlohách práce s textom. Stretávame sa s problémom, kedy máme informáciu alebo odpoved' na nejakú otázku, napísanú niekde v dokumentoch, ale máme také veľké množstvo dokumentov, že je problémom nájsť ten správny dokument, v ktorom sa daná informácia nachádza. Na vyhľadávanie v dokumentoch, sa využíva preto vyhľadávanie informácií. Ak však chceme tieto informácie automaticky spracovať a ďalej z nich odvodzovať alebo ich využívať automatizované, potrebujeme ich z dokumentov extrahovať a na to slúži extraktia informácií. S extrakciou informácií sa stretávame najmä pri extrakcii určitých metadát. Napríklad pri digitálnej knižnici, kde pre podporu vyhľadávania extrahujeme informácie o autorovi, názve diela, roku vydania a podobne. IE nachádza uplatnenie aj pri získavaní informácií pre potreby štatistik, kde extrahujeme podľa šablóny a pravidel informácie, ktoré nás zaujímajú a následne výčíslime ich výskyt, alebo extrahujeme číselné údaje, ktoré ďalej spracovávame. Jednou z aplikácií je aj extraktia profilu človeka, kde dokážeme profilovať ľudí na základe ich komunikácie a zdieľaných informácií. Takéto použitie nachádza uplatnenie v národnej bezpečnosti ale aj rôznych sociálnych výskumoch. [26]

3.1 Koncept

„Extraktia informácií (Information extraction IE) je technológia, založená na analýze prirodzeného jazyka, s cieľom extrakcie kúskov informácií.“ [4]. Vstupom procesu je zvyčajne neštruktúrovaný textový dokument alebo záznam hovorenej reči a výstupom sú štruktúrované údaje. Tieto údaje môžeme priamo zobraziť používateľovi, uložiť do databázy, prípadne využiť na indexáciu dokumentov pre potreby vyhľadávania informácií (Information retrieval IR). Aj keď pri extrakcii informácií využívame rôzne technológie vyhľadávania informácií a spracovania prirodzeného jazyka, ide o rozličné metódy. Vyhľadávanie informácií slúži na vyhľadanie relevantného textu potenciálne zahŕňajúceho danú informáciu a jeho zobrazenie používateľovi. Avšak aplikácie využívajúce extrakciu informácií, analyzujú text a prezentujú len špecifické informácie, ktoré používateľa zaujímajú. Ak si zoberieme príklad dopytu „Filip Bednárik“, systém vyhľadávania informácií nám vráti všetky dokumenty, v ktorých sa hľadaný výraz nachádza. Systém extrakcie informácií nám ale môže vrátiť vlastnosti osoby ako vek, bydlisko, vzťahy s inými osobami a podobne, ktoré získal analýzou dokumentov, kde sa zmienka o osobe nachádzala. Aké informácie nám systém extrakcie informácií vráti, nie je presne stanovené, a preto je užšie viazaný na určitú doménu a spôsob použitia

takéhoto systému. IE v porovnaní s IR je tiež výpočtovo náročnejšia^[4] a vyžaduje väčšie množstvo času na analýzu textu. Výhodu má však pri veľkom počte dokumentov, kde dokáže sprehľadniť a zvýrazniť dôležité informácie, ktoré nás zaujímajú a hlavne v prípade, keď nemáme čas si jednotlivé časti dokumentov čítať.

Extrakcia informácií je oblasť, ktorá rieši množstvo problémov a delí sa na viacero úloh. Na MUC (z angl. message understanding conferences) konferenciách bolo definovaných 5 základných úloh IE (z angl. information extraction).

- Rozpoznávanie pomenovaných entít (Named entity recognition, NER). Extrahuje a klasifikuje osoby, mestá, temporálne výrazy a podobne. Príklad pomenovanej entity je “Fakulta informatiky a informačných technológií”.
- Extraktia relácií (Coreference resolution, CO). Nájdenie a identifikovanie vzťahov medzi entitami extrahovanými pomocou NER. Príkladom je vzťah “FIIT” a “Fakulta informatiky a informačných technológií”, kde ide o rovnakú entitu.
- Konštrukcia šablónových prvkov (Template elements, TE). Pridáva popisné informácie ku entitám z NER za pomoci CO.
- Konštrukcia šablónových relácií (Template relations, TR). Vyhladávanie vzťahov medzi entitami a ich vlastnosťami identifikovanými v TE.
- Produkcia scenárových šablón (Scenario template, ST). Umožňuje napasovanie výsledkov z TE a TR na špecifickú problémovú doménu.

3.2 Špecifika slovenského jazyka

Slovenský jazyk patrí medzi západoslovanské jazyky (slovenčina, čeština, polština a lužická srbčina)^[9]. Slovenčina na rozdiel od angličtiny patrí medzi vysoko flektívne jazyky^[8]. Pravidlá morfológie sú pomerne komplikované a nie vždy je jednoduché určiť koreň slova.

Pri lematizácii je najväčší problém s tvarovými homonymami, napr. slovo “je” môže byť vyčasované sloveso byť alebo jest, preto aj v prípade dostupnosti slovníka nevieme určiť jednoznačnú lému samostatne stojaceho slova. Pri rozhodovaní musíme využiť kontext slova, jeho postavenie vo vete, čas alebo pád v ktorom sa nachádza a tiež musí sedieť do vety významovo. Pri vyhľadávaní informácií, nie je tento problém až taký markantný no pri extrakcii informácií môže spôsobiť veľké rozdiely.

Pri stemmingu sa stretávame s problémom pri slovesách, kde nemôžeme jednoducho odstrániť sufixy vyčasovaných slovies, lebo tieto koncovky sa bežne vyskytujú na konci

podstatných mien. Taktiež je preto vhodné využiť informáciu o slovnom druhu a čase slova, ktoré chceme stemovať.

Výhodou slovenčiny oproti germánskym jazykom pri extrakcii informácií, je využívanie sufixov pri priezviskách osôb, a teda je jednoduchšie identifikovať pohlavie osoby podľa mena a priezviska.

Na Slovensku sa využíva formát dátumu dd.MM.yyyy.

V slovenčine na rozdiel od germánskych jazykov, nehrá poloha slov až takú veľkú rolu pri identifikovaní slovného druhu.

3.3 Existujúce nástroje pre extrakciu informácií

Extrakcia informácií je všeobecný pojem a široká oblast' pokrývajúca rôzne problémy. Pre vyriešenie komplikovanejších problémov, ako napr. rozpoznávanie pomenovaných entít, je potrebné využiť celú radu nástrojov pre prácu s prirodzeným jazykom. Pre potreby riešenia zložitejších problémov preto vznikli súpravy nástrojov pre prácu s prirodzeným jazykom, ako CoreNLP, OpenNLP, NLTK (Natural Language Toolkit) alebo GATE.

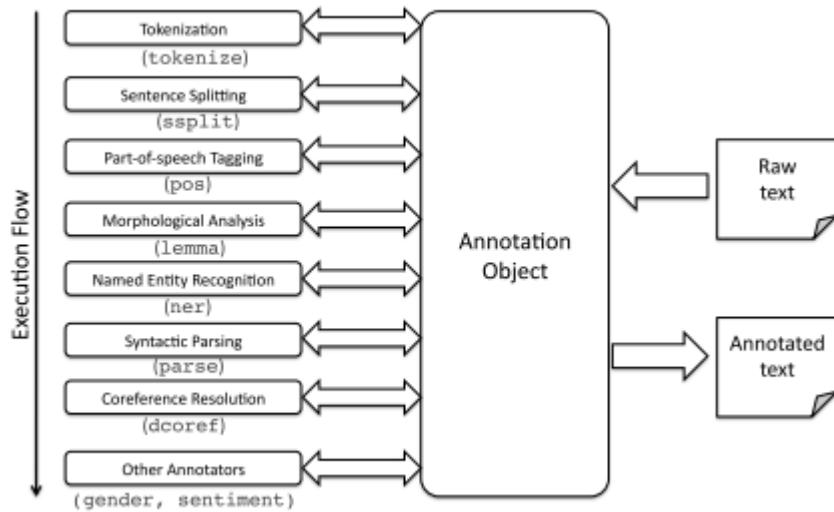
3.3.1 GATE Developer

GATE Developer je integrované vývojové prostredie (IDE) pre prácu s jazykom. Je súčasťou rodiny produktov GATE, ktoré umožňujú kolaboratívnu prácu s jazykom, poskytujú rozhrania a knižnice pre prácu s jazykom, vysoko škálovateľné riešenia a indexový server. Aktuálna verzia GATE Developer vo verzii 8.1 predstavuje nástroj pre jednoduchú správu procesov a nastavení práce s prirodzeným jazykom. Nástroj poskytuje grafické používateľské rozhranie, v ktorom je možné nastaviť proces anotácie a využiť pritom existujúce knižnice pre extrakciu informácií. Taktiež umožňuje priamo prácu na ručnej anotácii dokumentov, vyhodnocovanie a vizualizáciu výsledkov. Gate prichádza v základnej verzii spolu s aplikáciou ANNIE a OpenNLP, ktoré obsahujú niekoľko nástrojov, ktoré sú použiteľné ako stavebné prvky pri riešení problémov extrakcie informácií. Pri konfigurácii projektu, je možné nastaviť postupnosť jednotlivých krokov ako aj nástrojov, ktoré sa pri týchto krokoch použijú.^[3].

3.3.2 CoreNLP

CoreNLP je produkt vyvíjaný na Stanfordskej univerzite v Kalifornii. Ide o sadu nástrojov pre prácu s prirodzeným jazykom. Jednotlivé nástroje sa dajú spustiť samostatne alebo ich je možné spustiť sekvenčne za sebou v zreteženom spracovaní textu. Ku väčšine nástrojov je dostupný tzv. anotátor, ktorého vstupom sú anotácie

a výstupom rozšírené anotácie. Každý anotátor predstavuje jeden nástroj pri práci s prirodzeným jazykom a má definované, aké anotácie produkuje a aké vyžaduje od predchádzajúcich anotátorov.



Obrázok 2. Architektúra CoreNLP [18]

Takéto riešenie umožňuje jednoduché dopĺňanie vlastných anotátorov, ako aj integráciu existujúcich. Medzi základné nástroje CoreNLP patria:

- TokenizerAnnotator – Tokenizér s možnosťou výberu implementácie. Základnou implementáciou je PTBTokenizer, ktorý je definovaný pravidlami vo formáte jflex¹².
- WordsToSentencesAnnotator – Lingvistický oddelovač viet na základe tokenov, ktoré identifikoval TokenizerAnnotator.
- POSTaggerAnnotator – Štatistický značkovač POS. Využíva model maximálnej entropie.
- MorphaAnnotator – Lingvistický lematizér anglického jazyka. Využíva definované pravidlá a slovník vo formáte jflex.
- TokensRegexNERAnnotator/RegexNERAnnotator/NERCombinerAnnotator – Lingvistický a štatistický NER. Lingvistický využíva vlastný jazyk podobný regulárnym výrazom a je možné ho používať na sekvencie. Štatistický NER využíva CRF model.
- ParserAnnotator – Syntaktický analyzátor a anotátor závislostí. Obsahuje rôzne implementácie na úrovni štatistického modelu, lingvistického modelu až po neurónové siete

¹² <http://jflex.de/>

- CorefAnnotator – Deterministický, štatistický a hybridný extraktor relácií.
- NumberAnnotator – Pravidlový klasifikátor počítateľných pomenovaných entít, času a dátumov.

3.3.3 OpenNLP

OpenNLP podobne ako CoreNLP poskytuje sadu nástrojov pre spracovanie prirodzeného jazyka. OpenNLP využíva pri väčšine problémov štatistický model MEMM. Nástroje, ktoré OpenNLP poskytuje:

- Sentence Detector – Štatistický oddeľovač viet, založený na štatistickom modeli maximálnej entropie. V princípe ide o binárny klasifikátor, ktorý určuje, či interpunkčné znamienko je hranicou vety alebo nie.
- Tokenizer – Oddeľovač slov. OpenNLP poskytuje tri implementácie tokenizéra. Whitespace Tokenizer oddeľuje jednotlivé tokeny pomocou medzery. Simple Tokenizer klasifikuje tak, že znaky a sekvencie rovnakej triedy sú tokenom. Learnable Tokenizer deteguje tokeny na základe štatistického modelu maximálnej entropie.¹³
- Name Finder – Štatistický klasifikátor pomenovaných entít. Využíva MEMM a umožňuje trénovať vlastné črty modelu a jednoduché nastavenie dostupných čít.
- Part-Of-Speech Tagger – Štatistický klasifikátor slovných druhov. Používa MEMM.
- Chunker a Parser – Nástroje na syntaktickú analýzu viet. Využívajú štatistický model MEMM.

3.3.4 NLTK

Platforma pre prácu s prirodzeným jazykom v jazyku Python. Poskytuje nástroje pre tokenizáciu, stemovanie, syntaktickú analýzu, značkovanie, sémantické odvodzovanie a obsahuje obalené implementácie CoreNLP. Výhodou NLTK je najmä jednoduchosť práce priamo v Python konzole a keďže ide o skriptovací jazyk, nie je potrebná komplikácia zdrojového kódu, a preto je NLTK vhodný na prototypovanie a učenie sa základov práce s prirodzeným jazykom.

¹³ <https://opennlp.apache.org/documentation/1.6.0/manual/opennlp.html>

4 Rozpoznávanie pomenovaných entít (NER)

Rozpoznávanie pomenovaných entít je jeden z najznámejších problémov IE. Systémy rozpoznávania pomenovaných entít sú určené na identifikáciu mien ľudí, miest, organizácií, dátumov, súm peňazí a podobne v texte^[4].

NER v anglickom jazyku je v súčasnosti možné vykonávať v niektorých doménach a na niektorých entitách až s 95 % presnosťou. Ľudský anotátor tiež robí pri anotácii chyby a nie je stopercentný, aj preto sa NER radí medzi oblasti, ktoré dosahujú výsledky ako ľudia. Ako referenčný zdroj uvedieme konferenciu MUC-7, kde vyhodnocovali úspešnosť NER systému v porovnaní s ľudským anotátorom a tu podľa metriky F-Measure dosiahol v niektorých prípadoch automatický systém 96,42 % a ľudský anotátor 96,68 %^[22].

Výhodou NER je tiež, že nie je príliš späť s doménou, a teda je ho možné použiť na príbuzné domény bez veľkých zmien.

V slovenčine je však NER na o niečo nižšej úrovni a dosahuje výsledky na úrovni 89%^[14].

4.1 Metódy podľa prístupu

Metódy extrakcie pomenovaných entít sa delia na tri základné skupiny. Lingvistické metódy, metódy strojového učenia (štatistické) a hybridné metódy.

4.1.1 Lingvistické metódy

Prvá skupina je založená na gramatike a pravidlách, sem patria najmä metódy využívajúce pravidlá, ontológie, slovníky, regulárne výrazy, rozhodovacie stromy a kontextové slová^[12]. Takýto prístup má svoje výhody a nevýhody.

Výhody:

- Rýchlosť - Metóda je zvyčajne veľmi rýchla a dokáže spracovať veľké množstvo textu v krátkom čase
- Vysoká presnosť - Metóda je veľmi presná a dosahuje nízke hodnoty nesprávne určených entít
- Jednoduchosť riešenia - Ak je úloha jednoduchá je jednoduché napísat regulárny výraz

Nevýhody:

- Nízka flexibilita - Dopĺňanie slovníkov môže byť komplikovaný a častý proces
- Čažšia udržateľnosť - S narastajúcim počtom pravidiel je čoraz komplikovanejšie udržiavať konzistenciu a prehľad nad jednotlivými pravidlami

Regulárne výrazy

Regulárne výrazy nám poskytujú mechanizmus pre výber špecifických reťazcov zo sady znakových reťazcov. Každý znak v regulárnom výraze je buď rozpoznávaný ako metaznak so špeciálnym významom alebo ako obyčajný znak s jeho doslovňím významom. Dokopy sa používajú na identifikáciu časti textu spĺňajúcej daný vzor.

Pri získavaní údajov, sa regulárne výrazy používajú na vyhľadávanie typických reťazcov, ktoré obsahujú neobyčajný znak alebo majú iný formát ako obyčajné slovo, prípadne sa využíva na vyhľadanie rôznych variácií jedného slova alebo reťazca. Príkladom vyhľadávania informácií podľa regulárneho výrazu môže byť vyhľadanie všetkých emailov na stránke, dátumov, alebo pri spracovaní číselných hodnôt s jednotkami, ktoré sú všeobecne známe.

Príklad regulárneho výrazu podľa RFC 5322¹⁴ na nájdenie emailu na stránke:

```
[a-zA-Z!#$%&'*+/?^`{|}~-]+(?:\.[a-zA-Z!#$%&'*+/?^`{|}~-]+)*@  
(?:[a-zA-Z](?:[a-zA-Z-]*[a-zA-Z])?\.\.)+[a-zA-Z](?:[a-zA-Z-]*[a-zA-Z])?
```

4.1.2 Metóda založená na štatistických modeloch

Metódy založené na štatistických modeloch predstavujú riešenie pomocou strojového učenia, najčastejšie ide o učenie s učiteľom. Ide o trénovanie štatistického modelu pomocou ručne anotovaného korpusu a následné využitie tohto modelu pri klasifikácii.

Výhody:

- Jednoduché škálovanie – Metóda dokáže nahradzať veľké množstvo pravidiel a neprestáva sa učiť a stále sa zlepšuje
- Nezávislosť riešenia na jazyku a doméne – Samotná metóda nie je závislá (resp. veľmi málo) na jazyku a doméne a je možné ju natrénovať na dostatočne veľkom korpuze pre akýkoľvek jazyk a doménu
- Inteligencia – Štatistické metódy nájdu aj také vzťahy a závislosti, ktoré si človek nevšimne a nie sú na prvý pohľad zjavné

Nevýhody:

- Potreba veľkého korpusu – Pre trénovanie štatistickej metódy je potrebný dostatočne veľký ručne anotovaný korpus
- Nižšia rýchlosť – Metóda je o niečo pomalšia ako jednoduché pravidlá alebo vyhľadávanie v slovníku alebo strome
- Nižšia presnosť – Presnosť závisí od toho ako sa štatistický model natrénuje

¹⁴ <https://tools.ietf.org/html/rfc5322>

Existujúce riešenia pre extrakciu pomenovaných entít využívajú tri najpoužívanejšie štatistické metódy pre klasifikáciu pomenovaných entít.

Skryté markovovské modely (HMM)

HMM sa radí medzi najpoužívanejšie generatívne modely, ktoré sa všeobecne používajú pri učení pravdepodobnostných modelov [6]. HMM sa vďaka svojej povahe radí medzi základné metódy využívané pri extrakcii informácií.

