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

PŘÍPRAVA VYHODNOCOVACÍ SADY PRO SLOŽITÉ PROBLÉMY ROZPOZNÁVÁNÍ A ZJEDNOZNAČŇOVÁNÍ POJMENOVANÝCH ENTIT POMOCÍ CROWDSOURCINGU

PREPARING EVALUATION SET FOR COMPLEX PROBLEMS OF RECOGNITION AND DISAMBIGUATION OF NAMED ENTITIES THROUGH CROWDSOURCING

DIPLOMOVÁ PRÁCE
MASTER'S THESIS

AUTOR PRÁCE **Bc. PETER PASTOREK**
AUTHOR

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

BRNO 2019

Vysoké učení technické v Brně
Fakulta informačních technologií

Ústav počítačové grafiky a multimédií (UPGM) Akademický rok 2018/2019

Zadání diplomové práce

Student: **Pastorek Peter, Bc.**
Program: Informační technologie Obor: Počítačová grafika a multimédia
Název: **Příprava vyhodnocovací sady pro složité problémy rozpoznávání a zjednoznačňování pojmenovaných entit pomocí crowdsourcingu**
Preparing Evaluation Set for Complex Problems of Recognition and Disambiguation of Named Entities through Crowdsourcing
Kategorie: Web

Zadání:

1. Seznamte se s metodami a principy využití crowdsourcingu a gamifikace pro anotaci lingvistických dat a s dostupnými nástroji pro rozpoznávání a zjednoznačňování pojmenovaných entit.
2. Na základě získaných poznatků navrhněte systém pro získávání dat s vyznačenými entitami a odkazy na Wikipedii.
3. Implementujte daný systém a vyhodnoťte výsledky sběru dat a kvalitu získaných anotací.
4. Vytvořte stručný plakát prezentující práci, její cíle a výsledky.

Literatura:

- Sil, A., Kundu, G., Florian, R., Hamza, W. Neural Cross-Lingual Entity Linking. AAAI Conference on Artificial Intelligence, North America, 2018.
- Sabou, M., Bontcheva, K., Derczynski, L., & Scharl, A. (2014, May). Corpus Annotation through Crowdsourcing: Towards Best Practice Guidelines. In *LREC* (pp. 859-866).

Při obhajobě semestrální části projektu je požadováno:

- Funkční prototyp systému

Podrobné závazné pokyny pro vypracování práce viz http://www.fit.vutbr.cz/info/szz/

Vedoucí práce: **Smrž Pavel, doc. RNDr., Ph.D.**
Vedoucí ústavu: Černocký Jan, doc. Dr. Ing.
Datum zadání: 1. listopadu 2018
Datum odevzdání: 22. května 2019
Datum schválení: 1. listopadu 2018

Abstrakt

Práca sa zaoberá prípravou vyhodnocovacej sady pre problémy rozpoznávania a zjednoznačňovania pomenovaných entít. Vyhodnocovacia sada je výsledkom automatizovaného spracovania a crowdsourcingu. Vyhodnocovacia sada môže byť použitá na testovanie pokročilých prípadov v rozpoznávaní a zjednoznačňovaní pomenovaných entít.

Abstract

This Master's Thesis prepares Evaluation Set for Problems of Recognition and Disambiguation of Named Entities. Evaluation Set is created using Automatization and Crowdsourcing. Evaluation Set can be used in testing Edge Cases in Recognition and Disambiguation of Named Entities.

Kľúčové slová

rozpoznávanie pomenovaných entít, zjednoznačňovanie pomenovaných entít, crowdsourcing

Keywords

Recognition of Named Entities, Disambiguation of Named Entities, Crowdsourcing

Citácia

PASTOREK, Peter. *Příprava vyhodnocovací sady pro složité problémy rozpoznávání a zjednoznačňování pojmenovaných entit pomocí crowdsourcingu*. Brno, 2019. Diplomová práce. Vysoké učení technické v Brně, Fakulta informačních technologií. Vedoucí práce doc. RNDr. Pavel Smrž, Ph.D.

Příprava vyhodnocovací sady pro složité problémy rozpoznávání a zjednoznačňování pojmenovaných entit pomocí crowdsourcingu

Prehlásenie

Prehlasujem, že som túto semestrálnu prácu vypracoval samostatne pod vedením pána doc. RNDr. Pavla Smrže, Ph.D. Ďalšie informácie mi poskytol pán Ing. Lubomír Otrusina. Uviedol som všetky literárne pramene a publikácie, z ktorých som čerpal.

.....
Peter Pastorek
21. mája 2019

Poďakovanie

Chcel by som sa poďakovať vedúcemu práce pánovi doc. RNDr. Pavlu Smržovi, Ph.D. za smerovanie pri práci a odbornú podporu.

Obsah

1 Úvod **2**

2 Úloha rozpoznávania a zjednoznačňovania pomenovaných entít **3**
2.1 Extrakcia pomenovanej entity 3
2.2 Rozpoznanie pomenovanej entity 4
2.3 Zjednoznačňovanie pomenovanej entity 4

3 Testovanie rozpoznávania a zjednoznačňovania pomenovaných entít **5**
3.1 Formáty vyhodnocovacích sád 5
3.2 Metriky používané pre vyhodnotenie algoritmov na rozpoznávanie a zjednoznačňovanie pomenovaných entít 6

4 Vytvorenie vyhodnocovacej sady z výstupu projektu Wikilinks **8**
4.1 Získanie reprezentujúcich odkazov 9
4.2 Získanie relevantných odkazov 10
4.3 Manipulovanie textu 10

5 Crowdsourcing a jeho použitie pri vytváraní vyhodnocovacej sady **12**
5.1 Typy aplikácií Crowdsourcingu 12
5.2 Príklady Crowdsourcingu 13
5.3 Hra používajúca vyhodnocovaciu sadu 16

6 Aplikovanie vyhodnocovacej sady **18**
6.1 Semantic Enrichment Component 18
6.2 LingPipe 23

7 Vyhodnocovacie sady **25**

8 Záver **27**

Literatúra **28**

Kapitola 1

Úvod

Táto diplomová práca je pokračovanie semestrálneho projektu[23].

Problémy rozpoznávania pomenovaných entít (Named Entity Recognition – NER) a zjednoznačňovania pomenovaných entít (Named Entity Disambiguation – NED) sú pod problémy spracovania prirodzeného jazyka (Natural Language Processing – NLP). Ich spojenie je často označené ako rozpoznávanie a zjednoznačňovanie pomenovaných entít (Named Entity Recognition and Disambiguation – NER/D) alebo linkovanie pomenovaných entít (Named Entity Linking – NEL). Linkovanie pomenovaných entít musí navyše od rozpoznávania a zjednoznačňovania vygenerovať odkaz do databáze pomenovaných entít (Knowledge Base – KB). Najznámejšia databáza pomenovaných entít je Wikipédia. Pokročilý systém na rozpoznávanie a zjednoznačňovanie pomenovaných entít by mal byť schopný nájsť pomenované entity v texte a priradiť k nim správny odkaz na článok Wikipédie.

Problémy rozpoznávania a zjednoznačňovania pomenovaných entít sú jednoduchšie pre človeka ale zložité pre stroj. Algoritmy na rozpoznávanie a zjednoznačňovanie pomenovaných entít sa stále vyvíjajú a preto sú potrebné nástroje na ich testovanie a porovnávanie. Táto práca sa zaoberá práve prípravou vyhodnocovacej sady na splnenie toho účelu. Vyhodnocovacia sada má obsahovať zaujímavé a pokročilé prípady kde sa môže algoritmus na rozpoznávanie a zjednoznačňovanie pomenovaných entít jednoducho pomýliť. To môžu byť rôzne entity s rovnakým menom alebo entity pre ktoré sa používa skratka.

Pretože problémy rozpoznávania a zjednoznačňovania pomenovaných entít sú jednoduchšie pre človeka je jedna z najefektívnejších možností použitie Crowdsourcingu. Crowdsourcing je zložený zo slov crowd (dav) a source (zdroj). Ide o systém kde stroj na pozadí spolupracuje s veľkým davom ľudí na vykonaní úlohy.

