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

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

Zadání bakalářské práce

24563

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

Kapitola 1

Úvod

S rozvojom informačnej spoločnosti sa čoraz viac informácií presúva do digitálneho priestoru. To otvára nové možnosti pre ich spracovanie a následné využitie. Veľká časť týchto informácií je však uložená v textovej podobe, čo ich spracovanie komplikuje. Ako jeden zo spôsobov získavania informácií z textu vznikol postup, založený na rozpoznávaní vlastných mien a ich bližšej špecifikácií — rozpoznávanie pomenovaných entít. Samotné rozpoznanie týchto pomenovaných entít však veľa informácií neprináša, a tak sa vyvinul proces zjednoznačňovania pomenovaných entít, ktorého cieľom je previazať rozpoznanú entitu s jej definíciou. Ako definícia entity sa typicky využíva odkaz na príslušnú stránku v niektorej internetovej encyklopédii alebo odkaz do špecializovanej databázy. Takéto postupy nachádzajú uplatnenie pri sémantickej analýze textu a môžu byť využité pri vytváraní konkrétnych aplikácií, ako je napríklad odporúčací systém, internetový vyhľadávač, konverzačný robot, strojový prekladač jazyka apod.

Cieľom tejto práce bolo vytvoriť základný systém obsahujúci všetky prerekvizity potrebné k zjednoznačňovaniu pomenovaných entít v slovenčine. Takýto systém by potom mohol slúžiť ako základ pre ďalší vývoj komplexného systému pre zjednoznačňovanie.

Prvá časť práce je vstupom do problematiky, snaží sa vysvetliť základné pojmy a popísať zaužívané prístupy k riešeniam. Zámerom je poskytnúť čitateľovi vedomosti potrebné na zorientovanie sa v danej problematike a zároveň mu umožniť lepšie porozumieť postupom použitým 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 priraďovania správneho významu pomenovaným entitám. Celok uzatvára kapitola 4, venujúca sa štruktúre a práci s textom Wikipédie.

Druhú časť (kapitola 5) tvorí návrh a implementácia systému. Je tu popísaný ako celkový návrh systému, tak aj implementačné a najmä funkčné detaily jeho vybraných častí. Konkrétne sa práca zameriava na vytvorenie znalostnej bázy a doplnenie morfologického slovníka pre slovenčinu.

V tretej časti (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 zlepšeniami.

Kapitola 2

Pomenované entity a spôsoby ich rozpoznávania

Podstatou celej tejto práce sú pomenované entity. Preto je hneď v úvode tejto kapitoly snaha o exaktnejšie vysvetlenie toho, čo vlastne pomenovaná entita je. Zvyšná časť sa potom zaoberá problematikou ich rozpoznávania a stručne popisuje niektoré používané postupy.

Úvodná časť tejto kapitoly čerpá z článkov [17] a [32]. Pojem pomenovaná entita (ang. named entity, ďalej len entita) označuje každý konkrétny objekt, ktorý možno pomenovať vlastným menom (môže byť viacslovné). Tento objekt môže existovať fyzicky alebo byť abstraktný a takisto môže mať viacero pomenovaní. Vo vedeckej komunite však neexistuje zhoda na jednoznačnej definícii entity. Samotný pojem bol zavedený na MUC-6 (šiesta zo série Message Understanding Conferences), kde sa zároveň určili 3 triedy delenia entít: organizácia, osoba, miesto (ang. organisation, person, location) [17]. Rovnako tam boli zavedené aj akési formálne pravidlá, ktoré mali zadefinovať, čo možno označiť ako pomenovanú entitu. Tie boli však prispôsobené angličtine, čo by komplikovalo ich aplikáciu na iné jazyky (napr. rozdiel v písaní veľkých písmen na začiatku názvov medzi angličtinou a slovenčinou), a zároveň s rozvojom nových postupov rozpoznávania, stratili svoj význam. Rovnako k zavedenému spôsobu triedenia sa v praxi pristupuje značne flexibilne a sada tried sa upravuje podľa potreby. Príklad ďalších tried: udalosť, časový údaj, peňažná čiastka, dielo apod.

Keďže sú entity zvyčajne nosičom kľúčovej informácie vo vete, môžu byt užitočným prostriedkom pre spracovávanie prirodzeného jazyka. Konkrétne sa môže jednať o oblasti ako strojový preklad [4], strojové odpovedanie na otázky [39], získavanie informácií [18] a pod. Rovnako ako ostatné časti extrakcie informácií z textu, aj entity so sebou prinášajú nové výzvy a jednou z nich je práve ich rozpoznávanie v texte. To sa následne vyčlenilo ako samostatná úloha v rámci extrakcie informácií s názvom **rozpoznávanie pomenovaných entít** (ang. named-entity recognition, ďalej NER).

NER typicky zahŕňa 2 časti: identifikáciu a klasifikáciu [32]. Pod pojmom identifikácia je zahrnuté nielen samotné rozpoznanie, že sa jedná o entitu, ale aj to, ktoré všetky slová sú súčasťou tej entity (ohraničenie entity). Klasifikácia potom označuje správne priradenie kategórie k entite. Zvyčajne sa obe tieto časti dajú riešiť (a aj riešia) súčasne. Podobne ako v ostatných disciplínach spracovávania prirodzeného jazyka, aj tu vzniká množstvo nejednoznačností. Príklady takých viet:

- **Most** bol zahalený hmlou. — Príklad na nejednoznačnosť v identifikácii, môže sa jednať všeobecne o hocijaký most alebo o mesto na severozápade ČR.

- **Fox** kritizoval **Washington**. — Príklad nejednoznačnosti v klasifikácii, Fox môže byť meno (James Fox) alebo organizácia (Fox News), takisto Washington môže označovať mesto alebo organizáciu (metaforické označenie vlády USA).

Ďalšou veľkou výzvou pri NER je zdrojový korpus. Aby mohol systém nejakú entitu rozpoznať, musí buď priamo vedieť o jej existencii, alebo sa k jej poznaniu musí nejako dopracovať. K tomu je potrebný zdrojový korpus (slovník, lexikón), obsahujúci aspoň minimálne množstvo entít pokrývajúcich celú požadovanú oblasť jazyka. Vytvorenie takéhoto korpusu však potrebuje značné zdroje a určitú odbornosť z danej jazykovej oblasti. Preto je efektívnejšie (a v praxi preferované), sa zamerať namiesto celého jazyka ako celku, len na jednu určitú oblasť (napr. novinové články, cestovanie, umenie apod.). Navyše, ľudský jazyk sa neustále vyvíja a keďže entity sú jeho súčasťou, je potrebné myslieť aj na pravidelné aktualizovanie zdrojového korpusu.

2.1 Rozpoznávanie na základe znalostnej bázy

Najpriamočiarejší prístup k rozpoznávaniu entít je pomocou znalostnej bázy (ang. knowledge base, ďalej len KB) [35]. KB je, zjednodušene povedané, centralizované úložisko informácií vzťahujúce sa ku konkrétnej aplikácii. V problematike NER, KB predstavuje zoznam všetkých rozpoznávaných entít spolu s príslušnými informáciami. Práve tvorba KB predstavuje v tomto prístupe najväčšiu výzvu a od jej kvality (kvantity) sa potom odvíja celková úspešnosť systému (systém nerozpozná entity, ktoré nie sú v KB).

2.2 Rozpoznávanie na základe pravidiel

Prvé prístupy k NER boli primárne založené na pravidlách (ang. rule-based). Väčšina takýchto systémov sa skladá z 3 hlavných komponentov: (1) súbor pravidiel extrakcie entít; (2) gazetteery[1] pre jednotlivé triedy entít; (3) extrakčný modul (ang. the extraction engine), ktorý aplikuje predchádzajúce dve časti na vybraný text. Súbor pravidiel a gazetteery sú pritom buď celé vyrobené ručne, alebo sú zostavené z niekoľkých ručne vyrobených častí. Úspešným príkladom tohto prístupu je [36]. Systém začína so súborom jednoduchých počiatočných pravidiel pre niekoľko známych entít, ako je Nikaragua, pozri tabuľku 2.1. Tieto pravidlá sa aplikujú a rozširujú pomocou techniky iteratívneho bootstrappingu, čoho výsledkom je extrahovanie nových entít (v príklade 2.1 San Sebastian).

Tabuľka 2.1: Príklad pravidiel použitých 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 — všeobecne označuje slovník obsahujúci zemepisné názvy; v problematike NER označuje slovník obsahujúci entity špecifickej oblasti.

Takéto systémy na základe pravidiel sú relatívne presné, ale zvyčajne poskytujú malé pokrytie jazyka a sú funkčné len pre menšie oblasti. Ich výkon do veľkej miery závisí od komplexnosti pravidiel a gazetteeru. Okrem toho, začlenenie hlbších znalostí (než sú len povrchové slová) do takéhoto systému vyžaduje pracné manuálne úsilie. A práve na tento problém ponúkol riešenie štatistický prístup NER, ktorý je flexibilnejší pri začleňovaní lingvistických znalostí (napr. syntaxe), čo umožňuje tvorbu robustnejších systémov. [32]

2.3 Štatistické rozpoznávanie

Rastúca popularita štatistických metód v spracovávaní prirodzeného jazyka spolu s rozšírením dostupných zdrojov dát zasiahli aj do výskumu v oblasti NER. Použitie takýchto metód by umožnilo vynechať najprácnejšiu časť vo vtedajších systémoch NER — ručnú tvorbu pravidiel a gazetteerov. Krátko na to už dokázali nové štatistické NER systémy, napr. [30] a [31], prekonať aj tie najlepšie systémy založené na pravidlách.

Princíp fungovania štatistických NER je založený na technike učenia sa s učiteľom (ang. supervised learning) a systémy sú teda tvorené dvomi hlavnými časťami:

1. Označené trénovacie dáta — korpus obsahujúci entity oanotované potrebnými informáciami.

2. Štatistický model — pravdepodobnostná reprezentácia trénovacích dát.

Štatistický model je tvorený parametrami, ktoré mapujú jazykové udalosti na pravdepodobnosť. Napríklad takýmito parametrami môže byť: pravdepodobnosť, že prvé slovo vety je entita; pravdepodobnosť, že konkrétne slovo označuje osobu; pravdepodobnosť, že za entitou osoba bude nasledovať ďalšia entita osoba apod. Tieto parametre sa model učí na oanotovaných trénovacích príkladoch. Správne rozpoznanie entity sa potom typicky docieli jedným zo spôsobov: (1) naučením sa zjednoznačňovacích pravidiel na základe diskriminačných (rozlišovacích) znakov; (2) naučením sa parametra predpokladaného rozloženia, ktoré maximalizuje pravdepodobnosť správnosti trénovacích dát [37]. Medzi konkrétne techniky strojového učenia používanými 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 fields, CRF) [29].