Metóda sa skladá zo skrytých stavov, počiatočného a koncového stavu, prechodov medzi stavmi a výstupov. Je založená na dvoch základných predpokladoch:

- Základný Markovov predpoklad

Stav v čase t_n (označovaný ako S_{t_n}), pre akúkoľvek platnú (nezápornú) hodnotu $n-1$ (musí existovať stav pred ním), závisí len na stave v t_{n-1} (označovaný ako $S_{t_{n-1}}$)

$$P(S_{t_3} = l | S_{t_2} = j, S_{t_1} = i) = P(S_{t_3} = k | S_{t_2} = j)$$

- Stacionárny predpoklad

Pravdepodobnosti predstavujúce prechody medzi stavmi sa nemenia v čase

$$P(S_{t_n} = i | S_{t_{n-1}} = j) = P(S_{t_{n+l}} = i | S_{t_{n+l-1}} = j), l \geq 0$$

HMM je v podstate stavový automat. Na základe učených vlastností prechádza stavmi cez prechody s najvyššou pravdepodobnosťou a vracia konečný stav [19].

Výhodou tejto metódy je jednoduchosť a vysoká účinnosť. Avšak metóda má svoje obmedzenia kvôli základnému predpokladu, ktorý nepredpokladá vzdialenejšie vzťahy a zároveň predpokladá vysokú nezávislosť jednotlivých stavov. Preto pri niektorých úlohách je efektívnejšie využiť diskriminačné modely [26].

Markovovské modely maximálnej entropie (MEMM)

MEMM sú formou diskriminačných modelov pre označovanie sekvenčných údajov. MEMM na rozdiel od HMM považuje jednotlivé pozorovania za podmienené, na rozdiel od generovaných v HMM. Vzhľadom na to, môže každý prechod medzi stavmi závisieť aj na iných vlastnostiach sekvencie pozorovania. V kombinácii s modelom maximálnej entropie, ktorý sa snaží dosiahnuť pravdepodobnostné rozdelenie s maximálnou možnou entropiou dosahuje metóda veľmi dobré výsledky pri trénovaní metód extrakcie informácií. Túto metódu využíva aj populárny systém OpenNLP [21]. MEMM vďaka svojej povahe umožňuje používateľovi definovať vlastnosti, ktoré sú sice od seba závislé, no obsahujú v sebe informáciu navyše. Toto sa dá využiť najmä

v označovaní slovných druhov alebo detekcii pomenovaných entít, kedy využívame aj to, že slová začínajú na veľké písmeno alebo v určitej sekcii textu. Tieto vlastnosti môžu sice byť závislé na doméne, ale model je možné náštrénovať pre každú doménu samostatne [26].

Podmienené náhodné polia (CRF z angl. Conditional Random Fields)

Podmienené náhodné polia predstavujú neriadený grafický model, ktorý sa trénuje pre maximalizáciu podmienenej pravdepodobnosti [26]. CRF oproti tradičným metódam berie do úvahy kontext. Umožňuje diskriminačné trénovanie a obojsmerný tok pravdepodobnostných informácií v rámci celej sekvencie. CRF pozostáva z podmieneného sekvenčného modelu, ktorý reprezentuje pravdepodobnosť skrytej sekvencie stavov na základe určitých pozorovaní [5]. Trénovanie CRF modelu pozostáva z definovania zoznamu čít sekvenčí a k optimalizácii ich hodnotenia pomocou minimalizácie (v CoreNLP QNMinimizer). Jednotlivé črty je možné definovať pri trénovaní a bežne sa využívajú črty ako:

- Slovo – Základná črta je slovo, táto črta má význam najmä pri zistovaní kontextu, vtedy je vhodné sledovať aj predchádzajúci a nasledujúci token.
- Trieda – Pri klasifikácii zohľadňuje početnosť danej triedy v trénovacnej množine.
- Sekvencie tried – Zohľadňuje sekvencie tried ako črtu pri trénovaní a klasifikácii
- Typové sekvencie a tvar slova – Zohľadňuje sekvencie typu slova (či začína na veľké, malé písmeno, celé kapitálkami a podobne)
- Črta výskytu – Táto črta zohľadňuje či sa okolité slová začínajú na veľké písmeno a zlepšuje najmä anotáciu sekvenčí, ktoré v strede obsahujú slovo s malým začiatocným písmenom
- Začiatok a pozícia vo vete – Črta využíva segmentáciu viet
- Lémy – pri trénovaní štatistického modelu sa využije léma ako črta zistená v predchádzajúcim kroku
- POS značky – Užitočná črta, ktorá hovorí o slovnom druhu a páde daného tokenu
- Oddelovanie slov – Črta zabezpečujúca, že jednotlivé slová nemusia byť v konkrétnom poradí a zároveň zachováva smer. Je možné nastaviť do akej vzdialenosť má disjunkcia fungovať.

- NGramy – Jedna zo základných čít rieši pomerne komplikovaný problém morfológie jazyka, rozdelením tokenov na menšie časti a zároveň umožňuje trénovanie robustnejšieho modelu voči preklepom. Štandardne je možné definovať maximálnu dĺžku NGramov. (minimálna dĺžka je väčšinou nemenná a to 2). Je tiež možné zakázať stredové NGramy tzv. midNGramy, odstránenie diakritiky NGramov alebo prevod na malé písmená abecedy.
- Gazzetteery – Črta hovorí o tom, či sa daný token nachádza v preddefinovanom slovníku a tiež hovorí o tom akú má triedu.

CRF model sa využíva okrem iného na rozpoznávanie menných entít, rozpoznávanie objektov a segmentáciu obrázka. Ako jadro pri spracovaní pomenovaných entít využíva túto metódu CoreNLP.

4.2 Metódy podľa typu extrahovanej entity

Informácie, ktoré chceme extrahovať, majú spravidla rôznu povahu. Typické entity zahrňujú dátumy, sumy, mená a priezviská, firmy adresy a podobne. Pri jednotlivých entitách využívame rôzne spôsoby spracovania. V tejto sekcii porovnávame prístupy a metódy. V sekcii experimentov uvádzame porovnanie praktického nasadenia jednotlivých technológií v praxi.

4.2.1 Všeobecne použiteľné dostupné riešenia

Existuje pomerne veľké množstvo aplikácií a technológií pre spracovanie prirodzeného jazyka. Výber testovaných aplikácií sme urobili na základe požiadaviek na testovanie (možnosť testovať aplikácie zadarmo) a možnú podporu slovenského jazyka. Porovnávali sme najmä najznámejšie aplikácie a technológie, ktoré sú aktuálne dostupné a stále aktuálne.

OpenNLP Name Finder¹⁵

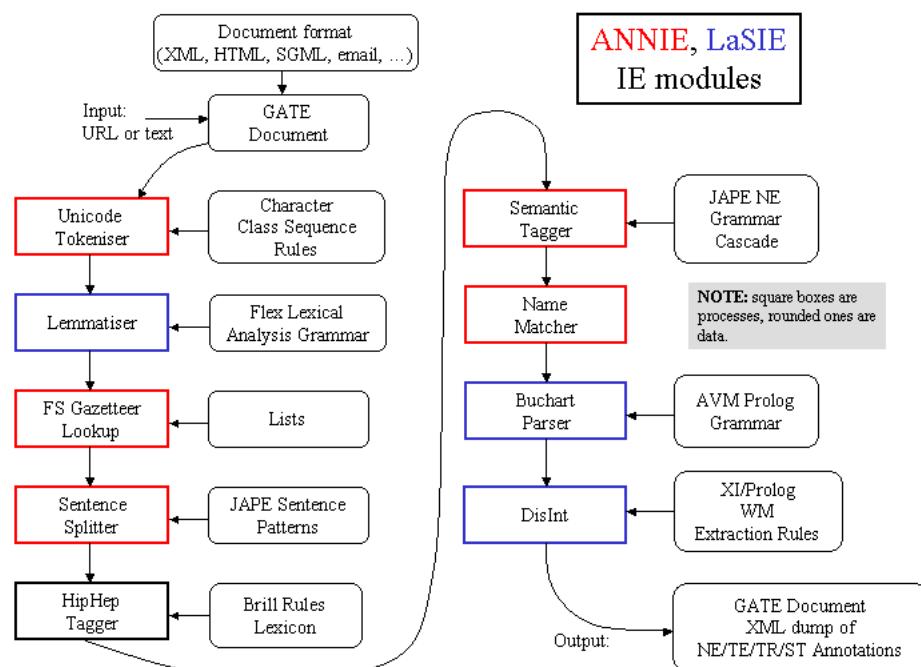
OpenNLP Name Finder využíva štatistické metódy pre trénovanie a klasifikáciu pomenovaných entít. Ručnou anotáciou dostatočne veľkého korpusu, vieme naučiť štatistický model, na základe ktorého môžeme klasifikovať pomenované entity. Nástroj využíva základnú sadu sledovaných ukazovateľov, ktoré je možné si upraviť a zlepšiť tak doménovo špecifické rozpoznávanie. Toto riešenie je použiteľné na všetky problémy v tejto sekcii, avšak často s nižšou úspešnosťou ako riešenia, ktoré využívajú pravidlá, slovníky a regulárne výrazy. Najväčšou nevýhodou takéhoto riešenia je

¹⁵ <https://opennlp.apache.org/documentation/1.6.0/manual/opennlp.html>

potreba veľkého anotovaného korpusu, ktorý je väčšinou aj doménovo špecifický. Trénovanie modelu je založené na použití klasifikátora maximálnej entropie.

ANNIE NE Transducer a ANNIE Gazetteer

ANNIE NE Transducer a ANNIE Gazetteer sú súčasťou nástroja GATE developer. ANNIE (a Nearly-New Information Extraction System) sa skladá z viacerých komponentov, ktoré spolu tvoria systém pre extrakciu informácií. Jednotlivé komponenty využívajú pri extrakcii pomenovaných entít kombináciu slovníkov, regulárnych výrazov a pravidiel vo formáte JAPE. Riešenie v porovnaní s OpenNLP nevyžaduje trénovanie na veľkom korpuse a dosahuje lepšie výsledky pri jednoduchých úlohách.



Obrázok 3. Componenty ANNIE¹⁶

CoreNLP CRFClassifier

Súbor nástrojov CoreNLP obsahuje implementácie dvoch rôznych prístupov ku rozpoznávaniu entít. Pri štatistickom prístupe je možné využiť CRFClassifier, ktorý je možné natrénovať na základe ručne anotovaného korpusu a potom vďaka triede NERClassifierCombiner ho využiť pri anotovaní entít. Pomocou NERCombinerAnnotator je ho možné využiť aj v zreteženom spracovaní. Využíva pritom štatistickú metódu podmienených náhodných polí (Conditional Random Fields

¹⁶ <https://gate.ac.uk/sale/tao/splitch6.html#chap:annie>

CRF¹⁷), kde je možné nakonfigurovať veľké množstvo parametrov, ktoré sa využívajú ako čerty pre trénovanie štatistického modelu.

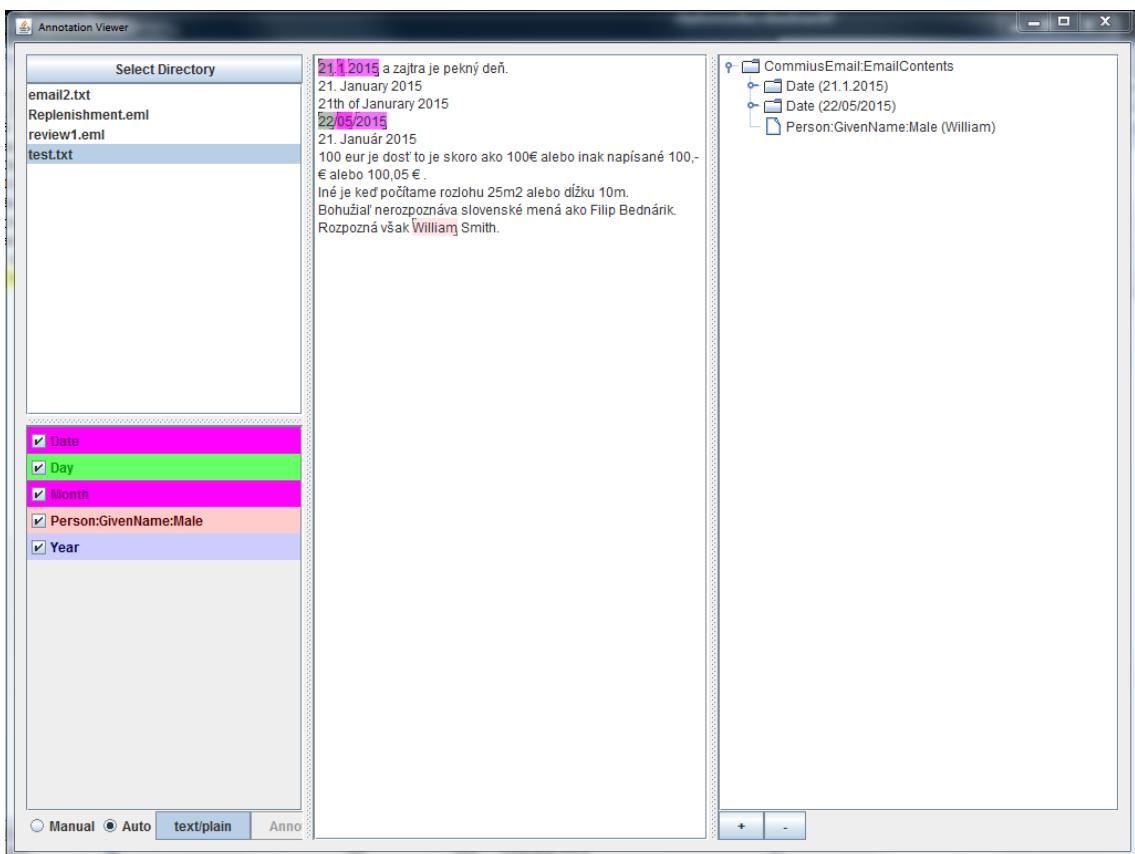
CoreNLP RegexNERAnnotator

Druhým prístupom pri rozpoznávaní pomenovaných entít v CoreNLP je lingvistický (pravidlový) prístup. RegexNERAnnotator využíva jednoduché pravidlá a slovníky pre označenie triedy daného slova. Pre komplikovanejšie pravidlá je možné využiť implementáciu TokenRegexNERAnnotator, ktorá umožňuje tvorbu pravidiel nad sekvenciami tokenov a pri tom využívať anotácie predchádzajúcich krovok, ako lému slova alebo jeho značku, prípadne regulárne výrazy nad jedným slovom alebo sekvenciou slov.

Ontea

Ontea je riešenie z dielne slovenskej akadémie vied. Aktuálne vo verzii 1.0. Toto riešenie je podobne ako ostatné aplikácie v tomto porovnaní skôr knižnicou ako finálnym riešením. Aplikácie poskytuje GUI rozhranie a jednoducho sa používa. Je pomerne dobre konfigurovateľná a metódu extrakcie informácií zakladá na pravidlách, vzoroch a slovníkoch. Ontea bohužiaľ vo verejne dostupnej verzii nepodporuje slovenský jazyk, avšak pri doplnení slovníkov, vzorov a pravidiel je možné toto riešenie využiť aj na anotovanie slovenčiny.

¹⁷ <http://nlp.stanford.edu/software/CRF-NER.shtml>



Obrázok 4. Používateľské rozhranie aplikácie Ontea 1.0

Bc. Lubomír Sedlář - Hybridní systém pro detekci pojmenovaných entit v českém textu¹⁸

Projekt diplomovej práce napísaný v jazyku Python, určený pre český jazyk, ktorý je veľmi príbuzný slovenskému jazyku. Aj keď názov práce hovorí o hybridnom systéme, systém je založený na pravidlách a slovníkoch. Stiahnutelná verzia neobsahuje slovníky a webová verzia nie je funkčná (otestované v decembri 2015). Porovnávali sme preto len prístup k riešeniu problémov a nie finálnu aplikáciu.

Ondrej Kaššák – Extraktcia pomenovaných entít zo slovenského textu

Projekt bakalárskej práce študenta na FIIT STU. Aplikácia je dostupná online (december 2015) ¹⁹. Metóda pre extrakciu entít využíva korpus Wikipédie, ktorý obsahuje veľké množstvo entít, ktoré sú zaradené v taxonómii. Pri spracovaní číselných hodnôt využíva rozhodovací strom a na spracovanie dátumov regulárne výrazy. Pri spracovaní prvých slov vo vetách si pomáha Slovenským národným korpusom.

¹⁸ http://is.muni.cz/th/359719/fi_m/

¹⁹ <http://mus.fiit.stuba.sk/ner/>

4.2.2 Extraktia dátumov a iných časových údajov

Dátumy sú z pohľadu extrakcie informácie dôležitou informáciou. Vnášajú totiž do údajov temporálnu zložku, pri ich rozpoznaní môžeme vylepšiť vyhľadávanie (vieme poskytovať novšie informácie za predpokladu, že tie staršie postupe strácajú relevanciu), ďalej ich využívame pri extrakcii udalostí a ich zasadenia do časovej osi, ale aj ako základný poznatok o rôznych entitách. Na Slovensku spolu s menom, priezviskom a miestom narodenia ich využívame, aj ako spôsob identifikácie osoby.

V prirodzenom jazyku sa stretávame s rôznym vyjadrením časových údajov. Ako príklad si môžeme uviest' dátum *24.12.2015*. Ekvivalentom takého vyjadrenia je *štodrý deň v roku 2015*. Častokrát musíme vychádziať z kontextu a pri slovných vyjadreniach časového údaju sa predpokladá znalosť kontextu. To znamená, že ak v texte uvádzame: *páchatel sa dopustil trestného činu 21.1.2014 a na druhý deň opustil krajinu*. Dostávame dva časové údaje, prvý v číselnom vyjadrení (21.1.2014) a druhý v slovnom vyjadrení (na druhý deň), ktorý na základe kontextu vyhodnotíme ako dátum 22.1.2014.

OpenNLP Name Finder

OpenNLP je možné natrénovať, aby rozpoznával aj dátumy. Na takúto pomerne jednoduchú úlohu, však potrebuje veľkú trénovaciu množinu a stále je možné, že nebude dosahovať dostatočné výsledky. Riešenie sa dá samozrejme prispôsobiť, ale jediná výhoda oproti vlastnému riešeniu, je podpora niektorých základných operácií samotným nástrojom.

