VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
BRNO UNIVERSITY OF TECHNOLOGY

FAKULTA INFORMAČNÍCH TECHNOLOGIÍ
FACULTY OF INFORMATION TECHNOLOGY

ÚSTAV POČÍTAČOVÉ GRAFIKY A MULTIMÉDIÍ
DEPARTMENT OF COMPUTER GRAPHICS AND MULTIMEDIA

ZJEDNOZNAČŇOVÁNÍ POJMENOVANÝCH ENTIT VE SLOVENŠTINĚ

NAMED ENTITY DISAMBIGUATION IN SLOVAK

BAKALÁŘSKÁ PRÁCE
BACHELOR'S THESIS

AUTOR PRÁCE **SAMUEL KRIŽAN**
AUTHOR

VEDOUCÍ PRÁCE **doc. RNDr. PAVEL SMRŽ, Ph.D.**
SUPERVISOR

BRNO 2022

Vysoké učení technické v Brně
Fakulta informačních technologií

Ústav počítačové grafiky a multimédií (UPGM) Akademický rok 2021/2022

Zadání bakalářské práce

Student: **Križan Samuel**
Program: Informační technologie
Název: **Zjednoznačňování pojmenovaných entit ve slovenštině**
Named Entity Disambiguation in Slovak
Kategorie: Umělá inteligence

Zadání:

1. Seznamte se s metodami vyznačování a zjednoznačňování zmínek o pojmenovaných entitách v textu na základě strojového učení.
2. Navrhněte a implementujte systém pro automatickou extrakci typů a základních atributů pojmenovaných entit z exportu dat slovenské Wikipedie.
3. Vytvořte systém pro automatické doplňování tvarů jmen a ověřování úplnosti dat vzhledem k zadanému korpusu.
4. Vyhodnoťte výsledky systému na reprezentativním vzorku dat.
5. Vytvořte stručný plakát prezentující práci, její cíle a výsledky.

Literatura:

- Manning, C. D., Schütze, H., Foundations of Statistical Natural Language Processing, MIT Press, 1999, ISBN 0-262-13360-1.

Pro udělení zápočtu za první semestr je požadováno:

- funkční prototyp řešení

Podrobné závazné pokyny pro vypracování práce viz https://www.fit.vut.cz/study/theses/

Vedoucí práce: **Smrž Pavel, doc. RNDr., Ph.D.**
Vedoucí ústavu: Černocký Jan, doc. Dr. Ing.
Datum zadání: 1. listopadu 2021
Datum odevzdání: 11. května 2022
Datum schválení: 1. listopadu 2021

Abstrakt

Práca sa zaoberá rozpoznávaním a zjednoznačňovaním pomenovaných entít. V rámci práce bol vytvorený základný systém obsahujúci všetky prerekvizity potrebné pre zjednoznačňovanie pomenovaných entít v slovenčine. Súčasťou systému je vytvorenie znalostnej bázy z exportu slovenskej Wikipédie. Tá bola následne porovnávaná so znalostnou bázou z Wikidát, čím sa zistilo, že hlavným prínosom použitia znalostnej bázy z Wikipédie pre slovenčinu je väčšie pokrytie entitami s odkazom na slovenskú Wikipédiu a lepšie určovanie tried entít. Okrem toho bola vykonaná aj aktualizácia morfologického slovníka výskumnej skupiny KNOT@FIT, ktorá priniesla zlepšenie v rozsahu 33-39 %. Práca predpokladá možné využitie v spojitosti s rozšírením systému o zjednoznačňovací modul a zlepšením pokrytia alternatívnych pomenovaní.

Abstract

Thesis deals with the topic of named entity recognition and disambiguation. A basic system was created which includes all prequisitions necessary for named entity disambiguation in Slovak language. Part of the system is building of a knowledge base out of an export from Slovak Wikipedia. This was subsequently compared to knowledge base obtained from Wikidata, which revealed that the main contribution of Wikipedia knowledge base for Slovak language is greater coverage of entities with link to Slovak Wikipedia and better determination of entity classes. Apart from that, morfological dictionary of KNOT@FIT research group was updated, which yielded an improvement by 33-39 %. This work presumes possible utilization in relation to system extention by a disambiguation modul and enhancement of alternative names coverage.

Kľúčové slová

pomenovaná entita, rozpoznávanie pomenovaných entít, zjednoznačňovanie pomenovaných entít, znalostná báza, Wikipédia, extrakcia informácií

Keywords

named entity, named entity recognition, named entity disambiguation, knowledge base, Wikipedia, information extraction

Citácia

KRIŽAN, Samuel. *Zjednoznačňování pojmenovaných entit ve slovenštině*. Brno, 2022. Bakalářská práce. Vysoké učení technické v Brně, Fakulta informačních technologií. Vedoucí práce doc. RNDr. Pavel Smrž, Ph.D.

Zjednoznačňování pojmenovaných entit ve slovenštině

Prehlásenie

Prehlasujem, že som túto bakalársku prácu vypracoval samostatne pod vedením pána doc. RNDr. Pavla Smrža, Ph.D. Uviedol som všetky literárne pramene a publikácie, z ktorých som čerpal.

.....
Samuel Križan
9. mája 2022

Poďakovanie

Rád by som poďakoval môjmu vedúcemu práce doc. RNDr. Pavlovi Smržovi Ph.D. za poskytnutú odbornú pomoc a vedenie pri vypracovaní tejto práce. Tiež by som chcel poďakovať Ing. Janovi Doležalovi za poskytnuté dáta pre vyhodnotenie.

Obsah

1 Úvod 2

2 Pomenované entity a spôsoby ich rozpoznávania 3
2.1 Rozpoznávanie na základe znalostnej bázy 4
2.2 Rozpoznávanie na základe pravidiel 4
2.3 ’tatistické rozpoznávanie 5
2.3.1 Skrytý Markovov model 5
2.3.2 Model zalo^{o}ený na princípe maximálne entropie 6
2.4 Rozpoznávanie hlbokým u£ením 7

3 Postupy zjednozna£¬ovania pomenovaných entít 8
3.1 Vz´ah medzi rozpoznávaním a zjednozna£¬ovaním 8
3.2 Výber kandidátov 9
3.3 Ohodnotenie kandidátov 9
3.4 Problém chýbajúcej entity pre priradenie 10
3.5 Zjednozna£¬ovanie na základe textovej reprezentácie 10
3.6 Zjednozna£¬ovanie na základe grafovej reprezentácie 11

4 Získavanie dát z Wikipédie 12
4.1 ’truktúra textu £lánkov 12
4.2 Prístup cez API 13

5 Návrh a implementácia systému 15
5.1 Tvorba znalostnej bázy 16
5.1.1 Získavanie informácií z Wikipédie 19
5.1.2 Aktualizácia znalostnej bázy 21
5.2 Tvorba morfologického slovníka 22

6 Vyhodnotenie práce 24
6.1 Porovnanie znalostnej bázy z Wikipédie so znalostnou bázou z Wikidát . . 24
6.2 Vyhodnotenie rozpoznaných entít 27

7 Záver 29

Literatúra 30

A Pravidlá pre rozpoznanie tried entít 34

B Plagát 36

Kapitola 1

Úvod

S rozvojom informa£nej spolo£nosti sa £oraz viac informácií presúva do digitálneho priestoru. To otvára nové mo$nosti pre ich spracovanie a následné vyu$itie. Ve©ká £as´ týchto informácií je v²ak ulo$ená v textovej podobe, £o ich spracovanie komplikuje. Ako jeden zo spôsobov získavania informácií z textu vznikol postup, zalo$ený na rozpoznávaní vlastných mien a ich bli$²ej ²peci kácií rozpoznávanie pomenovaných entít. Samotné rozpoznanie týchto pomenovaných entít v²ak ve©a informácií nepriná²a, a tak sa vyvinul proces zjednozna£¬ovania pomenovaných entít, ktorého cie©om je previaza´ rozpoznanú entitu s jej de níciou. Ako de nícia entity sa typicky vyu$íva odkaz na príslu²nú stránku v niektorej internetovej encyklopédii alebo odkaz do ²pecializovanej databázy. Takéto postupy nachádzajú uplatnenie pri sémantickej analýze textu a môu by´ vyu$ité pri vytváraní konkrétnych aplikácií, ako je napríklad odporú£ací systém, internetový vyh©adáva£, konverza£ný robot, strojový prekla da£ jazyka apod.

Cie©om tejto práce bolo vytvori´ základný systém obsahujúci v²etky prerekvizity potrebné k zjednozna£¬ovaniu pomenovaných entít v sloven£ine. Takýto systém by potom mohol slú$i´ ako základ pre ¤al²í vývoj komplexného systému pre zjednozna£¬ovanie.

Prvá £as´ práce je vstupom do problematiky, sna$í sa vysvetli´ základné pojmy a popísa´ zau$ívané prístupy k rie²eniam. Zámerom je poskytnú´ £itate©ovi vedomosti potrebné na zorientovanie sa v danej problematike a zárove¬ mu umo$ni´ lep²ie porozumie´ postupom pou$itým v tejto práci. Kapitola 2 sa zaoberá pomenovanými entitami (ang. named entities) a technikami na ich rozpoznávanie. Kapitola 3 rozoberá sémantickú stránku pomenovaných entít a techniky prira¤ovania správneho významu pomenovaným entitám. Celok uzatvára kapitola 4, venujúca sa ²truktúre a práci s textom Wikipédie.

Druhú £as´ (kapitola 5) tvorí návrh a implementácia systému. Je tu popísaný ako celkový návrh systému, tak aj implementa£né a najmä funk£né detaily jeho vybraných £astí. Konkrétne sa práca zameriava na vytvorenie znalostnej bázy a doplnenie morfologického slovníka pre sloven£inu.

V tretej £asti (kapitola 6) sa nachádza vyhodnotenie práce. To pozostáva z porovnania vytvorenej znalostnej bázy so znalostnou bázou vytvorenou z Wikidat v rámci inej práce a vyhodnotenia vplyvu aktualizácie morfologického slovníka na rozpoznávanie entít. Následne sú diskutované hlavné nedostatky rozpoznávania týmto systémom spolu s navrhnutými zlep²eniami.

Kapitola 2

Pomenované entity a spôsoby ich rozpoznávania

Podstatou celej tejto práce sú pomenované entity. Preto je hne¤ v úvode tejto kapitoly snaha o exaktnej²ie vysvetlenie toho, £o vlastne pomenovaná entita je. Zvy²ná £as´ sa potom zaoberá problematikou ich rozpoznávania a stru£ne popisuje niektoré pou°ívané postupy.

Úvodná £as´ tejto kapitoly £erpá z £lánkov [17] a [32]. Pojem pomenovaná entita (ang. named entity, ¤alej len entita) ozna£uje ka°dý konkrétny objekt, ktorý mo°no pomenova´ vlastným menom (mô°e by´ viacslovné). Tento objekt mô°e existova´ fyzicky alebo by´ abstraktný a takisto mô°e ma´ viacero pomenovaní. Vo vedeckej komunite v²ak neexistuje zhoda na jednozna£nej de nícii entity. Samotný pojem bol zavedený na MUC-6 (²iesta zo série Message Understanding Conferences), kde sa zárove¬ ur£ili 3 triedy delenia entít: organizácia, osoba, miesto (ang. organisation, person, location) [17]. Rovnako tam boli zavedené aj akési formálne pravidlá, ktoré mali zade nova´, £o mo°no ozna£i´ ako pomenovanú entitu. Tie boli v²ak prispôsobené angli£tine, £o by komplikovalo ich aplikáciu na iné jazyky (napr. rozdiel v písaní ve©kých písmen na za£iatku názvov medzi angli£tinou a sloven£inou), a zárove¬ s rozvojom nových postupov rozpoznávania, stratili svoj význam. Rovnako k zavedenému spôsobu triedenia sa v praxi pristupuje zna£ne exibilne a sada tried sa upravuje pod©a potreby. Príklad ¤al²ích tried: udalos´, £asový údaj, pe¬a°ná £iastka, dielo apod.

Ke¤°e sú entity zvy£ajne nosi£om k©ú£ovej informácie vo vete, mô°u byt u°ito£ným prostriedkom pre spracovávanie prirodzeného jazyka. Konkrétne sa mô°e jedna´ o oblasti ako strojový preklad [4], strojové odpovedanie na otázky [39], získavanie informácií [18] a pod. Rovnako ako ostatné £asti extrakcie informácií z textu, aj entity so sebou priná²ajú nové výzvy a jednou z nich je práve ich rozpoznávanie v texte. To sa následne vy£lenilo ako samostatná úloha v rámci extrakcie informácií s názvom rozpoznávanie pomenovaných entít (ang. named-entity recognition, ¤alej NER).

