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

EXTRAKCE INFORMACÍ Z WIKIPEDIE

INFORMATION EXTRACTION FROM WIKIPEDIA

BAKALÁŘSKÁ PRÁCE
BACHELOR'S THESIS

AUTOR PRÁCE — **RUDOLF JURIŠICA**
AUTHOR

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

BRNO 2024

VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
FAKULTA INFORMAČNÍCH TECHNOLOGIÍ

Zadání bakalářské práce

154486

Ústav: Ústav počítačové grafiky a multimédií (UPGM)
Student: **Jurišica Rudolf**
Program: Informační technologie
Název: **Extrakce informací z Wikipedie**
Kategorie: Informační systémy
Akademický rok: 2023/24

Zadání:

1. Seznamte se s metodami extrakce informací z textu na základě strojového učení.
2. Navrhněte a implementujte systém pro automatickou extrakci typů a základních atributů pojmenovaných entit z exportu dat anglické, české a slovenské Wikipedie.
3. Vytvořte systém pro pravidelné aktualizace znalostní báze, který označí typy změn, k nimž došlo.
4. Vyhodnoťte výsledky systému na reprezentativním vzorku dat.
5. Vytvořte stručný plakát prezentující práci, její cíle a výsledky.

Literatura:

- Raaijmakers, Stephan. *Deep Learning for Natural Language Processing*. Simon and Schuster, 2022, ISBN 978-1617295447.

Při obhajobě semestrální části projektu 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, prof. Dr. Ing.
Datum zadání: 1.11.2023
Termín pro odevzdání: 9.5.2024
Datum schválení: 21.12.2023

Abstrakt

Cílem práce je snížit počet neznámých odkazovaných entit ve článcích české Wikipedie. Dosáhnuto toho bylo jednak za využití pomocných, již existujících řešení, tvořených výzkumnou skupinou KNOT na VUT FIT, a dále pak vytvořením sady programů. Tyto programy se automaticky spouští každý měsíc při vydání nové verze Wikipedie. Automaticky doplní znalostní bázi o nová jména, vygeneruje jejich odvozené tvary, a upraví samotné články přímo na Wikipedii.

Abstract

The goal of this thesis is to reduce the number of unknown referenced entities in Czech Wikipedia articles. This has been achieved by using some existing solutions, created by the KNOT research group at FIT BUT, and then by creating a set of programs. These programs are automatically run every month, when a new version of Wikipedia is released. They will automatically add new names to the knowledge base, generate their derived forms, and edit the articles themselves directly on Wikipedia.

Klíčová slova

Wikipedie, extrakce informací, morfologie, analýza, Pywikibot, znalostní báze, přirozený jazyk, Wikidata, strojové učení, cizí jména, přídavná jména

Keywords

Wikipedia, information extraction, morphology, analysis, Pywikibot, knowledge base, natural language, Wikidata, machine learning, foreign names, adjectives

Citace

JURIŠICA, Rudolf. *Extrakce informací z Wikipedie*. Brno, 2024. Bakalářská práce. Vysoké učení technické v Brně, Fakulta informačních technologií. Vedoucí práce doc. RNDr. Pavel Smrž, Ph.D.

Extrakce informací z Wikipedie

Prohlášení

Prohlašuji, že jsem tuto bakalářskou práci vypracoval samostatně pod vedením pana doc. Smrže a uvedl všechny literární prameny a publikace, ze kterých jsem čerpal.

.....
Rudolf Jurišica
8. května 2024

Poděkování

Děkuji svému vedoucímu, panu doc. RNDr. Pavlu Smržovi, Ph.D., za odborné vedení práce a rychlé reakce na dotazy a dále pánům inženýrům Janu Doležalovi a Tomáši Volfovi.

Obsah

Seznam obrázků

Kapitola 1

Úvod

Schopnost strojů porozumět lidské řeči je jedním z klíčových pilířů v oblasti informačních technologií. Její přínos spočívá především v nabídnutí nového způsobu, jakým komunikujeme s našimi elektronickými zařízeními. Přitom nezáleží, zda jde o písemnou, nebo mluvenou komunikaci.

Důležitým aspektem je rozpoznávání pojmenovaných entit v textu, neboli Named Entity Recognition (NER). Tato dovednost umožňuje strojům identifikovat specifické entity v textu, jako jsou třeba jména osob, lokací či organizací. Jeho neustálým vývojem se zabývají odvětví strojového učení a umělé inteligence. Ke zlepšování schopností strojů rozpoznávat pojmenované entity se využívá trénování na rozsáhlé sadě dat. Zapotřebí je také znalostní báze obsahující pojmenované entity, díky čemuž je možné je v textu rozpoznat. Detailnější popis bude uveden v kapitole 2.3.4.

Cílem této práce bylo vylepšit stávající znalostní bázi a schopnost programů rozpoznávat pojmenované entity v textu. K tomuto účelu byla využita data z největší internetové encyklopedie – Wikipedie. Nabízí veřejně dostupná data, spravovaná dobrovolníky z celého světa, což z ní dělá ideální zdroj pro získání rozsáhlého množství informací. Pracovalo se s tzv. dump soubory – soubory obsahující všechna data z Wikipedie (články, rozcestníky, infoboxy) ve formátu XML pro každou jazykovou variantu.

V textech, které má NER zpracovat, se mohou objevovat jména jak v základním tvaru (jako třeba „Itálie“), tak i jejich odvozené tvary (např. „italský“ – odkazuje na „Itálie“) či alternativní jména, jako třeba „Ivan IV., řečený Hrozný“. V případě „Hrozný“ se nejedná o příjmení, nýbrž o přídomek. Takovéto případy se pro NER jeví jako nejednoznačné, a nedokáže si s nimi poradit. Zlepšení znalostí (zjednoznačení) přináší speciální stránky „rozcestníky“ na Wikipedii, které agregují články se stejným názvem, a přidávají stručný popis ohledně odkazované entity. Díky rozcestníkům a souvisejícímu textu se dá lépe určit, o jakou entitu se jedná.

Extrakcí informací z Wikipedie se zabývalo již několik prací. Některé se týkaly tvorby znalostní báze z dump souborů Wikipedie ([28]), jiné třeba určováním typů entit ([21]). Tato práce je primárně zaměřená na morfologii. Pracovalo se s daty z české Wikipedie.

Nejprve bude v kapitole 2 představena daná problematika – Wikipedie, Wikidata a extrakce informací včetně způsobu měření úspěchu extrakce. Dále budou v kapitole 3 představeny programy řešící daný problém. Mezi nimi je např. „spaCy“ (3.2), což je knihovna napsaná v jazyce Python pro pokročilé zpracování přirozeného jazyka, včetně rozpoznávání pojmenovaných entit. Jeho nevýhodou je chybějící podpora pro češtinu. Další z představených řešení jsou programy vyvíjené na VUT FIT skupinou KNOT (3.4), se kterými navržené řešení pracuje, a vylepšuje jejich dovednosti (např. pro NER 3.4.4).

V kapitole 4 bude představen návrh jednotlivých modulů řešících daný problém, včetně využití již existujících řešení. V kapitole 4.5 bude popsáno zakomponování jednotlivých navržených modulů do systému, který je spouštěn každý měsíc při vydání nového dump souboru české Wikipedie. V kapitole 5 bude popsána samotná implementace daných modulů.

V poslední kapitole 6 budou prezentovány výsledky dosažené převážně za poslední půlrok – zpracované rozcestníky, získaná jména a k nim vytvořené odvozené tvary, a oprava hodnot na Wikipedii (míst a dat narození a úmrtí osob).

Kapitola 2

Rozbor řešené problematiky

V této kapitole bude rozebrána problematika týkající se extrakce informací z Wikipedie. Nejprve budou popsány a vysvětleny portály a nástroje týkající se Wikimedia – Wikidata, Wikipedia a její specifické prostředky, které byly v práci také zpracovávány – rozcestníky a infoboxy.

Dále bude rozebráno samotné zpracování přirozeného jazyka a jeho počást – extrakce informací. Popsány budou nejčastěji používané techniky pro extrakci informací. Pokračovat bude představení současných jazykových modelů zpracovávajících přirozený jazyk, včetně principu jejich trénování. Nakonec budou ukázány metriky pro měření úspěchu extrakce informací.

Kvalitně rozebrané téma „zpracování přirozeného jazyka“, včetně představení některých z jazykových modelů, a reálných implementačních ukázek, lze najít v doporučené literatuře ([27]).

2.1 Wikimedia

Nadace **Wikimedia** (oficiálně Wikimedia Foundation, Inc.) je nadace spravující wiki[1] projekty – Wikipedie, Wikislovník, Wikizdroje, Wikimedia Commons, Wikicesty, Wikidata, MediaWiki a další [37]. Cílem je podporovat otevřené wiki projekty a zajistit, že veškerý jejich obsah zůstane pro uživatele a čtenáře zdarma [37].

Jedním z projektů je MediaWiki – software, na kterém běží Wikipedie a další projekty Wikimedia[2]. Využívají jej desetitisíce webů a tisíce společností a organizací [13]. MediaWiki je svobodný, dostupný a bezplatný nástroj. Taktéž provozuje řadu API[3], které jsou využívány jak pro manuální práci s daty (úprava stránek, stahování obsahu stránek, přihlášení, vyhledávání, ...), tak pro automatizované programy či frameworky. Jedním z nich je Pywikibot (3.1), který je využit i v této práci.

2.1.1 Wikipedia

Wikipedia je mnohojazyčná online encyklopedie, na jejíž tvorbě pracují dobrovolní přispěvatelé z celého světa. Je vlastněna a provozována neziskovou organizací Wikimedia 2.1. Její

[1] Wiki – označení webů, které umožňují uživatelům přidávat a měnit stávající obsah [40].
[2] https://cs.wikipedia.org/wiki/MediaWiki
[3] https://www.mediawiki.org/wiki/API:Main_page

první spuštění se datuje na 15. ledna 2001. Zakladateli jsou Jimmy Wales[4] a Larry Sanger[5]. Zpočátku fungovala pouze anglická jazyková verze, brzy ale vznikly i další - česká verze byla spuštěna 3. května 2002. [41] Wikipedie obsahovala v roce 2021 více než 55 milionu článků, což ji činí největší internetovou encyklopedií[6].

Jedním z hlavních pilířů Wikipedie je tzv. crowdsourcing – získání služeb, pomoci a nápadů od velké skupiny lidí. Každý uživatel má možnost přispívat k obsahu Wikipedie tím, že edituje články, přidává nové informace nebo opravuje chyby.

Wikipedia je často terčem kritiky. Nejčastěji uváděným důvodem je, že žádná vydavatelská autorita neručí za správnost obsahu [44]. Další udávaný důvod je množství subjektivně orientovaných či velmi kontroverzních článků, jako např. George W. Bush, holokaust nebo anarchismus [4].

2.1.2 Wikidata

Projekt **Wikidata** úzce souvisí s Wikipedií. Jedná se o svobodnou, mnohojazyčnou, druhotnou databázi, která shromažďuje strukturovaná data jako podporu Wikipedie, Wikimedia Commons a dalších projektů hnutí Wikimedia a pro kohokoliv na světě [29].

Hlavní přednosti Wikidat (informace převzaty z Wikidata [29]):

- Svobodný projekt – data jsou zveřejněna pod licencí Creative Commons Public Domain Dedication 1.0[7], což umožňuje jejich využití a úpravu pro kohokoliv na světě. Data vkládají jak editoři Wikidat, tak automatizovaní boti[8].
- Druhotná databáze – Wikidata uchovávají údaje včetně jejich zdrojů a propojení s ostatními databázemi, což přináší zvýšenou rozmanitost znalostí a umožňuje jejich ověřitelnost.
- Strukturovaná data – data jsou ukládána v běžných souborových formátech (JSON, XML), což umožňuje snadnou práci s daty jak uživatelům, tak automatizovaným botům.
- Využití dat – Wikidata jsou využívána jak v celém spektru projektů Wikimedia, tak i projektech nesouvisejících s Wikimedia – např. Wikiskripta[9].

[4] https://cs.wikipedia.org/wiki/Jimmy_Wales
[5] https://cs.wikipedia.org/wiki/Larry_Sanger
[6] https://www.novinky.cz/clanek/internet-a-pc-wikipedia-se-za-dve-dekady-stala-jednou-z-nejpopularnejsich-webovych-stranek-40347686
[7] https://creativecommons.org/publicdomain/zero/1.0/
[8] https://www.wikidata.org/wiki/Wikidata:Bots
[9] https://www.wikiskripta.eu/w/Home

Ukázka obecné struktury Wikidat ve formátu JSON (používající specifikaci dle protokolu RFC-7159[10]):

```
{
  "id": "Q60",
  "type": "item",
  "labels": {},
  "descriptions": {},
  "aliases": {},
  "claims": {},
  "sitelinks": {},
  "lastrevid": 195301613,
  "modified": "2020-02-10T12:42:02Z"
}
```

2.1.3 Vztah Wikipedie a Wikidat

Wikipedie využívá řadu nástrojů Wikidat, čímž zajišťuje svoji důvěryhodnost a propojení napříč různými jazykovými verzemi Wikipedie. Příklad některých z nich:

- Interwiki odkazy – odkazy na stránky Wikipedie či jiných projektů Wikimedia napříč různými jazykovými verzemi Wikipedie. Příklad odkazu na hlavní stránku Wikipedie realizovaného pomocí Interwiki: `[[Wikipedia:Main Page]]` [14].
- Odkazy na autority – Wikidata obsahuje informace o autoritách, jako např. identifikátory osobností, knih či uměleckých děl.
- Infobox – informační tabulky na začátku článku Wikipedie, které shrnují základní údaje o tématu článku. 2.1.4

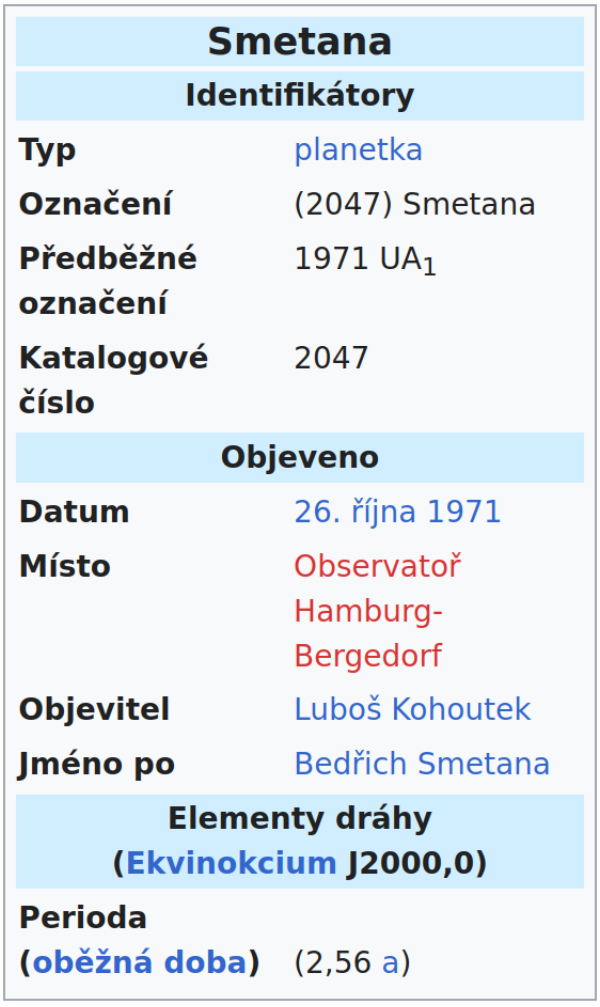
2.1.4 Infobox

Jedná se o informační tabulky na začátku článku Wikipedie, shrnující základní údaje o tématu článku. [30] Mohou se vztahovat k mnoha druhům entit: osoba, film, geografické entity, organizace, stavby, akce, vědy či další.