Vyhodnocovacie sady na rozpoznávanie a zjednoznačňovanie pomenovaných entít sú buď malé, platené alebo nekvalitné. Ak je vyhodnocovacia sada tvorená iba strojom, tak je nekvalitná. Ak je tvorba vyhodnocovacej sady podporovaná ľuďmi, tak je malá alebo platená, pretože čas ľudí väčšinou stojí peniaze.

Kapitola 2

Úloha rozpoznávania a zjednoznačňovania pomenovaných entít

Pomenovaná entita je objekt z reálneho sveta. Nemusia to byť len materiálne objekty ako ľudia alebo miesta, ale aj technológie alebo organizácie. Úloha rozpoznávania a zjednoznačňovania pomenovaných entít je jeden z problémov porozumenia textu. Ide o priradenie sémantického významu slovu alebo skupine slov.

Najčastejšie dva problémy sú:

- Určiť začiatok a koniec pomenovanej entity. Entita môže mať viacero mien akými sa môžu v teste odkazovať. Mená pomenovanej entity sa často prekrývajú. Jedno meno pomenovanej entity môže byť časťou iného mena rovnakej pomenovanej entity. Napríklad: Ing. Andrej Kiska môže byť odkazovaný ako Prezident Kiska alebo iba Kiska. Ak nebude celé meno entity vyznačené, tak sa algoritmus môže pokúsiť nevyznačené časti entity priradiť iným častiam vety. Tento problém sa volá: „one person, many names“[18].

- Priradiť meno správnej entite. Jedno meno môže byť priradené viacerým entitám. Pri takomto prípade nastane zjednoznačňovanie pomenovanej entity. Napríklad John Smith môže byť pofesor anatómie a chémie, architekt, vojak, politik, biskup, športovec alebo film. Tento problém sa volá: „one name, many people“ [18].

2.1 Extrakcia pomenovanej entity

Extrakcia pomenovanej entity sa stará o nájdenie potenciálneho mena entity a vymedzenie jeho rozsahu. Táto časť sa v niektorých článkoch kombinuje s rozpoznávaním pomenovanej entity. Často sa rieši ako jeden problém. Algoritmy, ktoré uľahčujú prácu rozpoznávaniu pomenovanej entity, sa nachádzajú v tejto časti programu na rozpoznávanie a zjednoznačňovanie pomenovaných entít. Takéto algoritmy môžu byť syntaktická a sémantická analýza textu alebo zistenie jazyka, v ktorom je text napísaný.

2.2 Rozpoznanie pomenovanej entity

Cieľom rozpoznania pomenovanej entity je z mena entity vyprodukovať zoznam pomenovaných entít, ktoré sa na dané meno môžu viazať. Entity v tomto zozname sa volajú kandidátne entity. Dobrý algoritmus na rozpoznávanie pomenovaných entít produkuje čo najmenší zoznam kandidátnych pomenovaných entít, v ktorom sa nachádza správna entita. Algoritmus na rozpoznávanie pomenovaných entít väčšinou obsahuje zoznam mien a ich mapovanie na entity. Buď vo forme slovníku mien alebo aj neurónovej siete.[27]

2.3 Zjednoznačňovanie pomenovanej entity

Cieľom zjednoznačňovania pomenovanej entity je zo zoznamu kandidátov vybrať správnu pomenovanú entitu, a tým priradiť časti textu odkaz do databáze pomenovaných entít. V prípade mnou vytvorenej vyhodnocovacej sady odkaz do databáze pomenovaných entít bude odkaz na článok z Wikipédie. Špeciálny prípad zjednoznačňovania pomenovanej entity je výber žiadnej entity alebo vybranie nulovej entity. Aj napriek tomu že bolo nájdené meno pomenovanej entity z databáze pomenovaných entít skutočná pomenovaná entita v databáze nemusí byť. Nulová entita reprezentuje všetky pomenované entity, ktoré nie sú v databáze pomenovaných entít. V texte môže byť nájdené meno človeka, ale nemusí byť dosť slávny na to aby mal vlastnú stránku Wikipédie. [27]

Údaje, ktoré by mali byť vo vyhodnocovacej sade, budú závisieť od údajov používaných na zjednoznačňovanie pomenovanej entity. Algoritmy na zjednoznačňovanie pomenovaných entít sa delia nasledovne:

- Bez kontextové algoritmy. Na zjednoznačňovanie používajú vlastnosti odkazu a pomenovanej entity.[27]
  - Tvar odkazu. Napríklad jednotné alebo množné číslo podstaného mena použitého ako odkaz na pomenovanú entitu.[27]
  - Popularita pomenovanej entity. Ak odkaz na pomenovanú entitu bude meno osoby, tak algoritmus vyberie najpopulárnejšiu osobu s daným menom.[27]

  Bez kontextové algoritmy by nemali byť efektívne pri pokročilých problémoch rozpoznania a zjednoznačňovania pomenovaných entít. Predpokladám ich úspešnosť medzi 50% až 60% pri kvalitnej vyhodnocovacej sade.

- Kontextové algoritmy. Na zjednoznačňovanie používajú vlastnosti textu.[27]
  - Slová okolo odkazu na pomenovanú entitu. Algoritmus si k pomenovanej entite ukladá slová vzťahujúce sa k pomenovanej entite. Porovnávať ich môže algoritmom Bag of words.[27]
  - Ostatné nájdené pomenované entity. Ak algoritmus nájde v texte hudobný album, tak predpokladá že nájdená osoba bude hudobník a nie športovec.[27]

  Na použitie kontextového algoritmu musí vyhodnocovacia sada obsahovať potrebný kontext. Kontextové algoritmy by mali byť úspešnejšie ako bez kontextové algoritmy. Predpokladám ich úspešnosť medzi 80% až 90% pri kvalitnej vyhodnocovacej sade.

Kapitola 3

Testovanie rozpoznávania a zjednoznačňovania pomenovaných entít

Prvé vyhodnocovacie sady určené pre testovanie a porovnávanie algoritmov na rozpoznávanie a zjednoznačňovanie pomenovaných entít boli vytvorené manuálnym anotovaním novinových článkov. Nevýhodou manuálneho anotovania je malé množstvo prvkov vo vyhodnocovacej sade. Výhodou je zaujímavosť daných položiek. Ak vstup nebude zaujímavý pre algoritmy na rozpoznávanie a zjednoznačňovanie pomenovaných entít s veľkou pravdepodobnosťou budú ignorované a nebudú pridané do vyhodnocovacej sady. Napríklad jedna z prvých vyhodnocovacích sád obsahovala ibe entity pod menom John Smith.[18]

Vyhodnocovacia sada sa dá vytvoriť aj premenovaním entít. Aj keď entity nemajú podobné mená ich mená v texte môžu byť nahradené rovnakým menom. Používa sa kombinácia mien originálnych entít. Napríklad mená entít „jablko“ a „slanina“ môžu byť nahradené menom „jablko-slanina“. Týmto spôsobom sa dá generovať väčšia vyhodnocovacia sada. Dajú sa na nej dobre testovať algoritmy na zjednoznačňovanie pomenovaných entít. Algoritmy na rozpoznávanie pomenovaných entít sa dajú na takejto vyhodnocovacej sade testovať iba čiastočne. Programy a algoritmy musia byť špeciálne pripravené na substitúciu použitú pri vytvorení vyhodnocovacej sady. Programy vytvorené na prácu s bežnými príkladmi na takejto vyhodnocovacej sade nebudú fungovať.

Manuálne anotovanie sa dá použiť na veľké množstvo dát s veľkým počtom ľudí. Na tieto účely sa zatiaľ používa Crowdsourcing platforma Mechanical Turk. Nevýhodou platformi Mechanical Turk je platenie užívateľou. Nájdu sa užívatelia, ktorý chcú dostať zaplatené bez toho aby prácu vykonaly.[18]

3.1 Formáty vyhodnocovacích sád

Nedá sa každá vyhodnocovacia sada spracovávať rovnakým spôsobom. Niektoré vyhodnocovacie sady majú pridané údaje, ktoré by mali uľahčiť spracovanie pre algoritmus rozpoznávania a zjednoznačňovania pomenovaných entít. Na takéto vyhodnocovacie sady je potrebné špeciálne rozšírenie, ktoré umožňuje ich spracovanie. Výhodou špeciálnych formátov je upravenie závislostí výsledkov testovania na rozličných častiach programu alebo môže donútiť algoritmus spracovávať len entity a mená, ktoré od neho vyhodnocovacia sada očakáva.