2.3.1 Skrytý Markovov model

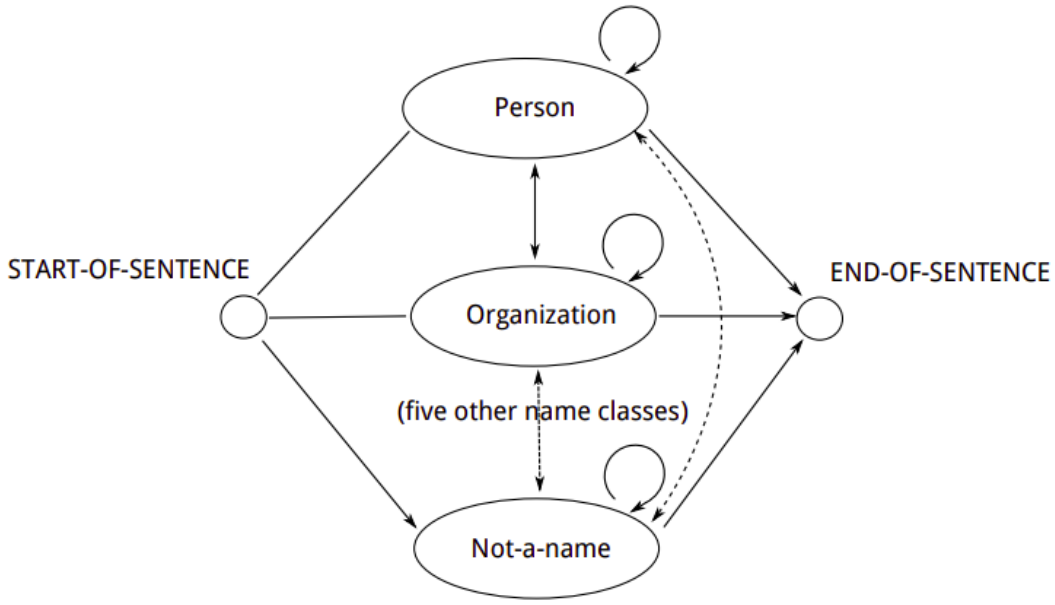
Skrytý Markovov model (ang. hidden Markov model, ďalej len HMM), bol prvý model aplikovaný na riešenie problému NER pre angličtinu — systém IdentiFinder [5]. IdentiFinder je navrhnutý tak, že každému slovu je možné na základe kontextu priradiť len jedno označenie triedy, do ktorej patrí alebo označenie NOT-A-NAME (nie meno), ak nereprezentuje žiadnu z požadovaných tried. Proces priraďovania tried slovám je znázornený na obr. 2.1.

Pri označovaní vety je úlohou nájsť najpravdepodobnejšiu postupnosť tried (NC) v danej postupnosti slov (W):

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

HMM je generatívny model, t.j. snaží sa generovať postupnosť slov (W) a označenia tried (NC) z distribučných 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].

Nakoľko je menovateľ (pravdepodobnosť postupnosti slov) konštanta, maximalizovanie pravej strany sa dosiahne maximalizovaním čitateľa (pravdepodobnosť prieniku postupnosti slov a postupnosti tried). Tým, že sa pracuje s maximalizovaním pravdepodobnosti prieniku dvoch javov, je docielené splnenie Markovovej vlastnosti. Pre efektívne prehľadávanie priestoru všetkých možných priradení tried a maximalizovanie P(W, NC) sa používa 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 všetky nasledujúce slová v tejto triede, kde každé z nich je podmienené jeho bezprostrednému predchodcovi.

Tieto kroky sa následne opakujú až dokým nie je vygenerovaná celá postupnosť skúmaných slov. Zároveň systém používa zvlášť ukončovací symbol, čo umožňuje počítať so stavom, že ktorékoľvek slovo bude posledným vo svojej triede. Na rozdiel od tradičného HMM, pravdepodobnosť vygenerovania konkrétneho slova pre každý stav slova v každej triede je 1. Tradičnejšie modely využívajú ešte sadu stavov v rámci každej triedy, kde môže každý predstavovať určitý sémantický význam (napríklad stavy pre triedu osoba by mohli byť: krstné meno, stredné meno, priezvisko), pričom každý z nich má určitú pravdepodobnosť spojenú s výskytom slova v slovnej zásobe. Z matematického hľadiska sú oba tieto prístupy platné, pokiaľ súčet všetkých prechodov z daného stavu je rovný 1. [5]

2.3.2 Model založený na princípe maximálne entropie

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

gujú tak, že sady znakov a trénovacích dát sa naučia váhy jednotlivých znakov, na základe ktorých potom vykonávajú klasifikáciu. Cieľom princípu maximálnej entropie, je maximalizovať entropiu dát, aby sa trénovacie dáta čo najviac zovšeobecnili. Každý znak je, spojený s parametrom λ, ktorý reprezentuje jeho váhu. Výpočet podmienenej pravdepodobnosti sa v práci [6] riadi nasledovným vzorcom:

$$P(f|h) = \frac{\prod_i \lambda_i^{g_i(h,f)}}{Z_\lambda(h)} \tag{2.3}$$

$$Z_\lambda(h) = \sum_f \prod_i \lambda_i^{g_i(h,f)}, \tag{2.4}$$

kde f sú priraďované značky; h sú slová, ktorým sa má priradiť značka; g sú jednotlivé diskriminačné znaky; i je počet diskriminačných znakov. Nakoniec sa na nájdenie cesty s najvyššou pravdepodobnosťou, ktorá reprezentuje požadovanú postupnosť značiek, použije Viterbiho algoritmus.

2.4 Rozpoznávanie hlbokým učením

Pod pojmom hlboké učenie (ang. deep learning, ďalej len DL) sa typicky myslí použitie neurónových sietí (ang. neural networks). V prípade NER sa využívajú najmä rekurentné neurónové siete. Tri hlavné výhody, ktoré prinášajú, pomenúva článok [24]. Prvou je nelineárna transformácia. DL modely sú schopné vďaka nelineárnej aktivačnej funkcii naučiť sa aj komplexné rozlišovacie znaky. Ďalšou výhodou je, že takéto modely vedia sami automaticky rozpoznať dôležité rozlišovacie znaky z nespracovaných dát. Tým pádom odpadá potreba ručného navrhovania týchto znakov. Treťou výhodou je možnosť trénovania spôsobom koniec-koniec (ang. end-to-end), čo umožňuje vytvoriť komplexnejšie systémy NER.

Prvé systémy NER na princípe DL boli založené na manuálne navrhnutých vektoroch rozlišovacích znakov, ktoré boli zväčša tvorené ortografickými znakmi (napr. veľké prvé písmeno alebo koncovka slova) [8]. Následne sa prešlo na spôsob, kde bol manuálne navrhovaný vektor nahradený dodatkami (ang. embeddings, ďalej je používaný anglický výraz), čo sú n-rozmerné priestory, typicky predučené na veľkej textovej sade niektorým algoritmom bez učiteľa (napr. skip-gram [44]) [9]. Vďaka tomu je systém schopný lepšie zachytávať sémantické a syntaktické vlastnosti slov.

Z hľadiska reprezentácie vstupu pre neurónovú sieť prevláda podľa [42] niekoľko rôznych prístupov. (1) na úrovni slov — slová sú samostatné celky reprezentované každé jedným embeddingom [44]; (2) na úrovni písmen — veta je postupnosť písmen, pričom každé písmeno je reprezentované jedným embeddingom [23]; (3) na úrovni podslov — jednotlivé časti slov sú reprezentované embeddingom [43]; (4) kombinácia predchádzajúcich prístupov. Pri reprezentácii na úrovni menších celkov než slov sa najprv predikuje označenie (trieda entity) pre každú časť (písmeno, morféma apod.) a tie sa prevedú na označenia celých slov až v postprocessingu. Výhodou takýchto prístupov je, že dokáže lepšie zohľadniť informácie obsiahnuté v jednotlivých morfémach (napr. predpony alebo prípony), a takisto si dokážu prirodzene poradiť aj so slovami mimo slovnej zásoby [24]. Ďalším zlepšením je zohľadnenie kontextu vety [3], kde sa pri reprezentácii na úrovni písmen využíva vetný kontext pre generovanie embeddingov celých reťazcov písmen. To má za následok, že rovnaké slová môžu byť, v závislosti na kontexte vety, reprezentované odlišnými embeddingami.

Kapitola 3

Postupy zjednoznačňovania pomenovaných entít

Pomenovaná entita, tak ako bola predstavená v predchádzajúcej kapitole, nenesie okrem svojho názvu a triedy žiadnu ďalšiu informáciu. Predstavuje teda len akúsi abstraktnú reprezentáciu entity bez poznania jej reálnych vlastností. To značne zmenšuje priestor jej reálneho využitia. Potenciál entít však spočíva práve v informáciách o vlastnostiach danej entity. Z tohto dôvodu sa začali rozvíjať rôzne postupy, ako tieto informácie k entitám pridať. Najpoužívanejším riešením je využitie znalostnej bázy (viac o KB v časti 2.1) tak, že sa rozpoznanej entite priradí záznam entity z KB [11], [25]. Tento proces sa nazýva zjednoznačňovanie pomenovaných entít (ang. named entity disambiguation, named entity linking, ďalej len NED).

NED so sebou prináša tri hlavné výzvy: (1) tvorba KB, (2) variácia mien, (3) nejednoznačnosť mien. Pri tvorbe KB predstavuje najväčší problém získanie dostatočne veľkého zdrojového korpusu, z ktorého by sa dali vyextrahovať entity pokrývajúce celú skúmanú oblasť jazyka. Pojem variácia mien predstavuje problém, že jedna entita môže byť označená viacerými rôznymi menami. Nejedná sa pritom len o prípady odlišných pomenovaní a prezývok, ale aj skracovanie slov, vynechanie časti pomenovania a pri ohybných jazykoch aj skloňovanie. Pri nejednoznačnosti mien sa rieši problém, kedy sa jednej rozpoznanej entite dá priradiť viacero entít z KB. To typicky nastáva pri vynechaní časti viacslovného pomenovania, ale bežne sa aj stáva (napr. u ľudí alebo geografických názvov), že viacero entít má rovnaké celé meno.

3.1 Vzťah medzi rozpoznávaním a zjednoznačňovaním

V typickom systéme na zjednoznačňovanie 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. Vďaka tomu je systém flexibilnejší a jednotlivé moduly môžu byť separátne vylepšované. Nevýhodou však je, že v prípade chybného výstupu NER, chyba zasiahne aj NED, ktorý sa z nej nevie zotaviť. Ďalšou takou nevýhodou izolácie je, že NER nemôže nijako zužitkovať informácie, s ktorými pracuje NED.

Existujú však aj riešenia, kde je NER a NED spojené v jednom modeli [28]. To umožňuje lepšie zachytiť vzájomné závislosti týchto dvoch procesov a vďaka zdieľaniu informácií potenciálne zlepšiť úspešnosť.