ANNIE NE Transducer

ANNIE rozpoznáva dátumy na základe regulárnych výrazov a pravidiel. Dokáže rozpoznávať aj výrazy ako „early October“ a podobne. Pravidlá sú napísané v pravidlovom jazyku JAPE²⁰. Bohužiaľ, v slovenčine neexistuje implementácia, ktorá by podporovala slovné vyjadrenia časových údajov a jednotlivé pravidla obsahujú natvrdo zabudované anglické výrazy.

CoreNLP SUTime

SUTime je temporálny značkovač pre rozpoznávanie a normalizáciu časových výrazov v anglickom teste^[11]. Je založený na jednoducho rozšíriteľných pravidlách. Pri anotácii časových výrazov využíva formu TIMEX3 značiek a všade, kde je možné dodržiava štandard ISO 8601. Časové entity dokáže normalizovať a reprezentovať

²⁰ <https://gate.ac.uk/sale/tao/splitch8.html#chap:jape>

v podobe Java tried. Nástroj tiež podporuje normalizáciu vzhľadom na referenčný dátum nájdený v texte. Rozlišuje základné triedy:

- Date - Dátum
- Time - Čas
- Duration - Trvanie
- Set – Súbor časových údajov. Ide najmä o opakujúce sa udalosti ako „ročne“ alebo „mesačne“.

Ontea

Ontea v základnej verzii obsahuje menšie množstvo pravidiel a malú znalostnú bázu. Nedokáže však rozpoznávať mesiace v dátumoch a slovné vyjadrenia dátumov.

Bc. Lubomír Sedlář - Hybridní systém pro detekci pojmenovaných entit v českém textu

Systém dokáže detegovať dátumy na základe regulárnych výrazov. Dokáže rozpoznávať aj mesiace v českom jazyku.

Ondrej Kaššák – Extraktcia pomenovaných entít zo slovenského textu

Riešenie dokáže spracovať dátumy v rôznych formátoch. Využíva pritom regulárne výrazy a dokáže rozpoznávať aj slovné pomenovania mesiacov. Toto riešenie však nedokáže spracovávať časové údaje uvedené len slovne, prípadne vypočítať základný tvar dátumu alebo času.

Iné riešenia

Veľmi pekne spracované a vysvetlené riešenie pre vzorové spracovanie vyvinul študent FIIT STU Rastislav Masaryk²¹, ktorý implementoval riešenie v rámci predmetu Vyhľadávanie informácií. Toto riešenie využíva podobne ako ostatné riešenia regulárne výrazy a ošetruje pomerne veľké množstvo spôsobov vyjadrenia časových údajov.

Časové údaje je možné vyadrovať aj čisto slovne a v tomto prípade je potrebné brat' do úvahy kontext a časový rámec, od ktorého sa slovný výraz odráža. Existujú knižnice, ktoré si dokážu so slovným vyadrovaním poradiť, ale sú dostupné len v angličtine. Ako príklad je možné uviesť knižnicu chrono²², ktorá dokáže spracovať časové vyjadrenia v prirodzenom jazyku.

²¹ <http://vi.ikt.ui.sav.sk/User:Rastislav.Masaryk?view=home>

²² <http://wanasit.github.io/pages/chrono/>

4.2.3 Číselné hodnoty

Pri spracovaní textu sa stretávame s rôznymi číselnými hodnotami ako sú sumy, poradia, veľkosti a dĺžky a podobne. Pri takýchto hodnotách sa najčastejšie využívajú regulárne výrazy na extrakciu z textu. Riešenia využívajú rovnaké postupy, ako pri spracovaní dátumov.

4.2.4 Osoby (mená a priezviská)

Pre extrakciu mien a priezvisiek sa najčastejšie používajú slovníky, ktoré sú rozšírené o pravidlá výskytu titulov. Tieto slovníky aplikácia využíva na klasifikovanie entity. Štandardným problémom pri rozpoznávaní osôb, je problém s prekryvaním osôb s adresami. Na základe analýzy registra adries a zoznamu mien a priezvisiek: 45% obcí a 50% ulíc obsahuje v názve slovo, ktoré je zhodné so stenom priezviska osoby.

Pre slovenčinu:

Bc. Lubomír Sedlář - Hybridní systém pro detekci pojmenovaných entit v českém textu

Riešenie využíva slovníky obsahujúce zoznamy všetkých českých mien a priezvisiek a titulov, ktoré sa môžu nachádzať pred menom alebo za menom. Systém ignoruje jednoslovné pomenovania osôb na začiatku vety, čím zvyšuje presnosť na úkor pokrytieia.

Ondrej Kaššák – Extraktia pomenovaných entít zo slovenského textu

Aplikácia má jednoduchý algoritmus a vyhľadáva krstné mená a tituly. Ak v ďalšej postupnosti nájde slová s veľkým písmenom označí ich za priezvisko a súčasť osoby.

Ontea

Ontea podporuje rozpoznávanie krstných mien pomocou slovníkov. Pre daný jazyk je potrebné mať zoznam všetkých dostupných krstných mien. Dostupné v základnej verzii sú anglické, španielske a talianske. Tieto slovníky je možné ukladať v hierarchii a teda, ak poskytneme zoznamy mužských a ženských mien, aplikácia dokáže túto informáciu využiť pri anotácii a určiť pohlavie osoby.

ANNIE NE Transducer a ANNIE Gazetteer

Riešenie je založené na pravidlach napísaných v jazyku JAPE a slovníkoch. Na základe slovníkových údajov dokáže rozpoznávať vo veľkom počte prípadov aj pohlavie osoby.

OpenNLP Name Finder

OpenNLP v tomto prípade začína ukazovať svoju silu, keďže získať kompletný zoznam všetkých mien a priezvisk je pomerne zložité a zároveň množstvo priezvisk je odvodených od bežne používaných slov a najmä profesií, je vhodné využiť kontext pre identifikáciu osoby. OpenNLP na základe natrénovania štatistickej metódy teda dokáže rozpoznávať aj osoby, ktoré by technológia založená na slovníkoch nerozpoznala.

Riešenie vyvinuté na FIIT STU na predmete Vyhladávanie informácií, ktoré prezentoval Bc. Michal Jesenský²³ využíva OpenNLP, pomocou ktorého autor natrénoval štatistický model pre rozpoznávanie osôb na základe SNK. Toto riešenie však nedosahuje dostatočne dobré výsledky a má nízke pokrytie.

CoreNLP CRFClassifier

CoreNLP CRFClassifier podobne ako OpenNLP Name Finder využíva štatistický model pre klasifikáciu osôb a ich extrakciu z textu. Rozdiel je v použitom druhu štatistického modelu (CoreNLP využíva CRF a OpenNLP MEMM).

4.2.5 Firmy

Pre vyhladávanie firiem sa využíva to, že firmy v názve častokrát obsahujú identifikátory ako s.r.o., a.s. a podobne. Tiež sa využívajú rôzne klúčové slová ako firma, spoločnosť alebo organizácia. Pri rozpoznávaní firiem je možné tiež využiť zoznam existujúcich firiem, prípadne službu pre dopytovanie obchodného registra. Avšak takýto prístup je viazaný na určitú krajinu.

Pre slovenčinu:

Ondrej Kaššák – Extraktia pomenovaných entít zo slovenského textu

Riešenie využíva kontextové slová ako firma, organizácia, s.r.o. a podobne pre vyhľadanie organizácií v texte.

Pre češtinu (čeština je veľmi príbuzný jazyk slovenčine a mnohé riešenia sú priamo aplikovateľné na slovenčinu prípadne je potrebných pári úprav).

Bc. Lubomír Sedlár - Hybridní systém pro detekci pojmenovaných entit v českém textu

Riešenie využíva kontextové slová ako organizácia, firma a pod., predložky často spájajúce sa s firmami a jednoduché pravidlá a regulárne výrazy. Využíva heuristiky, ktoré uprednostňujú presnosť pred pokrytím. Ak sa vo viacslovnom názve vyskytuje

²³ http://vi.ikt.ui.sav.sk/index.php?title=User:Michal.Jesensky/Projektov%C3%A9_zadanie

veľké písmeno na začiatku slova, druhé slovo začína s malým písmenom a tretie zase s veľkým, prvé dve slová budú odignorované.

Angličtina:

Ontea

Ontea podporuje rozpoznávanie firiem, avšak podľa našich skúseností, reálne vyhľadáva výskyty kľúčových slov ako s.r.o., ltd., a teda len typ spoločnosti a neanotuje spoločnosť samu. To ale pravdepodobne nie je veľký problém dorobiť. Pre svoju funkcionality využíva zoznam typov organizácií.

ANNIE NE Transducer

ANNIE využíva súbor pravidiel, ktoré sú určené len pre anglický jazyk, na základe ktorých vyhľadáva v texte organizácie. Využíva pritom podobne ako ostatné riešenia kontextové slová ako *company* (spoločnosť), *organisation* (organizácia) a podobne.

OpenNLP

Technológia OpenNLP dosahuje pomerne dobré výsledky. V prípade pridania novej črty štatistického modelu, ktorou by bol výskyt typu spoločnosti zo slovníka za názvom spoločnosti, by táto technológia mohla dosahovať vynikajúce výsledky.

CoreNLP CRFClassifier

CoreNLP využíva štatistickú metódu CRF, ktorú je potrebné natrénovať na dostatočne veľkom anotovanom korpusе. V základnej verzii je natrénovaná na MUC 6 a MUC 7 dátach v angličtine.

4.2.6 Adresy a geolokačné entity

Pod geolokačné entity patria najmä: krajinu a štáty, kraje, oblasti, okresy, mestá, dediny, ulice, popisné čísla domov, PSČ a GPS súradnice. Pri väčšine z týchto entít postačuje kvalitný zoznam alebo ontológia obsahujúca všetky entity, ktoré by sa mohli v doménovo špecifickom teste nachádzať. Niektoré aplikácie dokážu dokonca prekladať adresy na GPS súradnice a naopak, čo má využitie pri rôznych úlohách v IE (Google, GIS, ...).

Pre PSČ existujú regulárne výrazy a využíva sa aj skutočnosť proximity názvu okresu.

Ontea

Ontea sa v prípade adres spolieha na regulárne výrazy a čiastočne na slovník, ktorý obsahuje krajinu a mestá. Takýto prístup si poradí aj so slovenskými adresami.

Ondrej Kaššák

Autor využíva znalostnú bázu wikipedie, kde vyhľadáva jednotlivé entity a klasifikuje ich podľa meta informácií uložených na stránke wikipedie. Špeciálne ošetuje skrátené názvy a názvy oblastí s veľkými písmenami.

Bc. Lubomír Sedlár - Hybridní systém pro detekci pojmenovaných entit v českém textu

Riešenie využíva slovníky, ktoré autor získal zo zdroja Openstreetmap²⁴ a Českého štatistického úradu. Pre rozpoznanie PSČ využíva regulárny výraz. Takýto prístup však nedokáže vyhľadať ustálené neoficiálne názvy lokalít a oblastí, ktoré sa nenachádzajú v slovníkoch. Adresy však rozpoznáva pomerne spoľahlivo.

ANNIE NE Transducer a ANNIE Gazetteer

ANNIE využíva súbor krátkych pravidiel pre identifikáciu adresy a slovníky, ktoré obsahujú lokality.

4.2.7 Telefónne čísla a emaily

Telefónne čísla a emaily sa extrahujú zvyčajne pomocou regulárnych výrazov. Vo väčšine prípadov sa rozpoznávajú už pri tokenizácii.

4.2.8 Doménovo špecifické entity

Medzi doménovo špecifické entity patria najmä: zákony, evidenčné čísla, identifikačné čísla ako číslo poistky, číslo OP, VP a pod., čísla účtov, IČO, spisové značky, čísla listov vlastníctva a parciel a podobne. Väčšinu z týchto entít je možné rozpoznať pomocou regulárnych výrazov a jednoduchých pravidiel. V existujúcich riešeniach nie sú tieto entity podporované, no všetky riešenia poskytujú možnosť rozšírenia o nové pravidlá, ktoré by tieto entity zahŕňali.

²⁴ <https://www.openstreetmap.org>

Tabuľka 3. Porovnanie NER nástrojov

	OpenNLP	CoreNLP	ANNIE	Ontea	Ondrej	Lubomír
				Kaššák	Sedlár	
<i>Podpora slovenského alebo českého jazyka</i>	✗	✗	✗	✗	✓	✓
<i>Volne stiahnutelné</i>	✓	✓	✓	✓	✗	✓
<i>Webová služba</i>	✗	✗ ²⁵	✗ ²⁵	✗	✓	✗ ²⁶
<i>Rozpoznávanie dátumov</i>	✗ ²⁷	✓	✓	✓	✓	✓
<i>Rozpoznávanie osôb</i>	✗ ²⁷	✗ ²⁷	✗ ²⁸	✗ ²⁸	✓	✗ ²⁹
<i>Firmy</i>	✗ ²⁷	✓	✓	✗ ³⁰	✓	✓
<i>Adresy</i>	✗ ²⁷	✓	✓	✓	✓	✓
<i>Metóda</i>	štatistická lingv.	štatistická, lingv.	lingv.	lingv.	lingv.	lingv.
<i>Technológie</i>	MEMM	CRF, Slovník, Regexp, Pravidlá	JAPE, Slovník Regexp	Slovník, Regexp, Pravidlá	Wiki, Regexp, Slovník, Pravidlá	Regexp, Slovník, Pravidlá
<i>Programovací jazyk</i>	Java	Java ³¹	Java	Java	Ruby	Python

²⁵ Online verzia je značne limitovaná

²⁶ V čase testovania nefunkčné (december 2015)

²⁷ Je potrebné trénovanie štatistického modelu

²⁸ Dokáže rozlišovať aj pohlavie osoby

²⁹ Nerozpoznáva jednoslovné mená na začiatku vety

³⁰ Nerozpoznáva samotné firmy ale len skratku typu firmy

³¹ Dostupné aj v iných jazykoch

4.3 Otvorené problémy a nedostatky technológií

Väčšina technológií využíva iba jeden z prístupov pre extrakciu pomenovaných entít. Riešenia bud' využívajú lingvistický alebo štatistický prístup. Oba prístupy majú svoje výhody a nevýhody a ich kombináciou a správnym ohodnotením jednotlivých čít, je možné dosiahnuť vo väčšine prípadov lepšie výsledky.

Dôležitým krokom je aj predspracovanie textu, na ktorý sa nekladie dostatočný dôraz. Najmä v slovenčine je problém s nedostatkom dostupných riešení a jazykových modelov pre predspracovanie. Tokenizácia a oddelovanie viet sú základné prvky predspracovania, ktoré umožňujú segmentáciu textu na menšie časti a zlepšenie presnosti a pokrytie NER.

Problémom pri extrahovaní samotných entít je najmä rozlišovanie adries a mien osôb, ako aj problém pri extrahovaní názvov organizácií, ktoré môžu obsahovať rôzne tvary slov a znakov. Problémom v slovenčine je tiež extrakcia temporálnych entít.

5 Extraktia informácií pre anonymizáciu dát

5.1 Anonymizácia - charakteristika problému

V súčasnosti sa zvyšuje trend otvárania údajov verejnosti. Pri zdieľaní niektorých dokumentov, však dochádza ku možnému narušeniu súkromia osôb, o ktorých sa v dokumentoch píše. V takomto prípade je potrebné dokumenty prehľadat' a nahradit' všetky osobné údaje tak, aby nebolo možné osoby alebo ich osobné informácie identifikovať, ale aby dokument nestratil výpovednú hodnotu. Tento proces sa nazýva anonymizácia a je sprevádzaný rôznymi problémami.

Postup anonymizácie:

1. Vyhľadanie a anotácia všetkých osobných údajov, mien, dátumov, adres, identifikačných čísel a podobne.
2. Vyhodnotenie, či sa má entita anonymizovať alebo nie.
3. Nahradenie údaju v texte za nič nehovoriacu skratku alebo identifikátor.

5.2 Špecifická súdnych rozhodnutí

Súdne rozhodnutie vydáva príslušný súd na konci konania na danom súde. V súdnych rozhodnutiach sú informácie o výsledku konania, informácie o navrhovateľovi, odporcovi a iných detailoch konania, najmä tie podľa ktorých sa súd rozhodol. Povaha informácií závisí značne od oblasti právnej úpravy (trestné právo, občianske právo, rodinné právo, obchodné právo a správne právo) a forme rozhodnutia (uznesenie, rozsudok, platobný rozkaz, trestný rozkaz a ďalšie). Podľa inštrukcie 24/11 Ministerstva spravodlivosti Slovenskej republiky (viac v prílohe D), musia súdy a Ministerstvo zverejňovať súdne rozhodnutia na stránke do 15 dní od právoplatného ukončenia konania³². Pred tým, než však môžu takéto rozhodnutie zverejniť, je potrebné ho anonymizovať. Pri bežnom počte cca 40 000 rozhodnutí za mesiac³³, zabera anonymizácia veľké množstvo človekohodín, ktoré by súdy vedeli využiť pre skvalitnenie poskytovania služieb a zrýchlenie súdnych procesov. Automatizácia procesu anonymizácie by ušetrila veľké množstvo zdrojov a zjednodušila by prácu desiatkam pracovníkov.

Podľa inštrukcie sa anonymizácia musí držať istých pravidiel. Tieto pravidlá však majú svoje výnimky, ktoré sú viazané na isté situácie.

Z analýzy anonymizovaných súdnych rozhodnutí sme prišli na to, že práve tieto informácie sa **anonymizujú**:

³² http://www.justice.gov.sk/ip/ira/instr_24_2011.pdf

³³ <https://obcan.justice.sk/infosud/-/infosud/zoznam-rozhodnutie>

Údaje o fyzickej osobe:

- Meno a priezvisko (zväčša navrhovateľ a odporca; zástupca, správca, likvidátor, súdny úradník sa neanonymizuje)
- Dátum narodenia
- Dátum smrti (v niektorých prípadoch)
- Rodné číslo
- Číslo vodičského preukazu
- Číslo občianskeho preukazu
- Adresa trvalého, prechodného bydliska a pre doručovanie
- Miesto narodenia
- Spisová značka týkajúca sa fyzickej osoby (iba niektoré, špecifické)

Údaje o právnickej osobe:

- Názov firmy (v špeciálnych prípadoch)
- IČO (v špeciálnych prípadoch)

Všeobecné:

- Číslo účtu
- Číslo listu vlastníctva a číslo parcely
- Číslo rozhodnutia sociálnej poisťovne
- Číslo spotrebiteľského úveru
- Evidenčné číslo vozidla
- Výška pokuty

Čo sa neanonymizuje:

- Názov firmy (vo väčšine prípadov)
- Predseda senátu, zástupca, správcova, likvidátor, súdny úradník
- Dátum udalosti
- Názov banky
- Značka správcu
- Miesto a adresa činu (nie vždy)

Pravidlá anonymizácie:

- mená, priezviská a ostatný anonymizovaný text sa nahradia vybranými zamenenými písmenami
- číslica sa nahradí znakom „X“

Počas analýzy sme našli informácie, ktoré by nemali byť zverejnené a naopak niekoľko informácií, ktoré mohli byť zverejnené, ale neboli. Túto skutočnosť môžeme ďalej využiť pri vyhodnocovaní úspešnosti našej metódy voči ľudskému anotátorovi.