Príklady formátov použitých vo vyhodnocovacích sadách:

- Čistý text je najjednoduchší formát pre vyhodnocovaciu sadu. Čistý text sa približuje reálnej situácií. Jediné pridané údaje sú meno entity a odkaz na entitu do databáze pomenovaných entít. Môže k nim byť pridaný aj odkaz na začiatok mena entity v texte.

  Príklad:

  en.wikipedia.org/wiki/United_Nations U.N.
  en.wikipedia.org/wiki/Rolf_Ekéus Ekeus
  en.wikipedia.org/wiki/Baghdad Baghdad

  U.N. offcial Ekeus heads for Baghdad.
  Inšpitovaný príkladom z[30].

  Príklad sa skladá z hlavičky a tela. Hlavička a telo sú rozdelené prázdnym riadkom. V hlavičke sú pomenované entity a k nim priradené mená použité v tele. V tele je text, ktorý dostane algoritmus na rozpoznávanie a zjednoznačňovanie pomenovaných entít. V tomto texte sa pokúsi rozpoznať a zjednoznačniť pomenované entity.

- Text s anotáciami pri každom slove. V takomto texte je každé slovo na samostatnom riadku. Za daným slovom sú anotácie pre dané slovo. Jedna z užitočných anotácií pre rozpoznávanie a zjednoznačňovanie pomenovaných entít je príslušnosť slova k menu entity. Ďalšou anotáciou môže byť typ pomenovanej entity alebo aj syntaktický význam slova vo vete.[30]

  Príklad:

  U.N. NNP I-NP I-ORG
  offcial NN I-NP O
  Ekeus NNP I-NP I-PER
  heads VBZ I-VP O
  for IN I-PP O
  Baghdad NNP I-NP I-LOC
  . . O O
  Prevzatý z[30].

  Prvé slovo v riadku je slovo vo vete. Štvrté slovo v riadku určuje funkciu slova pri rozpoznávaní a zjednoznačňovaní pomenovaných entít.[30]

3.2 Metriky používané pre vyhodnotenie algoritmov na rozpoznávanie a zjednoznačňovanie pomenovaných entít

Pri aplikovaní algoritmu na rozpoznávanie a zjednoznačňovanie pomenovaných entít je potrebné sledovať a analyzovať jeho výsledky. Sledované atribúty a metriky môžu meniť štruktúru vyhodnocovacej sady.

V dokumentoch, ktoré riešili testovanie algoritmov na rozpoznávanie a zjednoznačňovanie pomenovaných entít, boli nasledovné sledované atribúty a metriky:

- Presnosť – je percento správne nájdených pomenovaných entít.[18, 30, 27]

- Počet kandidátov – je priemerná veľkosť zoznamu kandidátov pri rozpoznávaní pomenovanej entity.[18]
- Presnosť vyhladania kandidátov – je percento zoznamov kandidátov, ktoré obsahujú správneho kandidáta.[18]
- Presnosť vyhľadania prázdnych kandidátov – je percento prázdnych zoznamov kandidátov, keď nemá byť priradená pomenovaná entita.[18]
- Presnosť nájdenia prázdnych odkazov – je percento nájdených odkazov, ku ktorým nemá byť priradená pomenovaná entita.[18]
- Presnosť nájdenia neprázdnych odkazov – je percento nájdených odkazov, ku ktorým má byť priradená pomenovaná entita.[30, 27]
- Výkon: $F_1 = \frac{2*precision*recall}{precision+recall}$ kde *precision* je presnosť a *recall* je presnosť nájdenia neprázdnych odkazov.[30, 27]

Kapitola 4

Vytvorenie vyhodnocovacej sady z výstupu projektu Wikilinks

Táto kapitola sa zaoberá použitím výstupu projektu Wikilinks výskumnej skupiny KNOT na vytvorenie vyhodnocovacej sady pre problémy rozpoznávania a zjednoznačňovania pomenovaných entít. Projekt Wikilinks používa komprimované webové dáta. Na účeli tohoto projektu boli použité dáta z Common Crawl[1]. Projekt Wikilinks z daných dát vydoluje odkazy na články Wikipédie.[5]

Z výstupu projektu Wikilinks a Common Crawl je možné generovať viacero typov vyhodnocovacích sád:

- Výstup projektu Wikilinks obsahuje kontext odkazu. Kontext odkazu je reprezentovaný zopár slovami pred a po nájdenom odkaze. Kontext odkazu poskytovaný výstupom Wikilinks môže byť použitý na vytvorenie jednotiek vyhodnocovacej sady. Takáto vyhodnocovacia sada nebude vhodná na testovanie extrakcie pomenovanej entity, ale bude dobrá na testovanie algoritmov zjednoznačňovania pomenovanej entity používajúcich kontext k rozhodovaniu medzi kandidátmi. Algoritmy na zjednoznačňovanie pomenovaných entít, ktoré budú používať iba odkaz bez kontextu, fungovať budú, ale pri takto vytvorenej vyhodnocovacej sade nebudú efektívne. Algoritmy na zjednoznačňovanie pomenovaných entít, ktoré používajú iné aspekty dokumentu ako ostatné nájdené pomenované entity, fungovať nebudú.

- Kontext získaný aplikáciou projektu Wikilinks nie je dostatočný pre testovanie všetkých aspektov rozpoznávania a zjednoznačňovania pomenovaných entít, ale vďaka dobrej konštrukcii projektu Wikilinks zdroj odkazu sa dá vystopovať. Zo zdroja zaujímavého odkazu sa dá získať celý paragraf, v ktorom sa vyskytuje zaujímavý odkaz. Pridaním väčšieho množstva textu sa zaujímavý odkaz hľadá ťažšie, a preto sa kladie väčšie bremeno na extrakciu a správne rozpoznanie pomenovanej entity. Väčšina algoritmov na rozpoznanie a zjednoznačňovanie pomenovaných entít by mala byť schopná takúto situáciu spracovať, ak vytvorený príklad nie je chybný. Pri vytváraní jednotky vyhodnocovacej sady týmto spôsobom začne byť problematická relevancia celého textu.

- Jedno z praktických použití rozpoznávania a zjednoznačňovania pomenovaných entít je premieňanie Web of documents na Web of data. V tej situácií je vstup celý

[1] https://commoncrawl.org/

dokument a výstup je jeho zaradenie do siete. Vyhodnocovacia sada, ktorá má túto situáciu čo najbližšie reprezentovať, musí obsahovať celé dokumenty. Na to aby sa dal program na vyhodnocovacej sade testovať je potrebné hľadané odkazy vyčistiť. Takto upravené dokumenty simulujú reálnu situáciu a aj zložitý a komplexný problém.

- Na algoritmy zjednoznačňovania pomenovaných entít, ktoré používajú na zjednoznačnenie ostatné nájdené pomenované entity, sa dá vytvoriť zvlášť vyhodnocovacia sada. Namiesto kontextu bude obsahovať zoznam odkazov na pomenované entity. Odstránenie redundancie bude iné ako pri kontextovom prístupe. Môže sa stať že jedna stránka odkazuje rovnakým slovom na dve rôzne pomenované entity. Testy vytvorené z takejto stránky budú očakávať rôzne odpovede na takmer rovnaký vstup.

V projekte sa budem zaoberať najme vyhodnocovacou sadou používajúcou krátky kontext. Väčšina volne dostupných programov na rozpoznávanie a zjednoznačňovanie pomenovaných entít obsahuje práve algoritmy, ktoré sa pozerajú na kontext okolo slova.

4.1 Získanie reprezentujúcich odkazov

Projekt Wikilinks vydoluje zo stránok takmer všetky odkazy na stránky Wikipédie, ale nie všetky odkazy reprezentujú pomenované entity. Pri reprezentujúcom odkaze musí byť možné priradiť správny článok Wikipédie (pomenovanú entitu) iba zo znalosťou textu bez meta-dát. Príklad jednoduchého nereprezentujúceho odkazu je „[číslo]“ alebo „wiki“. Nereprezentujúce odkazy nedokážu byť extrahované alebo rozpoznané, a preto nie sú použiteľné pre účeli rozpoznávania a zjednoznačňovania pomenovaných entít. To ich robí nevhodné pre vyhodnocovaciu sadu, a preto musia byť odfiltrované.