Obsah infoboxu se vyplňuje dvěma způsoby: lokálně editací šablony přímo ve článku, nebo automatickým přebíráním dat z odpovídající položky ve Wikidatech. [30] Vzhledem k tomu, že Wikidata obsahují informace sdílené napříč různými jazykovými verzemi, data v infoboxu jsou stejná u článků daného tématu ve všech jazykových verzích.

V některých případech může dojít k nesouladu informací v samotném článku a v infoboxu - např. první věta článku o nějaké osobě obsahuje odlišné místo narození, než jaké je uvedeno v infoboxu. V tom případě je pravděpodobnější, že správná hodnota je v infoboxu (tedy ve Wikidatech), z důvodu jejich větší ověřenosti správnosti dat (více viz 2.1.2).

[10] https://datatracker.ietf.org/doc/html/rfc7159



Smetana	
Identifikátory	
Typ	planetka
Označení	(2047) Smetana
Předběžné označení	1971 UA_1
Katalogové číslo	2047
Objeveno	
Datum	26. října 1971
Místo	Observatoř Hamburg-Bergedorf
Objevitel	Luboš Kohoutek
Jméno po	Bedřich Smetana
Elementy dráhy (Ekvinokcium J2000,0)	
Perioda (oběžná doba)	(2,56 a)

Obrázek 2.1: Infobox u článku Smetana (planetka)

Textová verze infoboxu (ve zdroji článku):

```
{{Infobox - planeta
 | typ = [[planetka]]
 | typ barva = planetka
 | označení = (2047) Smetana
 | předběžné označení = 1971 UA<sub>1</sub>
 | číslo = 2047
 | název = Smetana
 | kdy = [[26. říjen|26. října]] [[1971]]
 | kde = [[Observatoř Hamburg-Bergedorf]]
 | kým = [[Luboš Kohoutek]]
 | pojmenováno po = [[Bedřich Smetana]]
 | oběžná doba a = 2,56
}}
```

2.1.5 Rozcestníky

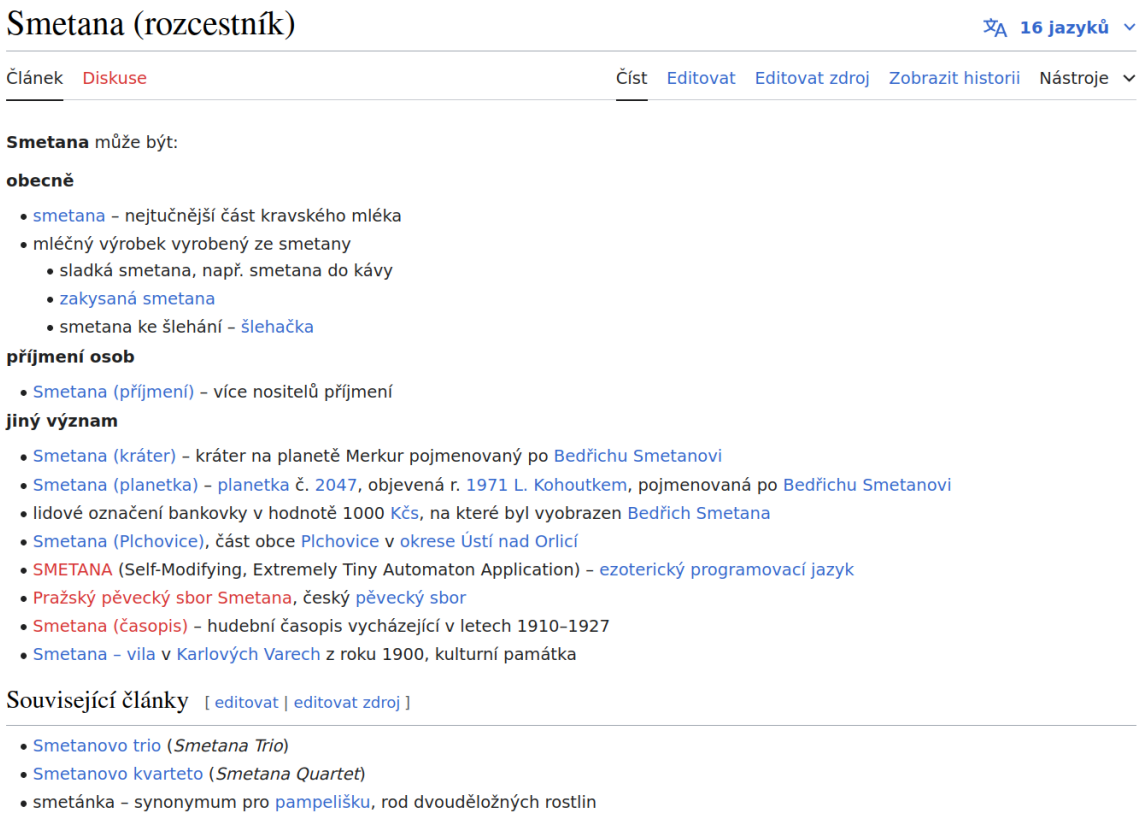
Speciálním typem stránky na Wikipedii jsou tzv. rozcestníky. Jedná se o stránky sloužící pro navigaci na všechny stránky z různých oborů, které mají stejný název. Například název liška označuje jak šelmu, tak houbu. [34]

Rozcestníky mají v názvu ve většině případů rozlišovač „rozcestník“, pomocí něhož lze odlišit stránku se článkem od rozcestníku (např. Smetana (rozcestník), viz obrázek 2.2). Rozcestníky obsahují jednotlivé odkazy na stránky včetně jejich specifikace. Seřazeny jsou většinou od nejtypičtějšího významu (např. ucho, jakožto lidský orgán, bude na prvním místě). Dále mohou obsahovat stručný popis dané entity, o co se vůbec jedná, případně její další názvy (lidové, latinské atd.). Jednotlivé položky rozcestníku mohou být, pro přehlednost, rozděleny do skupin podle významu (např. obecný význam, jména osob, názvy ulic, organizace a jiné).

Na rozcestníkových stránkách se také mohou objevovat odkazy na neexistující stránky (názvy jsou podbarveny červeně). Ty slouží pouze jako další výčet možných významů daného slova (jména), spolu se stručným popisem.

Při úpravě rozcestníků, nebo v souboru formátu XML obsahujícím extrahovanou Wikipedii, lze rozeznat rozcestníky od ostatních stránek pomocí identifikátoru uvedeného na konci stránky – `{{Rozcestník}}` pro českou Wikipedii, v případě slovenské pomocí `{{Rozlišovacia stránka}}` a v anglické `{{disambiguation}}`.

Ukázka rozcestníku „Smetana (rozcestník)“:



Obrázek 2.2: Rozcestník stránky Smetana

2.2 Zpracování přirozeného jazyka

Zpracování přirozeného jazyka (anglicky NLP - Natural language processing) je soubor technik na pomezí informatiky, matematiky, lingvistiky a strojového učení. Cílem je naučit počítače a jiné stroje porozumět lidské řeči ať už v psané nebo mluvené formě a dokázat z ní vstřebat informace, případně je naučit generovat vlastní text a řeč. Avšak při procesu zpracování se vyskytuje řada problémů:

- Lidská řeč není jednoznačná, tudíž počítačové výsledky nemusí vždy správně porozumět tomu, co autor daným tvrzením myslel.
- Pro pochopení některých informací je zapotřebí určitého kontextu, čímž se zvyšuje náročnost na pochopení textu/řeči.
- Mnoho jazyků má vlastní nářečí, slang, argot, což mohou být velmi obtížné překážky pro vytvoření trénovací sady, např. z důvodu lokálního použití daných útvarů (používá je jen pár desítek lidí na světě).

2.3 Extrakce informací

Extrakce informací je podmnožinou zpracování přirozeného jazyka. Soustředí se na extrakci informací z nestrukturovaných textových dat. Hlavním úkolem je identifikovat specifické informace v textu a extrahovat je do strukturovaného formátu, který je snadno strojově zpracovatelný.
Techniky a metody používané při extrakci informací: [25]

- **Tokenizace** – rozdělení textu na tokeny (slova, interpunkční znaménka)
- **Označení slovních druhů** (POS tagging)[11] – přiřazení slovního druhu ke každému slovu v textu
- **Rozpoznávání pojmenovaných entit** (NER)[12] – rozpoznávání a označení pojmenovaných entit v textu, především rozpoznání osob, organizací, geografických lokací a dat
- **Spojování pojmenovaných entit** (EL)[13] – propojení pojmenovaných entit v textu se záznamy ve znalostní bázi nebo databázi, především pro rozpoznání stejně pojmenovaných entit (např. kanoista Václav Havel a prezident Václav Havel)
- **Lemmatizace** – přiřazování základních tvarů slov - např. jednotná čísla k množným číslům (lidé → člověk), infinitiv k ostatním časům (hraje → hrát) a další
- **Klasifikace textu** – rozdělení (přiřazení) textu nebo části textu do kategorií podle jeho obsahu
- **Segmentace vět** (SBD)[14] – označení hranic vět, tedy začátku a konce vět v textu
- **Podobnost textů** – porovnání slov, textu a části textu na jejich podobnost

[11] POS – part-of-speech
[12] NER – named entity recognition
[13] EL – entity linking
[14] SBD – sentence boundary detection

- **Trénování** – zvětšování a aktualizování znalostí modelu, cílem je zlepšit schopnosti modelu s rozpoznáváním textu v určitých úlohách a předpovědích

2.3.1 Tokenizace

Tokenizace je proces, při kterém se vstupní text rozdělí na jednotlivé prvky, tzv. tokeny. Mohou to být jednotlivá slova, fráze, interpunkce, číslice nebo jiné významové jednotky. Tokenizace je jeden z prvotních procesů při extrakci informací a zpracování textu.

Text může být tokenizován několika způsoby (pro příklad uvažujme větu „Extrakce informací je zábava“) [9]:

- **Jednoslovné tokeny** (unigram token) – tokeny jsou rozděleny na jednotlivá slova, bez ohledu na okolní kontext. Jedná se o nejjednodušší formu tokenizace: „Extrakce“, „informací“, „je“, „zábava“
- **Víceslovné tokeny** (n-gram tokens) – rozdělení textu na tokeny sestávající z více slov. Tokenizace již bere v úvahu okolní kontext. Příklad tokenů:
  - Dvouslovné tokeny (bigram tokens) – „Extrakce je“, „je zábava“, „extrakce informací“
  - Trojslovné tokeny (trigram tokens) – „Extrakce informací je“, „extrakce je zábava“

2.3.2 Označení slovních druhů

Přiřazení slovních druhů (POS tagging) jednotlivým tokenům je velmi důležité pro pochopení kontektu a pro další manipulaci s daty. Nezpracovaný vstupní text se označuje jako „nestrukturovaná data“. Data, která jsou tokenizována a jsou jim přiřazeny slovní druhy, se již označují jako „strukturovaná data“. Díky POS tagging lze používat další techniky a metody pro zpracování textu.

V procesu přiřazování se používá řada algoritmů, včetně metod založených na pravidlech, statistických metod a technik strojového učení. Metody založené na pravidlech zkoumají text pomocí lingvistických pravidel vycházejících z gramatiky zpracovávaného jazyka. Statistické metody používají korpus[15] předem označených dat k identifikaci vzorů mezy slovy a značkami. Techniky strojového učení se používají k trénování systému tak, aby rozpoznával vzory v neoznačeném korpusu a následně ke každému slovu přiřadil správnou značku. [6]

2.3.3 Jazykové modely

Pojmem jazykový model označujeme v NLP strukturu nebo algoritmus, který je schopen analyzovat, porozumět nebo generovat lidský jazyk. Modely se učí pomocí trénování 2.3.4 na vlastnostech a charakteristikách daného jazyka, což má za následek přesnější predikce při generování textu a jeho porozumění.

[15] Korpus – sbírka textových dat používaná pro trénování a testování algoritmů a modelů strojového učení (např. novinové články, knihy, vědecké články . . .)

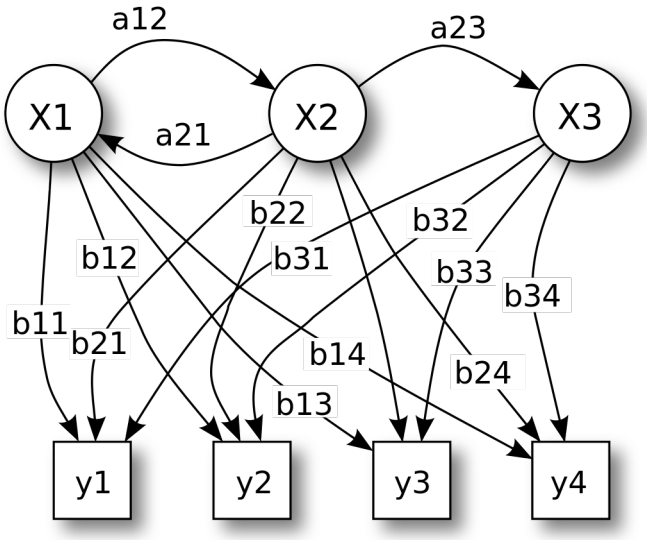
Modely lze dělit na několik typů:

- **Statistické modely** – zahrnují vývoj pravděpodobnostních modelů, které jsou schopny předpovědět slovo vzhledem k sekvenci slov, která mu předcházejí. Používají se např. pro sumarizaci textu, automatické návrhy při psaní zpráv či rozpoznání rukopisu (i když je hůře čitelný). [19]

  Příklady statistických modelů jsou Skrytý Markovův model (viz Obrázek 2.3) nebo n-gram (viz Tokenizace).