5.3 Existujúce riešenia pre anonymizáciu textu

Existujúce riešenia pre anonymizáciu textov, písaných v prirodzenom jazyku sú väčšinou určené pre anonymizáciu medicínskych záznamov a podporujú iba anglický jazyk. Existujú aj riešenia zaobrajúce sa anonymizáciou tabuľkových dát (napr. ARX³⁴), ktoré nemajú črty extrakcie informácií a v tejto práci nebudeme rozoberať ich prístupy.

5.3.1 Deid

Deid je nástroj vyvájaný študentmi na americkej MIT, napísaný v jazyku PERL a určený pre anonymizáciu lekárskych záznamov.

Aplikácia dokáže anonymizovať rôzne typy údajov: mená, adresy, dátumy týkajúce sa pacienta (neanonymizuje roky), vek pacienta (ak je vyšší než 89 rokov), telefónne čísla, emaily, zdravotné, sociálne a iné identifikačné čísla, čísla zdravotných záznamov, webové adresy a čísla účtov.

Aplikácia využíva jednoduché, najmä lingvistické metódy pre klasifikáciu pomenovaných entít.

Pri menách využíva tituly, slovníky a kontext. Využívajú taktiež databázu pacientov, ktorá im zjednodušuje identifikáciu osôb.

Pri dátumoch využíva regulárne výrazy. Autori tvrdia, že rozpoznávať typy dátumov je príliš komplikované, preto to nerobia a anonymizujú všetky dátumy. Dátumy anonymizujú pridávaním náhodných časových dĺžok, ktoré sú konzistentné v rámci pacienta, aby bolo možné sledovať rozdiely dátumov, ale nie dátum, kedy bol pacient liečený. Aplikácia tiež anonymizuje časové údaje vyjadrené sviatkom alebo významnou opakujúcou sa udalosťou.

Lokality sú anonymizované pomocou zoznamov. Keďže ide o voľný text, aplikácia berie do úvahy preklepy. Lokality, ktoré nie sú v slovníku, aplikácia rozpoznáva vďaka definovaným kontextovým výrazom.

Pre anonymizáciu telefónnych čísel, emailov a iných identifikačných alebo komunikačných čísel a identifikátorov s fixným formátom, riešenie využíva regulárne výrazy^[20].

³⁴ <http://arx.deidentifier.org/>

5.3.2 NLM-Scrubber

Nástroj NLM-Scrubber je napísaný v jazyku perl (zistené pomocou nástroja Resource Hacker) a skompilovaný do konzolovej aplikácie, dostupný pre Linux, Windows a Mac OS X. Pri vývoji, pomocou vizuálneho nástroja pre značkovanie, ručne označili väčšie množstvo dokumentov. Nástroj je opäť zameraný na rovnaké entity ako Deid^[16].

Pri anotácií mien na rozdiel od Deid označuje NLM-Scrubber aj typ osoby (pacient, príbuzný, zamestnávateľ, poskytovateľ zdravotnej starostlivosti a skupina ostatné). Pri anonymizácii osôb využíva Bayesov naivný klasifikátor a konečný stavový automat^[15].

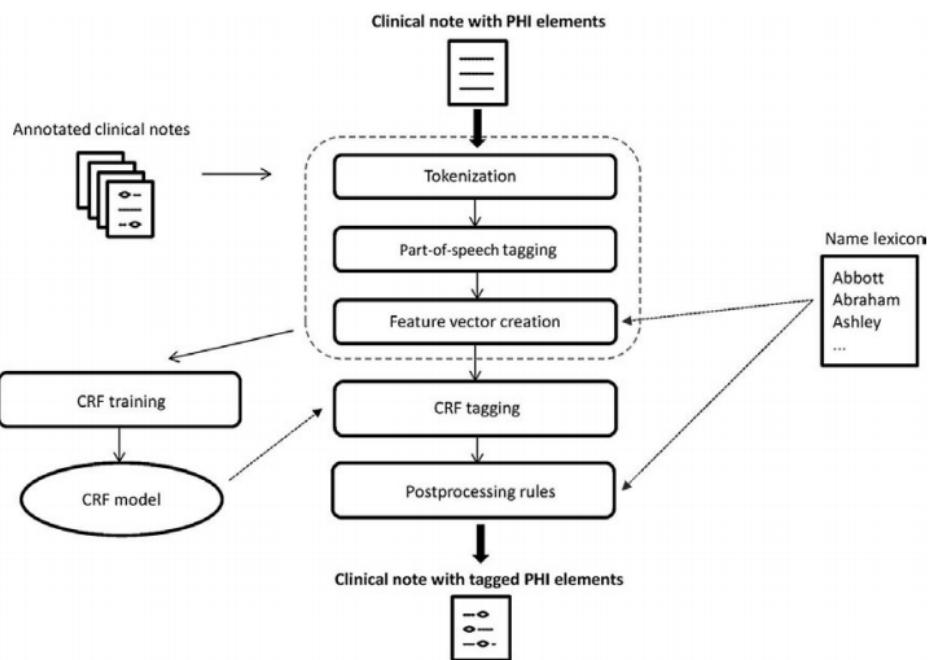
```
Date of Consult:  
January 1, 2020  
  
Reason for Consult:  
Fever of unknown origin  
  
History of Present Illness:  
SpongeBob SquarePants is an 18 month old Hispanic boy who presented to his pediatrician's office with a fever of 104.4 degrees. No known source of fever could be determined and the patient was admitted to the Pediatric Emergency Room at St. Children's Hospital earlier this afternoon. The patient was medicated with Tylenol at home at 9:30am.  
  
Past Medical History:  
SpongeBob was the product of a normal vaginal delivery at 39 weeks. SpongeBob is a healthy child with no history of illness or previous hospitalizations. The patient is up to date with his immunizations, with his last given at his pediatrician's office, Dr. Walt Disney in Disney World, Florida last Tuesday.  
  
Family History:  
SpongeBob's mother, Suzy, is a 32-yr old woman with a history of hypothyroidism. She is a homemaker. SpongeBob's father, Steven, is 38-yrs old with no medical problems. He is the current Director of Cartoons at Kindergarten Studios. SpongeBob has three siblings: a 12-yr old brother, Joey, and identical twin sisters, Sarah and Sara, five years of age, all of whom are healthy.  
  
Consult:  
Patient was seen in the Pediatric ER, Room 123-ABC. Vital signs are as follows: temp: 40.1 degrees Celsius, weight: 15.2 kgs, pulse ox: 99%; HR: 138; BP: 94/40. Head: PEARLA, tympanic membranes clear, copious amount of mucous noted in bilateral nares with drainage, no cough noted. Lungs: clear. Heart: normal rate and rhythm. Abd: soft and nontender.  
  
Plan:  
Admit patient to Pediatric Protocol #00-AA-0000. Arrange for accommodations for mother at The Sleeping Beauty Inn.  
  
Dr. Harry Potter, Chief Pediatrician pager# 001  
HP1  
  
Please send copy of consult to:  
Dr. Walt Disney, Animated Pediatrics Associates  
Cartoon Building  
123 Sesame Street, Suite #333  
Disney World, Florida 11111-1111
```

Obrázok 5. Anonymizovaný text pomocou nástroja NLM-Scrubber

5.3.3 MITRE Identification Scrubber Toolkit (MIST)

MIST je nástroj taktiež určený na anonymizáciu lekárskych záznamov. Ako základnú metódu využíva podmienené náhodné polia, ktoré ako štatistický model trénuje pomocou ľudskej manuálnej anotácie. Využívajú CARAFE engine pre prácu

s podmienenými náhodnými poľami a po procese značkovania používajú modul syntézy, ktorý nahradza osobné údaje anonymizovanými.



Obrázok 6. Postup krokov anonymizácie pomocou nástroja MIST

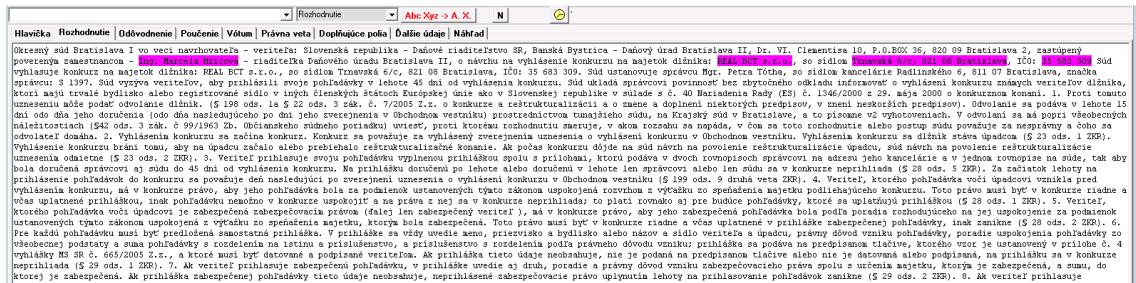
5.3.4 Riešenia pre anonymizáciu v slovenčine

V slovenskom jazyku nie sú voľne dostupné nástroje na anonymizáciu voľných textov. V doméne justície inštrukcia č. 24/2011 hovorí o použití aplikácie Súdny Manažment, ktorá obsahuje nástroj na anonymizáciu.

Anonymizáciou v slovenčine sa zaoberal aj autor diplomovej práce s názvom "Anonymizácia a ochrana dát", ktorá sa zaoberá najmä anonymizáciou štruktúrovaných údajov. Autor v práci opisuje právnu a sociálnu zložku problematiky a navrhuje spôsob merania anonymizácie [13]. Táto práca je zaujímavá skôr z pohľadu aplikácie anonymizácie ako z technického hľadiska.

Existujúce riešenie pre anonymizáciu súdnych rozhodnutí

Na súdoch používajú poverení zamestnanci aplikáciu Súdny Manažment, ktorá umožňuje okrem iného správu spisov a vydaných rozhodnutí. Po vyplnení rozhodnutia, používateľ Súdneho Manažmentu môže označiť osobné údaje v texte rozhodnutia vybratím časti textu a kliknutím na tlačidlo „Abc Xyz -> A. X.“. V používateľskom rozhraní sa takýto text označí fialovou farbou.



Obrázok 7. Snímka časti obrazovky aplikácie Súdny Manažment, určená pre písanie súdneho rozhodnutia

Po vyplnení textu rozhodnutia, môže používateľ zvoliť publikovanie rozhodnutia. V takomto prípade je vyžadovaná anonymizácia daného rozhodnutia. Pre tento účel aplikácia otvorí dedikované okno, v ktorom je možné anonymizáciu vykonávať kliknutím na tlačidlo alebo ručným prepísaním textu. V prípade, že používateľ označil v predchádzajúcim kroku osobné údaje, v okne anonymizácie je tento text anonymizovaný. Pomocou tlačidla preskoč, prechádza kurzor po slovách, ktoré obsahujú čísla alebo sa začínajú na veľké písmeno. Používateľ má možnosť aktuálne vysvetnené slovo anonymizovať kliknutím na tlačidlo nahrad’.



Obrázok 8. Snímka časti obrazovky aplikácie Súdny Manažment, určená pre anonymizáciu súdneho rozhodnutia

Nahradzovanie pozostáva zo zvolenia písmena z abecedy, na základe prvého písmena slova a nahradenie slova zvoleným písmenom. Nahradzovanie je v danom rozhodnutí konzistentné, a teda napr. slová „Andrea“ a „Adam“ budú nahradené za „B.“ a „B.“. Čísla sú vždy nahradené za znak „X“.

Aplikácia umožňuje ručnú úpravu textu, a teda anonymizácia nemusí byť konzistentná. Po skončení anonymizácie, je tento dokument označený na publikovanie a zverejnený na stránke rozhodnutí³⁵.

5.4 Zhrnutie existujúcich riešení a ich problémy

Existujúce riešenia pre anonymizáciu neštruktúrovaného textu, sú aktuálne stále vo vývoji a výskume, čo potvrdzuje aj aktuálnosť článkov o tejto problematike. Existuje niekoľko riešení, ktoré používajú rôzne technológie a majú rozdielnú úspešnosť. Nástroje sú dostupné najmä pre anglický jazyk, ale pre slovenčinu nie je aktuálne dostupný nástroj pre automatizovanú anonymizáciu voľného textu.

Tabuľka 4. Porovnanie nástrojov MIST, Deid a NLM-S^[15].

Entita	Zlatý počet	Systém	Pokrytie	Špecifickosť	Správnosť	Presnosť	F-Skóre	F2-Skóre
<i>Meno pacienta</i>	2388	NLM-S	0,9992	0,9785	0,9785	0,0884	0,1625	0,3265
		MITdeid	0,9393	0,9829	0,9829	0,1032	0,1860	0,3586
		MIST	0,7425	0,9969	0,9963	0,3305	0,4574	0,5943
<i>Alfanumerické ID</i>	4165	NLM-S	0,9995	0,9926	0,9926	0,3299	0,4960	0,7109
		MITdeid	0,3467	0,9984	0,9960	0,4404	0,3880	0,3621
		MIST	0,9822	0,9984	0,9984	0,6940	0,8133	0,9069
<i>Adresa</i>	292	NLM-S	0,8356	0,9970	0,9969	0,0658	0,1219	0,2501
		MITdeid	0,4418	0,9988	0,9986	0,0829	0,1395	0,2367
		MIST	0,8562	0,9990	0,9989	0,1756	0,2914	0,4823
<i>Dátum</i>	29134	NLM-S	0,9893	0,9993	0,9991	0,9753	0,9823	0,9865
		MITdeid	0,9472	0,9990	0,9977	0,9619	0,9545	0,9501
		MIST	0,9922	0,9978	0,9977	0,9220	0,9558	0,9773

Z výsledkov môžeme vidieť, že úspešnosť riešenia značne závisí od anonymizovanej entity a od povahy anonymizovaných údajov. Pri porovnávaní riešení autori zámerne neuvádzajú presnosť ich riešenia, keďže ju majú dosť nízku. Toto je spôsobené najmä veľkým počtom nesprávne označených tokenov pre anonymizáciu.

Je na zváženie používateľa anonymizačného nástroja, či použije radšej nástroj, ktorý anonymizuje viac než by mal a pokryje radšej väčšie množstvo osobných údajov alebo označí osobné údaje, ktoré majú a nemajú byť anonymizované presnejšie.

Jedným z problémov je používanie len jednej metódy (lingvistickej, či štatistickej) pri identifikovaní entít, pričom hybridná metóda by dosahovala lepšie výsledky.

³⁵ <https://obcan.justice.sk/infosud/-/infosud/zoznam/rozhodnutie>

6 Otvorené problémy

Pri analýze predspracovania, rozpoznávania entít a procesu anonymizácie sme identifikovali nasledujúce problémy:

- Predspracovanie slovenského textu. Slovenčina je komplikovaný jazyk a súčasné metódy predspracovania textu svojou nižšou úspešnosťou, znižujú úspešnosť aj metód extrakcie informácií, ktoré sa spoliehajú na kvalitné predspracovanie.
- Rozpoznávanie entít osôb. Problematické je najmä rozpoznávanie osôb, keď veľké písmená sa môžu vyskytovať na začiatku vety.
- Rozpoznávanie entít firm. Problémom je korektná detekcia firm a tiež ohraničenie názvov firm, keďže firmy môžu mať rôznu veľkosť počiatočných písmen slov ako aj rôzne znaky.
- Rozpoznávanie slovných časových údajov v slovenčine. V oblasti spracovávania časových údajov je priestor na zlepšenie existujúcich riešení.
- Rozpoznávanie lokalít. Až 50 % adres obsahuje meno alebo priezvisko osoby.
- Rozlišovanie entít pre potreby anonymizácie. Najväčší problém súčasných anonymizačných nástrojov je nízke využívanie znalostí vydolovaných z textu pri rozhodovaní o anonymizácii entity.
- Presnosť a pokrytie. Pri anonymizácii je dôležitá požiadavka na pokrytie anonymizačným nástrojom a aj jeho presnosť, čo nie každé súčasné riešenie spĺňa dostatočne.

7 Ciele

Hlavným cieľom práce je navrhnuť metódu a implementovať riešenie, ktoré bude schopné identifikovať pomenované entity a rozhodnúť sa, či pomenovanú entitu anonymizovať, alebo nie, na základe atribútov entít.

Ciele pri návrhu:

Navrhnuť metódu extrakcie informácií, ktorá:

- Dokáže predspracovať text tak, aby extrahovala dostatočné množstvo informácií z textu. S cieľom využiť tieto informácie pri extrahovaní pomenovaných entít.
- Má vysoké pokrytie a presnosť pri rozpoznávaní pomenovaných entít v danej doméne
- Má vysoké pokrytie a presnosť pri anonymizácii nájdených entít v danej doméne
- Umožní zámennu jazyka a domény

Ciele pri implementácii:

Implementovať softvérový prototyp, ktorý:

- Umožňuje použitie nástrojov extrakcie informácií spolu s navrhnutou metódou ako knižnicu
- Umožní jednoduchú ručnú anotáciu textu cez webové rozhranie aplikácie
- Umožní zadávať úlohy extrakcie informácií cez webové rozhranie aplikácie
- Poskytne integračné rozhranie pre zadávanie úloh extrakcie informácií

Ciele overenia:

- Overiť riešenie na základe metrík na netriviálnom množstve údajov
- Vyhodnotiť úspešnosť anonymizácie voči ľudskému anotátorovi

II. Metóda

8 Metóda

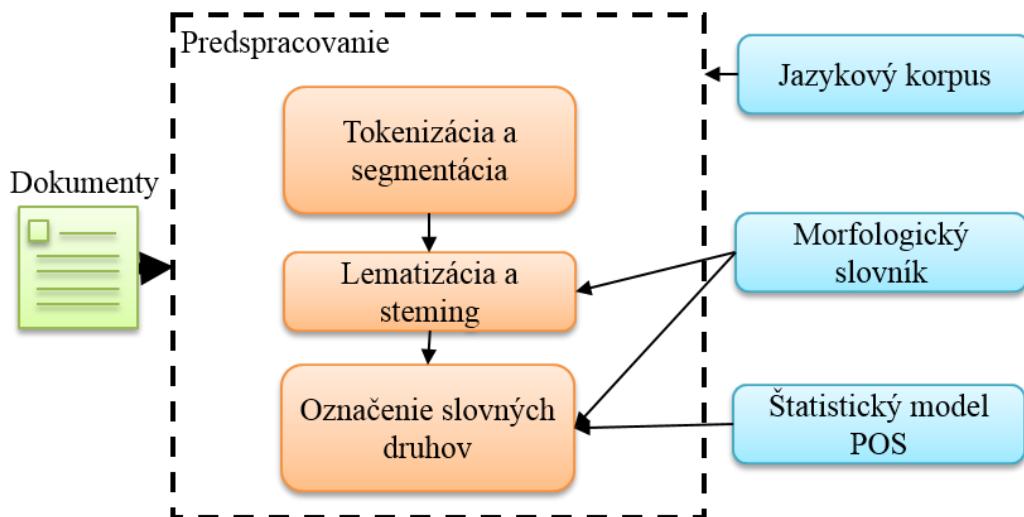
V analýze sme porovnali existujúce riešenia a identifikovali problémy spojené s extrakciou informácií pre účely rozpoznávania entít a anonymizácie. V tejto sekcií prezentujeme metódu, ktorá dokáže anonymizovať dokumenty napísané v prirodzenom jazyku.