NER typicky zah¯¬a 2 £asti: identi káciu a klasi káciu [32]. Pod pojmom identi kácia je zahrnuté nielen samotné rozpoznanie, °e sa jedná o entitu, ale aj to, ktoré v²etky slová sú súvisiace s danou entitou (ohrani£enie entity). Klasi kácia potom ozna£uje správne priradenie kategórie k entite. Zvy£ajne sa obe tieto £asti dajú rie²i´ (a aj rie²ia) sú£asne. Podobne ako v ostatných disciplínach spracovávania prirodzeného jazyka, aj tu vzniká mno°stvo nejednozna£ností. Príklady takých viet:

- Most bol zahalený hmlou. Príklad na nejednozna£nos´ v identi kácii, mô°e sa jedna´ v²eobecne o hocijaký most alebo o mesto na severozápade ƒR.

ˆ Fox kritizoval Washington . Príklad nejednozna£nosti v klasi kácii, Fox mô^{o}e by´ meno (James Fox) alebo organizácia (Fox News), takisto Washington mô^{o}e ozna£ova´ mesto alebo organizáciu (metaforické ozna£enie vlády USA).

„al²ou ve©kou výzvou pri NER je zdrojový korpus. Aby mohol systém nejakú entitu rozpozna´, musí bu¤ priamo vedie´ o jej existencii, alebo sa k jej poznaniu musí nejako dopracova´. K tomu je potrebný zdrojový korpus (slovník, lexikón), obsahujúci aspo¬ minimálne mno^{o}stvo entít pokrývajúcich celú po^{o}adovanú oblas´ jazyka. Vytvorenie takéhoto korpusu v²ak potrebuje zna£né zdroje a ur£itú odbornos´ z danej jazykovej oblasti. Preto je efektívnej²ie (a v praxi preferované), sa zamera´ namiesto celého jazyka ako celku, len na jednu ur£itú oblas´ (napr. novinové £lánky, cestovanie, umenie apod.). Navy²e, ©udský jazyk sa neustále vyvíja a ke¤^{o}e entity sú jeho sú£as´ou, je potrebné myslie´ aj na pravidelné aktualizovanie zdrojového korpusu.

2.1 Rozpoznávanie na základe znalostnej bázy

Najpriamo£iarej²í prístup k rozpoznávaniu entít je pomocou znalostnej bázy (ang. knowledge base, ¤alej len KB) [35]. KB je, zjednodu²ene povedané, centralizované úlo^{o}isko informácií vz´ahujúce sa ku konkrétnej aplikácii. V problematike NER, KB predstavuje zoznam v²etkých rozpoznávaných entít spolu s príslu²nými informáciami. Práve tvorba KB predstavuje v tomto prístupe najvä£²iu výzvu a od jej kvality (kvantity) sa potom odvíja celková úspe²nos´ systému (systém nerozpozná entity, ktoré nie sú v KB).

2.2 Rozpoznávanie na základe pravidiel

Prvé prístupy k NER boli primárne zalo^{o}ené na pravidlách (ang. rule-based). Vä£²ina takýchto systémov sa skladá z 3 hlavných komponentov: (1) súbor pravidiel extrakcie entít; (2) gazetteery[1] pre jednotlivé triedy entít; (3) extrak£ný modul (ang. the extraction engine), ktorý aplikuje predchádzajúce dve £asti na vybraný text. Súbor pravidiel a gazetteery sú pritom bu¤ celé vyrobené ru£ne, alebo sú zostavené z nieko©kých ru£ne vyrobených £astí. Úspe²ným príkladom tohto prístupu je [36]. Systém za£ína so súborom jednoduchých po£iato£ných pravidiel pre nieko©ko známych entít, ako je Nikaragua, pozri tabu©ku 2.1. Tieto pravidlá sa aplikujú a roz²irujú pomocou techniky iteratívneho bootstrappingu, £oho výsledkom je extrahovanie nových entít (v príklade 2.1 San Sebastian).

Tabu©ka 2.1: Príklad pravidiel pou^{o}itých v AutoSlog Information Extraction system. Prevzaté z [32].

Pattern	headquartered in <x>
Known locations	Nicaragua
New locations	San Miguel, Chapare region, San Miguel City
Pattern	to occupy <x>
Known locations	Nicaragua
New locations	San Sebastian neighborhood

[1] Gazetteer v²eobecne ozna£uje slovník obsahujúci zemepisné názvy; v problematike NER ozna£uje slovník obsahujúci entity ²peci ckej oblasti.

Takéto systémy na základe pravidiel sú relatívne presné, ale zvy£ajne poskytujú malé pokrytie jazyka a sú funk£né len pre men²ie oblasti. Ich výkon do ve©kej miery závisí od komplexnosti pravidiel a gazetteeru. Okrem toho, za£lenenie hlb²ích znalostí (ne º sú len povrchové slová) do takéhoto systému vyºaduje pracné manuálne úsilie. A práve na tento problém ponúkol rie²enie ²tatistický prístup NER, ktorý je exibilnej²í pri za£le¬ovaní lingvistických znalostí (napr. syntaxe), £o umoº¬uje tvorbu robustnej²ích systémov. [32]

2.3 'tatistické rozpoznávanie

Rastúca popularita ²tatistických metód v spracovávaní prirodzeného jazyka spolu s roz²írením dostupných zdrojov dát zasiahli aj do výskumu v oblasti NER. Pouºitie takýchto metód by umoºnilo vynecha´ najprácnej²iu £as´ vo vtedaj²ích systémoch NER ru£nú tvorbu pravidiel a gazetteerov. Krátko na to uº dokázali nové ²tatistické NER systémy, napr. [30] a [31], prekona´ aj tie najlep²ie systémy zaloºené na pravidlách.

Princíp fungovania ²tatistických NER je zaloºený na technike u£enia sa s u£ite©om (ang. supervised learning) a systémy sú teda tvorené dvomi hlavnými £as´ami:

1. Ozna£ené trénovacie dáta korpus obsahujúci entity oanotované potrebnými informáciami.
2. 'tatistický model pravdepodobnostná reprezentácia trénovacích dát.

'tatistický model je tvorený parametrami, ktoré mapujú jazykové udalosti na pravdepodobnos´. Napríklad takýmito parametrami môºe by´: pravdepodobnos´, ºe prvé slovo vety je entita; pravdepodobnos´, ºe konkrétne slovo ozna£uje osobu; pravdepodobnos´, ºe za entitou osoba bude nasledova´ ¤al²ia entita osoba apod. Tieto parametre sa model u£í na oanotovaných trénovacích príkladoch. Správne rozpoznanie entity sa potom typicky docieli jedným zo spôsobov: (1) nau£ením sa zjednozna£¬ovacích pravidiel na základe diskrimina£ných (rozli²ovacích) znakov; (2) nau£ením sa parametra predpokladaného rozloºenia, ktoré maximalizuje pravdepodobnos´ správnosti trénovacích dát [37]. Medzi konkrétne techniky strojového u£enia pouºívanými v problematiky NER patria napríklad: skrytý Markovov model (hidden Markov model, HMM) [5], rozhodovacie stromy (decision trees) [38], model maximálnej entropie (maximum entropy model, ME) [6], metóda podporných vektorov (support-vector machine, SVM) [12] alebo podmienené náhodné polia (conditional random elds, CRF) [29].

2.3.1 Skrytý Markovov model

Skrytý Markovov model (ang. hidden Markov model, ¤alej len HMM), bol prvý model aplikovaný na rie²enie problému NER pre angli£tinu systém IdentiFinder [5]. IdentiFinder je navrhnutý tak, ºe kaºdému slovu je moºné na základe kontextu priradi´ len jedno ozna£enie triedy, do ktorej patrí alebo ozna£enie NOT-A-NAME (nie meno), ak nereprezentuje ºiadnu z poºadovaných tried. Proces prira¤ovania tried slovám je znázornený na obr. 2.1.

Pri ozna£ovaní vety je úlohou nájs´ najpravdepodobnej²iu postupnos´ tried (NC) v danej postupnosti slov (W):

$$\max P(NC|W) \tag{2.1}$$

HMM je generatívny model, t.j. sna ºí sa generova´ postupnos´ slov (W) a ozna£enia tried (NC) z distribu£ných parametrov. Vychádza z Bayesovho pravidla:

$$P(NC|W) = \frac{P(W;NC)}{P(W)} \tag{2.2}$$

Obr. 2.1: Stavový diagram IdentiFinderu. Prevzaté z [5].

Nako©ko je menovate© (pravdepodobnos´ postupnosti slov) kon²tanta, maximalizovanie pravej strany sa dosiahne maximalizovaním £itate©a (pravdepodobnos´ prieniku postupnosti slov a postupnosti tried). Tým, ºe sa pracuje s maximalizovaním pravdepodobnosti prieniku dvoch javov, je docielené splnenie Markovovej vlastnosti. Pre efektívne preh©adávanie priestoru v²etkých moºných priradení tried a maximalizovanie P(W, NC) sa pouºíva Viterbiho algoritmus. Generovanie je potom rozdelené do 3 krokov:

1. Vyberie triedu nc, podmienenú predchádzajúcou triedou a predchádzajúcim slovom.
2. Vygeneruje prvé slovo v tejto triede, podmienené aktuálnou triedou a predchádzajúcou triedou.
3. Vygeneruje v²etky nasledujúce slová v tejto triede, kde kaºdé z nich je podmienené jeho bezprostrednému predchodcovi.

Tieto kroky sa následne opakujú aº dokým nie je vygenerovaná celá postupnos´ skúmaných slov. Zárove¬ systém pouºíva zvlá²´ ukon£ovací symbol, £o umoº¬uje po£íta´ so stavom, ºe ktoréko©vek slovo bude posledným vo svojej triede. Na rozdiel od tradi£ného HMM, pravdepodobnos´ vygenerovania konkrétneho slova pre kaºdý stav slova v kaºdej triede je 1. Tradi£nej²ie modely vyuºívajú e²te sadu stavov v rámci kaºdej triedy, kde môºe kaºdý predstavova´ ur£itý sémantický význam (napríklad stavy pre triedu osoba by mohli by´: krstné meno, stredné meno, priezvisko), pri£om kaºdý z nich má ur£itú pravdepodobnos´ spojenú s výskytom slova v slovnej zásobe. Z matematického h©adiska sú oba tieto prístupy platné, pokia© sú£et v²etkých prechodov z daného stavu je rovný 1. [5]

2.3.2 Model zaloºený na princípe maximálne entropie

Ako príklad diskriminatívneho modelu, je v tejto podkapitole popísaný princíp modelu maximálnej entropie (ang. maximum entropy, ¤alej len ME). Diskriminatívne modely fun-

gujú tak, ºe sady znakov a trénovacích dát sa nau£ia váhy jednotlivých znakov, na základe ktorých potom vykonávajú klasi káciu. Cie©om princípu maximálnej entropie, je maximalizova´ entropiu dát, aby sa trénovacie dáta £o najviac zov²eobecnili. Kaºdý znak je, spojený s parametrom , ktorý reprezentuje jeho váhu. Výpo£et podmienenej pravdepodobnosti sa v práci [6] riadi nasledovným vzorcom:

$$P(f\,|h) = \frac{\prod_i g_i(h;f_i)}{Z(h)} \tag{2.3}$$

$$Z(h) = \sum_f \prod_i g_i(h;f_i); \tag{2.4}$$

kde f sú prira¤ované zna£ky; h sú slová, ktorým sa má priradi´ zna£ka; g sú jednotlivé diskrimina£né znaky; i je po£et diskrimina£ných znakov. Nakoniec sa na nájdenie cesty s najvy²²ou pravdepodobnos´ou, ktorá reprezentuje poºadovanú postupnos´ zna£iek, pouºije Viterbiho algoritmus.

2.4 Rozpoznávanie hlbokým u£ením

Pod pojmom hlboké u£enie (ang. deep learning, ¤alej len DL) sa typicky myslí pouºitie neurónových sietí (ang. neural networks). V prípade NER sa vyuºívajú najmä rekurentné neurónové siete. Tri hlavné výhody, ktoré priná²ajú, pomenúva £lánok [24]. Prvou je nelineárna transformácia. DL modely sú schopné v¤aka nelineárnej aktiva£nej funkcii nau£i´ sa aj komplexné rozli²ovacie znaky. „al²ou výhodou je, ºe takéto modely vedia sami automaticky rozpozna´ dôleºité rozli²ovacie znaky z nespracovaných dát. Tým pádom odpadá potreba ru£ného navrhovania týchto znakov. Tre´ou výhodou je moºnos´ trénovania spôsobom koniec-koniec (ang. end-to-end), £o umoº¬uje vytvori´ komplexnej²ie systémy NER.

Prvé systémy NER na princípe DL boli zaloºené na manuálne navrhnutých vektoroch rozli²ovacích znakov, ktoré boli zvä£²a tvorené ortogra ckými znakmi (napr. ve©ké prvé písmeno alebo koncovka slova) [8]. Následne sa pre²lo na spôsob, kde bol manuálne navrhovaný vektor nahradený dodatkami (ang. embeddings, ¤alej je pouºívaný anglický výraz), £o sú n-rozmerné priestory, typicky predu£ené na ve©kej textovej sade niektorým algoritmom bez u£ite©a (napr. skip-gram [44]) [9]. V¤aka tomu je systém schopný lep²ie zachytáva´ sémantické a syntaktické vlastnosti slov.