Reprezentujúci odkaz musí spĺňať niekoľko podmienok. Sú dva typy podmienok: jedny odstraňujú podozrivé odkazy a druhé vyhľadávajú relevantné odkazy.

- Nesmie byť príliš krátky a ani príliš dlhý. Príliš krátke odkazy budú veľmi ťažko reprezentovať pomenovanú entitu. Príliš dlhé odkazy neodkazujú na pomenovanú entitu menom entity, ale vetou o entite. Veta o entite sa bude ťažko extrahovať a to nebude prispievať na vyhodnocovanie rozpoznávania a zjednoznačňovania pomenovaných entít. Pretože sa zaujímam aj o skratky, tak je minimálna dĺžka odkazu nastavená na dva znaky. Maximálna dĺžka odkazu je nastavená ako 120% dĺžky mena entity.

- Na to aby odkaz reprezentoval entitu, tak musí byť podobný menu entity. Podobnosť reťazcov sa dá merať viacerými algoritmami. Ako prvé musí odkaz obsahovať aspoň jedno slovo z mena entity. Tento spôsob je dosť dobrý a rýchly na odstránenie veľkého množstva irelevantných odkazov. Napríklad odkaz „here“ na Wikipédia stránku „Python_(programming_language)“, ktorý stránku nereprezentuje. Ale odkaz „party“ na Wikipédia stánku „The Party of Moderate Progress Within the Bounds of the Law“ takýmto filtrom prejde, a preto je vhodný rozšíriť o striktnejšie filtrovanie.

- Skratky sú vhodné meno entity, ale neprejdú predchádzajúcim filtrom založením na slovách. Je potrebné identifikovať skratky entít a cez filter ich prepustiť. Validné skratky môžu obsahovať iba jedno slovo. Ak link obsahuje iba jedno slovo, tak z mena entity sa vytvorí skratka a následne sa bude s textom odkazu porovnaná. Na vytvorenie skratky sa meno entity rozdelí na slová. Skratka sa vytvorí z prvých písmen slov entity, ale nie všetky prvé písmená slov sú vhodné pre skratku. Budú vytvorené dve skratky. Jedna bude vytvorená len z písmen a druhá bude vytvorená z písmen a aj

čísel. Ak sa link bude zhodovať na 90% s vytvorenými skratkami, tak bude braný ako správna skratka a filter ju prepustí.

4.2 Získanie relevantných odkazov

Všetky reprezentujúce odkazy by malo byť možné extrahovať, rozpoznať a zjednoznačniť, ale nie všetky sú vhodné do vyhodnocovacej sady. Na relevantné odkazy budú kladené ďalšie podmienky, ktoré ich budú robiť lepšie pre vyhodnocovaciu sadu.

- Vyhodnocovacia sada musí byť vytvorená na riešenie pokročilých problémov rozpoznávania a zjednoznačňovania pomenovaných entít, a preto by vyhodnocovacia sada nemala obsahovať jednoznačné príklady rozpoznávania a zjednoznačňovania pomenovaných entít. Na to aby bol príklad rozpoznávania a zjednoznačňovania pomenovaných entít nejednoznačný musí sa o odkaz zaujímať viacero pomenovaných entít. O odkaz sa musia zaujímať aspoň dve pomenované entity z dátovej sady, ale ak sa o odkaz zaujíma príliš veľa pomenovaných entít, tak je podozrivý na zlú reprezentáciu pomenovanej entity, ako napríklad odkaz „Political party“.

- Na zjednoznačnenie pomenovanej je potrebný kontext odkazu, a preto nedostatky v kontexte môže príklad urobiť nevhodný na priradenie do vyhodnocovacej sady. Ak je kontext okolo odkazu príliš krátky, tak zjednoznačnenie použitím takéhoto kontextu bude slabé.

  Projekt Wikilinks doluje odkazy z celej stránky, ale pre rozpoznávanie a zjednoznačňovanie pomenovaných entít sú najdôležitejšie odkazy nachádzajúce sa v prirodzenom texte. Odkazy z menu, hlavičiek alebo reklamy nie sú vhodné do vyhodnocovacej sady. Najlepší kontext sa skladá z prirodzených slov, a preto aspoň 70% kontextu sa musí skladať z malých písmen abecedy. Takýto prístup zároveň odstráni niektoré cudzie jazyky, ako napríklad ruština alebo čínština. Ak kontext obsahuje URL odkaz, tak je podozrenie že sa nejedná o prirodzený text, ale len zoskupenie odkazov. Počiatočný výstup projektu Wikilinks je dostatočne veľký aby boli odstránené všetky podozrivé prípady a vyhodnocovacia sada bude stále veľká.

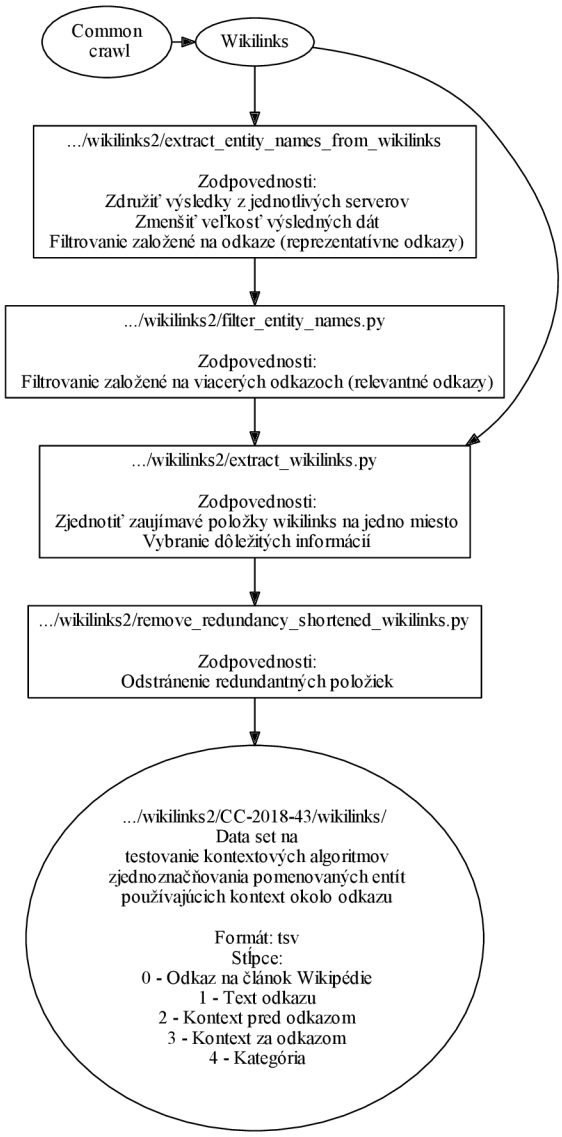
4.3 Manipulovanie textu

To že text má nejakú formu neznamená že vo vyhodnocovacej sade musí byť použitý v tej forme, ale jeho časti môžu byť podľa potreby upravované. Texty, ktoré sú podozrivé a cez filter neprešli, môžu byť upravené tak aby vyhovovali požiadavkom na vyhodnocovaciu sadu. Takýmto spôsobom sa dá získať väčšia vyhodnocovacia sada z rovnakého dátového základu. Nevýhodou tejto metódy je kvalita manipulovaných textov. Pretože manipulované texty vytvára počítač a nie človek, môžu byť čudné z pohľadu človeka. Pri manipulácií textu je overenie správnosti vyhodnocovacej sady dôležitejšie než len pri filtrovaní vyhodnocovacej. Pri manipulovaných textoch by mali byť výsledky Crowdsourcingu brané vážnejšie než pri ostatných textoch. Ale manipulovanie textu som nakoniec vo vyhodnocovacej sade nepoužil, pretože bola dostatočne len s filtrovanými textami.

- Zlepšenie reprezentatívnosti odkazu sa dá zariadiť manipulovaním textu. Nereprezentatívny odkaz môže byť nahradený odkazom, ktorý už reprezentatívny bude. Reprezentatívny odkaz sa dá získať z mena entity alebo môže byť použitý iný reprezenta-

tívny odkaz z vyhodnocovacej sady. Nahradený odkaz nemusí vždy presne pasovať do kontextu.