Obrázok 9. Diagram znázorňujúci kroky pri anonymizácii dokumentu

8.1 Predspracovanie

Predspracovanie sa skladá z viacerých, po sebe nasledujúcich krokov. Niektoré kroky na seba nadväzujú a je ich potrebné vykonávať v stanovenom poradí. Najprv prebieha sanácia textu, a teda odstránenie zbytočných medzier a neplatných znakov. Následne prebieha tokenizácia a segmentácia na vety. Potom nasleduje lematizácia a stemovanie. V rámci predspracovania sa ešte vykonáva označovanie slovných druhov a pádov. Po predspracovaní dokumentu, je dokument obohatený o extrahované informácie, ktoré boli počas predspracovania získané.



Obrázok 10. Diagram znázorňujúci kroky predspracovania

8.1.1 Sanácia dokumentu

V rámci sanácie dokumentu je potrebné nahradit' všetky neplatné znaky. Sanácia dokumentu je zvyčajne doménovo špecifická a vyplýva z povahy analyzovaného textu. V prípade html dokumentov v tomto kroku prichádza ku odstráneniu nadbytočných značiek a html dekódovanie. V niektorých prípadoch je zase potrebné nahradit'

tabulátory medzerami a odstrániť medzery medzi písmenami v slovách ako „s p á c h a l“ alebo „o d s u d z u j e“.

Tieto medzery odstránime v prípade, ak idú po sebe minimálne dve písmená oddelené medzerami a spolu tvoria slovo z morfológického slovníka a medzi jednotlivými slovami sú dve medzery. Špeciálne výnimky sú samostatne stojace spojenia „a s“, „a o“, „a u“, „a v“, „a k“, „a z“, spolu s variantom, kde na miesto „a“ dáme písmeno „i“. (Pozn.: „as“ je v slovníku).

8.1.2 Tokenizácia

Tokenizácia je dôležitou súčasťou metódy a je vhodné zvoliť tokenizáciu, ktorá ponecháva interpunkciu, korektnie spracuje skratky, dátumy, emaily a iné entity, ktoré je možné rozpoznať už pri tokenizácii. Trénovanie štatistického modelu a tvorba pravidiel je následne jednoduchšia o nutnosť tvorby pravidiel nad sekvenciami slov, čo má pozitívny vplyv na rýchlosť spracovania dokumentu.

8.1.3 Segmentácia viet

Segmentácia viet je dôležitá z pohľadu značkovania slovných druhov, ako aj rozpoznávania pomenovaných entít, kde záleží na poradí slov vo vete, a tiež je dôležitá informácia o hraniciach viet. Segmentáciu je vhodné použiť štatistickú alebo lingvistickú podľa konkrétnej domény. Porovnanie metód môžete nájsť v sekcii 2.1.2.

8.1.4 Lematizácia

Po tokenizácii a segmentácii nasleduje lematizácia. Lematizér využíva morfológický slovník, v ktorom sa nachádzajú základné tvary slova (lémy) a k nim prislúchajúce rôzne tvary slova. Pre zlepšenie úspešnosti navrhujeme postup pri lematizácii: Lematizér sa najprv pokúša nájsť slovo s pôvodnými veľkosťami jednotlivých písmen. V prípade, že sa mu nepodarí nájsť dané slovo, skúša opäť s malými písmenami (riešenie problému slov na začiatku riadku), ak sa mu nepodarí ani tak, transformuje slovo na začiatočné veľké písmeno a ostatné malé (riešenie problému slova „BRATISLAVA“).

8.1.5 Stemovanie

Stemer je vhodný najmä pre slová, ktoré nie sú v slovníku a práve pri rozpoznávaní pomenovaných entít je veľké množstvo tokenov, ktoré sa nenachádzajú v slovníku a nie je preto možné určiť lému slova. V takomto prípade využívame stemer. Slová ako „svojka“ alebo „páčik“, rôzne priezviská, mená a mnoho ďalších, nie sú spisovné slová a ešte sa nedostali do morfológického slovníka, no napriek tomu sa využívajú

v modernej konverzácií, alebo dokumentoch. Tieto slová sa teda nenachádzajú v morfológickom slovníku a nie je ich preto možné lematizovať.

Stemer využíva pravidlá gramatiky, aby odstránil prípony slov. V slovenčine je vhodné využiť pravidlá uvedené v Príručke slovenského pravopisu³⁶.

8.1.6 Označovanie slovných druhov

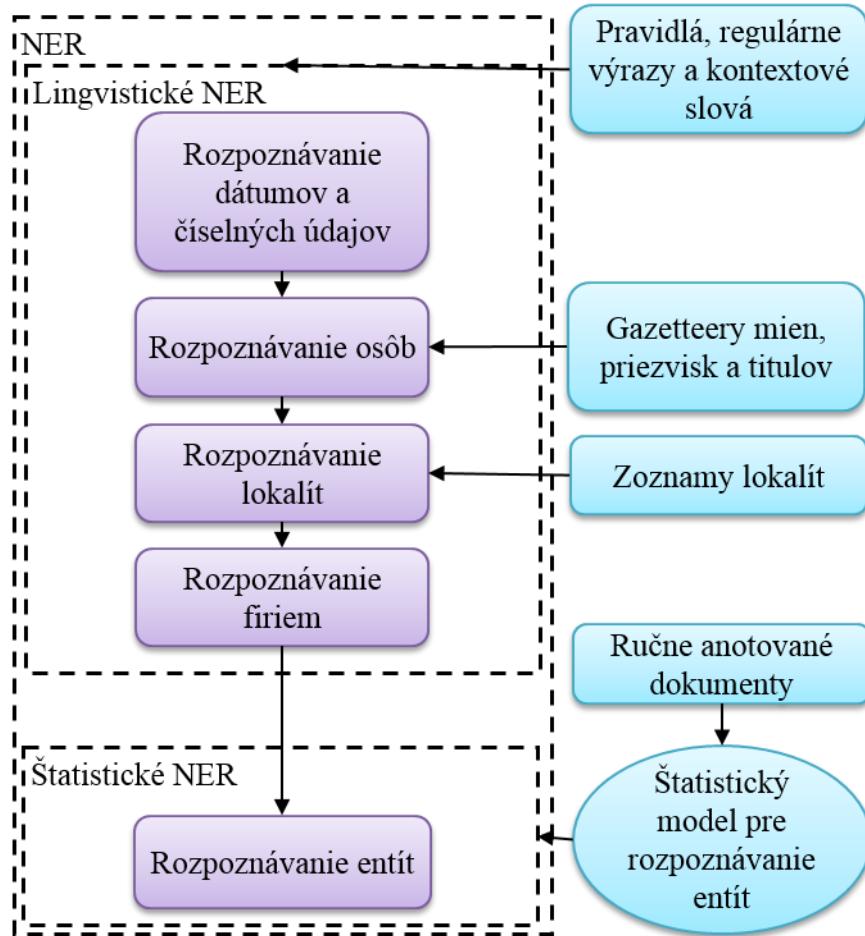
Pre podporu rozhodovania v prípade identifikácie menných entít a procesu anonymizácie, je vhodné vykonať označovanie slovných druhov v texte. Označovanie slovných druhov najlepšie reprezentuje jazykový model a dodáva informáciu o postavení daného slova v kontexte, preto je vhodnou črtou štatistického modelu pri rozpoznávaní pomenovaných entít.

8.2 Rozpoznávanie entít a extrakcia dodatočných informácií

V prvej fáze rozpoznávania entít vyhľadávame známe entity v texte. Známe entity získavame zo slovníkov a regulárnych výrazov. Následne aplikujeme pravidlá, ktoré pokrývajú špecifika jazyka a typické chyby pri anotácii pomocou slovníkov. Tieto pravidlá využívajú zistené entity, informácie z predspracovania, kontextové slová, ale aj regulárne výrazy a typické sekvencie slov.

V druhej fáze využívame všetky doteraz získané informácie, ako črty pri klasifikácii pomocou netrénovaného štatistického modelu. Výstup je potrebné skorigovať pravidlami, ktoré pokrývajú najčastejšie chyby štatistického modelu (tento problém je možné eliminovať väčším množstvom trénovaných údajov).

³⁶ <http://www.juls.savba.sk/ediela/psp2000/psp.pdf>



Obrázok 11. Diagram metódy rozpoznávania entít

8.2.1 Hybridné NER

V hybridnej metóde anotácie pomenovaných entít, je prvým krokom anotácia entít na základe slovníkov a pravidiel. Následne sa extrahované informácie použijú ako črty štatistického modelu.

Extrakcia časových údajov

Časové údaje extrahujeme najmä pomocou regulárnych výrazov. Pri dátumoch je potrebné ošetrovať rôzne spôsoby zápisu, keď je možné oddelovať dni, mesiace a roky lomkou, pomlčkou alebo bodkami. Štandardný formát na Slovensku je *dd.MM.yyyy*. Pri rozpoznávaní využívame zoznam slovných vyjadrení mesiacov a kontextové slová.

Extrakcia číselných údajov

V prípade číselných údajov ide najmä o počty, sumy, vzdialenosť, veľkosť a vek. V doméne súdnych rozhodnutí sme zatiaľ nenašli prípad použitia, v ktorom by bolo nutné anonymizovať uvádzané číselné hodnoty. Ich identifikácia ale môže byť užitočná pri iných problémoch. Sumy, vzdialenosť, plochy objemy a iné číselné hodnoty so

stanovenou veličinou rozpoznávame podľa jednotiek uvedených za číselnou hodnotou, prípadne kontextovým znakom alebo slovom.

Extrakcia kontaktných a osobných informácií

Príkladom takýchto informácií je napr. IČO firiem, emailové adresy, telefónne čísla, čísla občianskeho, zdravotného alebo vodičského preukazu, čísla účtov a kreditných kariet, evidenčné čísla, čísla listov vlastníctva a parciel. Extrakcia je vo väčšine prípadov na základe regulárneho výrazu nad sekvenciou slov s využitím kontextových slov ako „IČO“, „EČ“ a podobne.

Extrakcia osôb

Pri extrakcii osôb je možné použiť zoznamy mien a priezvisk. Jedným z najrozšíahlnejších verejne dostupných zoznamov je zoznam dostupný na stránkach Ministerstva vnútra Českej Republiky³⁷. Zoznam obsahuje väčšinu českých, slovenských, ale aj zahraničných priezvisk, vyskytujúcich sa na našom území.

Detekciu osôb je vhodné ošetriť aj podmienkami a pre lepšie rozpoznávanie zahraničných mien a priezvisk, je vhodné použiť natréovaný štatistický model, ktorý zabezpečí využitie informácií ako pozícia v texte, kontextové slová, ale tiež slovný druh alebo tvar slov.

Extrakcia firiem

Firmy sú identifikovateľné pomocou skratiek ako „s.r.o.“, „a.s.“, „k.s.“ a podobne. Keďže názvy firiem nemusia byť napísané veľkými písmenami, lingvistický prístup má pomerne nízku úspešnosť. Štatistický model na druhú stranu dokáže využiť všetky informácie z predspracovania a kontext pre identifikáciu celého názvu organizácie.

Pri rozpoznávaní firiem je tiež možné využiť integráciu na obchodný register, v ktorom je možné vyhľadať názov organizácie podľa IČO, ktoré sa vyskytuje v texte.

Extrakcia lokalít

Adresy sa rozpoznávajú pomocou slovníkov ulíc a obcí. Následne využívame pravidlá a regulárne výrazy nad sekvenciami tokenov pre identifikáciu adries. Celý proces je dopĺňaný štatistickým modelom, ktorý pomáha identifikovať lokality, ktoré nie sú uvedené v slovníkoch (napr. zahraničné lokality, ktoré nemáme v slovníkoch).

³⁷ <http://www.mvcr.cz/clanek/cetnost-jmen-a-prijmeni.aspx>

Extrakcia súdov

Súdy predstavujú doménovo špecifickú entitu pre oblasť justície. Súdy sú extrahované najmä na základe pravidiel, kontextových slov „krajský“, „okresný“, „súd“ a lokalita. Súdy tiež pokrýva štatistický model.

Extrakcia súdov ako samostatnej entity, aj keď ide o organizáciu má opodstatnenie pri identifikácii pôvodcu textu. Ukázalo sa, že rôzne súdy píšu rozhodnutia rôznym spôsobom. Informácia o súde je teda vhodnou črtou do štatistického modelu.

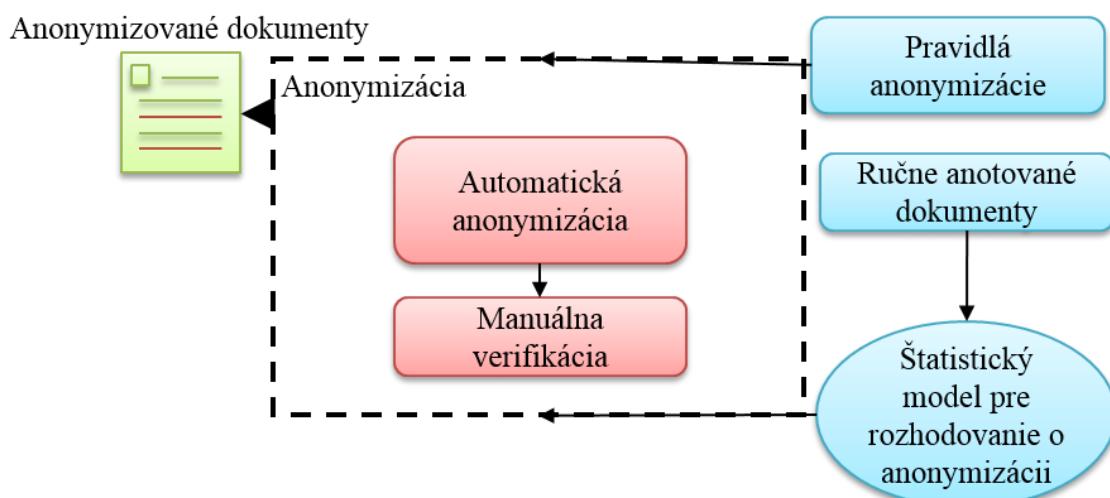
Vo väčšine prípadov je možné túto informáciu získať z meta údajov rozhodnutia.

8.3 Automatická anonymizácia

Po úspešnom rozpoznaní entít, je možné tieto entity anonymizovať. Anonymizácia prebieha v dvoch krokoch a vykonáva sa na základe natrénovaného štatistického modelu. Pri nižšom počte trénovacích dokumentov je vhodné vytvoriť pravidlá, ktoré pokrývajú základné chyby anotácie pomocou štatistického modelu.

V prvom kroku sa pomocou štatistického modelu a všetkých dostupných vlastností textu extrahovaných v predchádzajúcich krokoch anotujú všetky časti textu, ktoré je potrebné anonymizovať. Následne prebieha korekcia prípadných chýb pomocou pravidiel.

V druhom kroku sa na základe definovaných pravidiel, vykoná anonymizácia dokumentu. V našom prípade ide o nahradenie slov veľkým písmenom abecedy nasledované bodkou a čísla sú nahradzанé za písmeno X.



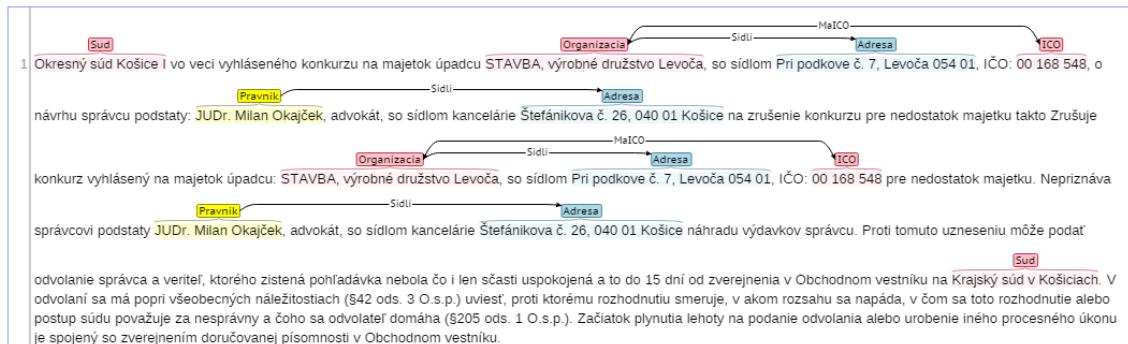
Obrázok 12. Diagram časti anonymizácie

8.4 Charakteristika dokumentov a doménové obmedzenia

Pri overovaní metódy budeme spracovávať súdne rozhodnutia. Keďže proces extrakcie informácií a rozpoznávania pomenovaných entít, ako aj samotná anonymizácia, je späť s doménou aplikácie, je potrebné sa podrobnejšie pozrieť na dokumenty, ktoré budeme spracovávať. Súdne rozhodnutia sú špecifické svojou štruktúrou, ale aj použitým jazykom a vyjadrovaním.

Pri súdnych rozhodnutiach sa využíva ustálená slovná zásoba, ktorá obsahuje spisovnú slovenčinu a určité odborné výrazy. Pri spracovaní dokumentov sme identifikovali špecifická súdnych rozhodnutí. Pri písaní súdnych rozhodnutí využívajú poverení zamestnanci aplikáciu Súdny Manažment, ktorá umožňuje písanie súdnych rozhodnutí len v textovej forme bez formátovania. Zvýraznenie dôležitých slov, teda vykonávajú oddelovaním písmen v slove medzerami. Keďže pri procese tokenizácie sa rozdeľujú slová podľa medzier a interpunkčných znamienok, takéto „zvýraznenie“ spôsobuje nekorektné rozdeľovanie slov na písmená. Je preto nutné vykonávať krok sanácie, v ktorom písmená spojíme do slova, a až potom text posunieme na tokenizáciu.