- **Pravidlové modely** (Rule-based models) – používají předdefinovaná pravidla nebo heuristiky, které jsou navrženy specialisty - např. lingvisty či odborníky na jazyk. Takováto sada pravidel může být buď vyjádřena pomocí Markovových řetězců nebo diferenciálními rovnicemi. [39]

- **Neuronové modely** – založeny na neuronových sítích. Překonávají nedostatky klasických modelů (např. n-gram), a používají se pro komplexní úlohy, jako je rozpoznávání řeči nebo strojový překlad. [22]



Obrázek 2.3: Pravděpodobnostní parametry skrytého Markovova modelu (příklad) [26]
X — stavy modelu
y — možná pozorování modelu
a — pravděpodobnost přechodu mezi stavy
b — pravděpodobnosti výstupů

Srovnání moderních jazykových modelů

V současné době (2024) existuje a používá se již řada jazykových modelů. Avšak jen některé z nich si získaly svoji popularitu, především díky přesnosti a neustálého zdokonalování. Níže jsou uvedeny a popsány některé z nich:

- **BERT** (Bidirectional Encoder Representations from Transformers) – vyvinutý společností Google v roce 2018 a doposud jeden z nejúspěšnějších modelů. Je založený na transformerech[16]. Jeho schopností, která jej odlišuje od ostatních modelů, je pracovat s kontextem zleva i zprava, tedy v celé větě (bidirectionality). Tato vlastnost je díky jeho trénování za pomoci dvou přístupů: maskování (MLM – masked language modeling) a predikce další věty (NSP – next sentence prediction). [8]

  Z modelu BERT vychází řada dalších modelů, které mohou jeho schopnosti ještě zlepšovat (RoBERTa, DeBERTa), a nebo se primárně zaměřují na výkonnost (ALBERT, DistilBERT, MobileBERT). [16]

- **GPT** (Generative Pre-trained Transformer) – série modelů od společnosti OpenAI. Tyto modely se staly známé především díky chatbotu ChatGPT, který byl bezplatně spuštěn firmou OpenAI na konci roku 2022. [15] První veřejně dostupná verze byla 3. Za ní následovala verze 3.5, která je v bezplatné verzi ChatGPT používána doposud. Model verze 3.5 byl trénovaný na datech do roku 2021. Nejnovější verze modelu je 4, která je dostupná v placené verzi ChatGPT, a na rozdíl od předchozích verzí má již přístup k internetu. [35]

  Stejně jako model BERT, je i GPT založen na transformerech. Avšak již nerozumí kontextu z obou stran, pouze zleva doprava, takže se označuje jako autoregresní[17]. Používá se především pro generování textu.

- **T5** (Text-To-Text Transfer Transformer) – model od společnosti Google, založený na bázi transformeru. Jeho trénovací sada je založena na text-to-text[18] úkolech, což mu umožňuje vynikat např. v překládání textů. Také je založen na transformerech. [20]

- **RoBERTa** (Robustly Optimized BERT approach) – vycházející z modelu BERT, vytvořen výzkumným týmem v organizaci Facebook AI. RoBERTa byl trénován na mnohonásobně větším korpusu, využíval kolem 160 GB dat, zatímco BERT byl trénován na sadě přibližně 10x menší. Při tréninku využíval techniku dynamického testování, což mu pomáhá naučit se robustnější a obecnější jazykové konstrukce. [43]

  Bylo prokázáno, že RoBERTa překonává BERT, XLNet a některé další moderní jazykové modely v mnoha částech zpracování přirozeného jazyka, jako je překlad textu, zodpovídání otázek nebo predikcích dalších vět (NSP). [43]

- **XLNet** – pokročilý model, který kombinuje prvky autoregresních modelů a transformerů. Pracuje s kontextem z obou směrů díky využití permutovaného způsobu tréninku – nepředpovídá následující tokeny ve vstupní sekvenci (jako např. BERT), ale předpovídá různé permutace tokenů. Dalším důvodem jeho úspěchu je dvoufázový trénink – nejprve se trénuje na permutovaných sekvencích a poté se ladí pomocí autoregresního tréninku. [7]

[16] Transformer (transformátor) – architektura neuronové sítě sloužící pro zpracování dlouhých sekvencí dat, využívající více hlavový mechanismus pozornosti, který dává různé váhy různým částem vstupních dat [33]

[17] Autoregresní model – nově generovaná data jsou závislé na předchozích datech [5]

[18] Text-to-text – převod jednoho textu na druhý

2.3.4 Trénování

Trénování v oblasti zpracování přirozeného jazyka zahrnuje proces učení modelů na základě připravených dat, aby byly schopny porozumět, zpracovávat a generovat lidský jazyk. Tento proces je klíčový pro samotné NLP pochody, jako je rozpoznávání entit, generování, porovnání textu a další.

Dílčí kroky při trénování modelů (inspirováno [11]):

1. **Sběr dat** – nejprve je potřeba sesbírat data, pomocí kterých se bude model učit. V závislosti na účelu modelu jsou vybrána příslušná data - texty, audio nebo obrázky. Nesmí se zapomenout vybrat taková data, která budou relevantní pro daný model, dostatečně rozmanitá a obsahující výrazy, se kterými se bude model nejčastěji setkávat. Tato data mohou být manuálně označená, tedy mohou obsahovat anotace, jako je značení slovních druhů, syntaktické značení nebo klasifikace entit.

2. **Předzpracování dat** – dalším krokem je zpracování dat do formátu, kterému daný model dokáže rozumět. To zahrnuje postupy jako jsou tokenizace, normalizace, lemmatizace nebo odstranění stop slov[19]. Předzpracování dat slouží ke snížení komplexity a víceznačnosti dat, a tedy vyzdvihnutí nejvýznamnějších specifik dat.

3. **Výběr modelu** – je potřeba vybrat správný model pro danou úlohu a podle typu dat. Existují různé typy modelů - statistické, pravidlové, neuronové nebo hybridní (více viz Jazykové modely). U modelů je potřeba brát ohled na jejich přesnost, rychlost, rozšiřitelnost a interpretovatelnost[20].

4. **Trénování modelu** – v tomto kroku se předávají data modelu, ze kterých se učí. Zahrnuje to rozdělení dat na trénovací, validační a testovací. Používají se různé učební algoritmy, jako je učení s učitelem (supervised learning), učení bez učitele (unsupervised learning) nebo zpětnovazební učení (reinforcement learning). Během těchto procesů model postupně zlepšuje svoje schopnosti porozumět a zpracovávat jazyk tím, že se učí vzory a vztahy mezi vstupními daty na základě zpětné vazby z dat a ztrátové funkce a snaží se minimalizovat chybu a maximalizovat výkon. Tento proces se často sleduje a vyhodnocuje pomocí metrik jako je přesnost, F-skóre[21] nebo perplexita[22].

5. **Optimalizace modelu** – jedná se o ladění a vylepšování modelu na základě výsledků a zpětné vazby z předchozího kroku. K tomu se používají různé techniky - regularizace, dávková normalizace nebo ořezávání gradientu. Ty pomáhají zabránit nadměrnému nebo nedostatečnému přizpůsobení, snížit rozptyl a lépe využít existující znalosti. Může se také experimentovat s různými architekturami modelů a porovnávat výsledky pomocí křížové validace nebo A/B testování.

6. **Testování** – nakonec je model testován na nezávislém korpusu, aby bylo vyzkoušeno jeho chování na nových, ještě nevyzkoušených datech. Díky tomu lze získat jeho výkonnost a spolehlivost při provádění konkrétní úlohy v reálném světě. Předchozí kroky

[19] Stop slova – slova bez skutečného/vnitřního významu - typicky předložky, spojky, v angličtině členy (a, an, the) [45]

[20] Interpretovatelnost modelu – schopnost člověka pochopit příčinu rozhodnutí modelu - čím vyšší je interpretovatelnost, tím snažší je pochopit předpovědi modelu [2]

[21] F-skóre – měřítko prediktivního výkonu

[22] Perplexita – míra, jak dobře rozdělení pravděpodobnosti nebo pravděpodobnostní model předikuje určitý vzorek dat - čím nižší perplexita, tím lépe model data predikuje [32]

mohou být několikrát opakovány, dokud se nedosáhne požadované úrovně výkonnosti modelu.

Po dokončení všech výše uvedených kroků je model nasazen do cílového prostředí a použit k jeho zamýšlenému účelu. Je potřeba také zajistit, aby splňoval určité etické a právní normy. V průběhu života modelu je potřeba jej neustále „přeškolovat“ za pomoci aktualizace jeho znalostní báze na základě zpětné vazby od uživatelů nebo nových dat.

2.3.5 Měření úspěchu extrakce informací

Měření úspěšnosti extrakce informací je klíčové pro posouzení výkonu a kvality systému zpracovávajícího přirozený jazyk. Níže jsou uvedeny některé z používaných metrik:

- **Přesnost** (Precision) – určuje, jaká část extrahovaných dat je skutečně relevantních. Vyjadřuje se jako podíl relevantních získaných dat a všech získaných informací. Ptáme se otázkou: „Kolik získaných dat je relevantních?“ [38, 42]

$$\text{Přesnost} = \frac{\text{získané relevantní informace}}{\text{všechny extrahované informace}}$$

- **Úplnost** (Recall) – určuje, jaká část skutečně existujících informací byla extrahována. Vyjadřuje se jako podíl relevantních získaných dat a celkový počet relevantních informací. Ptáme se otázkou: „Kolik relevantních dat je získaných?“ [38, 42]

$$\text{Úplnost} = \frac{\text{získané relevantní informace}}{\text{všechny relevantní informace}}$$

- **TP, FP, TN, FN** (True Positive, False Positive, True Negative, False Negative) – tyto metriky určují správně či špatně (relevantní/nerelevantní) určené entity. Slouží k výpočtu přesnosti a úplnosti. [38]
  - True Positive – relevantní, získané výsledky
  - False Positive – nerelevantní, získané výsledky
  - True Negative – nerelevantní, nezískané výsledky
  - False Negative – relevantní, nezískané výsledky

  Vzorce pro výpočet přesnosti a úplnosti za využití výše uvedených metrik jsou následující:

$$\text{Přesnost} = \frac{\text{TP}}{\text{TP} + \text{FP}} \qquad \text{Úplnost} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

  Vysvětlení na příkladu: Model rozpoznává psy na 20 obrázcích. 12 obrázků obsahuje psy (relevantní), 8 obsahuje kočky (nerelevantní). Model určí 10 obrázků jako výsledných - z toho 8 obrázků je se psy, 2 s kočkami. Výsledek je tedy - 8 TP (správně určené), 2 FP (špatně určené), 4 FN (zapomenuté, neurčené obrázky psů) a 6 TN (neurčené obrázky koček). [38] Z těchto údajů lze vypočítat přesnost a úplnost modelu:

$$\text{Přesnost} = \frac{8}{8+2} = \frac{8}{10} \qquad \text{Úplnost} = \frac{8}{8+4} = \frac{8}{12}$$

- **F-míra** (F-Score) – kombinuje metriky přesnost a úplnost do jednoho čísla od 0 do 1. Pokud je výsledek 1, model měl 100% úspěšnost, pokud 0, tak kompletně selhal buď v přesnosti, nebo úplnosti.

  Standardní měřítko je tzv. F_1 skóre, které se vypočítá jako harmonický průměr přesnosti a úplnosti: [1]

  $$F_1 = 2 * \frac{\text{přesnost} * \text{úplnost}}{\text{přesnost} + \text{úplnost}} = \frac{2 * \text{TP}}{2 * \text{TP} + \text{FP} + \text{FN}}$$

- **ROUGE** (Recall-Oriented Understudy for Gisting Evaluation) – míra, pomocí které se určuje kvalita generovaného textu, především sumarizace textů. Výsledek se porovná s referenčním textem, tvořeným lidmi. Měřítko pracuje s počtem shodujících se n-gramů, slovních sekvencí a dvojic slov v generovaném a referenčním textu. Výsledná hodnota se nachází v rozmezí 0 - 1, k jejímuž určení se používá F-Score. Čím vyšší hodnota, tím větší je shoda s referenčními daty. ROUGE tedy ve svém výpočtu závisí na přesnosti a úplnosti. Existuje několik variant ROUGE metrik: [10, 31]

  – ROUGE-N – týká se měření podobnosti n-gramů. „N“ v názvu označuje délku n-gramu - pokud se tedy jedná o ROUGE-1, porovnávají se unigramy, v případě ROUGE-2 se porovnávají bigramy atd. [10, 31]
  – ROUGE-L – zabývá se nejdelší společnou podsekvencí (LCS - Longest Common Subsequence). Míra podobnosti je podle této metriky určena délkou společné podsekvence vzhledem k délce referenčního textu. [10, 31]
  – ROUGE-W – vylepšený LCS algoritmus. Přiřazuje jednotlivým sekvencím váhy, které jsou závislé na délce společných podsekvencích - tedy upřednostňuje po sobě jdoucí podsekvence. [10, 31]

  Příklad LCS: Mějme 2 sekvence: [18]

  $$X = (A, B, C, A, D, F, B) \qquad Y = (B, C, J, D, B, H, A)$$

  Výsledná nejdelší společná posloupnost: (B, C, D, B)

- **BLEU** (Bilingual Evaluation Understudy) – navržen pro ohodnocení kvality přeloženého textu, používá se ale i na měření kvality ostatních generovaných textů. Stejně jako ROUGE, i zde se generovaný text porovnává s referenčním, lidmi připraveným textem. Jeho měřítko používá váhování n-gramů. [3, 17]

  Výpočet BLEU:

  Oříznutá přesnost (Clipped Precision) – vzhledem k tomu, že modely mají tendenci generovat větší množství stejných slov, než je nutné, je potřeba počty některých duplikovaných slov snížit. [3, 17]

  $$CP_n = \frac{\begin{array}{c}\text{počet oříznutých n-gram v generovaném textu} \\ \text{vyskytujících se v referenčním textu}\end{array}}{\text{celkový počet n-gram v generovaném textu}}$$

  Příklad (výpočet pro unigram):

  Generovaný text: Kočka je je je doma celá bílá.
  Referenční text: Kočka je doma a běhá po zahradě.

Slovo „je“ je v generovaném textu 3×, ale v referenčním textu pouze 1×. Takže ořízнutý počet slova „je“ bude 1. Slova „Kočka“ a „doma“ jsou v referenčním textu pouze 1×, tudíž jim přiřadíme 1.

Výslednou ořízнutou přesnost spočítáme:

$$CP = \frac{1+1+1}{7} = \frac{3}{7}$$

Vážený geometrický průměr přesnosti: slouží k vyvážení přesnosti generovaného textu vzhledem k délkovým rozdílům generovaného a referenčního textu. [3, 17]

$$\text{Vážený geometrický průměr přesnosti} = \prod_{n=1}^{N} CP_n^{w_n}$$

$$N = \text{váha gramu (unigram - 1, bigram - 2, } \dots) \qquad w_n = \frac{1}{N}$$

Trest za stručnost (BP – Brevity Penalty) – tímto výpočtem se penalizují texty, které jsou kratší, než je referenční text. [3, 17]

$$\text{trest za stručnost} = \begin{cases} 1, & \text{pokud } c > r \\ e^{(1-\frac{r}{c})}, & \text{pokud } c \leq r \end{cases}$$

kde r – délka referenčního textu, c – délka generovaného textu

Výsledné BLEU se vypočítá kombinací výše uvedených: [3, 17]

$$\text{BLEU} = \text{trest za stručnost} * \text{vážený geometrický průměr přesnosti}$$

Kapitola 3

Existující řešení

Wikipedie měla již v roce 2021 přes 55 milionů článků, což ji činí největší internetovou encyklopedií[1]. Je tedy zřejmé, že prací, zabývajících se extrakcí dat z Wikipedie a jejich následnou manipulací, vzniklo již mnoho a mnoho ještě vznikne. V této kapitole jsou tedy předvedeny programy, které řeší daný nebo obdobný problém, či jej řeší jen z části. Většinovou část tvoří programy vytvořené výzkumnou skupinou KNOT na VUT FIT[2].