3.2 Výber kandidátov

Prvý krokom k zjednoznačneniu, priradeniu príslušnej entity z KB, je nájdenie takej entity v KB. Takéto entity sú v práci ďalej nazývané ako kandidátne entity a entita, ktorej sa hľadajú kandidátne entity, je ďalej nazývaná skúmaná entita. Najpriamočiarejším spôsobom hľadania kandidátnych entít (nazývané aj generovanie kandidátov), je mať v KB pre každú entitu uložených čo najviac pomenovaní a následne vyhľadávať medzi nimi bežnými technikami na porovnávanie textových reťazcov [7], [47].

Veľkú časť prípadov odlišnosti názvov skúmanej a kandidátnej entity tvoria prípady, keď skúmaná entita je v dokumente pomenovaná skráteným názvom (skratkou, vynechaním niektorých slov z názvu, akronymom apod.). Pričom je bežné, že sa v jednom texte popri skrátenom pomenovaní vyskytne aj jedno kompletné pomenovanie pre tú istú entitu [41]. Preto sa rozšírili postupy, ktorých cieľom je rozgenerovať takéto skrátené pomenovania na kompletný tvar. Napríklad sa využíva prehľadávanie kontextu skúmanej entity (entity nachádzajúce sa v okolí skúmanej entity) alebo celého dokumentu. Ak sa tam nachádza entita, ktorá by mohla byť dlhším tvarom skúmanej entity, skúmaná entita sa naň rozgeneruje [7], [10]. Pri rozhodovaní o tom, ktorý tvar je vhodným vzorom pre rozgenerovanie, sa dajú využiť aj techniky strojového učenia, napr. metóda podporných vektorov (SVM) [45].

Odlišný prístup potom predstavuje využitie webových vyhľadávačov na identifikáciu kandidátnych entít. Napríklad [11] vyhľadáva názov skúmanej entity cez Google vyhľadávač a odkazy na Wikipédiu, ktoré sa nachádzajú v prvých 20 výsledkoch, považuje za kandidátne entity. Skúmanú entitu priamo na Wikipédii vyhľadáva 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 požadovaná. To prebieha tak, že sa každá kandidátna entita ohodnotí podľa určitých kritérií a tá s najvyšším ohodnotením sa vyberie ako pravdepodobne správna.

Opäť, ako pri výbere kandidátnych entít, najpriamočiarejším spôsobom je použiť techniky zisťujúce podobnosť textových reťazcov. Jedná sa napr. o techniky ako: editačná (Levenštejnova) vzdialenosť [27], Sørensenov–Diceov koeficient [33] alebo Hamminhova vzdialenosť [11].

Ďalším spôsobom ohodnotenia je popularita kandidátnej entity. Ten vychádza z konceptu, že čím je väčšia popularita entity, tým je väčšia pravdepodobnosť, že autor skúmaného textu myslel práve túto entitu. Popularita však nereprezentuje jednu konkrétnu metriku, a tak sa k nej dá pristupovať rôzne. Typicky sa ako metriky využívajú: návštevnosť príslušnej stránky na Wikipédii [13], dĺžka stránky na Wikipédii [11] alebo PageRank[1] hodnotenie odkazu na príslušnú stránku na Wikipédii [11].

Komplexnejšie spôsoby ohodnotenia využívajú okrem samotnej entity aj jej kontext. Z hľadiska skúmanej entity sa do kontextu môžu zahrnúť všetky alebo časť entít nachádzajúce sa v skúmanom dokumente. Z hľadiska kandidátnej entity kontext predstavujú entity spomínané na danej stránke na Wikipédii (alebo často len z prvého odstavca) a prípadne môže byť doplnený aj o ďalšie informácie získané z tej stránky (napr. kategórie, šablóny, 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 spoločnosti Google slúžiaci na číselné ohodnotenie významnosti stránok na webe.

(ang. bag-of-words model) [7], [47]. Cieľom je nájsť reprezentáciu kandidátnej entity, ktorá bude mať najväčšiu podobnosť s reprezentáciou skúmanej entity. To sa dosiahne bežnými vektorovými operáciami (napr. skalárny súčin, kosínová podobnosť) [10], [22] alebo štatistickými metódami (napr. Jaccardov index) [22]. Pre zistenie takejto podobnosti sa takisto dajú použiť aj techniky hlbokého učenia [16].

Okrem predchádzajúcich spôsobov, ktoré uvažujú zjednoznačňovanie pre každú entitu samostatne, existuje aj kolektívne zjednoznačňovanie, ktoré zohľadňuje vzájomnú previazanosť viacerých rozhodnutí NED a zjednoznačňuje tak všetky entity naraz [15]. Tento prístup vychádza z úvahy, že všetky entity v jednom dokumente by mali spolu sémanticky súvisieť.

3.4 Problém chýbajúcej entity pre priradenie

Veľa riešení uvažuje, že majú dokonalú KB so všetkými entitami, a tak problém chýbajúcej entity pre priradenie ignorujú [10], [15], [22]. V praxi je však veľmi nepravdepodobné, že by nejaká KB obsahovala úplne všetky entity. Niektoré práce k tomuto problému pristupujú už počas generovania tak, že umožňujú vygenerovať nula kandidátnych entít [7].

Ďalší spôsob je rozhodnúť o entite s najvyšším ohodnotením, či je správne ju priradiť. Ak nie, predpokladá sa, že kandidátne entity s nižším ohodnotením, by takisto nebolo správne priradiť. To sa dosahuje napríklad prahom pre minimálne potrebné ohodnotenie kandidátnej entity [34], [26] alebo binárnym klasifikátorom, ktorý využíva už spomínané techniky ohodnocovania (pozri 3.3) [47], [33].

Práca [11] rieši tento problém tým, že uvažuje chýbajúcu entitu ako bežnú kandidátnu entitu, tj. prejde procesom ohodnotenia ako ostatné kandidátne entity. A ak získa najvyššie ohodnotenie, znamená to, že príslušná správna entita chýba.

3.5 Zjednoznačňovanie na základe textovej reprezentácie

Jednou z prvých prác zaoberajúcich sa problematikou NED, ktorá doteraz slúži ako odrazový mostík pre ostatné práce, je [10]. Cieľom systému v tejto práci je vyhľadať v ľubovolnom texte entity a priradiť im hypertextový odkaz na príslušný článok na Wikipédii. Tento proces je tvorený dvomi fázami. V prvej sa extrahujú entity z Wikipédie spolu príslušnými informáciami — názov entity a všetky označenia, ktorými je daná entita pomenovaná v danom článku; trieda entity (osoba, miesto, organizácia, rôzne); kontext výskytu (slová a entity, ktoré popisujú alebo sa vyskytujú pri danej entite); označenia kategórií (kategórie, ktoré sú uvádzané v článku Wikipédie). V druhej fáze sa pomocou techník NER (použitý hybridný model využívajúci postupy založené na pravidlách, štatistikách a názvov z prvej fázy) identifikujú entity a priradí sa im trieda. Následne je zahájený proces zjednoznačňovania. [10]

Samotné zjednoznačňovanie je založené na princípe porovnávania vektorových reprezentácií. Skúmanej entite sa vyberú kandidátne entity na základe textovej zhody spolu s rozgenerovaním krátkych tvarov pomenovania, pričom musí byť splnená zhoda tried skúmanej a kandidátnej entity. Z častí kontextov (kandidátnych entít), ktoré sa vyskytli v texte a kategórií (takisto kandidátnych entít) sa potom vytvorí vektor dokumentu. Ten je následne porovnávaný s entitným vektorom (vytvoreným z kontextu a kategórií danej kandidátnej entity) každého zjednoznačnenia. Entita, ktorej vektor bude pri porovnaní s vektorom dokumentu dávať najväčšiu podobnosť (kosínová podobnosť), bude potom priradená danému pomenovaniu. [10]

3.6 Zjednoznačňovanie na základe grafovej reprezentácie

Ďalším z možných prístupov k zjednoznačňovaniu je využitie grafovej reprezentácie. V tomto prístupe sú entity (skúmaná a kandidátne) a vzťahy medzi nimi reprezentované formou grafu. Uzly sú tvorené kontextom okolo skúmanej entity (entity okolo skúmanej entity) a kandidátnymi entitami, pričom hrany znázorňujú referenčnú závislosť medzi uzlami. Pre každý kandidátny uzol sa potom vypočíta hodnota a najhodnotnejší z nich sa priradí skúmanej entite.

V [14] prebieha vyhodnotenie v troch stupňoch. V prvom sa uvažujú hrany smerujúce z kandidátnych uzlov do kontextových, v druhom sa uvažujú hrany z kontextových do kandidátnych uzlov a v treťom ich vážená kombinácia. Hľadané zjednoznačnenie potom udáva hrana vedúca do najhodnotnejšieho uzla. Pri porovnaní úspešnosti týchto stupňov vyšlo, že druhý je v priemere najúspešnejší, pričom kombinované riešenie dosahovalo najlepšie výsledky, ak sa v KB nachádzala príslušná entita, ak nie, výsledky boli naopak najhoršie [14].

Ďalšie značné zlepšenie priniesol kolektívny prístup, ktorý je dosiahnutý jednou grafovou reprezentáciou, v ktorej sú spolu znázornené ako vzťahy medzi cieľovými a kandidátnymi entitami, tak aj vzťahy medzi dvojicami kandidátnych entít, medzi ktorými je nejaká sémantická súvislosť [15].

Kapitola 4

Získavanie dát z Wikipédie

Keďže súčasťou tejto práce je aj tvorba znalostnej bázy z dát Wikipédie, nasledujúca kapitola je venovaná popisu štruktúry dát v nej a ich získavaniu. Na začiatku je nutné spomenúť, že nakoľko je Wikipédia otvorenou online encyklopédiou založenou na wiki redakčnom systéme, a teda články na nej môže tvoriť a upravovať ktokoľvek, žiadne z nižšie uvedených pravidiel nemusí byť stopercentne dodržané a je treba počítať s prípadmi, kedy sa v praxi článok značne líši od predpísanej štruktúry.

4.1 Štruktúra textu článkov

Táto podkapitola čerpá z oficiálnej stránky Wikipédie [2]. Najpriamočiarejším získavaním dát z Wikipédie je ich priame vyextrahovanie z textu článkov. Wikipédia využíva pre formátovanie textu vlastný značkovací jazyk, ukážku použitia tohto značkovacieho jazyka zachytáva obr. 4.1. Celková štruktúra článku sa potom okrem samotného tela textu skladá z nasledujúcich častí, pričom nie vždy článok obsahuje všetky z nich.

Úvod — nachádza sa na začiatku každého článku a je tvorený niekoľkými vetami, prípadne odsekmi. Obsahuje tučne zvýraznený názov článku a typicky v ňom býva najväčšia hustota informácií. Príklad toho, ako vyzerá úvod v zdrojovom kóde článku, zachytáva obr. 4.1.