Z h©adiska reprezentácie vstupu pre neurónovú sie´ prevláda pod©a [42] nieko©ko rôznych prístupov. (1) na úrovni slov slová sú samostatné celky reprezentované kaºdé jedným embeddingom [44]; (2) na úrovni písmen veta je postupnos´ písmen, pri£om kaºdé písmeno je reprezentované jedným embeddingom [23]; (3) na úrovni podslov jednotlivé £asti slov sú reprezentované embeddingom [43]; (4) kombinácia predchádzajúcich prístupov. Pri reprezentácii na úrovni men²ích celkov neº slov sa najprv predikuje ozna£enie (trieda entity) pre kaºdú £as´ (písmeno, morféma apod.) a tie sa prevedú na ozna£enia celých slov aº v postprocessingu. Výhodou takýchto prístupov je, ºe dokáºe lep²ie zoh©adni´ informácie obsiahnuté v jednotlivých morfémach (napr. predpony alebo prípony), a takisto si dokáºu prirodzene poradi´ aj so slovami mimo slovnej zásoby [24]. „al²ím zlep²ením je zoh©adnenie kontextu vety [3], kde sa pri reprezentácii na úrovni písmen vyuºíva vetný kontext pre generovanie embeddingov celých re´azcov písmen. To má za následok, ºe rovnaké slová môºu by´, v závislosti na kontexte vety, reprezentované odli²nými embeddingami.

Kapitola 3

Postupy zjednozna£¬ovania pomenovaných entít

Pomenovaná entita, tak ako bola predstavená v predchádzajúcej kapitole, nenesie okrem svojho názvu a triedy ºiadnu ¤al²iu informáciu. Predstavuje teda len akúsi abstraktnú reprezentáciu entity bez poznania jej reálnych vlastností. To zna£ne zmen²uje priestor jej reálneho vyuºitia. Potenciál entít v²ak spo£íva práve v informáciách o vlastnostiach danej entity. Z tohto dôvodu sa za£ali rozvíja´ rôzne postupy, ako tieto informácie k entitám prida´. Najpouºívanej²ím rie²ením je vyuºitie znalostnej bázy (viac o KB v £asti 2.1) tak, ºe sa rozpoznanej entite priradí záznam entity z KB [11], [25]. Tento proces sa nazýva zjednozna£¬ovanie pomenovaných entít (ang. named entity disambiguation, named entity linking, ¤alej len NED).

NED so sebou priná²a tri hlavné výzvy: (1) tvorba KB, (2) variácia mien, (3) nejednozna£nos´ mien. Pri tvorbe KB predstavuje najvä£²í problém získanie dostato£ne ve©kého zdrojového korpusu, z ktorého by sa dali vyextrahova´ entity pokrývajúce celú skúmanú oblas´ jazyka. Pojem variácia mien predstavuje problém, ºe jedna entita môºe by´ ozna£ená viacerými rôznymi menami. Nejedná sa pritom len o prípady odli²ných pomenovaní a prezývok, ale aj skracovanie slov, vynechanie £asti pomenovania a pri ohybných jazykoch aj sklo¬ovanie. Pri nejednozna£nosti mien sa rie²i problém, kedy sa jednej rozpoznanej entite dá priradi´ viacero entít z KB. To typicky nastáva pri vynechaní £asti viacslovného pomenovania, ale beºne sa aj stáva (napr. u ©udí alebo geogra ckých názvov), ºe viacero entít má rovnaké celé meno.

3.1 Vz´ah medzi rozpoznávaním a zjednozna£¬ovaním

V typickom systéme na zjednozna£¬ovanie entít modul NED nadväzuje na modul NER a tvoria tak komplexný systém, napr. [10]. V takomto systéme sú oba moduly navzájom nezávislé a aj ich trénovanie prebieha samostatne. V¤aka tomu je systém exibilnej²í a jednotlivé moduly môºu by´ separátne vylep²ované. Nevýhodou v²ak je, ºe v prípade chybného výstupu NER, chyba zasiahne aj NED, ktorý sa z nej nevie zotavi´. „al²ou takou nevýhodou izolácie je, ºe NER nemôºe nijako zuºitkova´ informácie, s ktorými pracuje NED.

Existujú v²ak aj rie²enia, kde je NER a NED spojené v jednom modeli [28]. To umoº¬uje lep²ie zachyti´ vzájomné závislosti týchto dvoch procesov a v¤aka zdie©aniu informácií potenciálne zlep²í´ úspe²nos´.

3.2 Výber kandidátov

Prvý krokom k zjednozna£neniu, priradeniu príslu²nej entity z KB, je nájdenie takej entity v KB. Takéto entity sú v práci ¤alej nazývané ako kandidátne entity a entita, ktorej sa h©adajú kandidátne entity, je ¤alej nazývaná skúmaná entita. Najpriamo£iarej²ím spôsobom h©adania kandidátnych entít (nazývané aj generovanie kandidátov), je ma´ v KB pre ka°dú entitu ulo°ených £o najviac pomenovaní a následne vyh©adáva´ medzi nimi be°nými technikami na porovnávanie textových re´azcov [7], [47].

Ve©kú £as´ prípadov odli²nosti názvov skúmanej a kandidátnej entity tvoria prípady, ke¤ skúmaná entita je v dokumente pomenovaná skráteným názvom (skratkou, vynechaním niektorých slov z názvu, akronymom apod.). Pri£om je be°né, °e sa v jednom texte popri skrátenom pomenovaní vyskytne aj jedno kompletné pomenovanie pre tú istú entitu [41]. Preto sa roz²írili postupy, ktorých cie©om je rozgenerova´ takéto skrátené pomenovania na kompletný tvar. Napríklad sa vyu°íva preh©adávanie kontextu skúmanej entity (entity nachádzajúce sa v okolí skúmanej entity) alebo celého dokumentu. Ak sa tam nachádza entita, ktorá by mohla by´ dlh²ím tvarom skúmanej entity, skúmaná entita sa na¬ rozgeneruje [7], [10]. Pri rozhodovaní o tom, ktorý tvar je vhodným vzorom pre rozgenerovanie, sa dajú vyu°i´ aj techniky strojového u£enia, napr. metóda podporných vektorov (SVM) [45].

Odli²ný prístup potom predstavuje vyu°itie webových vyh©adáva£ov na identi káciu kandidátnych entít. Napríklad [11] vyh©adáva názov skúmanej entity cez Google vyh©adáva£ a odkazy na Wikipédiu, ktoré sa nachádzajú v prvých 20 výsledkoch, pova°uje za kandidátne entity. Skúmanú entitu priamo na Wikipédii vyh©adáva napríklad [46].

3.3 Ohodnotenie kandidátov

Po výbere kandidátnych entít, ktorých býva typicky viac ako jedna, prichádza druhý krok rozhodnutie, ktorá z nich je tá po°adovaná. To prebieha tak, °e sa ka°dá kandidátna entita ohodnotí pod©a ur£itých kritérií a tá s najvy²²ím ohodnotením sa vyberie ako pravdepodobne správna.

Opä´, ako pri výbere kandidátnych entít, najpriamo£iarej²ím spôsobom je pou°í´ techniky zis´ujúce podobnos´ textových re´azcov. Jedná sa napr. o techniky ako: edita£ná (Leven²tejnova) vzdialenos´ [27], Sørensenov Diceov koe cient [33] alebo Hamminhova vzdialenos´ [11].

„al²ím spôsobom ohodnotenia je popularita kandidátnej entity. Ten vychádza z konceptu, °e £ím je vä£²ia popularita entity, tým je vä£²ia pravdepodobnos´, °e autor skúmaného textu myslel práve túto entitu. Popularita v²ak nereprezentuje jednu konkrétnu metriku, a tak sa k nej dá pristupova´ rôzne. Typicky sa ako metriky vyu°ívajú: náv²tevnos´ príslu²nej stránky na Wikipédii [13], d¨°ka stránky na Wikipédii [11] alebo PageRank[1] hodnotenie odkazu na príslu²nú stránku na Wikipédii [11].

Komplexnej²ie spôsoby ohodnotenia vyu°ívajú okrem samotnej entity aj jej kontext. Z h©adiska skúmanej entity sa do kontextu môºu zahrnú´ v²etky alebo £as´ entít nachádzajúce sa v skúmanom dokumente. Z h©adiska kandidátnej entity kontext predstavujú entity spomínané na danej stránke na Wikipédii (alebo £asto len z prvého odstavca) a prípadne môºe by´ doplnený aj o ¤al²ie informácie získané z tej stránky (napr. kategórie, ²ablóny, odkazy apod.). Tieto informácie sú potom reprezentované technikami ako vektorový priestorový model (ang. vector space model, VSM)[10], [46] alebo model vreca slov

[1] PageRank algoritmus od spolo£nosti Google slú°iaci na £íselné ohodnotenie významnosti stránok na webe.

(ang. bag-of-words model) [7], [47]. Cie©om je nájs´ reprezentáciu kandidátnej entity, ktorá bude ma´ najvä£²iu podobnos´ s reprezentáciou skúmanej entity. To sa dosiahne beºnými vektorovými operáciami (napr. skalárny sú£in, kosínová podobnos´) [10], [22] alebo ²tatistickými metódami (napr. Jaccardov index) [22]. Pre zistenie takejto podobnosti sa takisto dajú pouºi´ aj techniky hlbokého u£enia [16].

Okrem predchádzajúcich spôsobov, ktoré uvaºujú zjednozna£¬ovanie pre kaºdú entitu samostatne, existuje aj kolektívne zjednozna£¬ovanie, ktoré zoh©ad¬uje vzájomnú previaza-nos´ viacerých rozhodnutí NED a zjednozna£¬uje tak v²etky entity naraz [15]. Tento prístup vychádza z úvahy, ºe v²etky entity v jednom dokumente by mali spolu sémanticky súvisie´.

3.4 Problém chýbajúcej entity pre priradenie

Ve©a rie²ení uvaºuje, ºe majú dokonalú KB so v²etkými entitami, a tak problém chýbajúcej entity pre priradenie ignorujú [10], [15], [22]. V praxi je v²ak ve©mi nepravdepodobné, ºe by nejaká KB obsahovala úplne v²etky entity. Niektoré práce k tomuto problému pristupujú uº po£as generovania tak, ºe umoº¬ujú vygenerova´ nula kandidátnych entít [7].

„al²í spôsob je rozhodnú´ o entite s najvy²²ím ohodnotením, £i je správne ju priradí´. Ak nie, predpokladá sa, ºe kandidátne entity s ni²ím ohodnotením, by takisto nebolo správne priradi´. To sa dosahuje napríklad prahom pre minimálne potrebné ohodnotenie kandidátnej entity [34], [26] alebo binárnym klasi kátorom, ktorý vyuºíva uº spomínané techniky ohodnocovania (pozri 3.3) [47], [33].

Práca [11] rie²i tento problém tým, ºe uvaºuje chýbajúcu entitu ako beºnú kandidátnu entitu, tj. prejde procesom ohodnotenia ako ostatné kandidátne entity. A ak získa najvy²²ie ohodnotenie, znamená to, ºe príslu²ná správna entita chýba.

3.5 Zjednozna£¬ovanie na základe textovej reprezentácie

Jednou z prvých prác zaoberajúcich sa problematikou NED, ktorá doteraz slúºi ako odra-zový mostík pre ostatné práce, je [10]. Cie©om systému v tejto práci je vyh©ada´ v ©ubovol-nom texte entity a priradí´ im hypertextový odkaz na príslu²ný £lánok na Wikipédii. Tento proces je tvorený dvomi fázami. V prvej sa extrahujú entity z Wikipédie spolu príslu²-nými informáciami názov entity a v²etky ozna£enia, ktorými je daná entita pomenovaná v danom £lánku; trieda entity (osoba, miesto, organizácia, rôzne); kontext výskytu (slová a entity, ktoré popisujú alebo sa vyskytujú pri danej entite); ozna£enia kategórií (kategórie, ktoré sú uvádzané v £lánku Wikipédie). V druhej fáze sa pomocou techník NER (pouºitý hybridný model vyuºívajúci postupy zaloºené na pravidlách, ²tatistikách a názvov z prvej fázy) identi kujú entity a priradí sa im trieda. Následne je zahájený proces zjednozna£¬o-vania. [10]

Samotné zjednozna£¬ovanie je zaloºené na princípe porovnávania vektorových reprezen-tácií. Skúmanej entite sa vyberú kandidátne entity na základe textovej zhody spolu s roz-generovaním krátkych tvarov pomenovania, pri£om musí by´ splnená zhoda tried skúmanej a kandidátnej entity. Z £astí kontextov (kandidátnych entít), ktoré sa vyskytli v texte a ka-tegórií (takisto kandidátnych entít) sa potom vytvorí vektor dokumentu. Ten je následne porovnávaný s entitným vektorom (vytvoreným z kontextu a kategórií danej kandidátnej entity) kaºdého zjednozna£nenia. Entita, ktorej vektor bude pri porovnaní s vektorom do-kumentu dáva´ najvä£²iu podobnos´ (kosínová podobnos´), bude potom priradená danému pomenovaniu. [10]

3.6 Zjednozna£¬ovanie na základe grafovej reprezentácie

„al²ím z mo°ných prístupov k zjednozna£¬ovaniu je vyu°itie grafovej reprezentácie. V tomto prístupe sú entity (skúmaná a kandidátne) a vz´ahy medzi nimi reprezentované formou grafu. Uzly sú tvorené kontextom okolo skúmanej entity (entity okolo skúmanej entity) a kandidátnymi entitami, pri£om hrany znázor¬ujú referen£nú závislos´ medzi uzlami. Pre ka°dý kandidátny uzol sa potom vypo£íta hodnota a najhodnotnej²í z nich sa priradí skúmanej entite.