Štruktúra dokumentu je veľmi jednoduchá a obsahuje hlavičku dokumentu, ktorá obsahuje štruktúrované údaje, ktoré môžeme ďalej využiť pri extrakcii informácií a detekcii entít v texte. Ďalej nasleduje nadpis, ktorý hovorí o type rozhodnutia (rozsudok, platobný príkaz, uznesenie, rozhodnutie ...) . Informáciu o type rozhodnutia, oblasť právnej úpravy a povahu rozhodnutia máme dostupnú v rámci meta informácií, ktoré aplikácia spracuje spolu s dokumentom. Po nadpise nasleduje úvodný text, ktorý hovorí o tom, kto, kde, kedy vykonal rozhodnutie a spomína aj odporcu a pojednávanú vec. Nasleduje kľúčové slovné spojenie špecifické pre daný typ rozhodnutia a samotný výsledok rozhodnutia. Po výsledku je odôvodnenie, kde sa zvyčajne vymenúvajú zákony, paragrafy a vysvetľuje proces pri rozhodovaní a na konci je poučenie.



Obrázok 13. Príklad anonymizovaného súdneho rozhodnutia, v ktorom sme ručne farebne anotovali entity a vzťahy medzi nimi

Pred tým než je možné anonymizačný modul používať na anonymizáciu, je potrebné najprv natrénovala štatistický model. V našom prípade trénujeme model pomocou existujúcich ručne anotovaných súdnych rozhodnutí. Momentálne je dostupných 1 687 525 súdnych rozhodnutí a každý mesiac približne 40 000 súdnych rozhodnutí. Tieto rozhodnutia existujú v dvoch formách, v anonymizovanej a neanonymizovanej forme. Na neanonymizovanej verzii vykonávame rozpoznávanie a extrakciu entít a následne porovnávame rozdiely v oboch dokumentoch, aby sme identifikovali miesta, ktoré boli anonymizované. Potom anotujeme jednotlivé entity, podľa toho či boli alebo neboli anonymizované. Na základe tejto informácie trénujeme štatistický model, ktorý neskôr využívame pri rozhodovaní o anonymizácii entít.

Je nutné poznamenať, že aj ľudskí anotátori robia chyby. Pri trénovaní štatistického modelu, produkuje aplikácia záznamy o svojej činnosti a používateľ má potom možnosť skontrolovať, potencionálne chybné ľudské anotácie.

Po natrénovali štatistického modelu, je možné vykonávať automatizovanú anonymizáciu súdnych rozhodnutí. Pri takejto anonymizácii sa samozrejme stále prihliada na fixné pravidlá určené inštrukciou MSSR č. 24/11.

9 Implementácia a technológie

Pri implementácii sme kládli dôraz na dosiahnutie stanovených cieľov. Na základe analýzy existujúcich riešení a nástrojov, sme vybrali programovací jazyk, ktorý je multiplatformový, podporuje jednoduché rozširovanie knižnicami, nie je potrebné obaľovať existujúce knižnice nástrojov pre prácu s prirodzeným jazykom a zároveň sme kládli dôraz aj na výkon a rýchlosť. Tiež sme brali do úvahy požiadavku na webovú prezentáciu aplikácie ako softvérový prototyp. Ako najvhodnejší kandidáti pre programovací jazyk boli teda jazyk Java a Python a knižnice CoreNLP, OpenNLP a NLTK. Vzhľadom na povahu skriptovacieho jazyka Python, dosahuje Java lepšie výkonné výsledky pri práci so štatistickými modelmi a komplikovanými výpočtoch³⁸. Zvolili sme preto programovací jazyk Java. Správa závislostí aplikácie je založená na technológií Maven, ktorá poskytuje efektívny spôsob zdieľania knižníc a závislostí aplikácie.

Aplikáciu sme rozdelili na niekoľko komponentov a každý má samostatný Maven projekt. Aplikáciu sme pre zjednodušenie zapisovania názvu nazvali ASuR (Anonymizácia Súdnych Rozhodnutí). Aplikácia obsahuje komponenty:

- Decrete crawler – Nástroj pre získavanie anonymizovaných, neanonymizovaných súdnych rozhodnutí z webových zdrojov, ich porovnávanie a ukladanie do súborov.
- Brat – Nástroj pre ručnú anotáciu dokumentov, ktorý podporuje vizualizáciu anotácií entít a vzťahov.
- ASuR tools – Nástroje pre prácu pri trénovaní štatistických modelov, prácu so slovníkmi, vyhodnotenie úspešnosti metód a testovanie aplikácie. Obsahuje množstvo spustiteľných tried.
- NLP tools – Nástroje pre prácu s prirodzeným jazykom poskytované ako knižnica. Obsahuje modely, slovníky, pravidlá, ale aj automatizované testy pre predspracovanie.
- ASuR web – Webová aplikácia umožňujúca anonymizáciu súdnych rozhodnutí a prácu s prirodzeným jazykom. Aplikácia využíva knižnicu NLP tools.

9.1 Spoločná architektúra pri práci s prirodzeným jazykom

Pri práci s textom používame tzv. anotácie, ktoré obsahujú doplňujúce informácie o teste, ako aj extrahované informácie. Anotácie sú uložené v mape anotácií. Jednotlivé

³⁸ <https://benchmarksgame.alioth.debian.org/u64q/python.html>

anotácie sú vkladané do tejto mapy na základe interných a externých informácií získaných z textu a produkuje ich anotátor (angl. annotator). Tieto anotácie môžu obsahovať texty, mapy, zoznamy a predstavujú základnú štruktúru modelu pri extrakcii informácií z textu. Anotátor dopĺňa na základe štatistického modelu, pravidiel, slovníkov alebo vlastností textu anotácie do tohto modelu. Anotátory sú vykonávané za sebou a niektoré sú na sebe aj závislé, keďže využívajú informácie získané v predchádzajúcim kroku. Jednotlivé anotátory vkladáme do systému zrečazeného spracovania (pipeline), ktorý zabezpečuje vykonávanie anotácie v definovanom poradí.

9.2 Decrete crawler

Nástroj získava neanonimizované dokumenty z verejného zdroja³⁹ a vyhľadáva čísla spisov a dátum rozhodnutia v indexe anonymizovaných súdnych rozhodnutí, ktoré sú dostupné aj online⁴⁰ alebo vo formáte JSON⁴¹. Pre vytvorenie takéhoto indexu, je potrebné nainštalovať aplikáciu Solr a importovať JSON.

V prípade, že nástroj nájde rovnaký dokument v oboch systémoch, porovnáva tieto dokumenty a v prípade rozdielov ich ukladá na disk. Aplikácia obsahuje aj nástroj pre transformáciu nájdených rozdielov do formátu, ktorý je možný zobraziť v aplikácii Brat a tiež do trénovateľného formátu.

Aplikácia je implementovaná v jazyku Java 8 a využíva správu závislostí Maven. Pre získanie neanonimizovaných dokumentov využíva integráciu na Register úpadcov cez webové služby SOAP 1.1 a cez webové rozhranie. Pre získanie anonymizovaných dokumentov využíva integráciu cez REST rozhranie poskytované aplikáciou Solr.

Aplikácia využíva knižnice Selenium, JSoup, UniREST, pre získavanie údajov z webových zdrojov a NLP Tools pri transformáciách.

9.3 Brat

Nástroj Brat je existujúce riešenie, ktoré sme integrovali do našej aplikácie. Brat je nástroj pre rýchlu anotáciu, poskytuje intuitívne webové rozhranie pre anotáciu textu podporovanú spracovaním prirodzeného jazyka^[25]. Umožňuje využívanie NLP nástrojov, klávesových skratiek a jednoduchého používateľského rozhrania pre urýchlenie anotácie.

³⁹ <http://ru.justice.sk/>

⁴⁰ <https://obcan.justice.sk/infosud/-/infosud/zoznam/rozhodnutie>

⁴¹ <https://obcan.justice.sk/opendata>

Aplikáciu bolo nutné nainštalovať a nakonfigurovať pre naše potreby anotácie. Inštalácia je jednoduchá a vyžaduje len nakopírovanie súborov na server, nainštalovanie Python a spustenie skriptu standalone.py. Konfigurácia pozostáva z viacerých krovok.

9.3.1 Konfigurácia nástrojov

Brat podporuje využívanie nástrojov NLP pre urýchlenie anotácie. Tieto nástroje je potrebné konfigurovať. Pre naše potreby sme nakonfigurovali nástroje v súbore *tools.conf*:

```
Tokens tokenizer:ptblike
Sentences splitter:newline
Validation validate:none
Annotation-log logfile:<NONE>
```

Kedže v NLP tools využívame PTBTokenizer od CoreNLP, zvolili sme veľmi podobný tokenizér aj pri anotácii. Pri rozdeľovaní viet sme použili nový riadok a vyplňme validáciu anotácií. Anotačný log nie je v našom prípade potrebný.

9.3.2 Konfigurácia anotácie

V konfigurácii anotácií je možné nastaviť typy entít a ich hierarchiu ako aj udalosti, vzťahy a vlastnosti entít. Nastavuje sa v súbore *annotation.conf*.

```
[entities]
!Osoba
    !FyzickaOsoba
        Nepodnikatel
        Podnikatel
    Organizacia
    Pravnik
    PravnickaOrganizacia
Sud
Adresa
ICO
Suma

[relations]
<Osoba>=Podnikatel|Nepodnikatel|Organizacia|Pravnik|PravnickaOrganizacia
<FyzickaOsoba>=Podnikatel|Nepodnikatel
<Organizacie>=Organizacia|Podnikatel|PravnickaOrganizacia

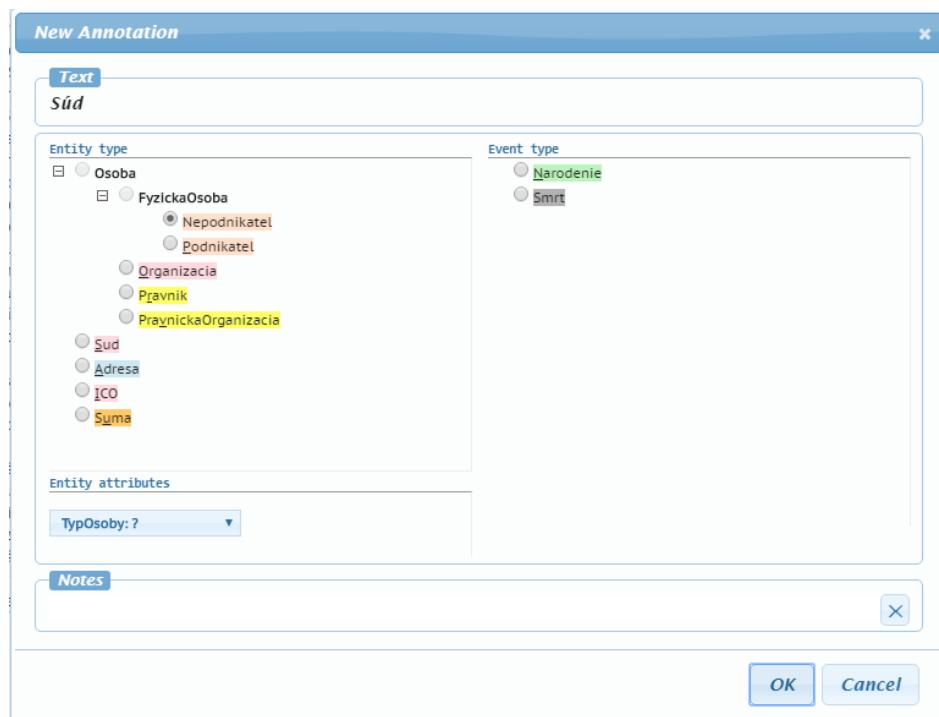
Zamestatny Ar1:<FyzickaOsoba>, Arg2:Organizacia
Sidli Ar1:<Osoba>, Arg2:Adresa
MaICO Ar1:ICO, Ar2:<Organizacie>, <REL-TYPE>:symmetric-transitive

[events]
<Osoba>=Podnikatel|Nepodnikatel|Organizacia
<FyzickaOsoba>=Podnikatel|Nepodnikatel|Pravnik

Narodenie Narodil:<FyzickaOsoba>
Smrt Zomrel:<FyzickaOsoba>
```

Konfigurácia obsahuje všetky relevantné entity pri anonymizácii, vzťahy, ktoré sú podstatné pri rozhodovaní o anonymizácii a udalosti, ktoré je vhodné anonymizovať.

Aplikácia po konfigurácii následne umožňuje anotáciu pomocou vysvietenia textu pre anotáciu a výberu typu entity alebo udalosti.



Obrázok 14. Obrazovka pridania novej anotácie v aplikácii Brat

9.3.3 Konfigurácia vizuálnych prvkov

Vizualizácia je podstatná súčasť anotačného nástroja, keď umožňuje v reálnom čase sledovať priebeh anotácie a jednoducho odhaliť chyby pri anotácii. Je preto vhodné zvoliť farebné odlišenie jednotlivých tried a udalostí pri anotácii. Takáto konfigurácia sa nastavuje v súbore *visual.conf*.

```
[labels]
[drawing]
SPAN_DEFAULT fgColor:black, bgColor:lightgreen, borderColor:darken
ARC_DEFAULT color:black, dashArray:-, arrowHead:triangle-5
FyzickaOsoba bgColor:#ffccaa
Podnikateľ bgColor:#ffccaa
Nepodnikateľ bgColor:#ffccaa
Pravnik bgColor:yellow
PravnickaOrganizacia bgColor:yellow
Organizacia bgColor:pink
Sud bgColor:pink
ICO bgColor:pink
Adresa bgColor:lightblue
Narodenie bgColor:lightgreen
Smrť bgColor:gray
Suma bgColor:orange
```

Formát definície farieb je rovnaký ako v css štýloch.

9.3.4 Konfigurácia klávesových skratiek

Klávesové skratky sú dôležitý prvok pri anotácii, keďže urýchľujú anotáciu jednotlivých dokumentov. V našom prípade obsahoval jeden dokument zvyčajne 12 anotácií, 4-5 vzťahov a jedna anotácia trvala priemerne približne 1 minútu.

Konfigurácia sa vykonáva v súbore *kb_shortcuts.conf*:

F	Nepodnikateľ
P	Podnikateľ
R	Pravnik
V	PravnickaOrganizacia
O	Organizacia
A	Adresa
I	ICO
N	Narodenie
S	Sud
D	Smrt
U	Suma

Naša konfigurácia dbá viac na sémantiku ako na rozloženie kláves.

9.4 ASuR tools

ASuR tools je projekt, ktorý obsahuje spustiteľné a konfigurovateľné Java triedy pre prácu so štatistickými modelmi, slovníkmi a anotovanými dokumentami. Odporúčame triedy spúšťať v integrovanom vývojárskom rozhraní IDE.

Obsahuje teda aj konfiguráciu čít štatistických modelov a využíva knižnicu NLP tools pre prácu s prirodzeným jazykom.

Po preskúmaní možností existujúcich knižníc pre prácu s prirodzeným jazykom, sme zvolili ako základ CoreNLP pre svoju komplexnosť a možnosť jednoduchého rozšírenia, podporu modelu anotácií a anotátorov, ako aj množstvu existujúcich implementácií, splňajúcich naše požiadavky na efektívnejšie dosiahnutie cieľov.

Ako rozšírenie využívame knižnicu Apache OpenNLP, ktorá nám poskytuje najmä prácu s modelmi maximálnej entropie. Pre efektívnejšiu prácu s dátovými štruktúrami a na sprehľadnenie zdrojového kódu, využívame knižnice Google Guava⁴² a Apache commons.

9.4.1 Nástroje pre prácu so slovníkmi a anotovanými dokumentmi

Pri práci so slovníkmi a veľkým množstvom dát, niekedy nepostačuje textový editor. V takomto prípade sme vytvorili niekoľko nástrojov, ktoré zjednodušujú túto prácu:

⁴² <https://github.com/google/guava>

- *AnnotationFilesUtils* – Obsahuje metódy pre ukladanie výsledkov anotácie do vizualizovačného, trénovateľného a vyhodnotiteľného formátu.
- *CopyOnlyNewFiles* – Nástroj vyberie súbory, ktoré ešte nie sú ručne anotované
- *DuplicateLinesRemover* – Nástroj pre odstránenie duplicitných riadkov v slovníkoch
- *FrequencyConsolider* – Nástroj normalizuje slová v slovníku, a spočítava ich frekvencie
- *GeoEntityUtils* – Nástroj pre spočítanie, koľko lokalít obsahuje v názve meno alebo priezvisko
- *LemmaStemGazetteCreator* – Nástroj lematizuje a stemuje slovníky a vytvára nové slovníky
- *PossibleAbbreviationExtractor* – Nástroj dokáže extrahovať možné skratky z textu na základe heuristiky, implementovanej pomocou regulárneho výrazu
- *POSTagDictionaryConverter* – Nástroj, ktorý konvertuje JÚLŠ formát morfologického slovníka do xml formátu použiteľného pri trénovaní OpenNLP maxent štatistického modelu.
- *PrepareExperiments* – Nástroj pre výber množín na testovanie a trénovanie pomocou metódy Monte Carlo
- *SuffixExtractor* – Nástroj pre extrakciu najpoužívanejších suffíxov v morfologickom slovníku

9.4.2 Nástroje pre trénovanie štatistických modelov

Pre trénovanie štatistických modelov sú pripravené triedy *TrainNER*, *TrainAnonymizer*, *POSTrainer* a *SSplitTrainer*. Upozornujeme, že je potrebné mať neoficiálnu upravenú verziu CoreNLP⁴³, ktorá podporuje použitie stackedNer pri trénovaní.