Šablóny — slúžia na vkladanie štruktúrovaných informácií a pomáhajú zjednotiť vzhľad a úplnosť informácií pre konkrétne druhy článkov. Do textu sa vkladajú ako samostatne vyznačené časti. Existuje niekoľko typov šablón. Medzi najpoužívanejšie šablóny sa radí **infobox šablóna**. Infobox obsahuje základné informácie o subjekte článku v tabuľkovom formáte a nachádza sa v pravej časti. Príklad šablóny infobox v zdrojovom kóde článku zachytáva obr. 4.2. Ďalšou šablónou je **navigačná šablóna**. Mal by slúžiť na rýchlu orientáciu čitateľa zrýchliť navigáciu medzi relevantnými článkami. Väčšinou obsahuje hlavné pojmy, objekty, osoby v danej téme a nachádza sa na spodnej strane. Príklad navigačných šablón v zdrojovom kóde článku, zachytáva obr. 4.1. Často používanou je potom aj **šablóna pre citáciu**, ktorá slúži na jednotné formátovanie referencií v článkoch. Okrem toho ešte existuje niekoľko inline tematických šablón, určené pre štruktúrované vloženie jednej konkrétnej informácie (napr. jazykový preklad daného slova).

Kategórie — sú hierarchicky usporiadané menné priestory. Spájajú články rovnakej témy, čo umožňuje tvorbu zoznamov na základe danej témy a ľahšiu navigáciu medzi súvisiacimi článkami. Článok sa označí kategóriou jednoduchým vložením danej kategórie

do článku. Jeden článok môže obsahovať viacero rôznych kategórií súčasne. Príklad kategórií v zdrojovom kóde článku, zachytáva obr. 4.2.

Záverečná sekcia — typicky obsahuje 5 častí: poznámky, referencie, literatúra, súvisiace články, externé odkazy. Má formu bežného textu v tvare nadpis a obsah. Obsah je potom tvorený zväčša šablónami a odkazmi.

```
{{Infobox Politik
| Meno = Tomáš Garrigue Masaryk
| Portrét = Tomáš Garrigue Masaryk, Bain News Service (Library of Congress, Bain Collection)
crop.jpg
| Veľkosť portrétu = 230px
| Podpis = T. G. Masaryk Signature.svg
| Úrad = [[Zoznam prezidentov Česko-Slovenska|prezident Česko-Slovenska]]
| Poradie = 1.
| Koniec obdobia = [[14. december]] [[1935]]
| Začiatok obdobia = [[14. november]] [[1918]]
| Viceprezident =
| Premiér = [[Karel Kramář]] <br /> [[Vlastimil Tusar]] <br /> [[Jan Černý]] <br /> [[Edvard
Beneš]] <br /> [[Antonín Švehla]] <br /> [[František Udržal]] <br /> [[Jan Malypetr]] <br />
[[Milan Hodža]]
| Predchodca =
| Nástupca = [[Milan Hodža]] <small>(úradujúci)</small> <br /> [[Edvard Beneš]]
}}
'''Tomáš Garrigue Masaryk''' (skrátene '''T. G. M.''' alebo '''TGM''', v anglickom zápise aj
''Thomas G. Masaryk''; * [[7. marec]] [[1850]], [[Hodonín]]{{--}}† [[14. september]]
[[1937]], [[Lány (okres Kladno)|Lány]]) bol [[Česko-Slovensko|česko-
slovenský]]<ref>{{Citácia elektronického dokumentu|priezvisko=televize|meno=Česká|-
titul=Masaryk - symbol, vizionář a opravdový Čechoslovák|url=https://ct24.ceskatelevize.cz/-
archiv/1147357-masaryk-symbol-vizionar-a-opravdovy-cechoslovak|dátum prístupu=2019-01-03|-
vydavateľ=ČT 24 - Česká televize|jazyk=cs}}</ref> [[filozof]], [[sociológ]], [[učiteľ|-
pedagóg]], [[politik]], štátnik a novinár, jeden zo zakladateľov ČSR, prvý [[prezident]]
[[Česko-Slovensko|ČSR]] (14. november 1918{{--}}14. december 1935) a predtým predseda
Dočasnej vlády Česko-Slovenska (14. október{{--}}14. november 1918), jeden z troch
vrcholných predstaviteľov [[Česko-slovenská národná rada|Česko-slovenskej národnej rady]]
(13. február 1916{{--}}14. október 1918) a poslanec rakúskej Ríšskej rady (4. marec
1891{{--}}10. október 1893, 17. júl 1911{{--}}25. september 1917).

== Život ==
Narodil sa ako Tomáš Masaryk v Hodoníne na [[Morava (región)|Morave]]. Jeho [[otec]], Jozef
Maszárik, bol pôvodom [[Slováci|Slovák]] z [[Kopčany (obec)|Kopčian]], chudobný kočiš na
statku Nathana Redlicha v Hodoníne. Jeho matka bola Terezie Kropáčková, panská kuchárka z
[[Hustopeče|Hustopečí]]. Masaryk neskôr uviedol, že mu v tomto chudobnom prostredí nikdy
nenapadlo, že raz bude niečím iným ako chudobným [[robotník]]om.
```

Obr. 4.1: Príklad zdrojového kódu článku na Wikipédii v značkovacom jazyku. Prvý odsek tvorí šablóna infobox, ďalší odsek je úvodná sekcia a posledný odsek je jedna kapitola bežného textu. Kód je prebraný z https://sk.wikipedia.org/wiki/Tomáš_Garrigue_Masaryk .

4.2 Prístup cez API

Táto podkapitola čerpá z oficiálnej dokumentácie Wikipédia API [1]. Ďalší z možných spôsobov, ako získať dáta z Wikipédie, je priame dotazovanie sa cez API (application programming interface, slov. rozhranie pre programovanie aplikácií). Wikipédia síce nemá vlastné oficiálne API, no MediaWiki (softvér, na ktorom je postavená Wikipédia) poskytuje REST API s prístupom priamo na Wikipédiu. Týmto spôsobom sa dajú s článkami vykonávať plnohodnotné operácie[1] ako pridávanie, úprava a mazanie článkov. Okrem toho sa takto dajú potrebné dáta z článkov aj sťahovať, či už formou stiahnutia celého zdrojového kódu

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

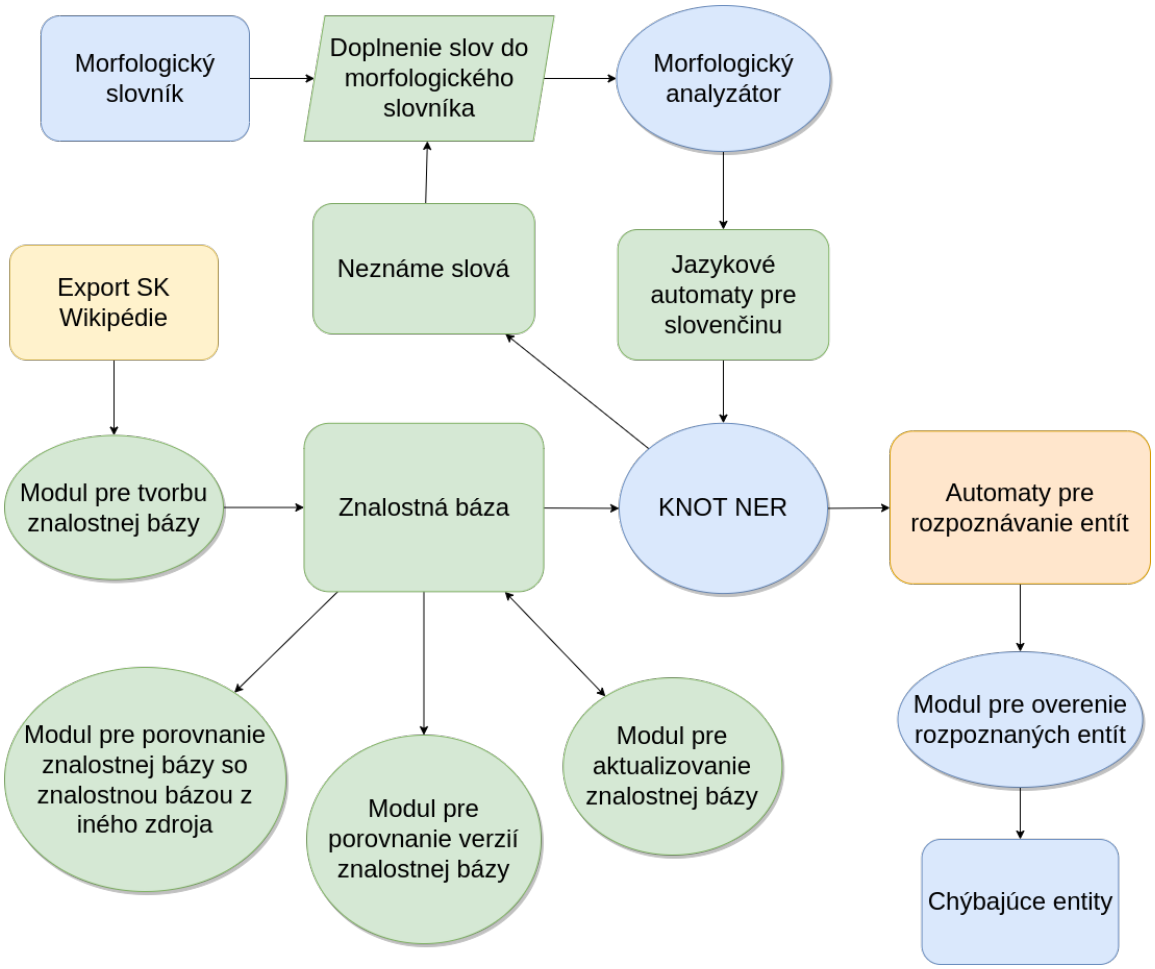
```
== Externé odkazy ==
* [http://old.hrad.cz/president/Masaryk/index.html TGM na stránke Pražského hradu]
* [http://www.utgm.cz/ Ústav T. G. Masaryka]

{{Prezidenti ČSR}}
{{Predsedovia vlád Česko-Slovenska}}
{{Autoritné údaje}}

{{DEFAULTSORT:Masaryk, Tomas Garrigue}}
[[Kategória:Českí filozofi]]
[[Kategória:Českí politici]]
[[Kategória:Prezidenti Česko-Slovenska]]
[[Kategória:Absolventi Karlovej univerzity]]
[[Kategória:Nositelia Radu čestnej légie]]
[[Kategória:Osobnosti z Hodonína]]
[[Kategória:České osobnosti slovenského pôvodu]]
[[Kategória:Osobnosti na českých bankovkách]]
[[Kategória:Osobnosti na česko-slovenských poštových známkach]]
[[Kategória:Osobnosti na českých poštových známkach]]
[[Kategória:Osobnosti na poštových známkach USA]]
[[Kategória:České osobnosti prvej svetovej vojny]]
[[Kategória:Signatári Pittsburskej dohody]]
[[Kategória:Osobnosti maďarsko-česko-slovenskej vojny]]
```

Obr. 4.2: Príklad zdrojového kódu článku na Wikipédii v značkovacom jazyku. Prvý odsek je koniec záverečnej sekcie, ďalší odsek sú navigačné šablóny, posledný odsek sú kategórie. Kód je prebraný z https://sk.wikipedia.org/wiki/Tomáš_Garrigue_Masaryk .