V [14] prebieha vyhodnotenie v troch stup¬och. V prvom sa uva°ujú hrany smerujúce z kandidátnych uzlov do kontextových, v druhom sa uva°ujú hrany z kontextových do kandidátnych uzlov a v tre´om ich vá°ená kombinácia. H©adané zjednozna£nenie potom udáva hrana vedúca do najhodnotnej²ieho uzla. Pri porovnaní úspe²nosti týchto stup¬ov vy²lo, °e druhý je v priemere najúspe²nej²í, pri£om kombinované rie²enie dosahovalo najlep²ie výsledky, ak sa v KB nachádzala príslu²ná entita, ak nie, výsledky boli naopak najhor²ie [14].

„al²ie zna£né zlep²enie priniesol kolektívny prístup, ktorý je dosiahnutý jednou grafovou reprezentáciou, v ktorej sú spolu znázornené ako vz´ahy medzi cie©ovými a kandidátnymi entitami, tak aj vz´ahy medzi dvojicami kandidátnych entít, medzi ktorými je nejaká sémantická súvislos´ [15].

Kapitola 4

Získavanie dát z Wikipédie

Ke¤ºe sú£as´ou tejto práce je aj tvorba znalostnej bázy z dát Wikipédie, nasledujúca kapitola je venovaná popisu ²truktúry dát v nej a ich získavaniu. Na za£iatku je nutné spomenú´, ºe nako©ko je Wikipédia otvorenou online encyklopédiou zaloºenou na wiki redak£nom systéme, a teda £lánky na nej môºe tvorí´ a upravova´ ktoko©vek, ºiadne z niº²ie uvedených pravidiel nemusí by´ stopercentne dodrºané a je treba po£íta´ s prípadmi, kedy sa v praxi £lánok zna£ne lí²i od predpísanej ²truktúry.

4.1 ’truktúra textu £lánkov

Táto podkapitola £erpá z o ciálnej stránky Wikipédie [2]. Najpriamo£iarej²ím získavaním dát z Wikipédie je ich priame vyextrahovanie z textu £lánkov. Wikipédia vyuºíva pre formátovanie textu vlastný zna£kovací jazyk, ukáºku pouºitia tohto zna£kovacieho jazyka zachytáva obr. 4.1. Celková ²truktúra £lánku sa potom okrem samotného tela textu skladá z nasledujúcich £astí, pri£om nie vºdy £lánok obsahuje v²etky z nich.

Úvod nachádza sa na za£iatku kaºdého £lánku a je tvorený nieko©kými vetami, prípadne odsekmi. Obsahuje tu£ne zvýraznený názov £lánku a typicky v ¬om býva najvä£²ia hustota informácií. Príklad toho, ako vyzerá úvod v zdrojovom kóde £lánku, zachytáva obr. 4.1.

’ablóny slúºia na vkladanie ²truktúrovaných informácií a pomáhajú zjednoti´ vzh©ad a úplnos´ informácií pre konkrétne druhy £lánkov. Do textu sa vkladajú ako samostatne vyzna£ené £asti. Existuje nieko©ko typov ²ablón. Medzi najpouºívanej²ie ²ablóny sa radí infobox ²ablóna . Infobox obsahuje základné informácie o subjekte £lánku v tabu©kovom formáte a nachádza sa v pravej £asti. Príklad ²ablóny infobox v zdrojovom kóde £lánku zachytáva obr. 4.2. „al²ou ²ablónou je naviga£ná ²ablóna . Mal by slúºi´ na rýchlu orientáciu £itate©a zrýchli´ navigáciu medzi relevantnými £lánkami. Vä£²inou obsahuje hlavné pojmy, objekty, osoby v danej téme a nachádza sa na spodnej strane. Príklad naviga£ných ²ablón v zdrojovom kóde £lánku, zachytáva obr. 4.1. ƒasto pouºívanou je potom aj²ablóna pre citáciu , ktorá slúºi na jednotné formátovanie referencií v £lánkoch. Okrem toho e²te existuje nieko©ko inline tematických ²ablón, ur£ené pre ²truktúrované vloºenie jednej konkrétnej informácie (napr. jazykový preklad daného slova).

Kategórie sú hierarchicky usporiadané menné priestory. Spájajú £lánky rovnakej témy, £o umoº¬uje tvorbu zoznamov na základe danej témy a ©ah²iu navigáciu medzi súvisiacimi £lánkami. ƒlánok sa ozna£í kategóriou jednoduchým vloºením danej kategórie

do £lánku. Jeden £lánok môºe obsahova´ viacero rôznych kategórií sú£asne. Príklad kategórií v zdrojovom kóde £lánku, zachytáva obr. 4.2.

Závere£ná sekcia typicky obsahuje 5 £astí: poznámky, referencie, literatúra, súvisiace £lánky, externé odkazy. Má formu beºného textu v tvare nadpis a obsah. Obsah je potom tvorený zvä£²a ²ablónami a odkazmi.

Obr. 4.1: Príklad zdrojového kódu £lánku na Wikipédii v zna£kovacom jazyku. Prvý odsek tvorí ²ablóna infobox, ¤al²í odsek je úvodná sekcia a posledný odsek je jedna kapitola beº-ného textu. Kód je prebraný z https://sk.wikipedia.org/wiki/Tomá²_Garrigue_Masaryk .

4.2 Prístup cez API

Táto podkapitola £erpá z o ciálnej dokumentácie Wikipédia API [1]. „al²í z moºných spôsobov, ako získa´ dáta z Wikipédie, je priame dotazovanie sa cez API (application programming interface, slov. rozhranie pre programovanie aplikácií). Wikipédia síce nemá vlastné o ciálne API, no MediaWiki (softvér, na ktorom je postavená Wikipédia) poskytuje REST API s prístupom priamo na Wikipédiu. Týmto spôsobom sa dajú s £lánkami vykonáva´ plnohodnotné operácie[1] ako pridávanie, úprava a mazanie £lánkov. Okrem toho sa takto dajú potrebné dáta z £lánkov aj s´ahova´, £i uº formou stiahnutia celého zdrojového kódu

[1] Pre celý zoznam spôsobov vyuºitia pozri dokumentáciu na https://www.mediawiki.org/wiki/API:Main_page

Obr. 4.2: Príklad zdrojového kódu £lánku na Wikipédii v zna£kovacom jazyku. Prvý odsek je koniec závere£nej sekcie, ¤al²í odsek sú naviga£né ²ablóny, posledný odsek sú kategórie. Kód je prebraný z https://sk.wikipedia.org/wiki/Tomá²_Garrigue_Masaryk .

£lánku, celého u^{o} naformátovaného £lánku alebo len niektorých konkrétnych dát, prípadne metadát. Podporované formáty odpovede sú: json, jsonfm, none, php, phpfm, rawfm, xml, xmlfm.

Kapitola 5

Návrh a implementácia systému

Obr. 5.1: Diagram vytvoreného systému. Elipsy znázor¬ujú funk£né moduly, obd¨ºniky znázor¬ujú dátové vstupy a výstupy, kosod¨ºnik znázor¬uje proces. šltou farbou je znázornený po£iato£ný vstup, zelenou komponenty, ktoré boli vytvorené v rámci tejto práce, modrou externé komponenty vyuºívané touto prácou a hnedou poºadovaný výstup.

Zámerom tejto práce bolo vytvori´ základný systém pre rozpoznávanie entít v sloven£ine, ktorý by bol vhodným východiskovým systémom pre následné roz²írenie o zjednozna£¬ova-

nie. K tomu boli vyuºité existujúce rie²enia v rámci výskumnej skupiny KNOT@FIT, ktoré boli vhodne doplnené o ¤al²ie moduly. Schéma takéhoto systému je znázornená na obr. 5.1. Hlavným cie©om tohto systému je privies´ potrebné dáta do modulu pre rozpoznávanie a zjednozna£¬ovanie entít.

V práci je tento modul tvorený nástrojom [20] (¤alej nazývaný NER@KNOT), ktorý vygeneruje automaty ur£ené k rozpoznávaniu a zjednozna£neniu entít. Ten na vstupe vyºaduje znalostnú bázu (podrobnej²iemu popisu KB sa venuje podkapitola 5.1). Ke¤ºe sa práca zaoberá rozpoznávaním entít v sloven£ine, ktorej sú£as´ou je aj ohýbanie slovných druhov, pre lep²ie rozpoznávanie entít je nutné prida´ do systému gramatiku, ktorá by takéto ohýbanie vygenerovala. K tomu práca vyuºíva nástroj [40] (¤alej nazývaný MA@KNOT), vytvorený vrámci jednej bakalárskej práce, ktorý vygeneruje potrebné jazykové automaty na ohýbanie slov. Tieto automaty sú vytvárané na základe morfologického slovníka pre daný jazyk. Pre sloven£inu síce KNOT@FIT má uº existujúci morfologický slovník, no ten obsahuje len obmedzenú slovnú zásobu a zna£né mnoºstvo slov je do¬ potrebné doplni´. Slová chýbajúce v slovníku vie potom automaticky rozpozna´ nástroj NER@KNOT.

Ako zdrojový korpus pre tvorbu KB (znalostnej bázy) je v práci pouºitý export slovenskej verzie Wikipédie. Wikipédia bola zvolená preto, aby sa overil jej prínos z h©adiska mnoºstva ponúkaných informácií v porovnaní s inými rie²eniami (napr. Wikidáta alebo DBpédia). KB je zostavená samostatným modulom, ktorý je vytvorený pre ú£ely tejto práce, na základe podobného modulu z práce [19] (¤alej nazývané entity_kb_czech3), ktorý je ur£ený na tvorbu KB z £eskej verzie Wikipédie.

Sú£as´ou systému sú aj moduly ur£ené na manipuláciu s vytvorenou KB. Jedným z nich je modul na aktualizáciu KB, ktorý vychádza z výsledkov porovnania dvoch £asovo rôznych verzií KB. Toto porovnanie umoº¬uje ¤al²í samostatný modul. Tretí modul je ur£ený na porovnanie KB vytvorenej v rámci tejto práce s KB vytvorenou v rámci inej práce z iného zdrojového korpusu.

Vyhodnotenie výsledného systému umoº¬uje nástroj na vyh©adávanie nerozpoznaných entít. Tento nástroj pracuje na princípe porovnávania entít rozpoznaných v £lánkoch Wikipédie s odkazmi v týchto £lánkoch. Nástroj bol v £ase tvorby tejto práce ur£ený k privátnemu pouºitiu, preto zoznam nerozpoznaných entít bol sprostredkovaný autorom tohto nástroja.

5.1 Tvorba znalostnej bázy

Znalostná báza pouºitá v tejto práci obsahuje okrem odkazov do Wikipédie aj základné informácie o daných entitách. Ke¤ºe je vytváraná z dát Wikipédie, ktorú tvorí z ve©kej vä£²iny ne²trukturovaný text, najvä£²ou výzvou pri jej tvorbe je extrakcia informácií. Vytvorenie znalostnej bázy v tejto práci je zaloºené na predchádzajúcej práci entity_kb_czech3, ktorej cie©om bolo vytvori´ KB z £eskej verzie Wikipédie. Tento proces prebieha v troch fázach.

V prvej fáze sa vytvorí KB z dumpu vo formáte ²peci ckom pre daný systém. Na vstupe berie súbor so zdrojovými textami v²etkých stránok v slovenskej verzii Wikipédie (dump Wikipédie) a súbor, v ktorom je zoznam presmerovávaných stránok na Wikipédii spolu s cie©ovými stránkami. Dump je potom iteratívne spracovávaný po jednotlivých £lánkoch, pri£om sa najprv od ltrujú v²etky stránky, ktoré nepojednávajú o entitách. Zostávajúce stránky predstavujú kaºdá jednu entitu. Následne celý systém funguje na princípe mechanického spracovávania textu pomocou extrakcie k©ú£ových slov a regulárnych výrazov (tento postup je bliº²ie popísaný v podkapitole 5.1.1). Výsledná KB rozpoznáva nasledujúce triedy entít: osoba, ktívna osoba, skupina osôb, ²tát, bývalý ²tát, sídlo, vodný tok, vodná plo-

cha, reliéf, vodopád, ostrov, polostrov, kontinent. Informácie o entitách sú potom rozdelené do jednotlivých atribútov ²peci ckých pre danú triedu. Pre atribúty spolo£né pre v²etky triedy pozri tabu©ku 5.1. Pre triedne ²peci cké atribúty pozri tabu©ku 5.2 a tabu©ku 5.3. Oproti pôvodnému rie²eniu entity_kb_czech3 je do KB ku ka°dej entite pridaný atribút Wikidata ID [1], ktorý mô°e slú°i´ ako jednozna£ný univerzálny identi kátor entity aj mimo tejto KB.