TrainNER

Nástroj má za úlohu nájsť všetky anotované súbory v zadanom priečinku, urobiť predspracovanie, lingvistickej NER a vyprodukovať trénovateľné súbory. Na základe trénovateľných súborov vo formáte .tsv následne natrénovať CRF štatistický model s takýmito parametrami:

⁴³ <https://github.com/drndos/CoreNLP>

- "map":"word=0,stackedNer=1,tag=2,lemma=3,answer=4" – Mapovanie črt z trénovateľného súboru tsv.
- "maxLeft":1 – Maximálny počet predchádzajúcich tried použitých ako črta na aktuálny token.
- "useClassFeature": true – Pri klasifikácii zohľadňuje početnosť danej triedy v trénovacej množine.
- "useWord":true - Základná črta je slovo, táto črta má význam najmä pri zisťovaní kontextu, vtedy je vhodné sledovať aj predchádzajúci a nasledujúci token.
- "useNGrams":true – Použitie NGramov.
- "maxNGramLeng":6 – Maximálna dĺžka NGramov, minimálne je predvolene 2.
- "usePrev":true – Využíva vlastnosti predchádzajúceho tokenu ako črtu.
- "useNext":true - Využíva vlastnosti nasledujúceho tokenu ako črtu.
- "useDisjunctive":true - Črta zabezpečuje, že jednotlivé slová nemusia byť v konkrétnom poradí a zároveň zachováva smer. Je možné nastaviť do akej vzdialenosť má disjunkcia fungovať.
- "useSequences":true - Zohľadňuje sekvencie tried ako črtu pri trénovaní a klasifikácii.
- "usePrevSequences":true – Zohľadňuje aj predchádzajúce sekvencie.
- "useTypeSeqs":true, "useTypeSeqs2":true, "useTypeySequences":true – Zohľadňuje sekvencie typu slova (či začína na veľké, malé písmeno, celé kapitálkami a podobne).
- "useOccurrencePatterns":true - Táto črta zohľadňuje či sa okolité slová začínajú na veľké písmeno a zlepšuje najmä anotáciu sekvencií, ktoré v strede obsahujú slovo s malým začiatočným písmenom.
- "useBeginSent", true - Črta využíva segmentáciu viet a začiatky viet.
- "wordShape":"chris2useLC" – Algoritmus určujúci tvar slova.
- "useLemmas":true – Použitie lém ako črty.
- "useTags":true – Použitie POS značiek ako črty.
- "usePosition":true – Použitie pozície vo vete ako črtu.

Výstup trénovania modelu je binárny súbor s NER modelom, ktorý sa následne využíva v komponente NER tools.

TrainAnonymizer

Nástroj vyhľadá všetky anotované dokumenty v zadanom priečinku. Urobí predspracovanie a hybridné NER a následne transformuje všetky tieto údaje spolu s informáciami z anotovaných dokumentov do .tsv formátu. Pri trénovaní využíva štatistický model s parametrami (parametre sú vysvetlené vyššie v sekcii TrainNER):

- "map":"word=0,stackedNer=1,tag=2,lemma=3,answer=4"
- "maxLeft":1
- "useClassFeature":true
- "useWord":true
- "useNGrams":true
- "maxNGramLeng":6
- "usePrev":true
- "useNext":true
- "useDisjunctive":true
- "useSequences":true
- "usePrevSequences":true
- "useTypeSeqs":true
- "useTypeSeqs2":true
- "useTypeySequences":true
- "useOccurrencePatterns":true
- "useBeginSent":true
- "wordShape":"chris2useLC"
- "useLemmas":true
- "usePosition":true
- "useTags":true

Výstup trénovania anonymizéra je štatistický model v binárnom formáte. Štatistický model sa využíva v komponente NLP tools.

POSTrainer

Nástroj využíva toky dát (stream) pre transformáciu kvázi XML formátu korpusu prim.6.1-all od JÚLŠ, do trénovateľného formátu OpenNLP a trénuje štatistický model maximálnej entropie. Využíva pritom parametre:

- **ALGORITHM_PARAM:"MAXENT"** - Štatistický model MEMM
- **ITERATIONS_PARAM:100** – Maximálny počet iterácií pri optimalizácii

- CUTOFF_PARAM:5 – Minimálny počet výskytov črty na to, aby sa dostala do modelu

Popri tom využíva *POSTokenFactory*, ktorá číta z xml súboru morfologického slovníka a obmedzuje natrénovanie nemožných značiek pre známe slová.

Pre účely trénovania určovania slovných druhov, bolo nutné zrekonštruovať pôvodný text, použitím prvého slova z každého riadku a doplnením znaku „_“ a informácií o slovnom druhu, páde a vzore. Po každom dokumente sme nechali prázdný riadok.

Príklad:

```
Spolu_Dx s_Eu7 manželom_SSms7 sa_R krátky_AAis4x čas_SSis4 venovala_VLjscf+  
aj_T nakrúcaniu_SSns3 filmov_SSip2 . .
```

Výsledkom je binárny súbor štatistického modelu, ktorý sa využíva v komponente NLP tools.

SSplitTrainer

Nástroj využíva toky dát (stream) pre transformáciu kvázi XML formátu korpusu prim.6.1-all z JÚLŠ, do toku tokenov, ktoré následne detokenizuje po vetách na základe jednoduchých pravidiel a posúva do trénovania štatistického modelu maximálnej entropie. Využíva pri tom rovnaké parametre ako POSTrainer.

Pre účely trénovania štatistického modelu segmentácie, bolo nutné zrekonštruovať pôvodný text použitím prvého slova v každom riadku a oddelením jednotlivých viet pomocou nového riadku. Po každej desiatej vete alebo novom dokumente, sme podľa odporúčaní pri trénovaní modelu nechali jeden prázdný riadok. Príklad:

```
Spolu s manželom sa krátky čas venovala aj nakrúcaniu filmov.\n  
Doľava k mestskej veži vedie zo Štúrovho námestia úzka a rušná časť pešej  
zóny - Sládkovičova ulica. \n
```

Výsledkom je binárny súbor štatistického modelu, ktorý sa využíva v komponente NLP tools.

9.4.3 Nástroje pre vyhodnocovanie úspešnosti riešení

Pri vyhodnocovaní a porovnávaní riešení a metód sme implementovali nástroje pre vyhodnocovanie. Medzi tieto nástroje patria:

- *StemmerRanker* – Nástroj, ktorý vyhodnocuje úspešnosť a rýchlosť stemerov na testovacích dátach morfologického slovníku.
- *POSTaggerRanker* – Nástroj pre vyhodnocovanie POSTagger riešení. Vyhodnocuje úspešnosť na menšej testovacej vzorke automatizované anotovaných tokenoch európskych dekrétov, ktoré boli ručne skontrolované a opravené.

- *EvaluateNER* – Nástroj vyhodnocuje úspešnosť NER metódy na základe porovnávania rozpoznaných a zlatých tried z tsv súborov, vytvorených pri spúšťaní nástroja *AsurNER*, ktorý vytvára súbory anotácie a tsv súbory pre vyhodnotenie.
- *EvaluateAnonymizer* – Nástroj, ktorý vyhodnocuje úspešnosť anotácie pri anonymizácii na základe porovnávania rozpoznaných a zlatých tried z tsv súborov, vytvorených pri spúšťaní nástrojov *StatAnonymizer* a *ManualAnonymizer*, ktoré vytvárajú súbory anotácie a tsv súbory pre vyhodnotenie.

9.5 NLP tools

NLP tools je sada nástrojov pre prácu s prirodzeným jazykom, využiteľná ako Java knižnica v iných aplikáciach. Aplikácia je závislá najmä na knižniciach CoreNLP, OpenNLP, Guava a Lucene. Nástroje využívajú model anotácií a anotátorov. Každý nástroj predstavuje jeden anotátor, ktorý je možné použiť v zrečazenom spracovaní.

9.5.1 SpaceHighlightingSanitizer

Takáto heuristika je implementovaná vo forme anotátora a je využitá v systéme zrečazeného spracovania ako prvý anotátor. Využíva sa pri kroku sanácia. Popis fungovania je napísaný v sekcií 8.1.1.

9.5.2 PTBTokenizerUtils

V práci sme zvolili lingvistickú tokenizáciu, pre dobré dosahované výsledky nástroja PTBTokenizer z knižnice CoreNLP v anglickom jazyku, ale tiež jednoduchosť úpravy pravidiel tejto tokenizácie a robustnosť riešenia. Pravidlá sú definované vo formáte jflex. Tieto pravidlá sú prekladané do jazyka Java. Pri preklade zároveň dochádza ku odvodzovaniu nových pravidiel a tvorbe uzáverov. Pravidlá pre anglický jazyk sme prispôsobili slovenskému jazyku.

Nástroj umožňuje jednoduchšiu prácu s tokenizérom.

9.5.3 SSplitUtils

Knižnica CoreNLP poskytuje lingvistickú segmentáciu viet na základe informácií, získaných z kroku tokenizácie, kedy sa vyhodnocujú hranice vety.

Pre možnosť zmeny tejto implementácie sme natrénovali štatistický model maximálnej entropie, ktorý poskytuje knižnica OpenNLP na korpuse v slovenskom jazyku s viac než 300 000 000 tokenov. Implementovali sme tiež anotátor, ktorý je možné využiť v zrečazenom spracovaní.

9.5.4 LemmatizerUtils

Nástroj poskytuje možnosť využitia slovníkového lematizéra a pravidlového stemera.

9.5.5 SlovakStemmer

Slovenský stemer, ktorý dokáže získať koreň slova na základe pravidiel. Jednotlivé pravidlá sme vytvorili na základe pravidiel slovenského pravopisu⁴⁴. Tieto pravidlá sme rozšírili vďaka extrakcii najčastejšie používaných prípon z morfologického slovníka. Proces optimalizácie stemera je možné vidieť na grafe v prílohe C.

9.5.6 TaggerUtils

Nástroj využíva implementovaný anotátor, ktorý obaľuje OpenNLP nástroj pre značkovanie. Pri svojej práci využíva natrénovaný štatistický model na 100 000 000 tokenoch zo SNK.

9.5.7 NERUtils

Nástroj poskytuje možnosť výberu implementácie rozpoznávania pomenovaných entít. Implementovaná je lingvistická metóda, ktorá využíva SUTime, NumberAnnotator, slovníky (ktoré obsahujú skoro 24 000 mien a 258 000 priezvisiek, všetky slovenské obce a ulice a tituly) a pravidlá nad sekvenciami tokenov. Štatistická metóda využíva CRF model natrénovaný na 760 ručne anotovaných dokumentoch. Hybridná metóda využíva kombináciu lingvistickej metódy, ako črtu pre štatistický model, natrénovaný na 760 ručne anotovaných dokumentoch, ktoré boli navyše automatizované lingvisticky anotované.

Priklad pravidla:

```
( [{ner:TITUL}] [{ner:MENO}] [{ner:PRIEZVISKO}|{word:/[\p{L}]+/}] ", "  
[ {ner:TITUL_ZA} ] ) Osoba
```

Medzi lingvistickým a štatistickým NER je potrebný ešte jeden anotátor, ktorý konvertuje *NamedEntityTagAnnotation* na *StackedNamedEntityTagAnnotation*, túto anotáciu je možné potom použiť v štatistickom modeli ako črtu.

9.5.8 AnonymizerUtils

Nástroj využíva natrénovaný CRF štatistický model a súbor pravidiel, na základe ktorých sa anonymizátor rozhoduje, či danú entitu spracuje alebo nie. Následne využije *ScrubberAnnotator*, ktorý na základe jednoduchých pravidiel nahradí anonymizované slová a čísla za nič nehovoriace znaky. Príklad pravidla, ktorý opravuje chyby štatistického modelu:

⁴⁴ <http://www.juls.savba.sk/ediela/psp2000/psp.pdf>

([{ner:Osoba}] [{ner:Adresa}] [{ner:Osoba}]) Osoba

9.6 NLP web

NLP web je webová aplikácia, ktorá umožňuje využívanie nástrojov cez používateľské rozhranie. Aplikácia umožňuje využívanie integračných rozhraní pre integráciu externých systémov, ako aj používateľské rozhranie pre testovanie používateľom.

The screenshot shows the NLP Nástroje (Tools) application. On the left, a sidebar menu lists various tools: Úvod, Tokenizácia, Segmentácia viet, Lematizácia, Stemovanie, POS značkovanie, Rozpoznávanie pomenovaných entít (which is selected and highlighted in blue), and Anonymizácia. The main content area is titled 'NER' and contains a text input field with placeholder text: 'Prosím zadajte text, v ktorom chcete rozpoznať entity:'. Below the text input is a text area containing a legal document snippet from a court judgment. At the bottom of the text area, there are three radio buttons for entity recognition methods: 'Hybridná metóda' (selected), 'Štatistická metóda', and 'Lingvistická metóda'. A blue 'Odoslat' (Send) button is located below the method selection. The results section at the bottom displays the analyzed text with entities highlighted and labeled: 'Sud', 'Osoba', 'Narodenie', and 'Adresa'. The analyzed text is: 'Okresný súd Trnava v právnej veci navrhovateľa - dlužníka: Vendelin Kiss, nar. 10.09.1979, bytom Poľovnícka ulica 661/24, 930 28 Okoč, o vyhlásenie konkurzu na jeho majetok, taktô Súd vyhlasuje konkúr na majetok dlužníka: Vendelin Kiss, nar. 10.09.1979, bytom Poľovnícka ulica 661/24, 930 28 Okoč. Súd uznáva konkúr za malý. Súd ustanovuje do funkcie správca JUDr. Petra Sopka, so sídlom Kancelárie Paulínska 24, Trnava. Súd ukladá správcovi vypracovať a predložiť súdu v lehote 15 dní od ustanovenia do funkcie podrobného písomného správu o stave zisťovania a zabezpečovania majetku úpadcu a vykonaných úkonoch, najmä o stave súpisu a zoznamu pohľadávok, v lehote 35 dní od ustanovenia do funkcie druhého podrobného písomného správu, v lehote 50 dní od ustanovenia do funkcie tretího podrobného písomného správu o týchto skutočnostiach a v lehote najneskôr 5 dní pred konaním prvej schôdze veriteľov štvrtú podrobnú písomnú správu o týchto skutočnostiach. Súd vyzýva veriteľov, aby si prihlásili svoje pohľadávky v lehote 45 dní od vyhlásenia konkúru. Súd ukladá'. Below the results, there is a note in Slovak: '1. Okresný súd Trnava v právnej veci navrhovateľa - dlužníka: Vendelin Kiss, nar. 10.09.1979, bytom Poľovnícka ulica 661/24, 930 28 Okoč. Súd uznáva konkúr za malý. Súd ustanovuje do funkcie správca JUDr. ustanovenia do funkcie podrobného písomného správu o stave zisťovania a zabezpečovania majetku úpadcu a vykonaných úkonoch, n lehote 50 dní od ustanovenia do funkcie tretího podrobného písomného správu o týchto skutočnostiach a v lehote najneskôr 5 dní pred k svoje pohľadávky v lehote 45 dní od vyhlásenia konkúru. Súd ukladá správcovi povinnosť bezodkladne informovať o konkúrnom členskych štátov Európskej únie, o spôsobe prihlásovania pohľadávok v súlade s článkom 40 Nariadenia Rady (ES) č. 1346/2000 v konkúre uplatňuje prihláškou (čl. 28 ods. 1 ZKR). 2. Veriteľ prihlásuje svoju pohľadávku vyplnenou prihláškou spolu s prílohami. I ustanovenia zákona č. 7/2005 Z. z. o konkúre a vyrovnaní a o zmene a doplnení niektorých zákonov. 3. Prihláška sa podáva v jednom konkúre; v jednom rovnopise veriteľ doručí prihlášku aj na súd. (čl. 28 ods. 2 ZKR). 4. Ak veriteľ doručí správcovi prihlášku neskôr, pohľadávku. Právo na pomerne uspokojenie veriteľa tým nie je dotknuté; môže byť však uspokojený len z výtažku zaradeného do správcovi. Zapisanie takejto pohľadávky do zoznamu pohľadávok správca zverejni v Obchodnom vestníku s uvedením veriteľa a p včas uplatniť aj zabezpečovacie právo, a to v základnej prihlásovacej lehote 45 dní od vyhlásenia konkúru, inak zanikne. (čl. 28 ods. 2 ZKR).

Copyright © Filip Bednárik 2016 & Spracovanie textu na FIIT 2015 – 2016 & essential data, s.r.o. Verzia: 0.0.1

Obrázok 15. Používateľské rozhranie NLP web so zobrazením vizualizácie výsledkov NER.

Aplikácia sa skladá z úvodnej stránky a z nástrojov dostupných v menu.

9.6.1 Tokenizácia

Stránka pozostáva z formulára pre zadanie textu pre tokenizáciu. Výstupom tokenizácie sú tokeny oddelené medzerou.

9.6.2 Segmentácia viet

Stránka pozostáva z formulára pre zadanie textu pre segmentáciu a konfiguračné nastavenie metódy segmentácie. Výstupom je zoznam viet oddelených novým riadkom.

9.6.3 Lematizácia

Na stránke je formulár pre zadanie textu pre lematizáciu. Výstupom sú lémy.

9.6.4 Stemovanie

Rovnako ako pri lematizácii je možné zadať text a výstupom je sekvencia stemov.

9.6.5 POS značkovanie

Pri značkovanií je možné zadať vstupný text a výstupom je formát OpenNLP, kde sú tokeny oddelené medzerami a jednotlivé značky sú oddelené znakom „_“ od tokenu.

Príklad:

```
Prosím_T zadajte_VMdpb+ text_SSis4 ,_Z ktorý_PAis4 chcete_VKepb+
označovať_VId+ ._Z
```

9.6.6 Rozpoznávanie pomenovaných entít

Pri rozpoznávaní pomenovaných entít je možné zvoliť metódu rozpoznávania. Dostupné sú tri implementácie: lingvistická, štatistická a hybridná. Po zadaní textu, výbere metódy a potvrdení formuláru sa zobrazí vizualizácia anotovaného textu, ako je možné vidieť na obrázku: Obrázok 15.

9.6.7 Anonymizácia

Anonymizácia využíva predspracovanie, hybridné NER a štatistický model pre rozpoznanie entít pre anonymizáciu a následné nahradenie znakov v zadanom texte. Výstupom je anonymizovaný text.

III. Vyhodnotenie

10 Experiments

Počas analýzy sme identifikovali niekoľko problémov existujúcich riešení a stanovili ciele, ktoré sa snažíme touto prácou dosiahnuť. V tejto časti práce opisujeme experimenty, ktorými overujeme jednotlivé časti riešenia.

10.1 Experiment č.1 – Predspracovanie textu v slovenskom jazyku

Prvou časťou metódy je predspracovanie. Od úspešnosti predspracovania závisí aj úspešnosť metódy pre extrakciu pomenovaných entít, ako aj rozhodovanie o anonymizácii entity. V tomto experimente sme porovnávali úspešnosť stemera na vzorke 3 294 081 slov morfologického slovníka. Príklad záznamu v morfologickom slovníku:

Abesínčan	Abesínčania	SSmp1
-----------	-------------	-------

Pri tomto experimente vyhodnocujeme miery presnosti, pokrytie a F-skóre.