článku, celého už naformátovaného článku 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 znázorňujú funkčné moduly, obdĺžniky znázorňujú dátové vstupy a výstupy, kosodĺžnik znázorňuje proces. Žltou farbou je znázornený počiatočný vstup, zelenou komponenty, ktoré boli vytvorené v rámci tejto práce, modrou externé komponenty využívané touto prácou a hnedou požadovaný výstup.

Zámerom tejto práce bolo vytvoriť základný systém pre rozpoznávanie entít v slovenčine, ktorý by bol vhodným východiskovým systémom pre následné rozšírenie o zjednoznačňova-

nie. K tomu boli využité existujúce riešenia v rámci výskumnej skupiny KNOT@FIT, ktoré boli vhodne doplnené o ďalšie moduly. Schéma takéhoto systému je znázornená na obr. 5.1. Hlavným cieľom tohto systému je priviesť potrebné dáta do modulu pre rozpoznávanie a zjednoznačňovanie entít.

V práci je tento modul tvorený nástrojom [20] (ďalej nazývaný NER@KNOT), ktorý vygeneruje automaty určené k rozpoznávaniu a zjednoznačneniu entít. Ten na vstupe vyžaduje znalostnú bázu (podrobnejšiemu popisu KB sa venuje podkapitola 5.1). Keďže sa práca zaoberá rozpoznávaním entít v slovenčine, ktorej súčasťou je aj ohýbanie slovných druhov, pre lepšie rozpoznávanie entít je nutné pridať do systému gramatiku, ktorá by takéto ohýbanie vygenerovala. K tomu práca využíva nástroj [40] (ďalej 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 slovenčinu síce KNOT@FIT má už existujúci morfologický slovník, no ten obsahuje len obmedzenú slovnú zásobu a značné množstvo slov je doň potrebné doplniť. Slová chýbajúce v slovníku vie potom automaticky rozpoznať nástroj NER@KNOT.

Ako zdrojový korpus pre tvorbu KB (znalostnej bázy) je v práci použitý export slovenskej verzie Wikipédie. Wikipédia bola zvolená preto, aby sa overil jej prínos z hľadiska množstva ponúkaných informácií v porovnaní s inými riešeniami (napr. Wikidáta alebo DBpédia). KB je zostavená samostatným modulom, ktorý je vytvorený pre účely tejto práce, na základe podobného modulu z práce [19] (ďalej nazývané entity_kb_czech3), ktorý je určený na tvorbu KB z českej verzie Wikipédie.

Súčasťou systému sú aj moduly určené na manipuláciu s vytvorenou KB. Jedným z nich je modul na aktualizáciu KB, ktorý vychádza z výsledkov porovnania dvoch časovo rôznych verzií KB. Toto porovnanie umožňuje ďalší samostatný modul. Tretí modul je určený na porovnanie KB vytvorenej v rámci tejto práce s KB vytvorenou v rámci inej práce z iného zdrojového korpusu.

Vyhodnotenie výsledného systému umožňuje nástroj na vyhľadávanie nerozpoznaných entít. Tento nástroj pracuje na princípe porovnávania entít rozpoznaných v článkoch Wikipédie s odkazmi v týchto článkoch. Nástroj bol v čase tvorby tejto práce určený k privátnemu použitiu, preto zoznam nerozpoznaných entít bol sprostredkovaný autorom tohto nástroja.

5.1 Tvorba znalostnej bázy

Znalostná báza použitá v tejto práci obsahuje okrem odkazov do Wikipédie aj základné informácie o daných entitách. Keďže je vytváraná z dát Wikipédie, ktorú tvorí z veľkej väčšiny neštrukturovaný text, najväčšou výzvou pri jej tvorbe je extrakcia informácií. Vytvorenie znalostnej bázy v tejto práci je založené na predchádzajúcej práci entity_kb_czech3, ktorej cieľom bolo vytvoriť KB z českej verzie Wikipédie. Tento proces prebieha v troch fázach.

V prvej fáze sa vytvorí KB z dumpu vo formáte špecifickom pre daný systém. Na vstupe berie súbor so zdrojovými textami všetkých stránok v slovenskej verzii Wikipédie (dump Wikipédie) a súbor, v ktorom je zoznam presmerovávaných stránok na Wikipédii spolu s cieľovými stránkami. Dump je potom iteratívne spracovávaný po jednotlivých článkoch, pričom sa najprv odfiltrujú všetky stránky, ktoré nepojednávajú o entitách. Zostávajúce stránky predstavujú každá jednu entitu. Následne celý systém funguje na princípe mechanického spracovávania textu pomocou extrakcie kľúčových slov a regulárnych výrazov (tento postup je bližšie popísaný v podkapitole 5.1.1). Výsledná KB rozpoznáva nasledujúce triedy entít: osoba, fiktívna osoba, skupina osôb, štát, bývalý štát, sídlo, vodný tok, vodná plo-

cha, reliéf, vodopád, ostrov, polostrov, kontinent. Informácie o entitách sú potom rozdelené do jednotlivých atribútov špecifických pre danú triedu. Pre atribúty spoločné pre všetky triedy pozri tabuľku 5.1. Pre triedne špecifické atribúty pozri tabuľku 5.2 a tabuľku 5.3. Oproti pôvodnému riešeniu entity_kb_czech3 je do KB ku každej entite pridaný atribút Wikidata ID[1], ktorý môže slúžiť ako jednoznačný univerzálny identifikátor entity aj mimo tejto KB.

Tabuľka 5.1: Tabuľka pre systémovo špecifickú KB, znázorňujúca atribúty na daných indexoch, ktoré sú rovnaké pre všetky triedy. Pozn. atribúty s príznakom „{m}“ môžu obsahovať viacero hodnôt.

index	atribút	popis
0	ID	identifikačné číslo entity unikátne v rámci danej KB
1	trieda	trieda entity
2	meno	meno entity nemusí byť unikátne
3	aliasy {m}	ďalšie pomenovania tejto entity
4	presmerovania {m}	stránky na Wikipédii, ktoré sú presmerovávané na túto entitu
5	popis	prvá veta článku na Wikipédii
6	názov na Wikipédii	názov článku 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	jednoznačný identifikátor entity používaný vo Wikidatach (Q-item)

Tabuľka 5.2: Tabuľka pre systémovo špecifickú KB, znázorňujúca rozdielne atribúty tried na daných indexoch, 1. časť. Pozn. atribúty s príznakom „{m}“ môžu obsahovať viacero hodnôt.

index	osoba, fiktívna osoba, skupina osôb	štát, bývalý štát	sídlo	vodný tok	vodná plocha
10	dátum nar.	zem. šírka	štát	kontinent {m}	kontinent {m}
11	miesto nar.	zem. dĺžka	zem. šírka	zem. šírka	zem. šírka
12	dátum úmr.	rozloha	zem. dĺžka	zem. dĺžka	zem. dĺžka
13	miesto úmr.	počet obyv.	počet obyv.	dĺžka	rozloha
14	zamestnanie {m}			rozloha	
15	národnosť {m}			prietok	
16				prameň	

[1]Wikidata ID je v tejto práci pomenovanie pre identifikátor „Q-item“ používaný v projekte Wikidata (https://www.wikidata.org/wiki/Wikidata:Main_Page). Na rozdiel od ostatných potenciálnych identifikátorov, je v čase nemenný.

Tabuľka 5.3: Tabuľka pre systémovo špecifickú KB, znázorňujúca rozdielne atribúty tried na daných indexoch, 2. časť. Pozn. atribúty s príznakom „{m}“ môžu obsahovať viacero hodnôt.

index	reliéf	vodopád	ostrov	polostrov	kontinent
10	kontinent {m}	kontinent {m}	kontinent {m}	zem. šírka	zem. šírka
11	zem. šírka	zem. šírka	zem. šírka	zem. dĺžka	zem. dĺžka
12	zem. dĺžka	zem. dĺžka	zem. dĺžka.		rozloha
13		výška	rozloha		
14			počet obyv.		

V druhej fáze sa z externého zdroja[2] stiahnu štatistiky stránok na Wikipédii (počet spätných odkazov, počet návštev článku, údaj, či má článok pre dané heslo primárny význam) pre každú entitu. Z týchto štatistík sa následne vypočíta ohodnotenie entity, ktoré je možné neskôr použiť pri zjednoznačňovaní. Tieto údaje sa následne doplnia do KB. Modul využitý v tejto fáze je prebraný z práce entity_kb_czech3 [19].

V tretej fáze sa táto verzia KB pretransformuje do univerzálneho formátu používaného v rámci výskumnej skupiny KNOT@FIT. V ňom sú rozpoznávané triedy: osoba, skupina osôb, umelec, geografická entita. Príslušné atribúty sú znázornené v tabuľke 5.4. Modul pre transformáciu je tiež prebraný z práce entity_kb_czech3 [19], pričom je mierne pozmenený.

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

Tabuľka 5.4: Tabuľka pre univerzálnu KB, znázorňujúca triedy a k nim príslušné atribúty. Každá entita obsahuje atribúty zo stĺpca „spoločné“, potom atribúty svojej triedy a na záver atribúty zo stĺpca „metriky“ Pozn. atribúty s príznakom „{m}“ môžu obsahovať viacero hodnôt.

spoločné	osoba	skupina	umelec	geo. entita	metriky
ID	pohlavie	indiv. mená {m}	formy umenia {m}	zem. šírka	spätné odkazy
trieda	dátum nar.	pohlavia {m}	vplyvy {m}	zem. dĺžka	návštevy
meno	miesto nar.	dátum nar. {m}	ULAN ID	typ {m}	primárny význam
jednoznač. meno	dátum úmr.	miesto nar. {m}	ďalšie url {m}	štát	skóre wiki
aliasy {m}	miesto úmr.	dátum úmr. {m}		počet obyv.	skóre metriky
popis	národnosť {m}	miesto úmr. {m}		nadmorská výška	istota
roly {m}		národnosť {m}		rozloha	
fiktívnosť				časová 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 časť zostavenia KB. Tento proces prebieha vo viacerých moduloch prebraných z entity_kb_czech3 [19], no nakoľko musel byť upravený na spracovávanie inej jazykovej verzie Wikipédie, rovnaká zostala len logická štruktúra modulov a princípy získavania informácií. Postupy pre získavanie konkrétnych informácií sú odlišné. Podrobnejší opis celého procesu spracovávania dát z Wikipédie je popísaný vo zvyšku tejto podkapitoly.