3.1 Pywikibot

Pywikibot[3] je framework napsaný v jazyce Python, který interaguje s různými API vydávanými MediaWiki[4]. Jeho cílem je především automatizovat práci s MediaWiki stránkami (například Wikipedií).

Nabízí nepřeberné množství funkcí pro práci s MediaWiki - stahování obsahu stránek, extrakce názvů článků, aktualizace článků a další. V této práci byl framework použit pro automatickou úpravu článků na Wikipedii, více viz kapitola 4.2.

Výhodou frameworku je jeho neustálý vývoj a pravidelné aktualizace.

3.2 spaCy

SpaCy je open-source knihovna pro pokročilé zpracování přirozeného jazyka. Je napsána v jazycích Python a Cpython. Byla vytvořena vývojáři Matthew Honnibal a Ines Montani, kteří ji v roce 2015 vydali. Od té doby je neustále vyvíjena a aktualizována. Aktuálně (duben 2024) je její nejnovější verze 3.7. Knihovna byla vytvořena především pro produkční využití, kde je nápomocna při zpracování velkého množství textu, primárně pro extrakci informací, NLP, anebo před-zpracování textu pro hluboké učení. [25]

Nabízí již předtrénované modely a podporu tokenizace a trénování pro více jak 70 jazyků. V současné době není čeština podporována. Mezi nabízené funkce patří tokenizace, lemmatizace, rozpoznávání pojmenovaných entit (NER), spojování pojmenovaných entit, klasifikace textu, podobnost textů nebo serializace.

SpaCy je velmi jednoduché k použití, problém by s tím neměli mít ani lidé zrovna začínající s problematikou NLP. Také není problém využívat ji v end-to-end produkčních

[1] https://www.novinky.cz/clanek/internet-a-pc-wikipedia-se-za-dve-dekady-stala-jednou-z-nejpopularnejsich-webovych-stranek-40347686

[2] https://www.fit.vut.cz/research/group/knot/.cs

[3] https://www.mediawiki.org/wiki/Manual:Pywikibot/Overview

[4] https://www.mediawiki.org/wiki/API:Main_page

aplikacích, především díky jejímu efektivnímu využití CPU i GPU. [23]

Ukázka programu v spaCy napsaném v jazyce Python (upravený z [24]):

```
import spacy

# Load English tokenizer, tagger, parser and NER
nlp = spacy.load("en_core_web_sm")

# Process whole documents
text = ("As you can see, spaCy is really easy to use. We can just insert "
        "some text and let it process it."
       "It can also recognize specific entities, such as "
       "Google, America or Rudolph.")
doc = nlp(text)

# Analyze syntax
print("Noun phrases:", [chunk.text for chunk in doc.noun_chunks])
print("Verbs:", [token.lemma_ for token in doc if token.pos_ == "VERB"])

# Find named entities, phrases and concepts
for entity in doc.ents:
    print(entity.text, entity.label_)
```

Ve výstupu jsou vypsána všechna podstatná jména, slovesa, a pojmenované entity (organizace, geografické entity a osoby):

```
Noun phrases: ['you', 'spaCy', 'We', 'some text', 'it', 'it',
               'It', 'specific entities', 'Google', 'America', 'Rudolph']
Verbs: ['see', 'use', 'insert', 'let', 'process', 'recognize']
Google ORG
America GPE
Rudolph PERSON
```

3.3 Hugging Face

Hugging Face je francouzsko-americká společnost a open-source komunita zabývající se vývojem algoritmů, knihoven a modelů pro strojové učení a zpracování přirozeného jazyka. Založena byla v roce 2016 francouzskými podnikateli: Clément Delangue, Julien Chaumond, a Thomas Wolf. [12, 36]

Hugging Face slouží primárně jako úložiště jazykových modelů, které jsou zdarma pro využití. Uživatelé mohou stávající modely upravovat, nebo přidávat svoje vlastní. Každý model je určen pro jiné využití, a podle toho se odvíjí jeho architektura. Ta se může týkat např. maximální kapacity odpovědi (1bit - ano/ne), velikosti vstupního textu (např. 7B), nebo maximální kapacity paměti obsahující předchozí odpovědi.

Platforma má v nabídce taky chatbota HuggingChat, který je určitou alternativou k velmi populárnímu ChatGPT[5]. Pro běh HuggingChat si uživatelé mohou vybrat, jaký model bude chatbot využívat, a tak mohou rovnou daný model vyzkoušet. [6]

3.4 Programy vyvíjené skupinou KNOT

Na VUT FIT působí výzkupná skupina znalostních technologií (KNOT)[7], která za svoji dobu působení přinesla již mnoho výsledků a vytvořila řadu programů. Některé z nich zde budou zmíněny, protože byly využity v rámci práce.

Dané programy nemají veřejné publikace, jsou dostupné pouze členům dané skupiny (KNOT), umístěné na platformě GitHub.

3.4.1 Tvorba znalostní báze z české Wikipedie

Účelem projektu bylo extrahovat informace ze souboru s obsahem české Wikipedie (dump Wikipedie) a vytvořit znalostní bázi entit na základě těchto informací. Autorem projektu je Michal Planička.

Vstupem programu je dump soubor ve formátu XML, ze kterého se extrahují základní údaje o entitách - především typ (osoba, geografická entita, stát atd.), jméno, stručný popis entity (typicky první řádek na Wikipedii) a URL odkaz na článek. Pro různé typy entit jsou k dispozici další informace - pro osoby místo a datum narození, pohlaví; pro státy velikost a množství populace, pro geografické entity (vodopád, reliéf, ostrov ...) délka, výška, velikost, množství populace a další.

Výsledný soubor je ve formátu TSV, který je na rozdíl od vstupního XML souboru přehledný a lehce čitelný. Jeho údaje jsou stěžejní pro ostatní vyvíjené moduly - generování tvarů jmen (3.4.2), morfologie, doplňování chybějících entit a také pro tuto samotnou práci.

Ukázka záznamu ve výstupním souboru pro entitu "John Francis":

```
ade9a4ab08  person  John Francis  John Francis  John Francis
    (15. února 1924 - 20. března 2012) byl skotský profesor a esperantista.
    0  https://cs.wikipedia.org/wiki/John_Francis  M  1924-02-15  2012-03-20
```

3.4.2 Generování tvarů jmen

Program sloužící pro generování tvarů jmen osob, lokací a událostí (**Namegen**). Autorem je Ing. Martin Dočekal.

Modul je schopný skloňovat jména a vytvářet k nim odvozené tvary - přídavná jména, ženské podoby jmen a zdrobněliny, které jsou taktéž vyskloňované.

Vstupem programu je TSV soubor obsahující jméno entity, druh (geografická entita, událost, osoba - muž, žena), případně URL odkaz na článek na Wikipedii obsahující danou entitu.

Výstupní soubor je opět ve formátu TSV, který již obsahuje vyskloňovaná jednotlivá jména ze vstupu. V případě víceslovného jména je každé slovo skloňováno zvlášť. Dále jsou ke jménům přděleny informace: slovní druh, pád, rod, mluvnické číslo a druh slova (křestní

[5] https://chat.openai.com/
[6] https://huggingface.co/chat/
[7] https://www.fit.vut.cz/research/group/knot/.cs

jméno, příjmení, název lokace, číslovka, spojka ...).

Příklad výstupu pro jméno "Karel":

```
Karel  cs  P:::M  Karel[k1gMnSc1]#jG|Karla[k1gMnSc2]#jG|Karlu[k1gMnSc3]#jG/
    Karlovi[k1gMnSc3]#jG|Karla[k1gMnSc4]#jG|Karle[k1gMnSc5]#jG|
    Karlu[k1gMnSc6]#jG/Karlovi[k1gMnSc6]#jG|Karlem[k1gMnSc7]#jG
    https://cs.wikipedia.org/wiki/Karel
```

3.4.3 Odhadovač vzorů jmen

Cílem programu je odhadovat, pomocí kterých vzorů by se dané jméno dalo skloňovat. Autory jsou členové skupiny KNOT.

Modul vybírá z rozsáhlé báze vzorů, a podle určitých kritérií vybere nejvhodnější (ve výsledku může zmínit i více vzorů). Kritériemi pro výběr jsou: přípona a koncovka jména, typ (křestní jméno, příjmení, geografické jméno), životnost, u osob pohlaví.

Na vstupu program očekává soubor ve formátu LNTRF – 3-4 sloupcový TSV soubor formátu Lemma-Note-TransformationRulePositive/Negative. Očekává všechna slova v 1. pádu.

Příklad vstupu:

```
Brno  jL  k1gMnSc1::
Karel  jG  k1gMnSc1::
```

Výstupem je soubor ve formátu LPN[8]:

```
Brno#torero#jL
Karel#filozof#jG
Karel#posel#jG
```

Pro jméno "Karel" našel program 2 různé vzory, pomocí kterých se může jméno skloňovat. V tomto případě se jedná o české jméno, a tedy správná varianta je posel. To, že program vrací více možností, je v pořádku, jelikož u cizích jmen nemusí platit stejná pravidla skloňování, jako u jmen českých. Pokud by se program omezil jen na jeden nejvhodnější výsledek, nemusel by se potom cizím jménům přiřadit správný vzor.

3.4.4 Rozpoznávání pojmenovaných entit

Modul pro rozpoznávání pojmenovaných entit v textu (**NER** – named entity recognition). Autory jsou Ing. Lubomír Otrusina, Ing. Jan Doležal a Ing. Tomáš Volf.

Program pro svůj běh potřebuje předem vytvořenou znalostní bázi (3.4.1) a automaty. Ty mají vygenerované skloňování a značkování jmen (3.4.2) ze znalostní báze.

Automaty také obsahují transformovaná jména:

- zkracování jmen (`Adolf Born` → `A. Born`)
- příjmení na prvním místě (`Adolf Born` → `Born, Adolf`)

[8] LPN – lemma-paradigm-note (jméno-vzor-poznámka)

- pouze poslední příjmení na prvním místě (`Johann Gottfried Bernhard Bach` → `Bach, J.`)
- jména s výskytem svatý (`Svatý Jan / Sv.Jan /Sv Jan` → `Sv. Jan`).

Příklad výstupu po spuštění NER se vstupním textem „Praha je velké město v Česku“:

```
0       5       kb      Praha   5040
23      28      kb      Česku   44665
```

Program dokázal rozpoznat 2 jména ve vstupním textu. Výstup obsahuje následující informace: první a druhý sloupec značí umístění jména ve vstupním textu, 3. sloupec udává, že entita má přímý záznam ve znalostní bázi, 4. sloupec obsahuje nalezenou entitu v textu a 5. sloupec značí řádek ve znalostní bázi, na kterém se daná entita nachází. Po přechodu do znalostní báze můžeme dohledat dodatečné informace, jako je odkaz na Wikipedii, základní tvar jména, stručné informace o entitě a další (více viz Tvorba znalostní báze z české Wikipedie).

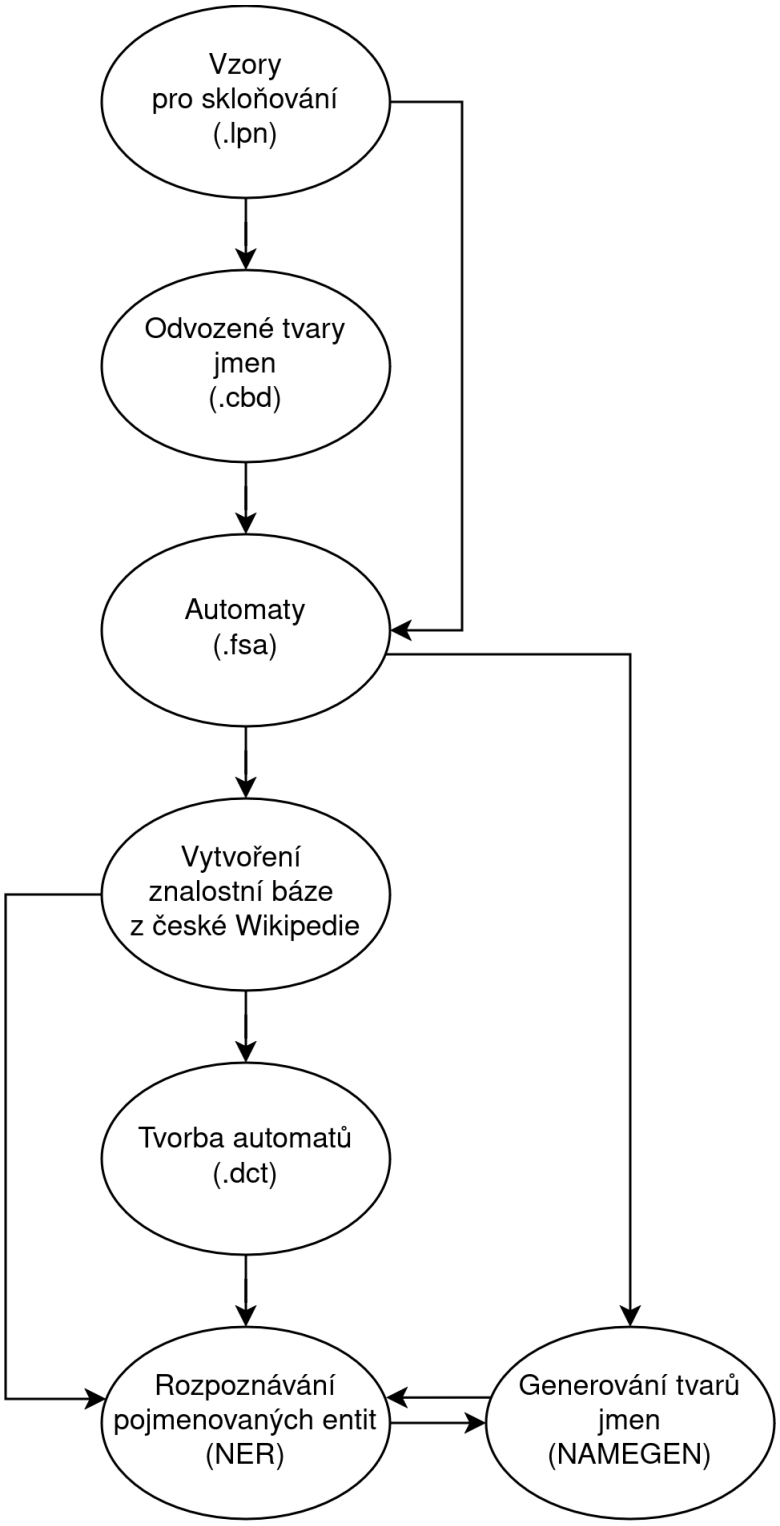
Spojení modulů pro spuštění rozpoznávání pojmenovaných entit

Program NER velmi závisí na datech, která mu jsou předána. Pro jeho samotné spuštění je potřeba tato data vygenerovat za pomoci výše zmíněných modulů:

1. Vygenerují se odvozené tvary jmen pro vstupní entity obsahující vzory pro skloňování (soubory formátu LPN). Výsledek se uloží do souborů ve formátu CBD[9], kde je obsažen typ odvození (např. přídavné jméno, zdrobnělina, příjmení ...), původní tvar ve formátu LPN, a nový tvar ve formátu LPN.
2. Z odvozených jmen a původních tvarů se vygenerují automaty ve formátu FSA[10]. Slouží k tomu, aby generátor tvarů jmen dokázal vygenerovat možné koreference na jména lidí v základním tvaru (např. „Karlův“ pro jméno „Karel“).
3. Dalším krokem je vytvoření znalostní báze z české Wikipedie, do které se bude NER odkazovat v případě rozpoznání některé pojmenované entity.
4. Za využití vytvořené znalostní báze se vygenerují automaty určené pro NER ve formátu DCT. Slouží pro rozpoznávání transformovaných jmen v textu (více viz Rozpoznávání pojmenovaných entit).
5. NER je připraven k použití. V průběhu svojí činnosti provolává Namegen (3.4.2), díky kterému dokáže rozpoznat i jména v jiném, než jen v 1. pádu.