- Zlepšenie relevantnosti odkazu sa dá zariadiť manipulovaním textu. Ak je odkaz jednoznačný, tak môže byť nahradený nejednoznačným odkazom. Napríklad URL na pomenovanú entity môže byť nahradený skráteným menom entity. Skrátené meno entity nezahŕňa zjednoznačňovacie prvky. Upravovanie kontextu je zložitejšie ako nahradenie odkazu.



Obr. 4.1: Proces vytvorenia vyhodnocovacej sady z výstupu projektu Wikilinks.

Kapitola 5

Crowdsourcing a jeho použitie pri vytváraní vyhodnocovacej sady

Crowdsourcing je zložený zo slov crowd (dav) a source (zdroj). Ide o systém kde stroj na pozadí spolupracuje s veľkým davom ľudí na vyriešení problému. Aby bol problém riešiteľný pomocou Crowdsourcingu by mal byť veľký a rozdeliteľný na veľké množstvo krátkych, jednoduchých a nezávislých častí. Na malé problémy nie je potrebný dav. Problém riešený pomocou Crowdsourcingu by mal byť časovo náročný ale logicky jednoduchý. V dave nie je dosť expertov na riešenie logicky zložitých problémov alebo úloh. Člen davu by mal mať potrebné znalosti na vyriešenie úlohy alebo by ich získanie nemalo trvať dlhšie ako desať minút. Výhoda davu je veľký paralelizmus. Veľké množstvo krátkych a nezávislých úloh lepšie využíva paralelizmus Crowdsourcingu. Problém Crowdsourcingu je slabá garancia spoľahlivosti výsledkov úloh. Ak výsledky jednej úlohy budú závislé od výsledkov druhej úlohy a v nich sa nájde chyba, tak sa bude propagovať systémom a môže znehodnotiť veľké množstvo získaných dát. Keď sa pri propagácií systémom ešte zväčší závislé úlohy nemusia byť riešiteľné. Pretože správne výsledky nie sú garantované, úlohy by mali byť vypočítané davom viac krát. Výsledky úloh je potrebné analyzovať. Nesprávne výsledky by mali byť vyradené a správne výsledky je potrebné spojiť do výsledku s najväčšou možnou kvalitou. Spracovanie výsledkov by malo byť možné vykonať strojom na pozadí Crowdsourcingu. Na spracovanie výsledkov Crowdsourcingu by nemalo byť potrebné vyrobiť ďalšiu Crowdsourcing úlohu.

Crowdsourcing musí brať do úvahy aj motiváciu davu. Dav by nemal byť motivovaný riešiť celý problém, ale motivácia davu by mala byť stavaná na jednotlivých úlohách. Jednotlivé úlohy by mali byť krátke práve kvôli zachovaniu motivácie. Je oveľa ťažšie motivovať dav na riešenie dlhých alebo otravných problémov.

Rozpoznávanie a zjednoznačňovanie pomenovaných entít je jednoduchšie pre človeka ako pre stroj, a preto je Crowdsourcing dobrý nástroj na podporu pri vytváraní vyhodnocovacej sady. Dáta poskytované Crowdsourcingom by mali viesť k vylepšeniu alebo zväčšeniu vyhodnocovacej sady.

5.1 Typy aplikácií Crowdsourcingu

Crowdsourcing má rôzne typy aplikácií závislé na jeho atribútoch. Pri Crowdsourcingu je potrebné rozpoznávať zložitosť úlohy, motivácia davu a vedľajšie prínosy vykonania úlohy.

- Nútená úloha je spôsob Crowdsourcingu, ktorej sa človek nezúčastní dobrovoľne. Predpokladá sa že užívateľ nie je veľmi motivovaný na vyriešenie úlohy správne, a preto musia byť implementované kvalitné mechanizmy na kontrolu údajou získaných touto metódou. Úloha musí byť jednoduchá a rýchlo vykonatelná. Ešte dôležitejší je vedľajší prínos vykonania úlohy, ktorý by mal byť okamžitý. Vedlajší prínos vykonania úlohy nemusí byť užitočný pre užívateľa, ktorý úlohu vykonáva, ale musí byť užitočný pre užívateľa, ktorý donútil vykonanie úlohy. Parametre nútenej úlohy spĺňa overenie že uživateľ je človek a nie stroj reCAPTCHA[6].

- Užitočná úloha je spôsob Crowdsourcingu, kde vykonanie úlohy musí byť užitočné pre užívateľa. Úloha môže byť aj komplexná, ale jej vykonanie musí mať vedľajšie prínosy pre užívateľa, ktorý úlohu vykonáva. Je v záujme užívateľa aby úlohu vyriešil správne, a preto sa predpokladá väčšia motivácia užívateľou. Parametre užitočnej úlohy spĺňa aplikácia na učenie jazyka Duolingo[6].

- Hra je spôsob Crowdsourcingu, ktorej sa užívateľ účastní dobrovoľne. Hra musí byť pre užívateľa zaujímavá alebo zábavná. Zaujímavosť a zábavnosť hry je motiváciou pre užívateľa na riešenie úlohy, ale užívateľ nemusí byť motivovaný úlohu vyriešiť úplne správne.

- Platená úloha je spôsob Crowdsourcingu, ktorej sa užívateľ zúčastňuje aby niečo získal. Motivácia je najčastejšie finančná. Takáto úloha nemusí mať žiadne vedľajšie prínosy alebo efekty. Motivácia správneho vyplnenia je priamoúmerná nástrojom na overenie správnosti vyplnenie. K úlohám na vyriešenie sa pridávajú úlohy, ktorých riešenie je už známe. Ak užívateľ vyplní známe úlohy správne, tak dostane odmenu. Parametre platenej úlohy spĺňa platforma Mechanical Turk[18].

5.2 Príklady Crowdsourcingu

Na základe vytvorenia vyhodnocovacej sady z výstupu projektu Wikilinks sa dá vytváranie vyhodnocovacej sady rozdeliť na nasledujúce kroky:

- Hľadanie reprezentatívnych odkazov a to je zároveň hľadanie validných mien pomenovaných entít.

- Hľadanie relevantných odkazov a to sú odkazy pokročilé problémy rozpoznávania a zjednoznačňovania pomenovaných entít.

- Hľadanie nezmyselných prvkov vyhodnocovacej sady. Prvok môže byť nezmyselný vďaka zlému kontextu. Prvky vyhodnocovacej sady, ktoré vznikli manipuláciou textu, sú náchylné na nezmyselnosť z pohľadu človeka.

- Hľadanie chybných prvkov vyhodnocovacej sady. Vstupné dáta môžu byť zlej alebo neznámej kvality, a preto niektoré odkazy môžu odkazovať na zlú entitu.

Návrhy aplikácií Crowdsourcingu sa budú týkať jedného alebo viacerých krokov a tým zväčšovať kvalitu vyhodnocovacej sady. Jedna Crowdsourcing aplikácia pravdepodobne nebude stačiť na pokrytie všetkých krokov vytvorenia vyhodnocovacej sady.

Aplikovanie Web of data

Web of data je jeden z výsledkov aplikovania rozpoznania a zjednoznačňovania pomenovaných entít. Ide o zoskupenie dokumentov a dát podľa pomenovaných entít. Zoskupenie dokumentov podľa pomenovaných entít sa dá použiť na zoskupenie dokumentov podľa ich témy. Programy na rozpoznanie a zjednoznačňovanie pomenovaných entít nie sú dosť dobré na prevedenie celého Web of documents na Web of data. Web of data má v niektorých prípadoch lepšie vlastnosti ako Web of documents? Web of data môže byť vhodný na vyhľadávanie dokumentov obsahujúcich dáta o špecifických entitách. Bežné vyhľadávače dokumentov nie sú dobré na rozpoznávanie medzi nejednoznačnými pomenovanými entitami. Napríklad ak užívateľ vyhľadáva informácie o LaTeXu ako nástroju na sadzbu textových dokumentov a nájde články o gume.

V tejto sekcii sa budem zameriavať na aplikáciu Web of data menšieho rozsahu a zároveň by aplikáciu mala byť užitočná pre užívateľa a pomáhať pri tvorbe vyhodnocovacej sady. Užitočnosť pre užívateľa bude slúžiť ako motivácia zúčastňovať sa Crowdsourcingu. Táto aplikácia Crowdsourcingu bude typu užitočná úloha.