Ako už bolo spomínané, prechádzanie jednotlivých článkov je založené na prechádzaní všetkých stránok v slovenskej verzii Wikipédie, uložených v jednom súbore v xml formáte (dump Wikipédie). Zo všetkých stránok sa najprv vyberú len tie, ktoré potenciálne pojednávajú o entitách a to konkrétne tak, že sa odstránia stránky, ktoré v nadpise obsahujú určité textové reťazce. Sú to nasledovné reťazce: „Wikipédia:“, „Redaktor:“, „Súbor:“, „MediaWiki:“, „Šablóna:“, „Pomoc:“, „Kategória:“, „Špeciálne:“, „Portál:“, „Modul:“, „Zoznam“, „Geografia“, „Spoločenstvo“. Zároveň sa preskočia aj presmerovávacie a rozlišovacie stránky. Každá zo zvyšných stránok potom reprezentuje jednu entitu.

Pri spracovávaní stránky sa najprv odstránia nepotrebné časti 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 položky „title“, presmerovania sa zoberú zo súboru s presmerovaniami, kde je kľúčom pri vyhľadávaní práve získaný názov. Z názvu sa potom pridaním prefixu „https://sk.wikipedia.org/wiki/“ a nahradením medzier za „_“ vytvorí url. Následne sa prejde k rozpoznávaniu triedy.

Trieda je určovaná na základe viacerých kľúčových znakov nachádzajúcich sa v texte. Ak k článku nie je priradená žiadna trieda, nepovažuje sa za entitu a preskočí sa. V základe sa osoby, skupiny osôb a fiktívne osoby rozpoznávajú všetky ako osoby a až dodatočne im je pridaný rozlišujúci príznak. Rovnako je to aj pri štátoch a bývalých štátoch. Využívajú sa pri tom 4 hlavné znaky: názov, kategórie, infobox/geobox, prvá veta článku (viac o štruktúre článku na Wikipédii pozri 4). Geobox je podobná šablóna ako infobox s tým, že je všeobecnejšie zameraná, obsahuje omnoho viac položiek a položky sú v anglickom jazyku. Niektoré stránky namiesto infoboxu používajú geobox. Konkrétne znaky rozlišované pri každej triede sú vypísané v tabuľkách v prílohe A.

Po rozpoznaní triedy sa entite pridá **meno** tak, že sa z názvu na Wikipédii odstránia popisné veci ako slová v zátvorkách apod. Ďalším pridaným atribútom je ID, ktoré je vygenerované zahashované poradové číslo entity. Následne sa volaním API Wikipédie získa **Wikidata ID**, **zemepisná dĺžka** a **zemepisná šírka**. Na API sa pošle dotaz s názvom stránky na Wikipédií. Vrátená odpoveď obsahuje viacero informácií v json formáte, takže je následne ďalej spracovaná a požadované informácie sú z nej vyextrahované. Nakoľko pri vytváraní KB dochádza k príliš veľkému počtu dotazov na API (pri naplnení kapacity API vracia chybu), modul implementuje cyklus, ktorý odosiela požiadavky na API, až dokiaľ nepríde kladná odpoveď (maximálne 100 krát, aby sa zamedzilo prípadnému zacykleniu). Aj z tohto dôvodu sa môže ojedinele stať, že sa dané atribúty touto cestou nezískajú.

V ďalšom kroku sa na základe obsahujúcich kategórií zistia príznaky pre oddelenie tried fiktívna osoba, skupina osôb a bývalý štát od tried osoba a štát.

Následne sa prechádza k získavaniu informácií z obsahu článku. To prebieha dvomi spôsobmi — extrakciou z infoboxu/geoboxu a extrakciou z prvej vety. V infoboxe aj geoboxe sú dáta uložené v štruktúrovanej podobe typu: „*položka = údaj*“ a väčšinou býva každá položka na novom riadku. Komplikáciou je, že infoboxov je veľa druhov a každý má odlišné položky alebo odlišne pomenované rovnaké položky. Oproti tomu geobox nemá viac druhov, a teda má vždy rovnaké položky. Táto všeobecnosť však spôsobuje príliš veľké množstvo položiek, ktoré sú si často navzájom podobné a potom je bežné, že rovnaký údaj je vo viacerých geoboxoch zapísaný v inej položke. Prvá veta je neštruktúrovaný text a informácie z nej sú získavané pomocou rôznych heuristík.

Infobox/geobox je spracovávaný po riadkoch a v nich sú regulárnymi výrazmi vyhľadávané kľúčové frázy zodpovedajúce požadovanej položke niektorého z typov infoboxov danej triedy alebo geoboxu. Takto získané údaje však bývajú uvedené v rôznych formátoch, preto je nutné ich následne transformovať do jednotnej podoby. Súčasťou tejto transformácie sú navyše procesy ako prevod formátu dátumov, rozdelenie viac hodnotových reťazcov na jednotlivé hodnoty alebo odstránenie prebytočných znakov. Týmto spôsobom sa dajú získať všetky zvyšné atribúty okrem popisu, ak sú príslušné dáta uvedené. Častým problémom však býva zlé formátovanie celých infoboxov, ktoré sa prejavuje napr. neodriadkovaním každej položky alebo zlým použitím značiek formátovacích značiek.

Prvá veta je z článku extrahovaná pomocou regulárnych výrazov a ďalej 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 očistená od formátova-

cích značiek a šablón. Aliasy sa získavajú pomocou viacerých regulárnych výrazov zameraných ako na celý kontext vety, tak aj na formátovacie značky a interpunkciu vety. Navyše sú využité aj šablóny pre cudzojazyčné preklady — „Lang“ a „V jazyku“. Informácie o narodení a úmrtí sú takisto extrahované z niekoľkých regulárnych výrazov, ktoré odzrkadľujú najčastejšie spôsoby uvádzania týchto informácií v prvej vete článku na Wikipédii. Pohlavie je získavané na základe tvaru slovies „byť“, „stať“, „patriť“ v spojení s koncovkou nasledujúceho prídavného mena.

5.1.2 Aktualizácia znalostnej bázy

Pre nájdenie najefektívnejšieho spôsobu aktualizácie KB, práca najprv porovnáva jednotlivé verzie KB vytvorené s mesačným rozostupom. Pre tieto činnosti bol vytvorený samostatný modul. Pri porovnávaní slúži ako kľúč atribút Wikidata ID (pre už spomínané vlastnosti). Ako sekundárny kľúč pri absencii Wikidata ID sa použije atribút Wikipédia url. Porovnávanie potom rozlišuje novo pridané entity, ako aj počty a typy zmien v atribútoch. Z porovnania verzií marec 2022 a apríl 2022 vyšlo, že nových entít je 148. Zmeny v atribútoch sú podrobnejšie popísané v tabuľke 5.5. Tieto štatistiky dokazujú, prečo nie je vhodné používať žiaden iný atribút ako kľúč k porovnávaniu — menia sa v čase, a tým pádom by sa mohla v KB vyskytnúť jedna entita viackrát, len s iným kľúčom. V predstavených štatistikách nastávajú zmeny aj vo Wikidata ID, to je z dôvodu spätného vytvárania KB zo staršieho dumpu — Wikidata ID sa z povahy API získava v aktuálnom čase na základe dát z dumpu. Ak dump nie je aktuálny (napr. niektoré názvy stránok na Wikipédii sa odvtedy zmenili), potom môže nastať situácia, že dotaz na API obsahuje názov stránky, ktorá je aktuálne už priradená k inej stránke, a teda API vráti iné Wikidata ID. Typický príklad je, že sa stránka na Wikipédii časom zmení na rozlišovaciu (rovnaký názov, iné Wikidata ID) a vytvorí sa nová stránka pre požadovanú entitu (iný názov, rovnaké Wikidata ID). Ak je KB vytváraná v čase aktuálnosti dumpu, tento prípad nenastane.

Pri porovnaní, najviac zmien nastalo v prvej vete článku, čo má následne vplyv aj na ďalšie atribúty z nej získavané. Keďže zmeny obsahovali, okrem pridaných a odstránených atribútov, aj úpravy v už existujúcich atribútoch, nedá sa jednoznačne určiť, že niektorá verzia obsahovala správnejšie atribúty. Aktualizačný modul pri svojej tvorbe však vychádza z predpokladu, že nové zmeny sú pravdepodobne k lepšiemu, preto sa pri aktualizácií berie ako základ novšia verzia. Pri otázke, či je vhodné doplniť atribúty zo staršej verzie, ktoré v novšej chýbajú, sa po bližšom preskúmaní konkrétnych prípadov zistilo, že väčšinu prípadov tvorili nesprávne uvedené informácie, takže ich doplnenie by bolo nežiadúce. Okrem atribútov však môžu byť odstránené aj samotné entity. Preskúmanie tejto možnosti sa docielilo spustením rovnakého skriptu, len s uvedením oboch KB na vstup v opačnom poradí. Tým sa zistilo, že v staršej verzii bolo 29 entít, ktoré sa nenachádzali v novšej.

Aktualizácia teda v konečnom dôsledku pozostáva z pridania týchto entít zo staršej verzie KB do novšej. Okrem samotného pridania im musí byť ešte zmenené ID, aby boli unikátne v rámci celej KB.

Tabuľka 5.5: Zmeny v spoločných 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 počet entít v novšej verzii			85703

5.2 Tvorba morfologického slovníka

Nakoľko je vyhľadávanie entít v tomto systéme založené na rozpoznávaní konkrétnych mien (a alternatívnych mien), slovenčina prináša výzvu v podobe ohýbania slov. Systém výskumnej skupiny KNOT@FIT k tomu využíva morfologické slovníky pretransformované na konečné automaty. Tieto slovníky sú, po vzore ľudského používania slovenčiny, založené na vzoroch (iných ako sú bežne používané ľuďmi), podľa ktorých sa generujú nové tvary slov. V aktuálnych verziách slovníkov sa nachádza určitá sada slov, no keďže novovytvorená KB obsahuje značné množstvo názvov, ktoré sa v týchto slovníkoch ešte nenachádzajú, je potrebné ich tam pridať.