Tabu©ka 5.1: Tabu©ka pre systémovo ²peci ckú KB, znázor¬ujúca atribúty na daných indexoch, ktoré sú rovnaké pre v²etky triedy. Pozn. atribúty s príznakom {m} mô°u obsahova´ viacero hodnôt.

index	atribút	popis
0	ID	identi ka£né £íslo entity unikátne v rámci danej KB
1	trieda	trieda entity
2	meno	meno entity nemusí by´ unikátne
3	aliasy {m}	¤al²ie pomenovania tejto entity
4	presmerovania {m}	stránky na Wikipédii, ktoré sú presmerovávané na túto entitu
5	popis	prvá veta £lánku na Wikipédii
6	názov na Wikipédii	názov £lánku tejto entity na Wikipédii, je unikátny v rámci jednej KB
7	obrázok {m}	cesta k obrázku na Wikimedia Commons
8	Wikipédia url	odkaz na stránku entity na Wikipédii
9	Wikidata ID	jednozna£ný identi kátor entity pou°ívaný vo Wikidatach (Q-item)

Tabu©ka 5.2: Tabu©ka pre systémovo ²peci ckú KB, znázor¬ujúca rozdielne atribúty tried na daných indexoch, 1. £as´. Pozn. atribúty s príznakom {m} mô°u obsahova´ viacero hodnôt.

index	osoba, ktívna osoba, skupina osôb	²tát, bývalý ²tát	sídlo	vodný tok	vodná plocha
10	dátum nar.	zem. ²írka	²tát	kontinent {m}	kontinent {m}
11	miesto nar.	zem. d¨°ka	zem. ²írka	zem. ²írka	zem. ²írka
12	dátum úmr.	rozloha	zem. d¨°ka	zem. d¨°ka	zem. d¨°ka
13	miesto úmr.	po£et obyv.	po£et obyv.	d¨°ka	rozloha
14	zamestnanie {m}			rozloha	
15	národnos´ {m}			prietok	
16				prame¬	

[1] Wikidata ID je v tejto práci pomenovanie pre identi kátor Q-item pou°ívaný v projekte Wikidata (https://www.wikidata.org/wiki/Wikidata:Main_Page). Na rozdiel od ostatných potenciálnych identi kátorov, je v £ase nemenný.

Tabu©ka 5.3: Tabu©ka pre systémovo ²peci ckú KB, znázor¬ujúca rozdielne atribúty tried na daných indexoch, 2. £as´. Pozn. atribúty s príznakom {m} môºu obsahova´ viacero hodnôt.

index	reliéf	vodopád	ostrov	polostrov	kontinent
10	kontinent {m}	kontinent {m}	kontinent {m}	zem. ²írka	zem. ²írka
11	zem. ²írka	zem. ²írka	zem. ²írka	zem. d¨ºka	zem. d¨ºka
12	zem. d¨ºka	zem. d¨ºka	zem. d¨ºka.		rozloha
13		výš²ka	rozloha		
14			po£et obyv.		

V druhej fáze sa z externého zdroja[2] stiahnu ²tatistiky stránok na Wikipédii (po£et spätných odkazov, po£et náv²tev £lánku, údaj, £i má £lánok pre dané heslo primárny význam) pre kaºdú entitu. Z týchto ²tatistík sa následne vypo£íta ohodnotenie entity, ktoré je moºné neskôr pouºi´ pri zjednozna£¬ovaní. Tieto údaje sa následne doplnia do KB. Modul vyuºitý v tejto fáze je prebraný z práce entity_kb_czech3 [19].

V tretej fáze sa táto verzia KB pretransformuje do univerzálneho formátu pouºívaného v rámci výskumnej skupiny KNOT@FIT. V ¬om sú rozpoznávané triedy: osoba, skupina osôb, umelec, geogra cká entita. Príslu²né atribúty sú znázornené v tabu©ke 5.4. Modul pre transformáciu je tieº prebraný z práce entity_kb_czech3 [19], pri£om je mierne pozmenený.

[2] ²tatistiky sú spracovávané v rámci KNOT@FIT a získavajú z interného úloºiska tejto výskumnej skupiny.

Tabu©ka 5.4: Tabu©ka pre univerzálnu KB, znázor¬ujúca triedy a k nim príslu²né atribúty. Ka^{o}dá entita obsahuje atribúty zo st¨pca spolo£né , potom atribúty svojej triedy a na záver atribúty zo st¨pca metriky Pozn. atribúty s príznakom {m} mô^{o}u obsahova´ viacero hodnôt.

spolo£né	osoba	skupina	umelec	geo. entita	metriky
ID	pohlavie	indiv. mená {m}	formy umenia {m}	zem. ²írka	spätné odkazy
trieda	dátum nar.	pohlavia {m}	vplyvy {m}	zem. d¨^{o}ka	náv²tevy
meno	miesto nar.	dátum nar. {m}	ULAN ID	typ {m}	primárny význam
jednozna£. meno	dátum úmr.	miesto nar. {m}	¤al²ie url {m}	²tát	skóre wiki
aliasy {m}	miesto úmr.	dátum úmr. {m}		po£et obyv.	skóre metriky
popis	národnos´ {m}	miesto úmr. {m}		nadmorská vý²ka	istota
roly {m}		národnos´ {m}		rozloha	
ktívnos´				£asová zóna {m}	
Wikipedia url				popisný kód	
Wikidata url				ID geonázvu {m}	
DBpedia url					
obrázok					

5.1.1 Získavanie informácií z Wikipédie

Získavanie informácií z dát Wikipédie tvorí základnú £as´ zostavenia KB. Tento proces prebieha vo viacerých moduloch prebraných z entity_kb_czech3 [19], no nako©ko musel by´ upravený na spracovávanie inej jazykovej verzie Wikipédie, rovnaká zostala len logická ²truktúra modulov a princípy získavania informácií. Postupy pre získavanie konkrétnych informácií sú odli²né. Podrobnej²í opis celého procesu spracovávania dát z Wikipédie je popísaný vo zvy²ku tejto podkapitoly.

Ako u^{o} bolo spomínané, prechádzanie jednotlivých £lánkov je zalo^{o}ené na prechádzaní v²etkých stránok v slovenskej verzii Wikipédie, ulo^{o}ených v jednom súbore v xml formáte (dump Wikipédie). Zo v²etkých stránok sa najprv vyberú len tie, ktoré potenciálne pojednávajú o entitách a to konkrétne tak, ^{o}e sa odstránia stránky, ktoré v nadpise obsahujú ur£ité textové re´azce. Sú to nasledovné re´azce: Wikipédia: , Redaktor: , Súbor: , MediaWiki: , ’ablóna: , Pomoc: , Kategória: , ’peciálne: , Portál: , Modul: , Zoznam , Geogra a , Spolo£enstvo . Zárove¬ sa presko£ia aj presmerovávacie a rozli²ovacie stránky. Ka^{o}dá zo zvy²ných stránok potom reprezentuje jednu entitu.

Pri spracovávaní stránky sa najprv odstránia nepotrebné £asti textu (napr. citácie, referencie, HTML poznámky apod.). Ako prvé atribúty sa vyextrahujú názov na Wikipédii , presmerovania a Wikipédia url . Názov sa zoberie priamo zo stránky z poloºky title , presmerovania sa zoberú zo súboru s presmerovaniami, kde je k©ú£om pri vyh©adávaní práve získaný názov. Z názvu sa potom pridaním pre xu https://sk.wikipedia.org/wiki/ a nahradením medzier za _ vytvorí url. Následne sa prejde k rozpoznávaniu triedy.

Trieda je ur£ovaná na základe viacerých k©ú£ových znakov nachádzajúcich sa v texte. Ak k £lánku nie je priradená ºiadna trieda, nepovaºuje sa za entitu a presko£í sa. V základe sa osoby, skupiny osôb a ktívne osoby rozpoznávajú v²etky ako osoby a aº dodato£ne im je pridaný rozli²ujúci príznak. Rovnako je to aj pri ²tátoch a bývalých ²tátoch. Vyuºívajú sa pri tom 4 hlavné znaky: názov, kategórie, infobox/geobox, prvá veta £lánku (viac o ²truktúre £lánku na Wikipédii pozri 4). Geobox je podobná ²ablóna ako infobox s tým, ºe je v²eobecnej²ie zameraná, obsahuje omnoho viac poloºiek a poloºky sú v anglickom jazyku. Niektoré stránky namiesto infoboxu pouºívajú geobox. Konkrétne znaky rozli²ované pri kaºdej triede sú vypísané v tabu©kách v prílohe A.

Po rozpoznaní triedy sa entite pridá meno tak, ºe sa z názvu na Wikipédii odstránia popisné veci ako slová v zátvorkách apod. „al²ím pridaným atribútom je ID, ktoré je vygenerované zahashované poradové £íslo entity. Následne sa volaním API Wikipédie získa Wikidata ID , zemepisná d¨ºka a zemepisná ²írka . Na API sa po²le dotaz s názvom stránky na Wikipédií. Vrátená odpove¤ obsahuje viacero informácií v json formáte, takºe je následne ¤alej spracovaná a poºadované informácie sú z nej vyextrahované. Nako©ko pri vytváraní KB dochádza k príli² ve©kému po£tu dotazov na API (pri naplnení kapacity API vracia chybu), modul implementuje cyklus, ktorý odosiela poºiadavky na API, aº dokia© nepríde kladná odpove¤ (maximálne 100 krát, aby sa zamedzilo prípadnému zacykleniu). Aj z tohto dôvodu sa môºe ojedinele sta´, ºe sa dané atribúty touto cestou nezískajú.

V ¤al²om kroku sa na základe obsahujúcich kategórií zistia príznaky pre oddelenie tried ktívna osoba, skupina osôb a bývalý ²tát od tried osoba a ²tát.

Následne sa prechádza k získavaniu informácií z obsahu £lánku. To prebieha dvomi spôsobmi extrakciou z infoboxu/geoboxu a extrakciou z prvej vety. V infoboxe aj geoboxe sú dáta uloºené v ²truktúrovanej podobe typu: poloºka = údaj a vä£²inou býva kaºdá poloºka na novom riadku. Komplikáciou je, ºe infoboxov je ve©a druhov a kaºdý má odli²né poloºky alebo odli²ne pomenované rovnaké poloºky. Oproti tomu geobox nemá viac druhov, a teda má vºdy rovnaké poloºky. Táto v²eobecnos´ v²ak spôsobuje príli² ve©ké mnoºstvo poloºiek, ktoré sú si £asto navzájom podobné a potom je beºné, ºe rovnaký údaj je vo viacerých geoboxoch zapísaný v inej poloºke. Prvá veta je ne²truktúrovaný text a informácie z nej sú získavané pomocou rôznych heuristík.

Infobox/geobox je spracovávaný po riadkoch a v nich sú regulárnymi výrazmi vyh©adávané k©ú£ové frázy zodpovedajúce poºadovanej poloºke niektorého z typov infoboxov danej triedy alebo geoboxu. Takto získané údaje v²ak bývajú uvedené v rôznych formátoch, preto je nutné ich následne transformova´ do jednotnej podoby. Sú£as´ou tejto transformácie sú navy²e procesy ako prevod formátu dátumov, rozdelenie viac hodnotových re´azcov na jednotlivé hodnoty alebo odstránenie prebyto£ných znakov. Týmto spôsobom sa dajú získa´ v²etky zvy²né atribúty okrem popisu, ak sú príslu²né dáta uvedené. ƒastým problémom v²ak býva zlé formátovanie celých infoboxov, ktoré sa prejavuje napr. neodriadkovaním kaºdej poloºky alebo zlým pouºitím zna£iek formátovacích zna£iek.

Prvá veta je z £lánku extrahovaná pomocou regulárnych výrazov a ¤alej sa z nej získava najmä popis , alternatívne mená (aliasy) a niektoré atribúty osôb dátum a miesto narodenia/úmrtia a pohlavie . Do popisu sa berie celá prvá veta o£istená od formátova-

cích zna£iek a ²ablón. Aliasy sa získavajú pomocou viacerých regulárnych výrazov zameraných ako na celý kontext vety, tak aj na formátovacie zna£ky a interpunkciu vety. Navy²e sú vyuºité aj ²ablóny pre cudzojazy£né preklady Lang a V jazyku. Informácie o narodení a úmrtí sú takisto extrahované z nieko©kých regulárnych výrazov, ktoré odzrkad©ujú naj£astej²ie spôsoby uvádzania týchto informácií v prvej vete £lánku na Wikipédii. Pohlavie je získavané na základe tvaru slovies by´ , sta´ , patrí´ v spojení s koncovkou nasledujúceho prídavného mena.