Aplikácia bude udržiavať sieť anotovaných dokumentov. Napríklad na podporu vytvárania diplomovej práce. Pri vytváraní diplomovej práce je potrebné preštudovať a použiť nájdenú literatúru. Do aplikácie môže užívateľ pridať nájdený dokument. Kombináciou automatickej anotácie programom na rozpoznanie a zjednoznačňovanie pomenovaných entít a manuálnej anotácie užívateľom môže byť dokument spracovaný na jednoduchšiu manipuláciu s nájdeným dokumentom v budúcnosti. Vďaka sieti dokumentov môže mať užívateľ lepší prehľad o dokumentoch použitých v diplomovej práci. Relevantné informácie z dokumentov v sieti budú pre užívateľa jednoduchšie dostupné.

Manuálna anotácia užívateľom dokáže opraviť nesprávne výsledky programu na rozpoznanie a zjednoznačňovanie pomenovaných entít. Opravy automatického spracovania užívateľom môžu byť aplikáciou skompilované a užívateľ ich môže darovať na podporu výskumu spracovania prirodzeného jazyka. Šanca že užívateľ dobrovoľne daruje dáta je v tejto situácií malá. Informácie sa dajú zbierať automaticky ale zároveň musí existovať prepínač na vypnutie automatického zasielania informácií. Šanca že užívateľ dobrovoľne daruje dáta na zlepšenie výskumu sa zvýši ak užívateľ bude profitovať z výskumu. Ak výskum bude poskytovať sieť dokumentov a informácií, z ktorých môže užívateľ čerpať, tak dobrovolné darovanie dát bude poskytovanú sieť vylepšovať a tým si užívateľ môže zjednodušiť získavanie zdrojov.

Nedobrovoľná úloha

Riešenie rozpoznania a zjednoznačňovania pomenovaných entít použitím počítača je nepresné. Robí to problém rozpoznania a zjednoznačňovania pomenovaných entít podobný konverzii obrázku na text alebo identifikácií objektov na obrázku. Problémy konverzie obrázku na text a identifikácie objektov na obrázku sa používajú na rozlišovanie medzi človekom alebo strojom a zároveň zbierajú dáta. Problém rozpoznania a zjednoznačňovania pomenovaných entít by malo ísť použiť podobným spôsobom.

Problém musí byť jednoduchý pre človeka a zložitý pre stroj. Nemalo by byť možné náhodou odhadnúť správny výsledok. Zároveň musia byť získané informácie na zväčšenie alebo vylepšenie vyhodnocovacej sady.

Pomenovaná entita musí byť rozšírená o popis, ktorý poskytne dosť informácií na jej rozlíšenie. Nemalo by byť potrebné vyhľadávať ďalšie informácie o pomenovanej entite.

Aby sa správny výsledok nedal náhodou odhadnúť, tak nemôže byť vyberaný jeden výsledok z viacerých možností. Na zhoršenie rozpoznávania strojom sa dá text previesť na obrázok so zašumením. Takto vytvorený obrázok sa bude strojom ťažko rozpoznávať a znemožní odhadovanie podporované algoritmom na rozpoznanie a zjednoznačňovanie pomenovaných entít.

Problém sa dá stavať dvoma spôsobmi:

- K textu priradiť pomenovanú entitu. Pomenovanú entitu bude potrebné vybrať z možností. Táto metóda je náchylná na náhodné vyberanie, a preto ju bude potrebné vykonať niekoľko krát. Úloha by mohla mať štyri príklady so známymi riešeniami a jeden bez známeho riešenia.

- K pomenovanej entite priradiť texty. V úlohe bude iba jedna pomenovaná entita. Bude prezentovaných niekoľko textov. Užívateľ musí vybrať, ktoré texty patria k pomenovanej entite. Pri niektorých textoch je známe že patria k pomenovanej entite, pri iných textoch je známe že nepatria k pomenovanej entite a pri jednom alebo dvoch textoch je príslušnosť k pomenovanej entite neznáma. Problém bude vybrať množstvo textu, ktoré bude užívateľovi zobrazené. Pri príliš veľkom množstve textu sa môže zmestiť na obrazovku iba jeden text. Táto metóda je odolnejšia proti náhodnému vyberaniu.

Pri pokuse o vytvorení rozpoznávania medzi človekom a strojom použitím rozpoznávania a zjednoznačňovania pomenovaných entít som našiel problémy, ktoré ho robia nevhodné na praktické použitie pri rozpoznávaní medzi človekom a strojom. Pri porovnaní s rozpoznávaním objektov na obrázku sa dajú jednoducho rozpoznať.

- Človek číta text sekvenčne slovo po slove a text po texte. Na to aby sa dalo spoľahnúť na rozpoznanie medzi človekom a strojom pomocou rozpoznávania a zjednoznačňovania pomenovaných entít je potrebné viacero textov. Čítanie prezentovaných textov aj s ich porozumením môže zabrať dve minúty až päť minút. Človek rozpoznáva obrázky paralelne. Hľadanie objektu na obrázku je oveľa rýchlejšie ako čítanie textu. Rozpoznávanie obrázkov použitých pri takýchto aplikáciách trvá menej ako jednu minútu. Toto pravidlo neplatí pre všetky obrázky. Existujú hry, kde človek hľadá na obrázku objekt. Takéto hry môžu trvať niekoľko minút a v niektorých zložitejších prípadoch aj hodín.

- Text dokáže zaberať viac priestoru než si človek než sa dá predpokladať. Pri dostatočnej veľkosti písma sa nemusia všetky texty zmestiť na malú obrazovku. Obrázky dokážu mať oproti textom väčšiu hustotu informácií, a preto sa dokážu zmestiť na menšiu plochu než ekvivalentné množstvo textu.

Vytváranie krížoviek a osemsmeroviek

Krížovky a osemsmerovky sú hry, ktoré používajú prirodzený jazyk. Na vytvorenie krížoviek a osemsmeroviek je potrebný slovník slov a k nim priradené prvky podľa ktorých sa dajú identifikovať. Slová v slovníku môžu byť validné mená pomenovaných entít a k nim priradený identifikátor ich môže jednoznačne priraďovať k pomenovanej entite. Konštrukcia slovníka môže byť vytvorená tak aby umožňovala vyhľadanie pomenovaných entít a ich validných mien. Slovník by mohol byť tvorený na štýl Wikipédie. Dostupný pre každého užívateľa a ktokoľvek by mohol pridávať nové slová alebo slová hodnotiť ako nekvalitné.

Ako základný prvok slovníka by mohla byť jednoznačná entita. K jednoznačnej entite sa budú viazať identifikátory a validné mená, ktoré ju môžu reprezentovať v krížovke,

osemsmerovke alebo texte. V prípade krížovky identifikátor bude krátky popis. V prípade osemsmerovky identifikátor môže byť obrázok. V krížovke a občas aj v osemsmerovke sa môžu vyskytovať aj slová, ktoré nie sú entity. Napríklad slovesá alebo spojky.

Časť slovníka môže byť napojená na Wikidata[1]. Wikidata obsahuje dáta o entitách. Entity sú jednoznačné a zároveň obsahuje aj krátky popis entity a pri niektorých obsahuje aj obrázok. Neobsahuje slovesá alebo spojky. Zdrojov použitého slovníka môže byť viac. Entity môže byť získané z Wikidata. Slovesá a spojky z inej databáze.

Na vytvorenie vyhodnocovacej sady je potrebné rozlišovať medzi pomenovanými entitami a entitami bez mien. Na to by bolo potrebné entity zaradiť do kategórií. Kategórie môžu byť použiteľné na anotovanie pomenovaných entít vo vyhodnocovacej sade. Aj pri tvorbe vyhodnocovacej sady sa pomenované entity môžu deliť na osoby, miesta alebo technológie.

5.3 Hra používajúca vyhodnocovaciu sadu

Na podporu tvorby vyhodnocovacej sady bola vytvorená hra. Pri tvorbe vyhodnocovacej sady nie je problém že je vyhodnocovacia sada malá, a preto nie je potrebné pomocou Crowdsourcingu pridávať príklady na rozpoznávanie a zjednoznačňovanie pomenovaných entít do vyhodnocovacej sady. Problém je že vyhodnocovacia sada obsahuje prvky, ktoré by v nej byť nemali. Cieľom Crowdsourcingu je vyčistiť vyhodnocovaciu sadu od nevhodných prvkov, a preto vytvorená hrabude hodnotiť prvky vyhodnocovacej sady a podľa hodnotenia prvkov bude možné nájsť prvky nevhodné do vyhodnocovacej sady. Na ohodnotenie prvku vyhodnocovacej sady je potrebné aby sa človek pokúsil rozpoznať a zjednoznačneniť prvok.