Rozpoznanie chýbajúcich slov z KB v slovníkoch umožňuje nástroj NER@KNOT. Pridávanie slov do slovníkov potom zabezpečujú skripty z nástroja MA@KNOT. Pre pridanie slova je nutné k nemu doplniť vzor. Vzor k danému slovu dokáže odhadnúť skript *lntrf2lpn.py* z MA@KNOT na základe informácií aspoň o jednom ohýbacom páde daného slova (napr. pre podstatné mená to zahŕňa informácie o slovnom druhu, rode, čísle a páde). Podstatné mená, ktorých je v KB drvivá väčšina, sú typicky automaticky v nominatíve, takže k nim stačí určiť rod a číslo. Skript umožňuje aj definovať gramatickú kategóriu nejednoznačne (napr. viac možností rodu pre jedno slovo), no tomu potom zodpovedá aj výsledná kvalita odhadu vzoru, ktorá s rastúcou nejednoznačnosťou stúpa. Preto je cieľom určiť gramatické kategórie čo možno najjednoznačnejšie. Časť z týchto slov takto dokázal určiť nástroj NER@KNOT už pri ich vyhľadávaní (približne tretinu). Zvyšným slovám je potrebné ručne priradiť rod a číslo. Keďže išlo o približne 35000 slov, bolo nutné využiť určité hromadné pravidlá. Niekoľko príkladov takýchto pravidiel použitých pri tejto práci:

- Osoba môže byť len mužského životného alebo ženského rodu.
- Geografická entita môže byť mužského neživotného, ženského alebo stredného rodu.
- Osoba končiaca na „ká“, „ná“, „ová“, alebo ak jej celý názov končí na tieto koncovky, je pravdepodobne ženského rodu.
- Osoba končiaca na „as“, „os“, „es“ je pravdepodobne mužského životného rodu.
- Entita končiaca na „ý“ je pravdepodobne mužského životného alebo neživotného rodu.

- Entita končiaca na „n“ je pravdepodobne mužského životného alebo mužského neživotného rodu.
- Entita končiaca na „o“ je pravdepodobne mužského životného, mužského neživotného alebo stredného rodu.
- Osoby sú pravdepodobne v jednotnom čísle.
- Geografické entity môžu byť v jednotnom aj množnom čísle.

Okrem týchto pravidiel bolo použitých ešte niekoľko ďalších, zameraných na vybrané podskupiny slov na základe ich koncoviek. Je zjavné, že pre každé z týchto pravidiel existuje veľké množstvo výnimiek. Niektoré boli následne ručne prerábané, no pre účely tejto práce postačia aj mierne nedokonalé slovníky. Celkovo je v rámci tejto prác pridaných do slovníkov 54000 slov. Tieto slovníky sa následne systémom MA@KNOT prevedú na konečné automaty, aby ich mohol pri svojej práci využívať systém NER@KNOT.

Kapitola 6

Vyhodnotenie práce

Nasledujúca kapitola obsahuje vyhodnotenie dvoch najdôležitejších 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 úspešnosť extrakcie atribútov, ako aj prínos Wikipédie ako zdrojového korpusu. Rozpoznávanie entít je hodnotené z pohľadu nerozpoznaných entít v odkazoch na Wikipédii, nasledovaných diskusiou o dôvodoch nerozpoznania entít a návrhom zlepšenia.

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 inšpirovaný porovnávacím modulom z 5.1.2. Využíva Wikidata ID ako kľúč pre spárovanie entít, len vo forme Wikidata url, respektíve Wikipedia url ako sekundárny kľúč. Ako už bolo spomínané, hlavným cieľom tohto porovnania je zistiť 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, že je KB niekoľko mesiacov stará, v tomto prípade nie je problém, pretože porovnávanie nebolo zamerané na konkrétne entity, ale na extrakciu atribútov. Rovnako bolo zámerom vykonať porovnanie na kvantitatívnej úrovni, preto pri porovnaní nebola uvažovaná správnosť extrahovaných atribútov.

Počet entít, ktoré obsahovala znalostná báza vytvorená v rámci tejto práce z Wikipédie, je znázornený v tabuľke 6.1, išlo o takmer 86000 entít. Prevzatá KB z Wikidát obsahovala okolo 8 miliónov záznamov. Medzi týmito dvomi KB sa podarilo spárovať entity v 16662 prípadoch. To znamená, že z 8 miliónov záznamov v KB z Wikidát reprezentuje entitu s vlastnou stránkou v slovenskej verzii Wikipédie len 16662 (prípadne mierne viac, ak uvažujeme nedokonalosť vytvorenej KB z Wikipédie). Ako bolo spomínané v kapitole 3.3, vo viacerých prístupoch k zjednoznačňovaniu sa popularita stránky entity na Wikipédii považuje za jednu z metrík určujúcich význam entity. Ak sa k tomu pripojí predpoklad, že popularita entity v slovenskej verzii Wikipédie odzrkadľuje jej význam v slovenčine, potom je možné považovať väčšie množstvo entít s prepojením na slovenskú verziu Wikipédie za značnú výhodu.

Tabuľka 6.1: Počet entít v KB z Wikipédie 04/2022.

celkovo	85703
geografická entita	55110
osoba	30579
skupina	14

Keďže entity obsahujú aj triedne špecifické atribúty, prvou vecou kontrolovanou pri porovnávaní entít bola zhoda tried. Tu sa zistilo, že 10960 zo spárovaných 16662 má v KB z Wikidát triedu „general“ — to znamená, že nebola zaradená do žiadnej z rozpoznávaných tried. Tým pádom tieto general entity neobsahujú žiadne triedne špecifické atribúty. Okrem toho má atribút „trieda“ značnú informatívnu hodnotu a môže byť využitý pri zjednoznačňovaní (napr. pri výbere kandidátnych entít pre priradenie, pozri 3.2). Zaujímavým je, že 6 entít malo rozdielne určené triedy, pričom išlo o prípady, keď sa dali obe priradené triedy považovať za správne. Napríklad entita „Tri rokliny“ bola určená ako geografická entita a ako organizácia. Alebo entita „Druhá česko-slovenská republika“ ako geografická entita a ako udalosť.

Porovnanie jednotlivých atribútov je zachytené v tabuľkách 6.2, 6.3, 6.4. Vyplýva z neho, že KB z Wikipédie dominuje v atribútoch: „popis“, „roly“, „fiktívnosť“, „obrázok“, „geotyp“. Pri atribúte popis má táto KB okrem kvantitatívnych aj kvalitatívne výhody — popisy sú dlhšie a obsahujú viac informácií (prvá veta článku proti 4-5 slovným popisom v druhej KB). „Roly“ v tomto prípade reprezentujú zamestnanie osôb a nachádzajú sa len u menšej časti entít. V druhej KB sa nenachádzajú vôbec. „Fiktívnosť“, „obrázok“ a „geotyp“ sa takisto v KB z Wikidát nenachádzajú vôbec vôbec. Oproti tomu, KB z Wikidát dosahuje pri ostatných triedne špecifických atribútoch rádovo lepšiu úspešnosť.

Špeciálnym atribútom sú „aliasy“, ktoré umožňujú samotné rozpoznávanie a zjednosnačňovanie. Z tabuľky 6.2 vyplýva, že KB z Wikipédie obsahuje približne dvakrát viac entít s aliasmi než KB z Wikidát. Zvyšných 321 entít, ktoré v oboch KB obsahovali aliasy, boli následne porovnávané s cieľom zistiť, v ktorej KB obsahujú jednotlivé entity viac unikátnych aliasov (aliasy nachádzajúce sa len v jednej z porovnávaných KB). Z tohto porovnania vyšlo, že v oboch KB je približne rovnaký počet takýchto entít.

Tabuľka 6.2: Porovnanie vybraných atribútov KB z Wikipédie (04/2022) s KB z Wikidát (10/2021). Stĺpec „len Wikipédia“ udáva počet entít KB z Wikipédie, ktoré obsahujú daný atribút, pričom príslušná entita KB z Wikidát ho neobsahuje. Stĺpec „len Wikidáta“ udáva opak a stĺpec „obe, ale rozdielne“ udáva počet 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
fiktívnosť	4906	0	0
Wikipedia url	41	0	10
obrázok	6943	0	0
počet porovnávaných entít		16662	

Tabuľka 6.3: Porovnanie atribútov entít triedy „osoba“ KB z Wikipédie (04/2022) s KB z Wikidát (10/2021). Stĺpec „len Wikipédia“ udáva počet entít KB z Wikipédie, ktoré obsahujú daný atribút, pričom príslušná entita KB z Wikidát ho neobsahuje. Stĺpec „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
národnosť	31	3792
počet porovnávaných entít		4604

Tabuľka 6.4: Porovnanie atribútov entít triedy „geografická entita“ KB z Wikipédie (04/2022) s KB z Wikidát (10/2021). Stĺpec „len Wikipédia“ udáva počet entít KB z Wikipédie, ktoré obsahujú daný atribút, pričom príslušná entita KB z Wikidát ho neobsahuje. Stĺpec „len Wikidáta“ udáva opak.

atribút	len Wikipédia	len Wikidáta
zemepisná šírka	0	282
zemepisná dĺžka	0	282
geotyp	1092	0
štát	0	602
počet obyvateľov	100	204
nadmorská výška	0	824
rozloha	22	335
časové zóny	0	683
popisný kód	0	382
ID geonázvu	0	938
počet porovnávaných entít		1098

6.2 Vyhodnotenie rozpoznaných entít

Pre vyhodnotenie rozpoznávania bol použitý výstup nástroja poskytnutý členom výskumnej skupiny KNOT@FIT. Tento nástroj prechádza odkazy v článkoch na Wikipédii a dáva ich ako vstup pre rozpoznávacie automaty (automaty vytvorené pomocou NER@KNOT). Následne porovnáva cieľovú 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, než ktorá nezodpovedá danému odkazu), nástroj ju pridá do výstupného súboru. Pre účely vyhodnotenia boli použité dve časovo odlišné 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 poznamenať, že 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 zapríčiňujú zlé rozpoznanie entít, bol prvý z súborov ručne analyzovaný formou náhodného výberu obsahujúcich entít. Týmto spôsobom bolo zistené, že medzi hlavné faktory zlého rozpoznania patria: (1) absencia vhodných aliasov, (2) nesprávne zjednoznačnenie entity, (3) nevygenerovanie správnych tvarov pre aliasy.

V tejto časti bolo cieľom vyhodnotiť 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 účel použité 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, že nastalo približne 33% zlepšenie. Pre rozdielne rozsahy KB sa však nedá jednoznačne určiť miera vplyvu doplnených slov do slovníka na zlepšenie — s vyšším počtom entít v KB sa zvyšuje aj počet porovnávaných odkazov, a tým pádom aj počet možných prípadov zlého rozpoznania entity. Ak by sme uvažovali lineárny vplyv počtu entít na počet zle rozpoz-

naných entít, po prirátaní spomínaných 6 % by sa dalo odhadnúť zlepšenie rozpoznávania, spôsobené aktualizáciou morfologického slovníka, na približne 39 %.