5.1.2 Aktualizácia znalostnej bázy

Pre nájdenie najefektívnej²ieho spôsobu aktualizácie KB, práca najprv porovnáva jednotlivé verzie KB vytvorené s mesa£ným rozostupom. Pre tieto £innosti bol vytvorený samostatný modul. Pri porovnávaní slúºi ako k©ú£ atribút Wikidata ID (pre uº spomínané vlastnosti). Ako sekundárny k©ú£ pri absencii Wikidata ID sa pouºije atribút Wikipédia url. Porovnávanie potom rozli²uje novo pridané entity, ako aj po£ty a typy zmien v atribútoch. Z porovnania verzií marec 2022 a apríl 2022 vy²lo, ºe nových entít je 148. Zmeny v atribútoch sú podrobnej²ie popísané v tabu©ke 5.5. Tieto ²tatistiky dokazujú, pre£o nie je vhodné pouºíva´ ºiaden iný atribút ako k©ú£ k porovnávaniu menia sa v £ase, a tým pádom by sa mohla v KB vyskytnú´ jedna entita viackrát, len s iným k©ú£om. V predstavených ²tatistikách nastávajú zmeny aj vo Wikidata ID, to je z dôvodu spätného vytvárania KB zo star²ieho dumpu Wikidata ID sa z povahy API získava v aktuálnom £ase na základe dát z dumpu. Ak dump nie je aktuálny (napr. niektoré názvy stránok na Wikipédii sa odvtedy zmenili), potom môºe nasta´ situácia, ºe dotaz na API obsahuje názov stránky, ktorá je aktuálne uº priradená k inej stránke, a teda API vráti iné Wikidata ID. Typický príklad je, ºe sa stránka na Wikipédii £asom zmení na rozli²ovaciu (rovnaký názov, iné Wikidata ID) a vytvorí sa nová stránka pre poºadovanú entitu (iný názov, rovnaké Wikidata ID). Ak je KB vytváraná v £ase aktuálnosti dumpu, tento prípad nenastane.

Pri porovnaní, najviac zmien nastalo v prvej vete £lánku, £o má následne vplyv aj na ¤al²ie atribúty z nej získavané. Ke¤ºe zmeny obsahovali, okrem pridaných a odstránených atribútov, aj úpravy v uº existujúcich atribútoch, nedá sa jednozna£ne ur£i´, ºe niektorá verzia obsahovala správnej²ie atribúty. Aktualiza£ný modul pri svojej tvorbe v²ak vychádza z predpokladu, ºe nové zmeny sú pravdepodobne k lep²iemu, preto sa pri aktualizácií berie ako základ nov²ia verzia. Pri otázke, £i je vhodné doplni´ atribúty zo star²ej verzie, ktoré v nov²ej chýbajú, sa po bliº²om preskúmaní konkrétnych prípadov zistilo, ºe vä£²inu prípadov tvorili nesprávne uvedené informácie, takºe ich doplnenie by bolo neºiadúce. Okrem atribútov v²ak môºu by´ odstránené aj samotné entity. Preskúmanie tejto moºnosti sa docielilo spustením rovnakého skriptu, len s uvedením oboch KB na vstup v opa£nom poradí. Tým sa zistilo, ºe v star²ej verzii bolo 29 entít, ktoré sa nenachádzali v nov²ej.

Aktualizácia teda v kone£nom dôsledku pozostáva z pridania týchto entít zo star²ej verzie KB do nov²ej. Okrem samotného pridania im musí by´ e²te zmenené ID, aby boli unikátne v rámci celej KB.

Tabu©ka 5.5: Zmeny v spolo£ných atribútoch entít medzi verziami KB z marca 2022 a z apríla 2022.

atribút	pridaný	odstránený	upravený
trieda	0	0	1
meno	0	0	18
aliasy	43	9	201
presmerovania	24	5	11
popis	5	1	212
názov na Wikipédii	0	0	20
obrázok	21	6	45
Wikipédia url	0	0	20
Wikidata ID	34	10	9
celkový po£et entít v nov²ej verzii			85703

5.2 Tvorba morfologického slovníka

Nako©ko je vyh©adávanie entít v tomto systéme zalo^{o}ené na rozpoznávaní konkrétnych mien (a alternatívnych mien), sloven£ina priná²a výzvu v podobe ohýbania slov. Systém výskumnej skupiny KNOT@FIT k tomu vyu^{o}íva morfologické slovníky pretransformované na kone£né automaty. Tieto slovníky sú, po vzore ©udského pou^{o}ívania sloven£iny, zalo^{o}ené na vzoroch (iných ako sú be^{o}ne pou^{o}ívané ©u¤mi), pod©a ktorých sa generujú nové tvary slov. V aktuálnych verziách slovníkov sa nachádza ur£itá sada slov, no ke¤^{o}e novovytvorená KB obsahuje zna£né mno^{o}stvo názvov, ktoré sa v týchto slovníkoch e²te nenachádzajú, je potrebné ich tam prida´.

Rozpoznanie chýbajúcich slov z KB v slovníkoch umo^{o}¬uje nástroj NER@KNOT. Pridávanie slov do slovníkov potom zabezpe£ujú skripty z nástroja MA@KNOT. Pre pridanie slova je nutné k nemu doplni´ vzor. Vzor k danému slovu doká^{o}e odhadnú´ skript lntrf2lpn.py z MA@KNOT na základe informácií aspo¬ o jednom ohýbacom páde daného slova (napr. pre podstatné mená to zah¯¬a informácie o slovnom druhu, rode, £ísle a páde). Podstatné mená, ktorých je v KB drvivá vä£²ina, sú typicky automaticky v nominatíve, tak^{o}e k nim sta£í ur£i´ rod a £íslo. Skript umo^{o}¬uje aj de nova´ gramatickú kategóriu nejednozna£ne (napr. viac mo^{o}ností rodu pre jedno slovo), no tomu potom zodpovedá aj výsledná kvalita odhadu vzoru, ktorá s rastúcou nejednozna£nos´ou stúpa. Preto je cie©om ur£i´ gramatické kategórie £o mo^{o}no najjednozna£nej²ie. ƒas´ z týchto slov takto dokázal ur£i´ nástroj NER@KNOT u^{o} pri ich vyh©adávaní (pribli^{o}ne tretinu). Zvy²ným slovám je potrebné ru£ne priradi´ rod a £íslo. Ke¤^{o}e i²lo o pribli^{o}ne 35000 slov, bolo nutné vyu^{o}i´ ur£ité hromadné pravidlá. Nieko©ko príkladov takýchto pravidiel pou^{o}itých pri tejto práci:

- Osoba mô^{o}e by´ len mu^{o}ského ^{o}ivotného alebo ^{o}enského rodu.
- Geogra cká entita mô^{o}e by´ mu^{o}ského ne^{o}ivotného, ^{o}enského alebo stredného rodu.
- Osoba kon£iaca na ká , ná , ová , alebo ak jej celý názov kon£í na tieto koncovky, je pravdepodobne ^{o}enského rodu.
- Osoba kon£iaca na as , os , es je pravdepodobne mu^{o}ského ^{o}ivotného rodu.
- Entita kon£iaca na ý je pravdepodobne mu^{o}ského ^{o}ivotného alebo ne^{o}ivotného rodu.

- ˆ Entita kon£iaca na n je pravdepodobne muºského ºivotného alebo muºského neºivotného rodu.
- ˆ Entita kon£iaca na o je pravdepodobne muºského ºivotného, muºského neºivotného alebo stredného rodu.
- ˆ Osoby sú pravdepodobne v jednotnom £ísle.
- ˆ Geogra cké entity môºu by´ v jednotnom aj mnoºnom £ísle.

Okrem týchto pravidiel bolo pouºitých e²te nieko©ko ¤al²ích, zameraných na vybrané podskupiny slov na základe ich koncoviek. Je zjavné, ºe pre kaºdé z týchto pravidiel existuje ve©ké mnoºstvo výnimiek. Niektoré boli následne ru£ne prerábané, no pre ú£ely tejto práce posta£ia aj mierne nedokonalé slovníky. Celkovo je v rámci tejto prác pridaných do slovníkov 54000 slov. Tieto slovníky sa následne systémom MA@KNOT prevedú na kone£né automaty, aby ich mohol pri svojej práci vyuºíva´ systém NER@KNOT.

Kapitola 6

Vyhodnotenie práce

Nasledujúca kapitola obsahuje vyhodnotenie dvoch najdôleºitej²ích procesov vykonaných v rámci tejto práce vytvorenie znalostnej bázy a rozpoznávanie pomenovaných entít. Znalostná báza je hodnotená formou porovnávania so znalostnou bázou z iného zdroja. Tým sa vyhodnotí, rozsah a úspe²nos´ extrakcie atribútov, ako aj prínos Wikipédie ako zdrojového korpusu. Rozpoznávanie entít je hodnotené z poh©adu nerozpoznaných entít v odkazoch na Wikipédii, nasledovaných diskusiou o dôvodoch nerozpoznania entít a návrhom zlep²enia.

6.1 Porovnanie znalostnej bázy z Wikipédie so znalostnou bázou z Wikidát

Pre porovnanie KB z Wikipédie s KB z Wikidát bol v rámci práce vytvorený samostatný modul. Tento modul je in²pirovaný porovnávacím modulom z 5.1.2. Vyuºíva Wikidata ID ako k©ú£ pre spárovanie entít, len vo forme Wikidata url, respektíve Wikipedia url ako sekundárny k©ú£. Ako uº bolo spomínané, hlavným cie©om tohto porovnania je zisti´ výhody a nevýhody tvorby KB z Wikipédie. Porovnávaná verzia KB z Wikipédie bola vytvorená z aprílového dumpu. KB z Wikidát bola prevzatá z práce [21] a bola vytvorená v októbri 2021. To, ºe je KB nieko©ko mesiacov stará, v tomto prípade nie je problém, pretoºe porovnávanie nebolo zamerané na konkrétne entity, ale na extrakciu atribútov. Rovnako bolo zámerom vykona´ porovnanie na kvantitatívnej úrovni, preto pri porovnaní nebola uvaºovaná správnos´ extrahovaných atribútov.

Po£et entít, ktoré obsahovala znalostná báza vytvorená v rámci tejto práce z Wikipédie, je znázornený v tabu©ke 6.1, i²lo o takmer 86000 entít. Prevzatá KB z Wikidát obsahovala okolo 8 miliónov záznamov. Medzi týmito dvomi KB sa podarilo spárova´ entity v 16662 prípadoch. To znamená, ºe z 8 miliónov záznamov v KB z Wikidát reprezentuje entitu s vlastnou stránkou v slovenskej verzii Wikipédie len 16662 (prípadne mierne viac, ak uvaºujeme nedokonalos´ vytvorenej KB z Wikipédie). Ako bolo spomínané v kapitole 3.3, vo viacerých prístupoch k zjednozna£¬ovaniu sa popularita stránky entity na Wikipédii povaºuje za jednu z metrík ur£ujúcich význam entity. Ak sa k tomu pripojí predpoklad, ºe popularita entity v slovenskej verzii Wikipédie odzrkad©uje jej význam v sloven£ine, potom je moºné povaºova´ vä£²ie mnoºstvo entít s prepojením na slovenskú verziu Wikipédie za zna£nú výhodu.

Tabu©ka 6.1: Po£et entít v KB z Wikipédie 04/2022.

celkovo	85703
geogra cká entita	55110
osoba	30579
skupina	14

Ke¤ºe entity obsahujú aj triedne ²peci cké atribúty, prvou vecou kontrolovanou pri porovnávaní entít bola zhoda tried. Tu sa zistilo, ºe 10960 zo spárovaných 16662 má v KB z Wikidát triedu general to znamená, ºe nebola zaradená do ºiadnej z rozpoznávaných tried. Tým pádom tieto general entity neobsahujú ºiadne triedne ²peci cké atribúty. Okrem toho má atribút trieda zna£nú informatívnu hodnotu a môºe by´ vyuºitý pri zjednozna£¬ovaní (napr. pri výbere kandidátnych entít pre priradenie, pozri 3.2). Zaujímavým je, ºe 6 entít malo rozdielne ur£ené triedy, pri£om i²lo o prípady, ke¤ sa dali obe priradené triedy považova´ za správne. Napríklad entita Tri rokliny bola ur£ená ako geogra cká entita a ako organizácia. Alebo entita Druhá £esko-slovenská republika ako geogra cká entita a ako udalos´.