[9] CBD: C – category of the relation (derivační třída - více viz 5.1), B – base (trojice ve formátu LPN, z něhož se vytváří odvozený tvar), D – derived (odvozený tvar ve formátu LPN)

[10] FSA – Finite State Automata – využívá se balík programů vytvořený Janem Daciukem z Gdaňské univerzity; soubory tohoto formátu vznikají vytvořením (minimálního) konečného automatu



Obrázek 3.1: Zakomponování jednotlivých modulů pro vygenerování a předání dat programu pro rozpoznání pojmenovaných entit

Kapitola 4

Návrh řešení

Zde bude představen návrh modulů, které byly v rámci projektu vypracovány. V kapitole 4.5 bude popsáno zakomponování veškerých modulů do systému, který automaticky, na začátku měsíce, zpracovává nejnovější dump Wikipedie, a jeho výstupem jsou nová jména včetně odvozených tvarů.

4.1 Doplňování odvozených tvarů jmen

Při rozpoznávání pojmenovaných entit (3.4.4) je potřeba mít vygenerované automaty, které obsahují skloňování jmen a transformované tvary jmem. Někdy se ale mohou vyskytovat odvozené tvary - v této práci jsem se zabýval těmito případy:

- příjmení vytvořené z křestního jména (Karel → Karlovi)
- ženská příjmení odvozená z mužských křestních jmen a příjmení (Adam → Adamová)
- přídavná jména pro geografické entity (Praha → pražský)
- přídavná jména přivlastňovací z křestních jmen a příjmení (Karel → Karlův)

Modul používá předem vytvořená derivační pravidla pro tvorbu odvozených tvarů. Hledí se na příponu (koncovku) jména, jeho typ (křestní jméno, příjmení, geografická entita) a vzor pro skloňování. Při tvorbě derivačních pravidel pro geografická jména jsem se řídil jazykovou příručkou pro tvorbu jmen domácích i cizích[1].

4.2 Úprava Wikipedie

Každý měsíc, při vydání nového dump souboru české Wikipedie, se může objevit řada nesrovnalostí. V samotných článcích na Wikipedii se objevují odlišné hodnoty v 1. větě článku (základní informace o popisované entitě) a ve Wikidatech (Infobox). Především se jedná o datum a místo narození a úmrtí (v případě textů o lidských entitách).

Tyto hodnoty je potřeba ověřit, zda se opravdu liší, nebo jde o tu stejnou hodnotu, akorát jinak zapsanou (např. `Fort Macleod, Alberta, Kanada` a `Fort McLeod, Kanada`). Taková data jsou brána jako rovnocenná a jsou ze souboru s nekonzistencemi smazána. Pokud se skutečně jedná o rozdílné hodnoty (např. datum úmrtí `1958-03-19` a `1955-03-19`), tak se vezme hodnota z Wikidat jako správná, jelikož se jedná o druhotnou databázi, která

[1] https://prirucka.ujc.cas.cz/?id=755

je spravovaná lidmi z celého světa napříč různými projekty (více viz 2.1.2). Hodnota na Wikipedii je platná pouze u daného článku, není spravovaná celosvětově. Tudíž je daleko méně pravděpodobné, že by se jednalo o správnou variantu.

Správná hodnota se tedy musí změnit v první větě daného článku přímo na Wikipedii - k tomu se využije framework pywikibot (3.1), který danou akci dokáže provést automatizovaně. Pro tuto akci je potřeba mít vytvořený účet na Mediawiki, případně přiznaný účet bota na Wikipedii.

4.3 Zpracování rozcestníků

Specifické stránky na Wikipedii „Rozcestníky“ (viz 2.1.5) mohou odkazovat na entity, které by se mohly stát alternativními jmény některých entit. Nemusí se jednat o položky, které známe, ale třeba o častá příjmení (např. „Smith“). Také se mohou vyskytovat odkazy na entity, které ještě neexistují na Wikipedii, a tudíž ani ve znalostní bázi (např. národopisec Václav Havel[2]). Z tohoto důvodu je důležité uchovávat si informaci, zda k odkazované entitě existuje článek na Wikipedii.

Navržený modul dokáže zpracovat rozcestníky v jazycích čeština, slovenština a angličtina.

Výsledný soubor je ve formátu TSV[3] obsahující 5 sloupců – název rozcestníku, název odkazované stránky, její stručný popis, kategorie/sekce na Wikipedii, do které spadá odkazovaná stránka, a informaci, zda článek existuje na Wikipedii (více viz 2.1.5). Poznámku, zda článek existuje, je potřeba uchávat z důvodu zajištění relevantnosti dat (informace o neexistující entitě může být smyšlená, případně entita nemusí v realitě existovat vůbec).

Ukázka zpracovaného rozcestníku pro stránku „Měsíc (rozcestník)“[4] (výstup byl lehce upraven a zkrácen):

```
Měsíc (rozcestník)  Měsíc  přirozená družice planety Země, vesmírné těleso  True
Měsíc (rozcestník)  měsíc (satelit)  přirozená družice (satelit) libovolné
    planety, druh vesmírného tělesa    True
Měsíc (rozcestník)  kalendářní měsíc  časová jednotka, v běžném kalendáři
    má délku 28-31 dní    True
Měsíc (rozcestník)  siderický měsíc  je doba oběhu Měsíce vzhledem ke hvězdám
    astronomické časové jednotky  True
Měsíc (rozcestník)  synodický měsíc  je doba mezi stejnými fázemi Měsíce
    astronomické časové jednotky  True
Měsíc (rozcestník)  tropický měsíc  je časová perioda (doba) mezi dvěma po sobě
    následujícími průchody Měsíce  astronomické časové jednotky  True
Měsíc (rozcestník)  anomalistický měsíc  je časová perioda (doba) mezi dvěma
    po sobě následujícími průchody Měsíce  astronomické časové jednotky  True
Měsíc (rozcestník)  drakonický měsíc  je časová perioda (doba) mezi dvěma po sobě
    následujícími průchody Měsíce  astronomické časové jednotky  True
Měsíc (rozcestník)  Měsíc (opera)  Měsíc (opera)  hudba  True
Měsíc (rozcestník)  Jiří Měsíc  pražský bezdomovec proslavený portálem
    YouTube a TV Nova  příjmení  False
```

[2] https://cs.wikipedia.org/wiki/Václav_Havel_(rozcestník)

[3] TSV – tab-separated values (tabulátorem oddělené hodnoty)

[4] https://cs.wikipedia.org/wiki/Měsíc_(rozcestník)

4.4 Generování tvarů jmen pro slova neznámého skloňování

NER (3.4.4) při generování automatů může narazit na jména, ke kterým nedokáže vygenerovat skloňování, odvozené tvary, nebo vygeneruje několik derivací, ale nedokáže určit, která je správná. Jedná se o slova, která morfologický analyzátor nezná. Taková slova vypisuje do souboru, včetně co nejvíce dodatečných informací, které dokázal o dané entitě určit (např. URL článku nebo typ entity).

V této práci jsem se zabýval slovy, u kterých není problém vygenerovat vzory pro skloňování. Jedná se o lidská jména končící na určitou příponu – např. ženská příjmení končící na „ová“, „ská“, nebo mužská příjmení končící na „on“, „en“, „ský“, „ič“ a další. Soubor s neurčitými jmény obsahuje informaci, zda se jedná o křestní jméno, příjmení, lokaci nebo jiný typ entity. Modul tedy pracuje pouze se slovy, u kterých je daná informace uvedena.

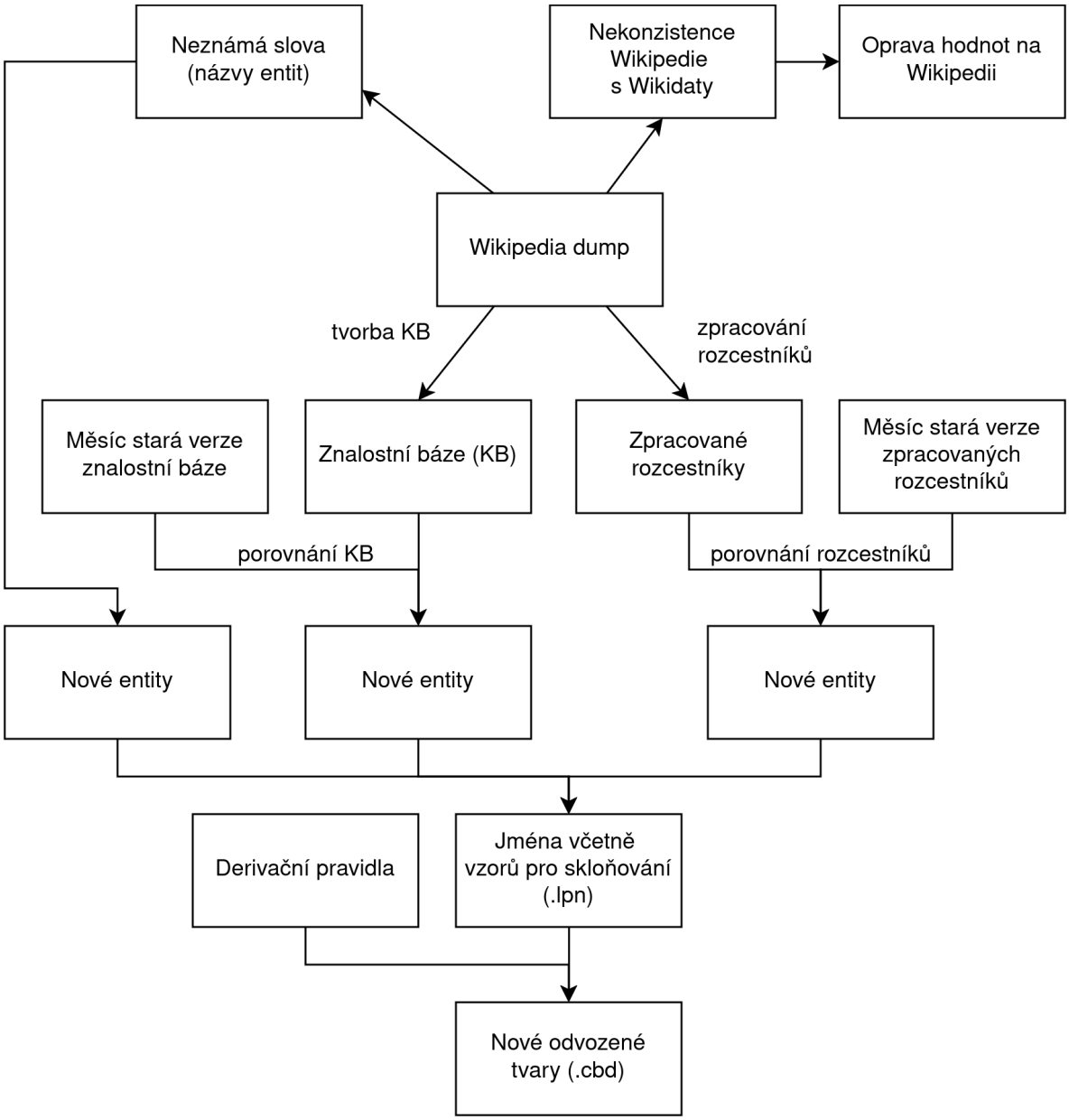
Tato získaná jména jsou následně předána modulu pro odhadování vzorů jmen (3.4.3). Výsledný soubor se jmény včetně vzorů může být použit pro aktualizaci schopností modulu NER.

4.5 Automatické generování odvozených tvarů jmen

Jak již bylo zmíněno výše, dump soubor Wikipedie je vydáván na začátku každého měsíce. Je ho tedy potřeba zpracovat, a vytvořit/aktualizovat znalostní bázi. K tomuto účelu je připraven automatický skript provádějící všechny potřebné úkony. Zakomponovány jsou všechny výše uvedené moduly připravené v této práci (grafické zpracování viz Obrázek 4.1):

1. Vygenerování nové znalostní báze (3.4.1) a zpracování rozcestníků z dump souboru (4.3)
2. Porovnání nové a staré znalostní báze – extrakce entit, které byly během posledního měsíce přidány na Wikipedii
3. Porovnání nového a starého zpracování rozcestníků – extrakce nově přidaných názvů entit (odkazů)
4. Extrakce slov neznámého skloňování (4.4)
5. Spojení výše uvedených extrahovaných slov, smazání duplicitních jmen
6. Spuštění generátoru tvarů jmen (3.4.2) a odhadovače vzorů (3.4.3) nad novými entitami, uložení výsledků ve tvaru LPN
7. Vygenerování nových odvozených tvarů (4.1)

Získané nové odvozené tvary jmen mohou být využity dalšími moduly, např. NER, pro zlepšení jejich funkcionality.



Obrázek 4.1: Propojení jednotlivých modulů do systému generujícího nové odvozené tvary jmen nově přidaných entit na Wikipedii

Kapitola 5

Implementace

V následujících sekcích bude popsána samotná implementace jednotlivých modulů představených v předchozí kapitole.

5.1 Doplňování tvarů jmen

Skript pro tvorbu odvozených tvarů jmen očekává na vstupu soubor se záznamy ve formátu LPN. Následně se ze souboru čtou data, a podle typu jména, přípony a vzoru se aplikují příslušná derivační pravidla uložená v souboru ve formátu TSV. Pro daný typ jména se vytvoří všechny možné odvozené tvary.

Pracuje se s následujícími typy jmen:

- jB – jména božstev, literárních postav, hrdinů bájí a historických postav
- jG – křestní jména
- jL – místní jména, označení geografických entit
- jS – příjmení

Derivační třídy jsou následující:

- 1102 – ženská podoba primárně mužského slova (Adam → Adamová)
- 1103 – jméno rodiny z mužského příjmení (Adam → Adamovi)
- 1201 – přídavné jméno přivlastňovací z podstatného jména (Adam → Adamův)
- 1202 – přídavné jméno z podstatného jména geografických entit (Praha → pražský)

Derivační pravidla
Jak již bylo uvedeno v části 4.1, program pracuje s předem nadefinovanými derivačními pravidly, která jsou uložena v souboru ve formátu TSV. Pravidla byla manuálně vytvořena ze vzorového souboru ve formátu CBD, který obsahoval primárně domácí, známá jména (Karel, Adam . . .). Derivační pravidla se liší pro každý typ odvození – přídavná jména, příjmení, ženská příjmení.