Hra bude modelovať ako súťaž medzi viacerými hráčmi. Hráči budú musieť rozpoznať a zjednoznačniť prvok vyhodnocovacej sady. Ak prvok vyhodnocovacej sady rozpozná správne, tak bude odmenený. Ak prvok vyhodnocovacej sady rozpozná nesprávne, tak bude odmenu nedostane alebo bude potrestaný. Dobré rozpoznávanie a zjednoznačňovanie pomenovaných entít dá hráčovi výhodu nad ostatnými hráčmi a zvýši jeho šance na výhru.

Pri mojej vyhodnocovacej sade pri rozpoznávaní a zjednoznačňovaní pomenovaných entít ide o priradenie odkazu s kontextom k článku Wikipédie. Odkaz s kontextom je reprezentovaný ako karta, kde odkaz je špeciálne označený vyznačený. Článok Wikipédie je reprezentovaný URL. URL na článok Wikipédie nestačí na rozpoznávanie a zjednoznačňovanie pomenovaných entít nestačí a je rozšírená o popis pomenovanej entity získaný z Wikipédie. Popisy všetkých URL sa nezmestia naraz na obrazovku, a preto sa zobrazí v na to určenej časti obrazovky popis URL článku nad ktorou prešla myš.

URL na články Wikipédie sú v tabuľke, ktorá je umiestnená v centre obrazovky. Tabuľka s URL na články Wikipédie bude zdielaná pre všetkých hráčov. Hra má limitovaný počet kariet, ktoré sa skladajú z odkazov s kontextom. Každý hráč má vlastnú ruku kariet, ktorá im je na začiatku hry priradená. Hráč má karty zobrazené na spodnej časti obrazovky. Cieľom hráča je zbaviť sa kariet vo svojej ruke. Prvý hráč, ktorý sa zbaví všetkých svojich kariet, hru vyhrá a tým ju ukončí pre všetkých hráčov.

Hráči sa striedajú postupne do kruhu striedajú v ťahoch. Keď je hráč na ťahu má dve možnosti:

- Hráč sa môže z tabuľky vybrať jeden článkov Wikipédie. Ak hráč vlastní karty, ktoré patria k vybranému článku Wikipédie, tak budú z jeho ruky odstránené. Ak hráč nemá ani jednu kartu v ruke, ktorá patrí vybranému článku, tak mu ako penalta pribudne

[1] www.wikidata.org

jedna ešte hráčovi nepriradená karta. Vybrať jeden z článkov Wikipédie je jediná možnosť ako si hráč môže zmenšiť počet kariet v ruke, ale zároveň hráč riskuje získanie karty. Rozpoznávanie a zjednoznačňovanie pomenovaných entít zväčšuje šancu že hráč sa karty zbaví a novú kartu nezíska.

- Hráč môže zahodiť kartu. Predpokladám že vyhodnocovacia sada nie je perfektná a obsahuje prvky, ktoré sa nedajú rozpoznať a zjednoznačniť. Zahodením karty sa hráč dokáže zbaviť vybranej karty, ale namiesto nej dostane novú kartu. Nezmení sa počet kariet v hráčovej ruke, a preto sa hráč ne nepriblíži bližšie k víťazstvu. Karta sa nedá zahodiť keď už nie sú volné karty na priradenie, ale vtedy aj keď si hráč odhadne článok Wikipédie a nebol odhadnutý správne novú kartu nedostane, pretože nie sú volné karty na priradenie. Zahodenie karty je zaujímavé z pohľadu Crowdsourcingu, pretože užívateľ označil kartu na dostatožne zlú že sa neoplatí rozpoznať a zjednoznačniť a tým riskovať potiahnutie novej karty.

Hráč má dve strategické možnosti s troma možnými výsledkami a pre Crowdsourcing sa zaznamenávajú tri parametre pre každý prvok. Koľko krát bol prvok správne rozpoznaný a zjednoznačnený, koľko krát bol nesprávne rozpoznaný a zjednoznačnený a koľko krát bola reprezentujúca prvok zahodená.

- Ak má prvok vyhodnocovacej sady veľa správnych rozpoznaní a zjednoznačnení, tak sa dá predpokladať že je ľahko rozpoznateľný a zjednoznačnitelný. Napríklad text: „it is driven directly from the engine's crankshaft via a belt or, in a two-stroke diesel engine, by spur“ kde „belt“ je odkaz na en.wikipedia.org/wiki/Belt_(mechanical)

- Ak má prvok vyhodnocovacej sady veľa nesprávnych rozpoznaní a zjednoznačnení, tak sa dá predpokladať že sa ťažko rozpoznáva a zjednoznačňuje alebo je v ňom chyba. Napríklad text: „contract to cover a particular planned event such as a wedding or graduation, or to illustrate an advertisement.“ kde „wedding“ je odkaz na en.wikipedia.org/wiki/Wedding_photography

- Ak má prvok vyhodnocovacej sady zahodenia, tak sa dá predpokladať že v ňom môže byť chyba alebo je v cudzom jazyku. Napríklad text: „Tina Turner in comeback-levyn nimibiisin 1980-luvun puolessa välissä, sain siitä Dire Straits-vibat. Syy tähän selkeni, kun kävi ilmi, että kappale ei“ kde „Dire Straits“ je odkaz na en.wikipedia.com/wiki/Dire_Straits

Kapitola 6

Aplikovanie vyhodnocovacej sady

Na zhodnotenie a testovanie vyhodnocovacej sady je najlepšie sa ju pokúsiť použiť. Sledovaním výsledkov programov na rozpoznanie a zjednoznačnenie pomenovaných entít sa dajú skúmať vlastnosti vyhodnocovacej sady, ale dajú sa skúmať aj vlastnosti programu na rozpoznávanie a zjednoznačňovanie pomenovaných entít.

Spojazdniť programy na rozpoznávanie a zjednoznačňovanie pomenovaných entít je zložité. Na väčšinu algoritmov, ktoré riešia rozpoznávanie a zjednoznačňovanie pomenovaných entít, je potrebná špecializovaná databáza. Ak sa jedná iba o rozpoznávanie a zjednoznačňovanie pomenovaných entít týkajúcich sa malého množstva tém, tak potrebná databáza nemusí byť veľká, ale vytvorená databáza je všeobecná, a preto je databáza potrebná na rozpoznávanie a zjednoznačňovanie veľká. Pokúšal som sa spojazdniť program na rozpoznávanie a zjednoznačňovanie pomenovaných entít DeepType[1]. Na jeho spojazdnenie bežným spôsobom je potrebné minimálne jeden terabajt volného miesta na disku, ale toľko nemám. DeepType sa dá spojazdniť manuálne iba s päťsto gigabajtmi volného miesta na disku, ale trvá to oveľa dlhšie. Zároveň spojazdnenie DeepType nie je šetrné ani na RAM pamäť. Pri spracovaní dát na vytvorenie databáze, podľa ktorej sa bude rozhodovať pri rozpoznávaní a zjednoznačňovaní pomenovaných entít, potrebuje šestnásť až dvadsať gigabajtov RAM pamäte, ale toľko nemám. To sa dá obísť rozšírením RAM pamäte o miesto na pevnom disku, ale po šiestich dňoch neustáleho zapisovania do disku a čítania z disku disk vydal čudný zvuk a spracovanie zmrzlo. Aj keď je databáza hotová, tak vyhľadávanie v nej trvá dlho. Už chápem prečo v niektorých súťažiach týkajúcich sa rozpoznávania a zjednoznačňovania pomenovaných entít dodávajú organizátori zvlášť sadu na učenie programov.