Presnosť: 0,9051

Pokrytie: 0,7608

F-skóre: 0,8267

Aplikácia je v porovnaní s ostatnými riešeniami aj najrýchlejšia, keďže dokázala celý slovník spracovať za ~3,61 sekundy.

Porovnanie riešení v tabuľkovom formáte:

Tabuľka 5. Vyhodnotenie experimentu č.1

Riešenie	Presnosť	Pokrytie	F-skóre	Čas (s)
Bednárik	0,9051	0,7608	0,8267	3,61
Pifková	0,6091	0,9351	0,7377	4,54
Kosorin	0,7534	0,8541	0,8006	8,64
Horváth	0,7437	0,9629	0,8392	N/A

Riešenie „Horváth“ sme testovali online, znížili sme počet požiadaviek za sekundu, aby sme príliš nezaťazili server a test trval niekoľko hodín. Riešenie je samozrejme oveľa rýchlejšie.

Ako vidíme na výsledkoch naše riešenie zatiaľ nedosahuje najvyššie F-skóre avšak vyhodnotenie pomocou tejto metriky nie je úplne objektívne ako opisuje aj článok [23].

10.2 Experiment č.2 – Určovanie slovných druhov (POS Tag)

Súčasťou predspracovania je aj značkovanie. V slovenčine existuje niekoľko dostupných riešení. Porovnali sme násť natrénovaný štatistický model s ďalšími

dostupnými nástrojmi. Porovnávanie prebiehalo na automatizovanom označkovanej textovej súťaži, ktorá obsahuje 371 tokenov. Tento korpus sme manuálne skontrolovali a opravili v ňom chybne označené pády. Úspešnosť sme počítali na základe toho, či sa zhodovala predpovedaná značka so značkou uvedenou v dokumente.

Tabuľka 6. Vyhodnotenie experimentu č. 2

Riešenie	počet tokenov	úspešnosť
Dagger	333	0,9550
Bednárik	371	0,9650
Naivný pravidlový korpusový pos tagger	326	0,6042

V priebehu vypracovania práce vyšlo ďalšie riešenie, ktoré dokáže značkovať text v slovenčine a je založené na CRF štatistickom modeli⁴⁵.

Na základe výsledkov vidíme, že štatistický prístup ku značkovaniu je účinnejší, ako lingvistický. Rozdiel medzi riešením Dagger a naším riešením je minimálny, čo vyplýva z toho, že obe riešenia boli trénované na rovnakom trénovacom korpuse s použitím podobného štatistického modelu.

10.3 Experiment č.3 – NER súdnych rozhodnutí

V tomto experimente ide o vyhodnotenie úspešnosti jednotlivých prístupov ku rozpoznávaniu pomenovaných entít v dokumentoch. Testovacia vzorka je ručne anotovaná pomocou nástroja Brat, vzorka obsahuje 760 dokumentov. Pri vyhodnotení budeme sledovať základné entity: Súd, IČO, Osoba, Narodenie, Adresa a Organizácia.

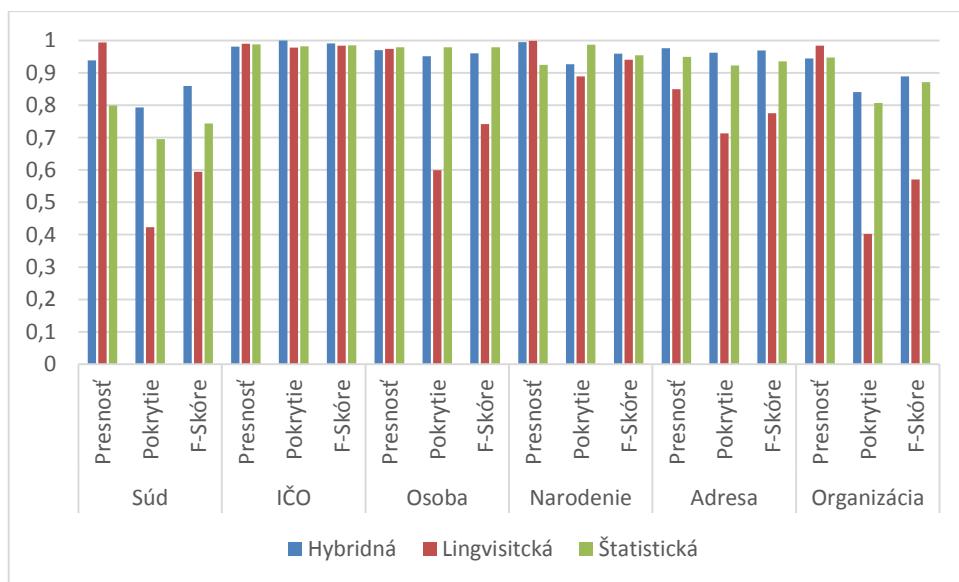
Vyhodnotenie hybridnej metódy sme vykonávali podľa metódy opakovanej overovania vzoriek, inak nazývanej aj krížová validácia Monte Carlo. Vyhodnotenie prebieha v iteráciách, v každej iterácii rozdeľujeme údaje na trénovacie a testovacie. Pre každú iteráciu natrénujeme štatistický model na trénovacej množine a vyhodnotíme ho na testovacej množine. V experimente sme opakovali rozdelenie množiny na 8/92 (testovacia/trénovacia) päť krát a vypočítali priemer.

Tabuľka 7. Vyhodnotenie experimentu č. 3

Entita	Metóda	Presnosť	Pokrytie	F-Skóre
Súd	Hybridná	0,9384	0,7930	0,8594
	Lingvistická	0,9940	0,4236	0,5940
	Štatistická	0,7989	0,6951	0,7434

⁴⁵ <https://github.com/Denrasill/SkCrfPosTagger>

<i>IČO</i>	Hybridná	0,9815	1	0,9906
	Lingvistická	0,9901	0,9784	0,9842
	Štatistická	0,9881	0,9823	0,9852
<i>Osoba</i>	Hybridná	0,9697	0,9518	0,9601
	Lingvistická	0,9743	0,5988	0,7417
	Štatistická	0,9787	0,9789	0,9788
<i>Narodenie</i>	Hybridná	0,9946	0,9271	0,9593
	Lingvistická	0,9988	0,8888	0,9406
	Štatistická	0,9244	0,9869	0,9546
<i>Adresa</i>	Hybridná	0,9764	0,9622	0,9692
	Lingvistická	0,8490	0,7133	0,7752
	Štatistická	0,9492	0,9229	0,9358
<i>Organizácia</i>	Hybridná	0,9448	0,8402	0,8892
	Lingvistická	0,9837	0,4020	0,5708
	Štatistická	0,9478	0,8068	0,8716



Obrázok 16. Grafové zobrazenie výsledkov experimentu.

Hybridná metóda dosahuje najlepšie výsledky vo všetkých prípadoch okrem osoby. V priemere dosahuje metóda v doméne justície f-skóre 0,9380, čo je pomaly porovnatelné s úspešnosťou človeka. Takúto vysokú úspešnosť si vysvetľujeme doménovými špecifikami a danou štruktúrou testovaných a trénovaných dokumentov.

10.4 Experiment č.4 Anonymizácia súdnych rozhodnutí

Tento experiment sme vykonávali na 1975 automaticky anotovaných dokumentoch, ktoré boli následne ručne skontrolované a opravené. Na základe rozdielov sme vypočítali úspešnosť ručnej anonymizácie. Experiment bol vykonávaný na trénovacej množine. Pri tomto experimente sledujeme aj F2-Skóre, ktoré dáva väčšiu váhu pokrytiu ako presnosti vzhladom na problematiku anonymizácie, kde sa snažíme maximalizovať najmä pokrytie.

Tabuľka 8. Vyhodnotenie experimentu č.4

Typ	TP	FP	FN	Presnosť	Pokrytie	F-Skóre	F2-Skóre
<i>Bednárik</i>	25707	8614	1817	0,7490	0,9340	0,8313	0,8900
<i>Ručná anonymizácia</i>	24387	158	3138	0,9936	0,8860	0,9367	0,9056

V slovenčine podľa našich zistení neexistuje verejne dostupný nástroj pre automatizovanú anonymizáciu textu. Pri porovnaní s ľudským anotátorom môžeme vidieť, že ľudský anotátor je menej pozorný pri označovaní entít, no keď už entitu označí, tak naozaj má byť anonymizovaná. V konečnom dôsledku naša metóda anonymizuje väčšie množstvo záznamov, ako by mala a preto dosahuje nižšiu presnosť.

V porovnaní s anglickými riešeniami môžeme vyhlásiť, že napriek tomu, že dosahujeme pomerne nízku presnosť, táto presnosť stále niekoľko násobne vyššia ako u nástrojov ako NLM-S alebo MITRE. Faktorom v tomto prípade, ale môže byť iná povaha testovacích údajov a špecifika slovenčiny a angličtiny.

11 Zhodnotenie

Cieľom našej práce bolo navrhnuť metódu, ktorá bude schopná identifikovať pomenované entity a rozhodnúť sa, či pomenovanú entitu anonymizovať, alebo nie na základe atribútov entít. V rámci analýzy sme porovnali existujúce prístupy ku riešeniu tohto problému a zároveň identifikovali problémy, ktoré existujú a sú spojené s prácou. Ukázali sme tiež súčasný stav práce s jazykom v slovenčine a porovnali existujúce riešenia. V práci navrhujeme hybridnú metódu extrakcie informácií, ktorá spĺňa predpoklady a rieši problémy identifikované počas analýzy problémovej oblasti. Navrhujeme metódu a nástroj, ktoré pokrývajú predspracovanie textu, rozpoznávanie pomenovaných entít a anonymizáciu osobných údajov na základe extrahovaných informácií.

V experimentoch predspracovania sa ukázalo, že naše metódy predspracovania v slovenčine sú úspešné a dosahujú porovnatelné výsledky ako konkurenčné riešenia. Tiež sme ukázali, že hybridná metóda dosahuje lepšie výsledky ako štatistická, či lingvistická metóda. V priemere o 17 % lepšie f-skóre ako lingvistická metóda a o 2,5 % vyššie f-skóre ako štatistická metóda. Tiež sme ukázali, že dosahujeme porovnatelné výsledky ako iné riešenia v inej doméne (porovnanie v rovnakej doméne nebolo vykonané kvôli nedostupnosti konkurenčných riešení v slovenskom jazyku v dobe vykonávania experimentov). Pri všetkých entitách v doméne dosahuje hybridná metóda vyššie f-skóre ako 85 % priemerne 93,8 %. Pre porovnanie človek bežne dosahuje úspešnosť 96 % a nástroj CoreNLP na doméne správ dosahuje f-skóre 86 %^[1].

Na základe experimentov môžeme tiež tvrdiť, že metóda dosahuje väčšie pokrytie ako ľudský anotátor, avšak nižšiu presnosť. Pri anonymizácii je dôležité pokryť čo najväčšie množstvo osobných údajov a preto sme pri vyhodnocovaní zvolili sledovanie F2-skóre, ktoré kladie väčší dôraz na pokrytie ako presnosť.

Riešenie má aj využitie v praxi, kde je potrebné anonymizovať 40 000 rozhodnutí mesačne, čo pri rýchlosti jedného rozhodnutia za 3 minúty predstavuje približne 83 dní čistého času anonymizácie mesačne.

Ďalším prínosom je opäťovná použiteľnosť nástrojov pre spracovanie prirodzeného jazyka v slovenčine. V súčasnosti existuje veľmi obmedzený počet voľne dostupných nástrojov pre prácu v slovenčine, čo spomaľuje vývoj softvéru schopného porozumieť textu a reči. Nástroje je možné využívať priamo na stránke <http://nlp.bednarik.top> a riešenie bude aj publikované na oficiálnej stránke fakulty, ktorá sa zaoberá spracovaním textu na adrese <http://text.fiit.stuba.sk/>.

V budúcej práci sa budeme venovať najmä zlepšovaniu úspešnosti anonymizéra, tvorbou nových pravidiel a hľadaním nových vhodných čítačiek pre štatistický model. Taktiež sa budeme snažiť zlepšiť proces extrakcie organizácií a osôb pomocou meta dát, ktoré sme v tejto fáze experimentovania, kvôli dôvernosti údajov nemali k dispozícii. Rozšírimo trénovaciu množinu na väčšie množstvo dokumentov, vzhľadom na skutočnosť, že sme mali dostupných iba 2000 dokumentov z celkového počtu 1 886 599. Priestor na rozšírenie je tiež pri využití závislostí entít ako črty štatistického modelu pri anonymizácii, metóda aj implementácia je na toto rozšírenie pripravená. V čase implementácie, ešte nebolo dostupné riešenie pre syntaktickú analýzu textu v slovenčine, ktorú pri tomto kroku potrebujeme. Ďalším námetom na zlepšenie je automatizácia procesu spúšťania trénovania štatistického modelu a automatická integrácia novo anonymizovaných dokumentov do trénovacej množiny.

12 Použité zdroje

- [1] AGARWAL, S. - SINGHAL, A. Autonomous Ontology Population from DBpedia based on Context Sensitive Entity Recognition. In *Fourth international joint conference on advances in engineering and technology, AET*. 2013. s. 580–589. .
- [2] BARCALA, F.M. et al. Tokenization and proper noun recognition for information retrieval. In *Proceedings - International Workshop on Database and Expert Systems Applications, DEXA*. 2002. Vol. 2002-Janua, s. 246–250. .
- [3] CUNNINGHAM, H. et al. Getting More Out of Biomedical Documents with GATE's Full Lifecycle Open Source Text Analytics. In *PLoS Computational Biology*. 2013. Vol. 9, no. 2. .
- [4] CUNNINGHAM, H. Information Extraction, Automatic. In *Encyclopedia of Language & Linguistics*. 2006. s. 665–677. ISBN 9780080448541.
- [5] FINKEL, J.R. et al. Incorporating non-local information into information extraction systems by gibbs sampling. In *in Acl [online]*. 2005. no. 1995, s. 363 – 370. Dostupné na internete:
[<http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.131.8904>](http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.131.8904).
- [6] GHAHRAMANI, Z. - JORDAN, M.I. Factorial hidden Markov models. In *Machine learning [online]*. 1997. Vol. 29, s. 245–273. Dostupné na internete:
[<http://link.springer.com/article/10.1023/A:1007425814087>](http://link.springer.com/article/10.1023/A:1007425814087).
- [7] HE, Y. - KAYAALP, M. A Comparison of 13 Tokenizers on MEDLINE December 2006. In *Building*. 2006. .
- [8] HLÁDEK, D. et al. Dagger: The Slovak morphological classifier. In *ELMAR, 2012 Proceedings [online]*. 2012. no. September, s. 12–14. Dostupné na internete:
[<http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=6338504&http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=6338504>](http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=6338504&http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=6338504).
- [9] HOCK, H.H. - JOSEPH, B.D. *Language History, Language Change, and Language Relationship: An Introduction to Historical and Comparative Linguistics* [online]. . [s.l.]: De Gruyter, 2009. ISBN 9783110214307.
- [10] HORÁK, A. et al. Slovak National Corpus. In SOJKA, P. et al. Ed. *Text, Speech and Dialogue SE - 12 [online]*. [s.l.]: Springer Berlin Heidelberg, 2004. s. 89–93. ISBN 978-3-540-23049-6 Dostupné na internete: [<http://dx.doi.org/10.1007/978-3-540-30120-2_12>](http://dx.doi.org/10.1007/978-3-540-30120-2_12).
- [11] CHANG, A.X. - MANNING, C.D. SUTime: A library for recognizing and normalizing time expressions. In *Lrec [online]*. 2012. no. iii, s. 3735–3740. Dostupné na internete: [<http://www-nlp.stanford.edu/pubs/lrec2012-sutime.pdf>](http://www-nlp.stanford.edu/pubs/lrec2012-sutime.pdf).
- [12] JAHAN, N. et al. Named Entity Recognition in Indian Languages Using Gazetteer Method and Hidden Markov Model : A Hybrid Approach. In *International journal of computer Science & Engineering Technology (IJCSET)*. 2012. Vol. 3, no. 12, s. 621–628. .
- [13] JUHÁSZ, P. Anonymizácia a ochrana dát. In . 2011. .
- [14] KAŠŠÁK, O. *Extrakcia pomenovaných entít zo slovenského textu*. 2014. .
- [15] KAYAALP, M. et al. A Report to the Board of Scientific Counselors September 2013 Clinical Text De-Identification Research. In . 2013. no. September. .
- [16] KAYAALP, M. et al. Challenges and Insights in Using HIPAA Privacy Rule for Clinical Text Annotation Lister Hill National Center for Biomedical Communications ,. In . 2015. .
- [17] LÁCLAVÍK, M. et al. Vyhľadávanie informácií. In . 2007. s. 1–25. .
- [18] MANNING, C.D. et al. The Stanford CoreNLP Natural Language Processing Toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational*

- Linguistics: System Demonstrations* [online]. 2014. s. 55–60. Dostupné na internete: <<http://aclweb.org/anthology/P14-5010>>.
- [19] MURTY, M.N. - DEVI, V.S. *Pattern Recognition: An Algorithmic Approach* [online]. . [s.l.]: Springer, 2011. ISBN 9780857294951.
- [20] NEAMATULLAH, I. et al. Automated de-identification of free-text medical records. In *BMC medical informatics and decision making*. 2008. Vol. 8, s. 32. .
- [21] NIGAM, K. et al. Using Maximum Entropy For Text Classification. In *Journal of Chemical Information and Modeling*. 1999. .
- [22] NRL, E.M. et al. MUC-7 EVALUATION OF IE TECHNOLOGY : Overview of Results MUC-7 Program Committee. In *Program*. 1998. no. April. .
- [23] POWERS, D.M.W. Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness & Correlation. In *Journal of Machine Learning Technologies*. 2011. Vol. 2, no. 1, s. 37 – 63. .
- [24] SCHMID, H. Improvements in Part-of-Speech Tagging with an Application to German. In *Proceedings of the ACL SIGDAT-Workshop*. 1995. .
- [25] STENETORP, P. et al. BRAT : a Web-based Tool for NLP-Assisted Text Annotation. In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics (EACL '12)* [online]. 2012. no. Figure 1, s. 102–107. Dostupné na internete: <<http://dl.acm.org/citation.cfm?id=2380921.2380942>>.
- [26] TANG, J. et al. Information Extraction: Methodologies and applications. In *Emerging Technologies of Text Mining: Techniques and Applications*. 2008. s. 1–33. .
- [27] WEBSTER, J.J. - KIT, C. Tokenization as the initial phase in NLP. In . 1992. .