U derivační třídy obsahující přídavná jména geografických entit (vyškovský, pražský) jsou pravidla vytvořena bez vzorů skloňování. U ostatních skupin se při derivování berou v potaz i vzory.

Ukázka záznamu ve formátu CBD:

```
1201  Adam#filozof#jG  Adamův#otcův#
```

Příklad pravidla pro tvorbu přídavných jmen geografických entit:

```
1202  jL  ika  ->  ický  africký  # Kostarika -> kostarický
```

Význam jednotlivých slov: `1202` – derivační třída, `jL` – typ jména, `ika` – očekávaná přípona jména, `ický` – nová koncovka jména a `africký` – nový vzor pro skloňování. Za symbolem „`#`“ je vyznačen příklad derivace. Výjimkou u této třídy je, že nový (derivovaný) tvar bude začínat malým písmenem.
Výstup derivace je následující:

```
1202  Kostarika#dýka#jL  kostarický#africký#sG
```

Druhý příklad (tvorba přídavných jmen přivlastňovacích):

```
1201  jG  ek  ->  kův  medvídek  otcův  # Slávek -> Slávkův
```

Struktura je stejná, jako u předchozího příkladu, pouze přibyl jeden sloupec navíc. Jedná se o 6. sloupec obsahující „`medvídek`“, což je vzor, pomocí kterého se skloňuje původní tvar slova. U této derivační třídy je tedy nutné kontrolovat jak příponu, tak vzor, před aplikováním pravidla. Narozdíl od skupiny 1202 se již v odvozeném tvaru nemění první písmeno na malé.
Výstup derivace:

```
1201 Slávek#medvídek#jG Slávkův#otcův#jG
```

5.2 Úprava znalostní báze

Modul pracuje se souborem ve formátu TSV obsahujícím odlišné hodnoty na Wikipedii a ve Wikidatech.

5.2.1 Kontrola dat narození a úmrtí

Řádky obsahující rozdílná data narození a úmrtí jsou označena hodnotou „`BIRTH DATE`“ nebo „`DEATH DATE`“. Samotné datum je ve formátu `YYYY-MM-DD`. Příklad záznamu:

```
Alfréd Meissner  person  BIRTH DATE  1871-04-19  1871-04-10
```

Daný záznam má odlišné hodnoty dat narození, tudíž se vezme hodnota z Wikidat (4. sloupec, 1871-04-19), kterým se přepíše hodnota na Wikipedii.

V záznamech se také mohou objevit nekompletní data, která místo čísla obsahují otazník (např. 1912-??-??). Takové případy jsou přeskakovány. Je to z toho důvodu, že jedno z dat mohlo být nesprávně extrahováno z Wikipedie (což je většinový případ). Typicky se jedná o nesprávně extrahované datum z infoboxu – např. období vlády dané osoby. Taková hodnota nesouvisí s datem narození ani úmrtí, tudíž není potřeba ji upravovat.

5.2.2 Kontrola míst narození a úmrtí

Záznamy obsahující rozdílná místa narození a úmrtí jsou označena hodnotou „`BIRTH PLACE`“ nebo „`DEATH PLACE`“. Lokace mohou obsahovat několik částí oddělených čárkou, např. `Fort Macleod, Alberta, Kanada`. Je potřeba zkontrolovat nejspecifičtější, tedy 1. část, zda není obsažena ve druhé části - pokud je, místa jsou totožná a není potřeba jej upravovat na Wikipedii. První hodnota bývá nejspecifičtější označení místa, protože se jedná o název města. Zbývající hodnoty jsou typicky jméno (jména) země. Příklad: `Budapešť, Uhersko` a `Budapešť`.

Předtím, než se začnou upravovat hodnoty přímo na Wikipedii, je potřeba zkontrolovat několik věcí, zda nejsou uvedená místa totožná. Jedná se především o:

- Místo je uvedeno v 6. pádu – `Vipiteno, Itálie` a `ve Vipitenu`. Skript zkontroluje, jestli jedna z hodnot na začátku neobsahuje předložku „v“, „ve“ nebo „na“. Pokud ano, vezme část uvedenou v 1. pádu (tedy druhou z hodnot), a předá ji modulu pro generování tvarů jmen (3.4.2). Ten vrátí tvary pro všech 7 pádů. Zkontroluje se, zda tvar místa s předložkou není uveden mezi vygenerovanými tvary, a v případě shody dané místo nebude opravovat.

- Název obsahuje slovo „City“ – velmi častý případ, kdy jedno místo obsahuje přídavek „City“ a druhé ne, např. `New York` a `New York City`. Program slovo „City“ odstraní a opět porovná místa.

- Název města není uvedený celý – místa jsou totožná, ale jedno je napsané ve zkráceném formátu - např. `Řepiště` a `Řepiště u Frýdku-Místku`. Ověří se, zda kratší z hodnot není obsažena ve druhém případě.

- Uvedené je pouze město – stejně jako v předchozím případě, i zde se ověří, jestli není kratší název obsažen v názvu druhého místa (např. `Budapešť, Uhersko` a `Budapešť`).

5.2.3 Úprava dat na Wikipedii

Potom, co jsou ověřeny všechny položky, a opravdu se liší, mohou se opravit přímo na Wikipedii. K tomu se využije framework Pywikibot (3.1) a jeho dostupné prostředky pro úpravu článků. Díky tomu, že každá položka obsahuje přesný název článku na Wikipedii, stačí pouze tuto hodnotu předat Pywikibotu jako název upravované stránky. Pomocí metody `Page` se získá celý obsah stránky, a v první větě se nalezne hodnota, která se má změnit. Ta je přepsána správnou hodnotou z Wikidat, a celý řádek se nyní může aktualizovat na Wikipedii pomocí metody `replace`.

Skript může být spuštěn automaticky. Pywikibot pouze potřebuje prvotní údaje, jako jsou přihlašovací údaje na Mediawiki, název metody, kterou má používat, a shrnutí provedené změny, které se uloží v historii změn.

5.3 Zpracování rozcestníků

Modul pracuje s dump soubory Wikipedie ve formátu XML. Výsledný soubor je ve formátu TSV obsahující 5 sloupců (jejich popis viz 4.3).

Pro čtení a zpracování vstupního souboru ve formátu XML je využita knihovna `lxml.etree`[1] pro jazyk Python. Informaci, zda se jedná o rozcestník, lze najít na posledním řádku před

[1] https://lxml.de/index.html

ukončovacím tagem `</text>` (ukázka rozcestníku viz příloha A). V případě české Wikipedie musí obsahovat označení `{{Rozcestník}}`, slovenské `{{Rozlišovacia stránka}}` a anglické `{{disambig}}` nebo `{{disambiguation}}`. Pokud se označení rozcestníku na daném místě nachází, může se přejít k samotnému zpracování stránky, tedy dat mezi tagy `<page>` a `</page>`.

Jednotlivé položky rozcestníku začínají symbolem „*“. Jejich název (název článku) je obsažen ve dvou hranatých závorkách (např. `[[syderický měsíc]]`). Toto značení se na Wikipedii projeví jako odkaz na článek, což nutně neznamená jeho existenci (může se odkazovat na neexistující článek, který je potřeba vytvořit). Také se mohou vyskytovat případy, kdy název není uveden v hranatých závorkách - potom se tedy vezme první část obsahu před pomlčkou (za ní následuje stručný popis entity) jako název. Případně je název obsažen ve speciální konstrukci, jako třeba `{{Položka rozcestníku|Měsíc (opera)|typ=dílo}}`. Tehdy se použije regulární výraz (regex) k tomu, aby se extrahoval text mezi svislými čarami, který se uloží jako název článku.

Název se může vyskytovat i v jiném, než 1. pádu, případně se na Wikipedii objevuje jiný název, než jaký má skutečně daný článek (např. `[[Železniční stanice|nádraží]]`). V první části (před svislou čarou) je skutečný název (odkaz) článku, který se uloží; druhá část je tvar, který se zobrazí na Wikipedii.

Zbývající část položky značí její stručný popis, který se uloží do třetího sloupce ve výstupním souboru. Pokud začíná pomlčkou, je zbytečné ji ukládat, takže je z textu odstraněna.

Obsah 4. sloupce (sekce na stránce, kam článek spadá) se zjišťuje pomocí zanoření. Nejvyšší úroveň začíná jedním symbolem „*“, druhý dvěma, třetí třemi atd. Název sekce se použije z názvu nadřazené položky, případně textu, který je uveden ve vyšší kategorii. Příklad ze stránky „Dráha“[2]:

```
** [[jízdní dráha]]
** [[závodní dráha]]
*** [[běžecká dráha]]
*** [[dostihová dráha]]
*** [[plochá dráha]]
```

Pro položky `běžecká dráha`, `dostihová dráha` a `plochá dráha` bude ve 4. sloupci uvedena kategorie `závodní dráha`.

Poslední sloupec obsahuje informaci, zda k odkazované entitě existuje na Wikipedii článek. Dosaženo je toho zavoláním HTTP metody GET na stránku `https://<cs/sk/en>.wikipedia.org/wiki/<název článku>`. Pokud je obdržen stavový kód 200[3], stránka existuje, v případě jiných (typicky 301[4], Wikipedie přesměruje na editační stránku pro vytvoření daného článku) neexistuje. Jelikož každý dotaz trvá přibližně 0,4 sekundy, zpracování 200 000 položek rozcestníku (přibližný počet odkazovaných stránek pro českou Wikipedii) by trvalo více jak 22 hodin. K urychlení je využit paralelismus (knihovna `multiprocessing`[5], který rozloží dotazy na všechna možná CPU na pracovním stroji.

[2] https://cs.wikipedia.org/wiki/Dráha
[3] 200 – Status OK
[4] 301 – Moved Permanently
[5] https://docs.python.org/3/library/multiprocessing.html

5.3.1 Extrakce nově přidaných jmen

Stejně jako při zpracování znalostní báze, i u rozcestníků se každý měsíc objeví nové odkazy na články. Řada z nich jsou lidská jména nebo názvy geografických entit, a pro ty je potřeba vygenerovat vzory pro skloňování a odvozené tvary.

Porovnají se dvě verze zpracovaných rozcestníků ve formátu .tsv (nejnovější a měsíc stará verze). Výsledkem jsou odkazy, které se neobjevovaly v předchozí verzi. Pro generování odvozených tvarů není potřeba řešit, zda se jedná pouze o přejmenovanou stránku.

Následně se prochází všechny nové odkazy, a používá několik metod pro zjištění, zda se jedná o lidskou entitu nebo geografickou. Pro určování se využívá souborů obsahujících všechny možné druhy lidských profesí nebo slov určujících, že se jedná o osobu – např. tituly, slova „jméno“, „příjmení“, hodnosti („major“, „plukovník“) a další. Pro geografické entity je využit soubor obsahující druhy lokací – např. „ostrov“, „moře“, „město“. Tyto soubory byly převzaty a upraveny z dřívějšího projektu zaměřeného na zpracování rozcestníků od Matúše Remeňa.

Pro určení typu entity se postupuje následovně:

1. Název článku nebo rozcestník obsahuje v závorkách některé ze slov označujících lidskou entitu (např. „`Zajíc (příjmení)`“)
2. Pokud neobsahuje, tak odkaz musí mít stručný popis, popřípadě se nachází v nějaké sekci na rozcestníku (více viz 4.3)
3. Prochází se, zda se některé ze slov z pomocných souborů nachází v popisu, případně v názvu sekce
4. Jestliže bylo určeno, že se jedná o lidskou i geografickou entitu, tak se nebude zapisovat na výstup, protože není nelze jednoznačně určit, o co se jedná (např. geografická entita mající v popisku text na způsob: „hřbitov, na kterém je pohřben podplukovník XX“, nebo osoba s popiskem „herec XX žijící na pobřeží YY“)

Následně se u všech získaných odkazů odstraní přídatky v závorkách (např. „`(příjmení)`“), čímž se získá samotné jméno. K těm se přidá typ jména (`jB, jL, jS`, viz 5.1), a může se přejít k samotnému generování odvozených tvarů.

5.3.2 Statistické vyhodnocení zpracování rozcestníků

Vyhodnocení zpracování rozcestníků probíhalo porovnáním dvou verzí rozcestníků (změn během jednoho měsíce) a určením počtu nově přidaných rozcestníků a odkazovaných stránek.

Nejdříve bylo potřeba zjistit, které stránky se liší s předchozí verzí. Z rozcestníků mohly být odstraněny, či přidány nové odkazy na články. Pro výsledný počet se neberou v úhavu položky rozcestníků, u kterých neexistuje odkaz. Pracuje se s počtem nově přidaných odkazů a dřívějších odkazů (těch, které nebyly nalezeny v nové verzi Wikipedia dumpu). Vzhledem k tomu, že některé odkazované články mohly být přejmenovány, tak se od počtu nově přidaných odkazů odečte počet starých odkazů. Vznikne číslo udávající, kolik stránek s odkazy bylo přidáno do rozcestníků.

Také bylo potřeba vyhodnotit, kolik rozcestníků bylo přidáno. Opět se porovnávají zpracované rozcestníky z verzí dumpu lišící se o měsíc. Pro získání finálního počtu nestačí pouze porovnat počty rozcestníků ve dvou verzích Wikipedie. Některé rozcestníky mohly

být smazané, nebo rozdělené na více (např. v únorové verzi 2024 z rozcestníku „`Zajíc (rozcestník)`“ vznikl nový – „`Zajíc (příjmení)`“), případně byl rozcestník rozdělen na více, a všechny se jmenují zcela jinak. V případě, že by se pracovalo pouze s rozdílem množství rozcestníků na Wikipedii, by nám u posledního případu vyšlo, že za poslední měsíc vznikl pouze jeden rozcestník, jenže ve skutečnosti vznikly 2 (např. při porovnání lednové a únorové verze 2024 by pouhým porovnáním množství rozcestníků vyšlo číslo 65, ale ve skutečnosti přibylo 76 nových rozcestníků).

V této práci je to řešeno tak, že se nejprve extrahují veškeré názvy rozcestníků, které se nenachází v předchozí verzi. Musí se ale počítat s možností přejmenování rozcestníků. Proto se ze všech názvů nových, i měsíc starých rozcestníků, odstraní (pokud existuje) přídavek „`(rozcestník)`“ (u slovenské Wikipedie „`(rozlišovacia stránka)`“). Následně se opět porovnají tyto upravené názvy rozcestníků s předchozí verzí, a výsledný počet opravdu se lišících rozcestníků se vezme jako finální počet nově přidaných rozcestníků.

5.4 Generování tvarů jmen pro slova neznámého skloňování

Na vstupu je očekáván soubor ve formátu LNTRF[6]. Modul daný soubor prochází a extrahuje jména odpovídající vzorům – slova končící na „ová“, „ská“, „ský“, „ič“, „an“ a další.