Porovnanie jednotlivých atribútov je zachytené v tabu©kách 6.2, 6.3, 6.4. Vyplýva z neho, ºe KB z Wikipédie dominuje v atribútoch: popis , roly , ktívnos´ , obrázok , geotyp . Pri atribúte popis má táto KB okrem kvantitatívnych aj kvalitatívne výhody popisy sú dlh²ie a obsahujú viac informácií (prvá veta £lánku proti 4-5 slovným popisom v druhej KB). Roly v tomto prípade reprezentujú zamestnanie osôb a nachádzajú sa len u men²ej £asti entít. V druhej KB sa nenachádzajú vôbec. Fiktívnos´ , obrázok a geotyp sa takisto v KB z Wikidát nenachádzajú vôbec vôbec. Oproti tomu, KB z Wikidát dosahuje pri ostatných triedne ²peci ckých atribútoch rádovo lep²iu úspe²nos´.

'peciálnym atribútom sú aliasy , ktoré umoº¬ujú samotné rozpoznávanie a zjednosna£¬ovanie. Z tabu©ky 6.2 vyplýva, ºe KB z Wikipédie obsahuje pribliºne dvakrát viac entít s aliasmi neº KB z Wikidát. Zvy²ných 321 entít, ktoré v oboch KB obsahovali aliasy, boli následne porovnávané s cie©om zisti´, v ktorej KB obsahujú jednotlivé entity viac unikátnych aliasov (aliasy nachádzajúce sa len v jednej z porovnávaných KB). Z tohto porovnania vy²lo, ºe v oboch KB je pribliºne rovnaký po£et takýchto entít.

Tabu©ka 6.2: Porovnanie vybraných atribútov KB z Wikipédie (04/2022) s KB z Wikidát (10/2021). St¨pec len Wikipédia udáva po£et entít KB z Wikipédie, ktoré obsahujú daný atribút, pri£om príslu²ná entita KB z Wikidát ho neobsahuje. St¨pec len Wikidáta udáva opak a st¨pec obe, ale rozdielne udáva po£et prípadov, kedy sa daný atribút nachádzal v oboch KB, ale nebol rovnaký.

atribút	len Wikipédia	len Wikidáta	obe, ale rozdielne
trieda	0	0	6
meno	32	0	451
aliasy	1514	786	321
popis	5304	20	11268
roly	605	0	0
ktívnos´	4906	0	0
Wikipedia url	41	0	10
obrázok	6943	0	0
po£et porovnávaných entít	16662		

Tabu©ka 6.3: Porovnanie atribútov entít triedy osoba KB z Wikipédie (04/2022) s KB z Wikidát (10/2021). St¨pec len Wikipédia udáva po£et entít KB z Wikipédie, ktoré obsahujú daný atribút, pri£om príslu²ná entita KB z Wikidát ho neobsahuje. St¨pec len Wikidáta udáva opak.

atribút	len Wikipédia	len Wikidáta
pohlavie	20	105
dátum narodenia	11	155
miesto narodenia	59	360
dátum úmrtia	28	108
miesto úmrtia	59	244
národnos´	31	3792
po£et porovnávaných entít	4604	

Tabu©ka 6.4: Porovnanie atribútov entít triedy geogra cká entita KB z Wikipédie (04/2022) s KB z Wikidát (10/2021). St¨pec len Wikipédia udáva po£et entít KB z Wikipédie, ktoré obsahujú daný atribút, pri£om príslu²ná entita KB z Wikidát ho neobsahuje. St¨pec len Wikidáta udáva opak.

atribút	len Wikipédia	len Wikidáta
zemepisná ²írka	0	282
zemepisná d¨ºka	0	282
geotyp	1092	0
²tát	0	602
po£et obyvate©ov	100	204
nadmorská vý²ka	0	824
rozloha	22	335
£asové zóny	0	683
popisný kód	0	382
ID geonázvu	0	938
po£et porovnávaných entít	1098	

6.2 Vyhodnotenie rozpoznaných entít

Pre vyhodnotenie rozpoznávania bol pouºitý výstup nástroja poskytnutý £lenom výskumnej skupiny KNOT@FIT. Tento nástroj prechádza odkazy v £lánkoch na Wikipédii a dáva ich ako vstup pre rozpoznávacie automaty (automaty vytvorené pomocou NER@KNOT). Následne porovnáva cie©ovú adresu odkazu s adresou, ktorú obsahuje rozpoznaná entita. Porovnávané sú len odkazy nachádzajúce sa pri niektorej z entít v KB. Ak je entita zle rozpoznaná (nie je vôbec rozpoznaná alebo odkazuje na inú entitu z KB, neº ktorá nezodpovedá danému odkazu), nástroj ju pridá do výstupného súboru. Pre ú£ely vyhodnotenia boli pouºité dve £asovo odli²né verzie takéhoto súboru. Prvý súbor bol vytvorený v januári 2022 na základe vtedy aktuálnej KB v dobe pred doplnením nových názvov do morfologického slovníka. Je nutné poznamena´, ºe KB vtedy obsahovala o asi 5000 entít menej (okolo 6 % KB). Druhý súbor bol vytvorený v máji 2022 z aprílovej verzie KB v dobe po aktualizácii slovníka. Aby sa zistilo, ktoré faktory zapríi¬ujú zlé rozpoznanie entít, bol prvý z súborov ru£ne analyzovaný formou náhodného výberu obsahujúcich entít. Týmto spôsobom bolo zistené, ºe medzi hlavné faktory zlého rozpoznania patria: (1) absencia vhodných aliasov, (2) nesprávne zjednozna£nenie entity, (3) nevygenerovanie správnych tvarov pre aliasy.

V tejto £asti bolo cie©om vyhodnoti´ generovanie správnych tvarov slov pomocou morfologického slovníka, ktorý bol v rámci práce doplnený o nové názvy (tento proces je popísaný v kapitole 5.2). Na základe uvedených zistení bolo pre tento ú£el pouºité kvantitatívne porovnanie dvoch spomenutých súborov (súbor pred aktualizáciou slovníka a súbor po aktualizácii slovníka). Prvý obsahoval 150761 zle rozpoznaných entít a druhý 100756. To znamená, ºe nastalo pribliºne 33% zlep²enie. Pre rozdielne rozsahy KB sa v²ak nedá jednozna£ne ur£i´ miera vplyvu doplnených slov do slovníka na zlep²enie s vy²²ím po£tom entít v KB sa zvy²uje aj po£et porovnávaných odkazov, a tým pádom aj po£et moºných prípadov zlého rozpoznania entity. Ak by sme uvaºovali lineárny vplyv po£tu entít na po£et zle rozpoz-

naných entít, po prirátaní spomínaných 6 % by sa dalo odhadnú´ zlep²enie rozpoznávania, spôsobené aktualizáciou morfologického slovníka, na pribliºne 39 %.

Následne bol aj druhý zo spomínaných súborov ru£ne analyzovaný s cie©om zisti´ konkrétne dôvody zlého rozpoznania entít. Naj£astej²ie dôvody boli:

- Nedostato£ne rozgenerované aliasy najviac prípadov, kedy zmienka o entite obsahovala len niektoré zo slov viacslovného názvu.
- Zlé priradená entita z KB £asto bola rozpoznaná len £as´ viacslovného názvu a na jej základe sa vytvorilo priradenie k entite, ktorá mala túto £as´ názvu ako alias.
- Chýbajúci alias prípady neextrahovaných aliasov (v zdrojovom £lánku sa nachádzali, ale v KB nie) a úplne chýbajúcich aliasov (chýbali aj v zdrojovom £lánku).
- Drobné odchýlky v zmienke o entite a aliase napr. vynechanie d¨º¬a, pouºitý iný typ mäk£e¬a apod.
- Chýbajúce sklo¬ovanie aliasov menej £asté prípady, ale stále sa vyskytovali.
- Uvádzané odkazy, ktoré by zrejme ani nemali by´ rozpoznané ako entita napr. názvy jazykov odkazujúce na ²táty.
- Chybný odkaz prípady, ke¤ dané slovo obsahovalo odkaz na zjavne nesprávnu stránku.

Hoci súbor po aktualizácii morfologického slovníka stále obsahuje prípady nerozpoznaných entít z dôvodu nevysklo¬ovania názvu, výrazne takýchto prípadov ubudlo. Zrejme by pomohlo niektoré slová v slovníku viac upresni´. Naj£astej²ie vyskytujúci sa dôvod nerozpoznania entity v²ak boli chýbajúce aliasy a vo ve©kej £asti prípadov i²lo o výskyt skrátených tvarov z viacslovných názvov, pri£om entity tieto skrátené tvary neobsahovali. V tomto prípade by prinieslo zna£né zlep²enie komplexnej²ie rozgenerovanie viacslovných názvov na skrátené tvary. ƒasto sa vyskytovali aj nesprávne priradené entity k daným zmienkam. To by malo vyrie²i´ pridanie vhodného modulu pre zjednozna£¬ovanie.

Kapitola 7

Záver

Zámerom práce bolo vytvorenie prerekvizít potrebných k zjednozna£¬ovaniu pomenovaných entít v sloven£ine. To bolo docielené vytvorením základného systému pre rozpoznávanie entít, zlo^{o}eného z vlastných modulov doplnených existujúcimi nástrojmi výskumnej skupiny KNOT@FIT. Navrhnutý systém najprv zostaví znalostnú bázu z exportu slovenskej Wikipédie. Následne sú jednotlivé názvy entít zo znalostnej bázy pridané do morfologického slovníka, z ktorého sú morfologickým analyzátorom vytvorené automaty umo^{o}¬ujúce sklo¬ovanie daných názvov. Vytvorená znalostná báza spolu s morfologickými automatmi potom slú^{o}i ako vstup pre nástroj na rozpoznávanie pomenovaných entít.

Modul pre zostavenie znalostnej bázy vytvorený v rámci tejto práce iteratívne prechádza jednotlivé £lánky Wikipédie a na princípe vyh©adávania k©ú£ových slov a regulárnych výrazov z nich extrahuje pomenované entity spolu s príslu²nými atribútmi. Doplnkami k tomuto procesu sú moduly ur£ené na aktualizáciu znalostnej bázy a na porovnanie znalostnej bázy s inou znalostnou bázou. Názvy chýbajúce v morfologickom slovníku sú zis´ované nástrojom NER od KNOT@FIT a následne sú do¬ ru£ne pridávané. Z dôvodu nutnosti ur£i´ gramatické kategórie týchto slov sa nepodarilo tento proces zautomatizova´.

Sú£as´ou rie²enia bolo porovnanie danej znalostnej bázy vytvorenej z Wikipédie so znalostnou bázou vytvorenou z Wikidát, s cie©om zisti´ prínosy a nedostatky pou^{o}itia Wikipédie ako zdrojového korpusu. Z porovnania vy²lo, ^{o}e znalostná báza z Wikipédie obsahuje podstatne viac entít s väzbou na slovenskú Wikipédiu a viac entít v nej má ur£enú triedu. Takisto priná²a lep²ie výsledky pri atribútoch popis, obrázok, ktívnos´ a povolanie osôb, podtyp pre geogra cké entity a mierne lep²ie výsledky pri obsiahnutých alternatívnych pomenovaniach. Naopak, znalostná báza z Wikidát dosahuje lep²ie výsledky pri v²etkých ostatných triedne ²peci ckých atribútoch. V ¤al²ej £asti bolo na odkazoch v £lánkoch Wikipédie hodnotené rozpoznávanie entít a vplyv aktualizácie morfologického slovníka. Pozorované zlep²enie bolo pribli^{o}ne v rozmedzí 33 a^{o} 39 %. Rovnako boli týmto procesom zistené aj hlavné nedostatky v rozpoznaných entitách, medzi ktorými dominovalo nedostato£né pokrytie alternatívnych pomenovaní a nesprávne prira¤ovanie entít zo znalostnej bázy.

V prípade budúceho roz²írenia o vhodný modul pre zjednozna£¬ovanie a zlep²enia rozgenerovávania alternatívnych mien, by vytvorený systém mohol dosahova´ solídne výsledky v oblasti zjednozna£¬ovania pomenovaných entít v sloven£ine. Práca zárove¬ predpokladá mo^{o}né vyu^{o}itie systému v rámci výskumnej skupiny KNOT@FIT ako roz²írenie k existujúcemu rie²eniu pre £e²tinu.

Literatúra

[1] API:Main page [online]. Wikimedia Foundation, 2001- [cit. 2022-22-4] Dostupné z: https://www :mediawiki :org/wiki/API:Main_page .

[2] Wikipedie:Vzhled a styl [online]. Wikimedia Foundation, 2001- [cit. 2022-21-4] Dostupné z: https://cs :wikipedia :org/wiki/Wikipedie:Vzhled_a_styl .

[3] Akbik , A., Blythe , D. a Vollgraf , R. Contextual string embeddings for sequence labeling. In: Proceedings of the 27th international conference on computational linguistics. 2018, s. 1638 1649.

[4] Babych , B. a Hartley , A. Improving machine translation quality with automatic named entity recognition. In: Proceedings of the 7th International EAMT workshop on MT and other language technology tools, Improving MT through other language technology tools, Resource and tools for building MT at EACL 2003 2003.