Následne bol aj druhý zo spomínaných súborov ručne analyzovaný s cieľom zistiť konkrétne dôvody zlého rozpoznania entít. Najčastejšie dôvody boli:

- Nedostatočne rozgenerované aliasy — najviac prípadov, kedy zmienka o entite obsahovala len niektoré zo slov viacslovného názvu.
- Zlé priradená entita z KB — často bola rozpoznaná len časť viacslovného názvu a na jej základe sa vytvorilo priradenie k entite, ktorá mala túto časť názvu ako alias.
- Chýbajúci alias — prípady neextrahovaných aliasov (v zdrojovom článku sa nachádzali, ale v KB nie) a úplne chýbajúcich aliasov (chýbali aj v zdrojovom článku).
- Drobné odchýlky v zmienke o entite a aliase — napr. vynechanie dĺžňa, použitý iný typ mäkčeňa apod.
- Chýbajúce skloňovanie aliasov — menej časté prípady, ale stále sa vyskytovali.
- Uvádzané odkazy, ktoré by zrejme ani nemali byť rozpoznané ako entita — napr. názvy jazykov odkazujúce na štáty.
- Chybný odkaz — prípady, keď dané slovo obsahovalo odkaz na zjavne nesprávnu stránku.

Hoci súbor po aktualizácii morfologického slovníka stále obsahuje prípady nerozpoznaných entít z dôvodu nevyskloňovania názvu, výrazne takýchto prípadov ubudlo. Zrejme by pomohlo niektoré slová v slovníku viac upresniť. Najčastejšie vyskytujúci sa dôvod nerozpoznania entity však boli chýbajúce aliasy a vo veľkej časti prípadov išlo o výskyt skrátených tvarov z viacslovných názvov, pričom entity tieto skrátené tvary neobsahovali. V tomto prípade by prinieslo značné zlepšenie komplexnejšie rozgenerovanie viacslovných názvov na skrátené tvary. Často sa vyskytovali aj nesprávne priradené entity k daným zmienkam. To by malo vyriešiť pridanie vhodného modulu pre zjednoznačňovanie.

Kapitola 7

Záver

Zámerom práce bolo vytvorenie prerekvizít potrebných k zjednoznačňovaniu pomenovaných entít v slovenčine. To bolo docielené vytvorením základného systému pre rozpoznávanie entít, zloženého 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 umožňujúce skloňovanie daných názvov. Vytvorená znalostná báza spolu s morfologickými automatmi potom slúži ako vstup pre nástroj na rozpoznávanie pomenovaných entít.

Modul pre zostavenie znalostnej bázy vytvorený v rámci tejto práce iteratívne prechádza jednotlivé články Wikipédie a na princípe vyhľadávania kľúčových slov a regulárnych výrazov z nich extrahuje pomenované entity spolu s príslušnými atribútmi. Doplnkami k tomuto procesu sú moduly určené na aktualizáciu znalostnej bázy a na porovnanie znalostnej bázy s inou znalostnou bázou. Názvy chýbajúce v morfologickom slovníku sú zisťované nástrojom NER od KNOT@FIT a následne sú doň ručne pridávané. Z dôvodu nutnosti určiť gramatické kategórie týchto slov sa nepodarilo tento proces zautomatizovať.

Súčasťou riešenia bolo porovnanie danej znalostnej bázy vytvorenej z Wikipédie so znalostnou bázou vytvorenou z Wikidát, s cieľom zistiť prínosy a nedostatky použitia Wikipédie ako zdrojového korpusu. Z porovnania vyšlo, že znalostná báza z Wikipédie obsahuje podstatne viac entít s väzbou na slovenskú Wikipédiu a viac entít v nej má určenú triedu. Takisto prináša lepšie výsledky pri atribútoch popis, obrázok, fiktívnosť a povolanie osôb, podtyp pre geografické entity a mierne lepšie výsledky pri obsiahnutých alternatívnych pomenovaniach. Naopak, znalostná báza z Wikidát dosahuje lepšie výsledky pri všetkých ostatných triedne špecifických atribútoch. V ďalšej časti bolo na odkazoch v článkoch Wikipédie hodnotené rozpoznávanie entít a vplyv aktualizácie morfologického slovníka. Pozorované zlepšenie bolo približne v rozmedzí 33 až 39 %. Rovnako boli týmto procesom zistené aj hlavné nedostatky v rozpoznaných entitách, medzi ktorými dominovalo nedostatočné pokrytie alternatívnych pomenovaní a nesprávne priraďovanie entít zo znalostnej bázy.

V prípade budúceho rozšírenia o vhodný modul pre zjednoznačňovanie a zlepšenia rozgenerovávania alternatívnych mien, by vytvorený systém mohol dosahovať solídne výsledky v oblasti zjednoznačňovania pomenovaných entít v slovenčine. Práca zároveň predpokladá možné využitie systému v rámci výskumnej skupiny KNOT@FIT ako rozšírenie k existujúcemu riešeniu pre češtinu.

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, č. 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 unified 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, č. 2, s. 155–170.

[13] Gattani, A., Lamba, D. S., Garera, N., Tiwari, M., Chai, X. et al. Entity extraction, linking, classification, and tagging for social media: a wikipedia-based approach. *Proceedings of the VLDB Endowment*. VLDB Endowment. 2013, zv. 6, č. 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, č. 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*. Navštívené 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, č. 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, č. 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 Paass, 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-specific 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] Černý, S. *Morfologický analyzátor pomocí konečných automatů.* Brno, CZ, 2008. Bakalářská práce. Vysoké učení technické v Brně, Fakulta informačních 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 Processing.* 1997, 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 affix 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, č. 8, s. 279–288.

[45] Zhang, W., Sim, Y.-C., Su, J. a Tan, C.-L. Entity linking with effective acronym expansion, instance selection and topic modeling. In: *Twenty-Second International Joint Conference on Artificial 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

Tabuľka A.1: Osoba — pravidlá pre rozpoznanie.

infobox	kategórie	prvá veta
Zoznam typov infoboxov osôb načítaný z vopred vytvoreného súboru.	Typické ľudské kategórie ako: narodenia, úmrtia, postavy, osobnosti, ľudia, bohovia apod.	Výrazy týkajúce sa ľudí napr. značky pre narodenie alebo úmrtie, slovesá byť, patriť, stať sa apod.

Tabuľka A.2: Štát — pravidlá pre rozpoznanie.

kategórie
„Štáty v [*názov kontinentu*]“, „Členovia [*názov medzinárodnej organizácie*]“, „Neuznané štáty“, „Zaniknuté štáty“.

Tabuľka A.3: Sídlo — pravidlá pre rozpoznanie.

názov	infobox/geobox	kategórie
Základ slova definujúceho 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“

Tabuľka A.4: Vodný tok — pravidlá pre rozpoznanie.

názov	infobox/geobox	kategórie
Obsahuje slová: „bystrina“, „potok“, „riečka“, „rieka“, „veľtok“.	Obsahuje slová: „potok“, „river“, „stream“, „creek“.	„potoky/rieky v/na“,

Tabuľka A.5: Vodná plocha — pravidlá pre rozpoznanie.

názov	infobox/geobox	kategórie
Obsahuje slová: „rybník“, „jazero“, „more“, „oceán“, „tôňa“, „vodné dielo“.	Obsahuje slová: „vodná plocha“, „more“, „oceány“, „lake“, „sea“, „reservoir“, „water“, „ocean“.	„Rybníky/Jazerá/Riečne jazerá/Vodné diela/Zálivy-/Rieky/Potoky v/na“

Tabuľka 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“, „Kaňony“, „Doliny“, „Nížiny“, „Priepasti“, „Sopky“, „Púšte“.

Tabuľka 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

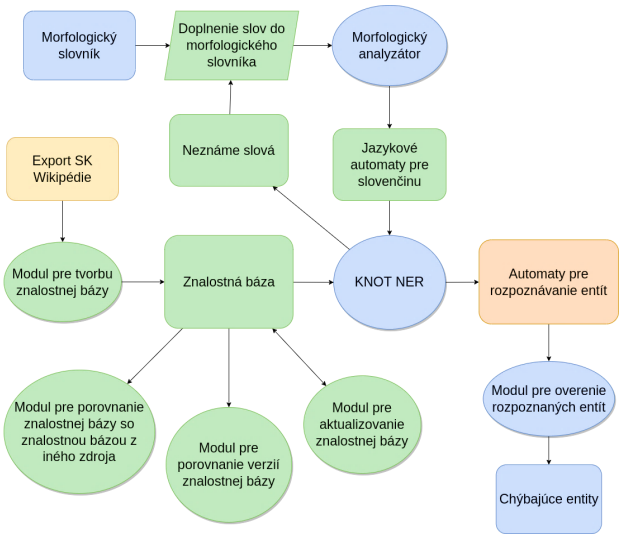
Plagát

Zjednoznačňovanie pomenovaných entít v slovenčine

Samuel Križan
Vysoké učení technické v Brně, 2022
bakalárska práca, vedúci: doc. RNDr. Pavel Smrž, Ph.D.

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.



Obr. 1: Diagram vytvoreného systému; žltá – počiatočný vstup, zelená – komponenty vytvorené v rámci práce, modrá – využívané externé komponenty, hnedá – požadovaný výstup.

Metodológia

Modul pre zostavenie znalostnej bázy iteratívne prechádza jednotlivé články Wikipédie a na princípe vyhľadávania kľúčových slov a regulárnych výrazov z nich extrahuje pomenované entity spolu s príslušnými atribútmi. Doplnkami k tomuto procesu sú moduly určené na aktualizáciu znalostnej bázy a na jej porovnanie s inou znalostnou bázou. Názvy chýbajúce v morfologickom slovníku sú zisťované nástrojom NER od KNOT@FIT a následne sú doň ručne pridávané na základe heuristík pre určenie gramatických kategórií.

Výsledky

Tab. 1: Počet entít v znalostnej báze (04/2022)

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

Tab. 2: Štatistiky aktualizácie morfologického slovníka

počet slov pridaných do slovníka	54 000
zlepšenie rozpoznávania entít	33-39 %

Záver

Z porovnania so znalostnou bázou z Wikidát vyšlo, že znalostná báza z Wikipédie obsahuje podstatne viac entít s väzbou na slovenskú Wikipédiu a viac entít v nej má určenú triedu. Takisto prináša lepšie výsledky pri atribútoch popis, obrázok, fiktívnosť a povolanie osôb, podtyp pre geografické entity a mierne lepšie výsledky pri obsiahnutých alternatívnych pomenovaniach. Naopak, znalostná báza z Wikidát dosahuje lepšie výsledky pri všetkých ostatných triedne špecifických atribútoch.
Z hľadiska rozpoznávania entít boli ako hlavné nedostatky systému zistené nedostačujúce pokrytie alternatívnych pomenovaní a nesprávne priraďovanie entít zo znalostnej bázy. V prípade ďalšieho pokračovania by preto bolo vhodné zlepšiť rozgenerovávanie alternatívnych pomenovaní a pridať vhodný modul pre zjednoznačňovanie.

Poďakovanie

Poďakovanie patrí 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.

VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ FAKULTA INFORMAČNÍCH TECHNOLOGIÍ