Příklad vhodného zápisu ve vstupním souboru:

```
cs  cs  Rybačuková  jS  k1gFnSc1::  P:::F  @  Ada Fedorivna Rybačuková
  https://cs.wikipedia.org/wiki/Ada_Rybačuková  all names have multiple
  derivations
```

Je zde uvedeno, o jaký typ entity se jedná (`jS` – příjmení), o ženské jméno (`P:::F`) a tvar 1. pádu skloňování (`k1gFnSc1::`). Pro takový záznam je snadné odhadnout vzor skloňování, takže se vezme a předá modulu pro odhadování vzorů (3.4.3). Výsledek je ve formátu LPN:

```
Rybačuková#vrátná#jS
```

Daný výstup se může předat dalším modulům pro zpracování, například generátoru odvozených tvarů jmen (5.1).

5.5 Automatické generování odvozených tvarů jmen

Nakonec mohou být jednotlivé moduly zakomponovány do jednoho systému.

Na začátku se vytvoří nová znalostní báze (3.4.1) z nejnovějšího dump souboru Wikipedie. Ta se porovná s měsíc starou verzí KB, čímž se získají nová jména entit, které za poslední měsíc přibyly na Wikipedii. Stejně tak se zpracují rozcestníky (4.3), a zpracovaný TSV soubor se porovná s měsíc starým zpracováním rozcestníků. Výsledkem jsou opět nová jména entit. Také se zpracuje soubor ve formátu LNTRF obsahující slova neznámého skloňování (4.4), z čehož se získají další jména.

Všechna získaná jména se uloží do souborů podle jejich typu (křestní jméno, příjmení, lokace). Následně se použije program pro generování tvarů jmen (3.4.2), který vrátí základní informace o daném slově pomocí značek. Např. `k1gFnSc1::`, kde `k1` – podstatné jméno, `gF`

[6]LNTRF – 3-4 sloupcový TSV soubor formátu Lemma-Note-TransformationRulePositive/Negative

– ženský rod, `nS` – jednotné číslo, `c1` – 1. pád. Pro další práci se jmény se extrahují pouze slova v 1. pádu.

Jakmile jsou ke jménům přiřazeny morfologické informace, mohou se předat programu pro odhadování vzorů (3.4.3). Výsledkem jsou soubory ve formátu LPN, ze kterých se již mohou vygenerovat odvozené tvary ve formátu CBD. Ty mohou být využity dalšími moduly, např. pro zlepšení schopností NERu.

Kapitola 6

Dosažené výsledky

Zpracování výsledků a vyhodnocení práce probíhalo na datech z dump souborů Wikipedie za poslední půlrok (říjen 2023 - duben 2024), není-li uvedeno jinak. Samotná zpracovaná Wikipedie je vydávána každý měsíc, takže rozdíly při porovnávání verzí Wikipedie (a tedy i samotné statistiky) jsou jeden měsíc.

6.1 Vygenerované odvozené tvary

V této sekci budou představeny výsledky generování odvozených tvarů jmen pro jména domácí, cizí i netradiční. V první části (6.1.1) jsou představeny výsledky na datech za půl roku (říjen 2023 - duben 2024), včetně detailních počtů nově přidaných entit osob a lokací na Wikipedii. Ve druhé části (6.1.2) jsou představeny celkové počty přidaných odvozených tvarů jmen pro cizí jména z celé české Wikipedie.

6.1.1 Výsledky za období říjen 2023 – duben 2024

Zde jsou představeny počty nově získaných tvarů jmen. Data byla získávána za poslední půlrok (říjen - duben 2024) z české Wikipedie. Jedná se o nově přidané odvozené tvary jmen pro jména osob (křestní jména a příjmení) a názvy geografických entit.

Zdrojem nových informací byla nová znalostní báze (porovnávala se vždy s měsíc starou bází) a nové zpracování rozcestníků (porovnávalo se s měsíc starou verzí zpracování rozcestníků). Výstupem porovnání znalostní báze byly nové entity, u rozcestníků nové odkazy na články (v tabulkách 2. a 3. sloupec). Nakonec bylo potřeba zkontrolovat, zda se daná jména nenachází mezi „známými“ jmény – tedy těmi, pro které jsou známy vzory pro skloňování (formát LPN), a tedy jednotlivé moduly s nimi již pracují. Taková slova není potřeba znovu zpracovávat. Jsou extrahována pouze nenalezená jména, ke kterým jsou odhadovány vzory. Úspěšné výsledky jsou uloženy ve formátu LPN (4. sloupec v tabulkách).

Posledním zdrojem pro získání nových jmen byl soubor obsahující slova, která morfologický analyzátor nezná (4.4). Počet jmen, pro která se podařilo odhadnout vzory skloňování, a tedy vytvořit výslednou trojici slov ve formátu LPN, je ve výsledných tabulkách uveden ve 4. sloupci 2. řádku. Protože se zpracovávala pouze jména osob, se kterými si odhadovač vzorů dokázal poradit (nebylo potřeba řešit množství nově přidaných entit), je v tabulkách vynechán 2. a 3. sloupec. Při prvním zpracování (říjen - listopad, Tabulka 6.1) se přidalo nejvíce jmen, protože byla nashromážděná za delší období. U dalších verzí je počet přidávaných jmen nižší, jelikož se jedná pouze o ta, která přibyla za poslední měsíc. Řadu jmen díky předchozímu měsíci již morfologický analyzátor zná.

Zdroj nových jmen	Přidaných entit osob	Přidaných geografických entit	Nových jmen včetně vzorů (LPN)
Jména neznámého skloňování	–	–	**204**
Znalostní báze	716	91	**67**
Rozcestníky	261	151	**42**
Celkový počet jmen v základním tvaru (LPN)			307
Celkový počet vygenerovaných odvozených tvarů jmen (CBD)			396
Celkový počet nových tvarů (CBD + LPN)			**688**

Tabulka 6.1: Přidané základní i odvozené tvary jmen osob a lokací z české Wikipedie za období říjen - listopad 2023

Zdroj nových jmen	Přidaných entit osob	Přidaných geografických entit	Nových jmen včetně vzorů (LPN)
Jména neznámého skloňování	–	–	**73**
Znalostní báze	717	144	**70**
Rozcestníky	502	122	**53**
Celkový počet jmen v základním tvaru (LPN)			163
Celkový počet vygenerovaných odvozených tvarů jmen (CBD)			233
Celkový počet nových tvarů (CBD + LPN)			**376**

Tabulka 6.2: Přidané základní i odvozené tvary jmen osob a lokací z české Wikipedie za období listopad - prosinec 2023

Zdroj nových jmen	Přidaných entit osob	Přidaných geografických entit	Nových jmen včetně vzorů (LPN)
Jména neznámého skloňování	–	–	**47**
Znalostní báze	637	167	**54**
Rozcestníky	267	144	**38**
Celkový počet jmen v základním tvaru (LPN)			137
Celkový počet vygenerovaných odvozených tvarů jmen (CBD)			234
Celkový počet nových tvarů (CBD + LPN)			**350**

Tabulka 6.3: Přidané základní i odvozené tvary jmen osob a lokací z české Wikipedie za období prosinec - leden 2024

Zdroj nových jmen	Přidaných entit osob	Přidaných geografických entit	Nových jmen včetně vzorů (LPN)
Jména neznámého skloňování	–	–	**65**
Znalostní báze	899	135	**72**
Rozcestníky	427	145	**63**
Celkový počet jmen v základním tvaru (LPN)			168
Celkový počet vygenerovaných odvozených tvarů jmen (CBD)			296
Celkový počet nových tvarů (CBD + LPN)			**451**

Tabulka 6.4: Přidané základní i odvozené tvary jmen osob a lokací z české Wikipedie za období leden - únor 2024

Zdroj nových jmen	Přidaných entit osob	Přidaných geografických entit	Nových jmen včetně vzorů (LPN)
Jména neznámého skloňování	–	–	**63**
Znalostní báze	791	104	**80**
Rozcestníky	290	96	**28**
Celkový počet jmen v základním tvaru (LPN)			170
Celkový počet vygenerovaných odvozených tvarů jmen (CBD)			273
Celkový počet nových tvarů (CBD + LPN)			**419**

Tabulka 6.5: Přidané základní i odvozené tvary jmen osob a lokací z české Wikipedie za období únor - březen 2024

Zdroj nových jmen	Přidaných entit osob	Přidaných geografických entit	Nových jmen včetně vzorů (LPN)
Jména neznámého skloňování	–	–	**89**
Znalostní báze	1 180	144	**48**
Rozcestníky	415	136	**35**
Celkový počet jmen v základním tvaru (LPN)			163
Celkový počet vygenerovaných odvozených tvarů jmen (CBD)			175
Celkový počet nových tvarů (CBD + LPN)			**327**

Tabulka 6.6: Přidané základní i odvozené tvary jmen osob a lokací z české Wikipedie za období březen - duben 2024

Z výše uvedených výsledků je patrné, že každý měsíc přibude na českou Wikipedii přibližně 170 nových jmen, pro která neznáme skloňování, a také nemáme dostupné žádné odvozené tvary. Je tedy výhodné provádět každý měsíc kontrolu nových dat, a generovat jejich odvozené tvary, čímž se zlepší schopnosti navazujících modulů (např. NER).

6.1.2 Celkový počet přidaných odvozených tvarů jmen

Odvozené tvary jmen bylo potřeba také vygenerovat pro netradiční a cizí jména (např. „A'Hearn“, „Fetfatzidis“). Zpracovávala se jména z Wikipedie a Wikidat získaná v průběhu

několika posledních let (tedy včetně výsledků z 6.1.1). Jednalo se o velké množství dat - 146 057 vstupních jmen. Celkově se k nim podařilo vygenerovat 385 075 odvozených tvarů.

Počet jmen	146 057
Vygenerovaných odvozených tvarů	385 075

Tabulka 6.7: Celkové množství vygenerovaných odvozených tvarů jmen

6.2 Oprava hodnot na Wikipedii

Zde jsou uvedeny výsledky, které bylo potřeba opravit na české Wikipedii, jelikož se jednalo o nesouhlasná data s těmi uvedenými ve Wikidatech (byl použit modul 4.2). Týká se to míst a dat narození či úmrtí osob, o nichž byl daný článek na Wikipedii.

Opravy hodnot probíhaly z dump souboru květnové verze Wikipedie 2024 (verze `20240501`). Nahlášených změn bylo 8 077, z toho různých míst bylo 6 383, a dat 1 694. Po provedení přezkoumání, zda se místa opravdu liší, nebo jde jen o doplnění hodnot, případně jsou místa napsaná v jiném formátu (např. v 1. a 6. pádu – jedná se tedy o totožná místa), bylo odstraněno 5 775 záznamů, celkově tedy zbylo 608 míst pro opravu na Wikipedii.

To stejné bylo provedeno i pro data narození či úmrtí osob. Po odstranění neúplných či nejednoznačných hodnot zbylo 1 029 dat, vyřazeno jich bylo tedy 665.

Nahlášené výsledky	Místa narození/úmrtí	Data narození/úmrtí
Původní počet záznamů	6 383	1 694
Odstraněné záznamy	5 775	665
Výsledný počet záznamů pro opravu	608	1 029

Tabulka 6.8: Množství nesrovnalostí s Wikipedií a Wikidaty z květnové verze české Wikipedie

6.3 Zpracování rozcestníků

Níže uvedené výsledky vycházejí ze zpracování popsaného v 5.3.2 pro českou, slovenskou a anglickou Wikipedii. Druhý sloupec zobrazuje množství nově přidaných rozcestníků během jednoho měsíce, a třetí sloupec množství nových odkazů na články, které se přidaly na veškeré rozcestníky.

Období srovnání	Počet nových rozcestníků	Počet nových odkazů
říjen - listopad 2023	66	521
listopad - prosinec 2023	104	895
prosinec - leden 2024	63	683
leden - únor 2024	76	691
únor - březen 2024	64	573
březen - duben 2024	73	655

Tabulka 6.9: Počty nově přidaných rozcestníků a odkazů v rozcestnících na české Wikipedii

Období srovnání	Počet nových rozcestníků	Počet nových odkazů
říjen - listopad 2023	81	266
listopad - prosinec 2023	25	148
prosinec - leden 2024	38	153
leden - únor 2024	34	138
únor - březen 2024	23	110
březen - duben 2024	23	113

Tabulka 6.10: Počty nově přidaných rozcestníků a odkazů v rozcestnících na slovenské Wikipedii

Období srovnání	Počet nových rozcestníků	Počet nových odkazů
říjen - listopad 2023	1 891	15 586
listopad - prosinec 2023	1 386	9 217
prosinec - leden 2024	1 764	19 779
leden - únor 2024	1 333	14 145
únor - březen 2024	1 371	17 094
březen - duben 2024	1 228	14 714

Tabulka 6.11: Počty nově přidaných rozcestníků a odkazů v rozcestnících na anglické Wikipedii

Kapitola 7

Závěr

V této práci byl představen systém, který dokáže na začátku každého měsíce zpracovat získaná data z dump souboru české Wikipedie. Následně nově získaná data porovná s výsledky z minulé verze (tedy o jeden měsíc starší), a nově přidané hodnoty zpracuje. Především extrahuje jména osob a lokací, a generuje k nim odvozené tvary (přídavná jména, ženské tvary jmen z mužských podob a jména rodin z mužského příjmení). Výsledkem je zlepšení schopností navazujících modulů (generování tvarů jmen, NER).

Na českou Wikipedii přibude každý měsíc přibližně 170 nových, neznámých jmen osob a geografických entit (ať už cizích, nebo domácích). Tato jména je vhodné pravidelně zpracovávat – generovat k nim odvozené tvary – aby mohly moduly zpracovávající přirozený jazyk pracovat se stále nejaktuálnějšími daty. Stejně tak je důležité opravovat hodnoty na Wikipedii v nesouladu s daty ve Wikidatech, aby při dalším zpracování dump souboru nedocházelo k nekonzistencím. Mohlo by docházet např. k nepřesnému určování pojmenovaných entit, jelikož by daná entita měla přiřazeno více míst nebo dat narození či úmrtí.

Práce přinesla pravidla pro tvorbu odvozených tvarů jmen, která se aplikují podle typu slova (jména osob či geografických entit), přípony, a případně vzoru skloňování. Po jejich aplikování na celkově 146 057 jmen získaných z dat české Wikipedie, k nim bylo vygenerováno 385 075 nových odvozených tvarů.

Rozšíření či vylepšení tohoto systému může být přidání nových pravidel pro tvorbu odvozených tvarů. Může se týkat ať už přidání nových vzorů skloňování, či samotné pracování s novými typy jmen – např. pseudonymy, patronymy, označení příslušníků národů a měst (Němec, Pražan), nebo jmen věcí – výrobků, lodí, svátků, vlaků, univerzit, hvězd a dalších.

Literatura

[1] Akre, K. *F-score* online. 2024. Dostupné z: https://www.britannica.com/science/F-score. [cit. 2024-04-04].

[2] AWS. *Model interpretability* online. Dostupné z: https://docs.aws.amazon.com/whitepapers/latest/ml-best-practices-healthcare-life-sciences/model-interpretability.html. [cit. 2024-04-04].