[5] Bikel , D. M., Schwartz , R. a Weischedel , R. M. An algorithm that learns what's in a name. Machine learning. Springer. 1999, zv. 34, £. 1, s. 211 231.

[6] Borthwick , A. E. A maximum entropy approach to named entity recognition. New York University, 1999.

[7] Chen , Z., Tamang , S., Lee , A., Li , X., Lin , W.-P. et al. CUNY-BLENDER TAC-KBP2010. Citeseer. 2010.

[8] Collobert , R. a Weston , J. A uni ed architecture for natural language processing: Deep neural networks with multitask learning. In: Proceedings of the 25th international conference on Machine learning 2008, s. 160 167.

[9] Collobert , R., Weston , J., Bottou , L., Karlen , M., Kavukcuoglu , K. et al. Natural language processing (almost) from scratch. Journal of machine learning research 2011, zv. 12, ARTICLE, s. 2493 2537.

[10] Cucerzan , S. Large-scale named entity disambiguation based on Wikipedia data. In: Proceedings of the 2007 joint conference on empirical methods in natural language processing and computational natural language learning (EMNLP-CoNLL) 2007, s. 708 716.

[11] Dredze , M., McNamee , P., Rao , D., Gerber , A., Finin , T. et al. Entity disambiguation for knowledge base population. In: Proceedings of the 23rd International Conference on Computational Linguistics. 2010.

[12] Ekbal , A. a Bandyopadhyay , S. Named entity recognition using support vector machine: A language independent approach. International Journal of Electrical, Computer, and Systems Engineering. 2010, zv. 4, £. 2, s. 155 170.

[13] Gattani , A., Lamba , D. S., Garera , N., Tiwari , M., Chai , X. et al. Entity extraction, linking, classi cation, and tagging for social media: a wikipedia-based approach. Proceedings of the VLDB Endowment. VLDB Endowment. 2013, zv. 6, £. 11, s. 1126 1137.

[14] Guo , Y., Che , W., Liu , T. a Li , S. A graph-based method for entity linking. In: Proceedings of 5th International Joint Conference on Natural Language Processing. 2011, s. 1010 1018.

[15] Han , X., Sun , L. a Zhao , J. Collective entity linking in web text: a graph-based method. In: Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval. 2011, s. 765 774.

[16] He , Z., Liu , S., Li , M., Zhou , M., Zhang , L. et al. Learning entity representation for entity disambiguation. In: Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). 2013, s. 30 34.

[17] Hirschman , L. The evolution of evaluation: Lessons from the message understanding conferences. Computer Speech & Language. Elsevier. 1998, zv. 12, £. 4, s. 281 305.

[18] Khalid , M. A., Jijkoun , V. a Rijke , M. d. The impact of named entity normalization on information retrieval for question answering. In: Springer. European Conference on Information Retrieval. 2008, s. 705 710.

[19] KNOT@FIT . Entity_kb_czech3. Prebraté 1.10.2021. Dostupné z: https://github :com/KNOT-FIT-BUT/entity_kb_czech3 .

[20] KNOT@FIT . NER. Prebraté 31.3.2022. Dostupné z: https://github :com/KNOT-FIT-BUT/NER.

[21] KNOT@FIT . Wikidata2. Nav²tívené na internej wiki KNOT 10.11.2021. Dostupné z: https://knot :fit :vutbr :cz/wiki/index :php/Wikidata2 .

[22] Kulkarni , S., Singh , A., Ramakrishnan , G. a Chakrabarti , S. Collective annotation of wikipedia entities in web text. In: Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining. 2009, s. 457 466.

[23] Kuru , O., Can , O. A. a Yuret , D. Charner: Character-level named entity recognition. In: Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers. 2016, s. 911 921.

[24] Li , J., Sun , A., Han , J. a Li , C. A Survey on Deep Learning for Named Entity Recognition. IEEE Transactions on Knowledge and Data Engineering. 2022, zv. 34, £. 1, s. 50 70. DOI: 10.1109/TKDE.2020.2981314.

[25] Li , Y., Tan , S., Sun , H., Han , J., Roth , D. et al. Entity disambiguation with linkless knowledge bases. In: Proceedings of the 25th international conference on world wide web. 2016, s. 1261 1270.

[26] Li , Y., Wang , C., Han , F., Han , J., Roth , D. et al. Mining evidences for named entity disambiguation. In: Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining 2013, s. 1070 1078.

[27] Liu , X., Li , Y., Wu , H., Zhou , M., Wei , F. et al. Entity linking for tweets. In: Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2013, s. 1304 1311.

[28] Luo , G., Huang , X., Lin , C.-Y. a Nie , Z. Joint entity recognition and disambiguation. In: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing 2015, s. 879 888.

[29] Luo , L., Yang , Z., Yang , P., Zhang , Y., Wang , L. et al. An attention-based BiLSTM-CRF approach to document-level chemical named entity recognition. Bioinformatics . Oxford Academic. 2018, zv. 34, £. 8, s. 1381 1388.

[30] Mikheev , A., Moens , M. a Grover , C. Named entity recognition without gazetteers. In: Ninth Conference of the European Chapter of the Association for Computational Linguistics. 1999, s. 1 8.

[31] Miller , S., Crystal , M., Fox , H., Ramshaw , L., Schwartz , R. et al. BBN: Description of the SIFT system as used for MUC-7. In: Seventh Message Understanding Conference (MUC-7): Proceedings of a Conference Held in Fairfax, Virginia, April 29-May 1, 1998 . 1998.

[32] Mohit , B. Named entity recognition. In: Natural language processing of semitic languages Springer, 2014, s. 221 245.

[33] Monahan , S., Lehmann , J., Nyberg , T., Plymale , J. a Jung , A. Cross-Lingual Cross-Document Coreference with Entity Linking. In: TAC . 2011.

[34] Pilz , A. a Paaÿ , G. From names to entities using thematic context distance. In: Proceedings of the 20th ACM international conference on Information and knowledge management 2011, s. 857 866.

[35] Radford , W., Carreras , X. a Henderson , J. Named entity recognition with document-speci c KB tag gazetteers. In: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing 2015, s. 512 517.

[36] Riloff , E. a Phillips , W. An introduction to the sundance and autoslog systems Technical Report UUCS-04-015, School of Computing, University of Utah, 2004.

[37] Sharnagat , R. Named entity recognition: A literature survey. Center For Indian Language Technology Indian Institute of Technology. 2014, s. 1 27.

[38] Szarvas , G., Farkas , R. a Kocsor , A. A multilingual named entity recognition system using boosting and c4. 5 decision tree learning algorithms. In: Springer. International Conference on Discovery Science 2006, s. 267 278.

[39] Toral , A., Noguera , E., Llopis , F. a Munoz , R. Improving question answering using named entity recognition. In: Springer. International Conference on Application of Natural Language to Information Systems 2005, s. 181 191.

[40] ƒerný , S. Morfologický analyzátor pomocí kone£ných automat· Brno, CZ, 2008. Bakalá°ská práce. Vysoké u£ení technické v Brn¥, Fakulta informa£ních technologií. Dostupné z: https://www :fit :vut :cz/study/thesis/7067/ .

[41] Wacholder , N., Ravin , Y. a Choi , M. Disambiguation of proper names in text. In: Fifth Conference on Applied Natural Language Processing1997, s. 202 208.

[42] Yadav , V. a Bethard , S. A survey on recent advances in named entity recognition from deep learning models.ArXiv preprint arXiv:1910.11470 . 2019.

[43] Yadav , V., Sharp , R. a Bethard , S. Deep a x features improve neural named entity recognizers. In: Proceedings of the seventh joint conference on lexical and computational semantics 2018, s. 167 172.

[44] Yao , L., Liu , H., Liu , Y., Li , X. a Anwar , M. W. Biomedical named entity recognition based on deep neutral network.Int. J. Hybrid Inf. Technol . 2015, zv. 8, £. 8, s. 279 288.

[45] Zhang , W., Sim, Y.-C., Su, J. a Tan , C.-L. Entity linking with e ective acronym expansion, instance selection and topic modeling. InTwenty-Second International Joint Conference on Arti cial Intelligence . 2011.

[46] Zhang , W., Su, J., Tan , C. L. a Wang , W. T. Entity linking leveraging automatically generated annotation. In: Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010) 2010, s. 1290 1298.

[47] Zhang , W., Tan , C. L., Su, J., Chen , B., Wang , W. et al. I2R-NUS-MSRA at TAC 2011: Entity Linking. In: Citeseer. TAC . 2011.

Príloha A

Pravidlá pre rozpoznanie tried entít

Tabu©ka A.1: Osoba pravidlá pre rozpoznanie.

infobox	kategórie	prvá veta
Zoznam typov infoboxov osôb na£ítaný z vopred vytvoreného súboru.	Typické ©udské kategórie ako: narodenia, úmrtia, postavy, osobnosti, ©udia, bohovia apod.	Výrazy týkajúce sa ©udí napr. zna£ky pre narodenie alebo úmrtie, slovesá by´, patri´, sta´ sa apod.

Tabu©ka A.2: 'tát pravidlá pre rozpoznanie.

kategórie
'táty v [názov kontinentu] , ƒlenovia [názov medzinárodnej organizácie] , Neuznané ²táty , Zaniknuté ²táty .

Tabu©ka A.3: Sídlo pravidlá pre rozpoznanie.

názov	infobox/geobox	kategórie
Základ slova de nujúceho sídlo napr. sídl , obec , dedin , mest apod.	Obsahuje výrazy: sídlo , obec , katastrálne územie , mesto , settlement .	mestá v/na , obce v/na , osady v/na , prímorské letoviská v/na

Tabu©ka A.4: Vodný tok pravidlá pre rozpoznanie.

názov	infobox/geobox	kategórie
Obsahuje slová: bystrina , potok , rie£ka , rieka , ve©tok .	Obsahuje slová: potok , river , stream , creek .	potoky/rieky v/na ,

Tabu©ka A.5: Vodná plocha pravidlá pre rozpoznanie.

názov	infobox/geobox	kategórie
Obsahuje slová: rybník , jazero , more , oceán , tô¬a , vodné dielo .	Obsahuje slová: vodná plocha , more , oceány , lake , sea , reservoir , water , ocean .	Rybníky/Jazerá/Rie£ne jazerá/Vodné diela/Zálivy-/Rieky/Potoky v/na

Tabu©ka A.6: Reliéf pravidlá pre rozpoznanie.

názov	infobox/geobox	kategórie
Obsahuje slová: hora , pohorie , vrch , priesmyk , sedlo .	Obsahuje slová: reliéf , vrch , priesmyk , pohorie , Mountain , Pass , Mountain Range , Volcano , Caldera , Valley .	Obsahuje slová: Pohoria , Vrchy , Ka¬ony , Doliny , Ní^{o}iny , Priepasti , Sopky , Pú²te .

Tabu©ka A.7: Polostrov, ostrov, vodopád, kontinent pravidlá pre rozpoznanie.

názov	infobox/geobox	kategórie
Obsahuje slová: polostrov , ostrov , vodopád , kontinent , .	Obsahuje slová: peninsula , ostrov , island , vodopád , waterfall , kontinent , continent ,.	Obsahuje slová: Polostrovy , Ostrovy , Súostrovia , Vodopády .

Príloha B

Plagát

[redacted]

Vysoké | kechnické v + ‡ 2022
bakalárska práca , vedúci : doc. RNDr . Pavel 0 Ô %Ph.D.

Abstrakt

[redacted]

Metodológia

[redacted]

Výsledky

Tab . 1 ¨ ® º ntít v znalostnej báze (04 /2022)

celkom	**85 730**
geografická entita	55 110
osoba	30 579
skupina	14

Tab . 2: · N • N ã @ ktá ia morfologického slovníka

5 ® º slov pridaných do slovníka	**54 000**
• ü º 5 D º ra ž poznávania entít	**33-39 %**

[redacted]

Záver

Z porovnania so znalostnou bázou z Wikidát n u D ü º k nalostná báza z Wikipédie obsahuje podstatne viac entít s väzbou na slovenskú Wikipédiu a viac entít v nej má V 8 ® º tv du . Takisto 5 8 ã • D ü º 5 D ã výsledky pri atribútoch popis, obrázok, Ô ã ø N å a po anie osôb, podtyp pre geografické entity a mierne ü º 5 D ã výsledky pri obsiahnutých alternatívnych pomenovaniach . Naopak, znalostná báza z Wikidát dosahuje ü º 5 D ã výsledky pri n D º N ø v stá ných triedne D 5 º ¬ ã Ô ã a ú ov .
Z Ý þ • ³ ã @ po z n ávania entít boli ako hlavné nedostatky systému zistené ne d @ N • ® V ô V k ie alternatívnych pomenovaní a nesprávne 5 8 ã 8 • µ n entít zo znalostnej bázy . V prípade µ • ü D ã 5 ø 8 • ® n • ã preto bolo vhodné • ü º 5 D ã generovávanie alternatívnych pomenovaní a 5 8 ã ³ • v hodný modul pre • ô º ³ • • ® n • . ã º

. 7 Õ

[redacted]