[3] Baeldung, M. S. *Bleu Score* online. 2024. Dostupné z: https://www.baeldung.com/cs/nlp-bleu-score. [cit. 2024-04-04].

[4] BBC. *Topics that spark Wikipedia 'edit wars' revealed* online. 2013. Dostupné z: https://www.bbc.com/news/technology-23354613. [cit. 2024-04-04].

[5] Deepchecks, A. *Autoregressive model* online. Dostupné z: https://deepchecks.com/glossary/autoregressive-model/. [cit. 2024-04-04].

[6] Fineproxy, A. *POS značkování* online. Dostupné z: https://fineproxy.org/cs/wiki/part-of-speech-pos-tagging/. [cit. 2024-04-04].

[7] Garrin McGoldrick, S. P. *Understanding XLNet* online. 2019. Dostupné z: https://www.borealisai.com/research-blogs/understanding-xlnet/. [cit. 2024-04-04].

[8] Hashemi Pour, C. *BERT language model* online. 2024. Dostupné z: https://www.techtarget.com/searchenterpriseai/definition/BERT-language-model. [cit. 2024-04-04].

[9] Kurama, V. *Information Extraction* online. 2024. Dostupné z: https://nanonets.com/blog/information-extraction/. [cit. 2024-04-04].

[10] Lin, C.-Y. ROUGE: A Package for Automatic Evaluation of Summaries. In: *Text Summarization Branches Out*. Barcelona, Spain: Association for Computational Linguistics, červenec 2004, s. 74–81. Dostupné z: https://aclanthology.org/W04-1013.

[11] LinkedIn, A. I. a komunita. *How to Train NLP Models* online. Dostupné z: https://www.linkedin.com/advice/0/how-do-you-train-natural-language-processing. [cit. 2024-04-04].

[12] Lutkevich, B. *What is Hugging Face?* online. 2023. Dostupné z: https://www.techtarget.com/whatis/definition/Hugging-Face. [cit. 2024-04-04].

[13] MediaWiki, P. *MediaWiki* online. 2023. Dostupné z: https://www.mediawiki.org/wiki/MediaWiki/cs. [cit. 2024-04-04].

[14] Mediawiki, P. *Interwiki* online. 2024. Dostupné z: https://www.mediawiki.org/wiki/Manual:Interwiki/cs. [cit. 2024-04-04].

[15] OpenAI, A. *ChatGPT* online. 2024. Dostupné z: https://openai.com/blog/chatgpt. [cit. 2024-04-04].

[16] Otten, N. V. *Top 20 Most Powerful Large Language Models For NLP Tasks & Transfer Learning In 2024* online. 2023. Dostupné z: https://spotintelligence.com/2023/04/18/large-language-models-nlp/. [cit. 2024-04-04].

[17] Papineni, K.; Roukos, S.; Ward, T. a Zhu, W.-J. Bleu: a Method for Automatic Evaluation of Machine Translation. In: Isabelle, P.; Charniak, E. a Lin, D., ed. *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*. Philadelphia, Pennsylvania, USA: Association for Computational Linguistics, červenec 2002, s. 311–318. Dostupné z: https://aclanthology.org/P02-1040.

[18] Programiz, A. *Longest common subsequence* online. 2023. Dostupné z: https://www.programiz.com/dsa/longest-common-subsequence. [cit. 2024-04-04].

[19] Proxet, A. *Fundamentals of Statistical Natural Language Processing* online. 2021. Dostupné z: https://www.proxet.com/blog/fundamentals-of-statistical-natural-language-processing. [cit. 2024-04-04].

[20] Raffel, C. *T5 Explained* online. 2019. Dostupné z: https://paperswithcode.com/method/t5. [cit. 2024-04-04].

[21] Rusiňák, P. *Určování typů entit na základě extrakce informací z Wikipedie*. 2017. Bakalářská práce. Vysoké učení technické v Brně. Fakulta informačních technologií. Dostupné z: https://dspace.vut.cz/server/api/core/bitstreams/41b322f8-d340-4040-80c0-933d090c4606/content.

[22] Sachdeva, N. *What are Language Models in NLP?* online. 2023. Dostupné z: https://insights.daffodilsw.com/blog/what-are-language-models-in-nlp. [cit. 2024-04-04].

[23] spaCy, A. *SpaCy* online. Dostupné z: https://spacy.io/usage/facts-figures. [cit. 2024-04-04].

[24] spaCy, A. *SpaCy* online. Dostupné z: https://spacy.io/. [cit. 2024-04-04].

[25] spaCy, A. *SpaCy* online. 2024. Dostupné z: https://spacy.io/usage/spacy-101. [cit. 2024-04-04].

[26] Tdunning. *Hidden Markov Model* online. 2012. Dostupné z: https://cs.wikipedia.org/wiki/Skryt%C3%BD_Markov%C5%AFv_model#/media/Soubor:HiddenMarkovModel.svg. [cit. 2024-04-04].

[27] Tuffery, S. *Deep Learning for Natural Language Processing*. John Wiley & Sons, Incorporated, 2022. ISBN 1119845017.

[28] Valušek, O. *Extrakce informací z Wikipedie*. 2018. Bakalářská práce. Vysoké učení technické v Brně. Fakulta informačních technologií. Dostupné z: https://dspace.vut.cz/server/api/core/bitstreams/8e6908b3-65d9-438b-9c80-12ce1467c06e/content.

[29] Wikidat, P. *Wikidata* online. 2024. Dostupné z: https://www.wikidata.org/wiki/Wikidata:Introduction/cs. [cit. 2024-04-04].

[30] Wikipedia, P. *Infobox* online. 2023. Dostupné z: https://cs.wikipedia.org/wiki/N%C3%A1pov%C4%9Bda:Infoboxy. [cit. 2024-04-04].

[31] Wikipedia, P. *ROUGE (metric)* online. 2023. Dostupné z: https://en.wikipedia.org/wiki/ROUGE_(metric). [cit. 2024-04-04].

[32] Wikipedie, P. *Perplexita* online. 2023. Dostupné z: https://cs.wikipedia.org/wiki/Perplexita. [cit. 2024-04-04].

[33] Wikipedie, P. *Transformátor (model strojového učení)* online. 2023. Dostupné z: https://cs.wikipedia.org/wiki/Transform%C3%A1tor_(model_strojov%C3%A9ho_u%C4%8Den%C3%AD). [cit. 2024-04-04].

[34] Wikipedie, P. *Wikipedie:Rozcestníky* online. 2023. Dostupné z: https://cs.wikipedia.org/wiki/Wikipedie:Rozcestn%C3%ADky. [cit. 2024-04-04].

[35] Wikipedie, P. *Generative pre-trained transformer* online. 2024. Dostupné z: https://cs.wikipedia.org/wiki/Generative_pre-trained_transformer. [cit. 2024-04-04].

[36] Wikipedie, P. *Hugging Face* online. 2024. Dostupné z: https://en.wikipedia.org/wiki/Hugging_Face. [cit. 2024-04-04].

[37] Wikipedie, P. *Nadace Wikimedia* online. 2024. Dostupné z: https://cs.wikipedia.org/wiki/Nadace_Wikimedia. [cit. 2024-04-04].

[38] Wikipedie, P. *Precision and recall* online. 2024. Dostupné z: https://en.wikipedia.org/wiki/Precision_and_recall. [cit. 2024-04-04].

[39] Wikipedie, P. *Rule-based modeling* online. 2024. Dostupné z: https://en.wikipedia.org/wiki/Rule-based_modeling. [cit. 2024-04-04].

[40] Wikipedie, P. *Wiki* online. 2024. Dostupné z: https://cs.wikipedia.org/wiki/Wiki. [cit. 2024-04-04].

[41] Wikipedie, P. *Wikipedia* online. 2024. Dostupné z: https://cs.wikipedia.org/wiki/Wikipedie. [cit. 2024-04-04].

[42] Wikisofia, P. *Pojem relevance, její vyhodnocování, druhy relevance* online. 2021. Dostupné z: https://wikisofia.cz/wiki/Pojem_relevance,_jej%C3%AD_vyhodnocov%C3%A1n%C3%AD,_druhy_relevance. [cit. 2024-04-04].

[43] Yinhan Liu, M. O. a. d. *RoBERTa: A Robustly Optimized BERT Pretraining Approach* online. 2019. Dostupné z: https://arxiv.org/pdf/1907.11692.pdf. [cit. 2024-04-04].

[44] Černý, M. *Online encyklopedie nejsou jen Wiki* online. 2008. Dostupné z: https://www.lupa.cz/clanky/online-encyklopedie-nejsou-jen-wiki/. [cit. 2024-04-04].

[45] Štráfelda, J. *Co je stop slovo* online. Dostupné z: https://www.strafelda.cz/stop-slovo. [cit. 2024-04-04].

Příloha A

Rozcestník z dump souboru formátu XML

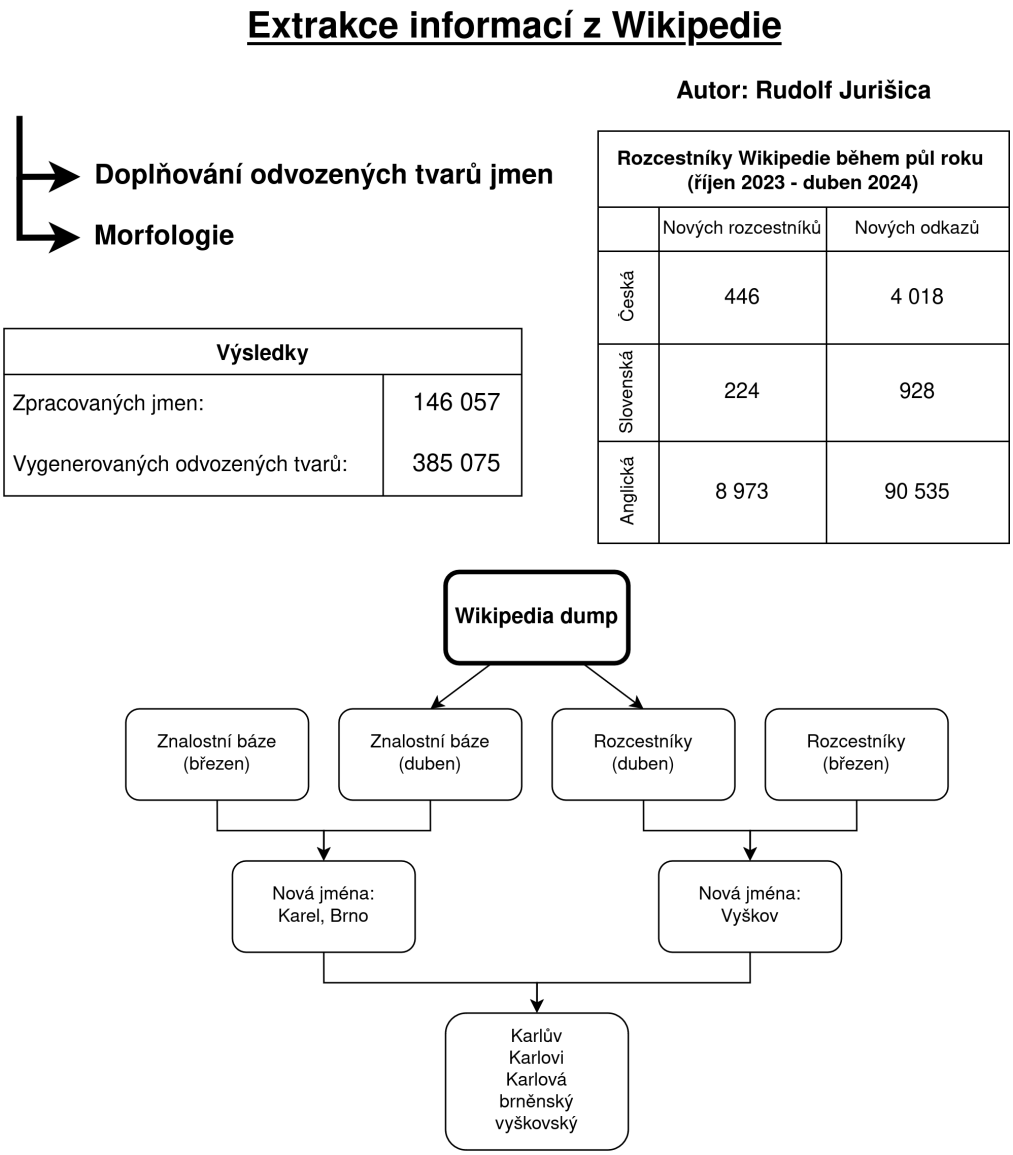
Ukázka rozcestníku z Wikipedie pro stránku „Měsíc (rozcestník)“ z únorové verze 2024:

```
  <page>
    <title>Měsíc (rozcestník)</title>
    <ns>0</ns>
    <id>76</id>
    <revision>
      <id>23351796</id>
      <parentid>22377369</parentid>
      <timestamp>2023-11-05T11:02:33Z</timestamp>
      <contributor>
        <username>Reing</username>
        <id>46874</id>
      </contributor>
      <comment>opravy významů &amp;quot;měsíc&amp;quot;</comment>
      <model>wikitext</model>
      <format>text/x-wiki</format>
      <text bytes="1382" xml:space="preserve">'''Měsíc''' má více významů:
* [[Měsíc]] - přirozená družice planety Země, vesm|írné těleso
* [[měsíc (satelit)]] - přirozená družice (satelit) libovolné planety,
    druh vesmírného tělesa
* [[kalendářní měsíc]] - časová jednotka, v běžném kalendáři má délku
    28-31 dní

; astronomické časové jednotky
* [[siderický měsíc]] je doba oběhu Měsíce vzhledem ke hvězdám
    (pravá oběžná doba)
* [[synodický měsíc]] je doba mezi stejnými fázemi Měsíce,
    jak se jeví ze Země
* [[tropický měsíc]] je časová perioda (doba) mezi dvěma
    po sobě následujícími průchody Měsíce [[jarní bod|jarním bodem]]
* [[anomalistický měsíc]] je časová perioda (doba) mezi dvěma po sobě
```

```
    následujícími průchody Měsíce [[Apsida (astronomie)|perigeem]]
* [[drakonický měsíc]] je časová perioda (doba) mezi dvěma po sobě
    následujícími průchody Měsíce [[Uzel (astrodynamika)|výstupným uzlem]]
    jeho dráhy

; hudba
* {{Položka rozcestníku|Měsíc (opera)|typ=dílo}}

; příjmení
* [[Jiří Měsíc]] - pražský bezdomovec proslavený portálem [[YouTube]]
    a [[TV Nova]]

== Související články ==
* [[Měsíček]]
* [[Moon]]
* [[Luna]]
* [[Portál:Měsíc]]

== Externí odkazy ==
* {{Wikicitáty|téma=Měsíc}}
* {{Wikislovník|heslo=Měsíc}}
* {{Wikislovník|heslo=měsíc}}

{{Rozcestník}}</text>
      <sha1>0xwkpmiozf3kciyl6urt58ith6w0f6q</sha1>
    </revision>
  </page>
```

Příloha B

Plakát



Obrázek B.1: Plakát prezentující práci s dosaženými výsledky