6.1 Semantic Enrichment Component

Semantic Enrichment Component[2] je jeden z projektov výskumnej skupiny KNOT. Semantic Enrichment Component (zkráceně SEC) poskytuje služby pre sémantické obohatenie textu[4]. Jedna z poskytovaných služieb je rozpoznávanie a zjednoznačňovanie pomenovaných entít. Vďaka prístupu na servery výskumnej skupiny KNOT som sa nemusel pokúsiť spojazdniť databázu pre fungovanie Semantic Enrichment Component, ale všetky potrebné prvky už boli pripravené na jednom z výskumných serverov. Nevýhoda bola nemožnosť editovať programy patriace Semantic Enrichment Component a tým ich pripraviť na spracovanie mnou vytvorenej vyhodnocovacej sady. Semantic Enrichment Component na roz-

[1] https://github.com/openai/deeptype
[2] http://sec.fit.vutbr.cz/sec_cs.html

poznávanie a zjednoznačňovanie pomenovaných entít používa Python skript. Na použitie Semantic Enrichment Component som vytvoril vlastný Python skript, ktorý používa Python skript Semantic Enrichment Component ako knižnicu. V mnou vytvorenom Python skripte som schopný volať funkciu zo skriptu Semantic Enrichment Component, ktorá rozpozná a zjednoznační zadaný text, a zároveň jednoducho nastaviť parametre pre rozpoznávanie a zjednoznačňovanie pomenovaných entít. Skript na rozpoznávanie a zjednoznačňovanie pomenovaných entít bol nastavený tak aby vrátil viacero ohodnotených pomenovaných entít namiesto jednej najlepšej pomenovanej entity. Výsledky rozpoznávania a zjednoznačňovania sa porovnajú s očakávaním výsledkom, a tým sa zistí miera úspešnosti. Pri kontextovom rozpoznávaní a zjednozančňovaní pomenovaných entít použitím Semantic Enrichment Component rozpoznávam tri vlastnosti:

- Prvok vyhodnocovacej sady bol úspešne extrahovaný. Algoritmus na rozpoznávanie a zjednoznačňovanie pomenovaných entít dostane celý text, z ktorého musí extrahovať, rozpoznať a následne zjednoznačniť pomenovanú entitu. Nemá odkaz na pomenovanú entitu predom označený. Extrahovanie prvku vyhodnocovacej sady sa považuje za úspešné keď algoritmus na rozpoznávanie a zjednoznačňovanie pomenovaných entít vráti aspoň jednu entitu asociovanú na presný text odkazu. Znamená to že textom odkazu očakával pomenovanú entitu a to sa považuje za úspešnú extrakciu pomenovanej entity.

- Prvok vyhodnocovacej sady bol úspešne rozpoznaný. Algoritmus na rozpoznanie a zjednoznačňovanie pomenovaných entít je nastavený tak aby vrátil zoznam ohodnotených entít. Zoznam entít sa v rozpoznávaní a zjednoznačňovaní pomenovaných entít volá zoznám kandidátnych entít. Ak sa očakávaná pomenovaná entita nachádza v zozname kandidátnych entít, tak sa rozpoznanie prvku vyhodnocovacej sady považuje za úspešné. Pri rozpoznaní prvku vyhodnocovacej sady nezáleží na ohodnotení pomenovaných entít.

- Prvok vyhodnocovacej sady bol úspešne zjednoznačňený. Na to aby bol prvok vyhodnocovacej sady úspešne zjednoznačňený musí byť úspešne rozpoznaný. Algoritmus na rozpoznanie a zjednoznačňovanie pomenovaných entít vracia zoznam ohodnotených entít. Pomenovaná entita s najväčším ohodnotením z kandidátnych entít je zjednoznačňená entita. Ak je zjednoznačňená entita rovnaká ako očakávania entita, tak je prvok vyhodnocovacej sady úspešne zjednoznačňený.

Semantic Enrichment Component zároveň obsahuje aj popularitu entity, a preto sa dá použiť aj zjednoznačňovanie bez kontextu. Na zjednoznačňovanie bez kontextu stačí iba text odkazu. Na text odkazu vráti zoznam kandidátnych pomenovaných entít zoradený podľa popularity pomenovaných entít. Dôležité je poradie očakávanej entity v zozname kandidátnych pomenovaných entít, ak očakávaná entita v zozname kandidátnych pomenovaných entít je. Ak je očakávanej entity v zozname kandidátnych pomenovaných entít, tak o nej Semantic Enrichment Component vie a nie len že ju má v databáze, ale zároveň rozpoznáva jej odkaz v prvku vyhodnocovacej sady ako validné meno očakávanej pomenovanej entity v prvku vyhodnocovacej sady.

Z výsledkov sa dá zistiť že zjednoznačnenie bez kontextu má lepšiu úspešnosť ako zjednoznačnenie s kontextom. To môže byť spôsobené viacerými faktormi. Jeden z nich môže byť nesprávna extrakcia pomenovanej entity. Rozpoznávanie a zjednoznačnenie s kontextom si musí odkaz na pomenovanú entitu extrahovať sám, ale rozpoznávanie a zjednoznačnenie

Kategória	Počet prvkov	Bez kontextu		S kontextom	
		Rozpoznávanie	Zjednoznačňovanie	Rozpoznávanie	Zjednoznačňovanie
Všetky	3612511	37.51%	26.94%	32.39%	14.73%
Other	2034327	2.76%	2.19%	2.43%	1.67%
location	676735	77.68%	48.63%	78.85%	33.21%
person	330750	81.39%	65.16%	76.28%	35.02%
nationality location	176184	95.52%	70.55%	90.2%	38.93%
artist	126265	90.85%	76.0%	88.36%	41.77%
mythology	78543	88.01%	79.49%	0.0%	0.0%
event	56053	79.51%	72.92%	0.0%	0.0%
visual_art_medium	36400	60.62%	57.65%	23.48%	22.34%
mythology person	14984	90.98%	15.89%	67.76%	35.32%
event location	11969	94.68%	84.8%	5.48%	4.34%
artist mythology	8622	95.07%	4.6%	92.1%	21.9%
location museum	5232	92.16%	14.68%	73.83%	33.41%
art_period_movement	5208	92.09%	88.82%	85.77%	85.23%
artist group	4831	98.68%	16.21%	94.02%	46.24%
location mythology	4826	90.97%	22.92%	65.81%	34.98%
artist person	4535	86.53%	63.02%	86.62%	27.92%
location person	4405	80.95%	64.49%	77.57%	47.83%
artwork	3096	88.28%	74.97%	20.41%	13.7%
nationality location person	2643	60.76%	5.22%	58.08%	8.1%
group	2566	60.56%	60.13%	0.0%	0.0%
location visual_art_medium	2504	74.16%	13.34%	79.87%	47.68%
visual_art_form	2455	60.73%	60.73%	17.6%	17.6%
group person	2295	98.34%	6.67%	95.12%	28.19%
event nationality location	2102	97.57%	96.76%	0.0%	0.0%
artist location	1955	90.64%	70.49%	76.42%	61.59%
visual_art_genre	1930	67.56%	58.81%	15.28%	13.21%
museum	1117	76.9%	69.56%	0.0%	0.0%
visual_art_form visual_art_medium	1038	91.04%	0.87%	4.91%	0.19%
artwork location	1028	94.26%	38.23%	36.28%	33.56%

Tabuľka 6.1: Tabuľka úspešnosti Semantic Enrichment Component na mnou vytvorenej vyhodnocovacej sade. Obsahuje úspešnosť pre bez kontextové aj kontextové riešenie. Kategórie pod 1000 prvkov nie sú zobrazené.

Kategória	Všetky prvky	Neúspešná Extrakcia	Neúspešné Rozpoznanie	Neúspešné Zjednoznačnenie	Pozitívne Zjednoznačnenie
Všetky	3612511	6.3%	1.41%	10.41%	3.18%
Other	2034327	0.36%	0.05%	0.61%	0.32%
location	676735	6.85%	2.03%	21.87%	8.37%
person	330750	6.39%	2.72%	30.51%	4.06%
location nationality	176184	6.42%	0.07%	37.5%	10.46%
artist	126265	5.93%	0.82%	34.26%	2.93%
mythology	78543	68.63%	19.35%	0.0%	0.0%
event	56053	71.11%	8.38%	0.0%	0.0%
visual_art_medium	36400	37.5%	0.21%	1.05%	0.11%
mythology person	14984	17.32%	7.7%	2.7%	31.68%
location event	11969	80.86%	9.35%	0.0%	4.33%
mythology artist	8622	4.98%	0.28%	3.42%	20.78%
location museum	5232	14.18%	5.47%	8.77%	32.45%
art_period_movement	5208	8.39%	2.28%	0.0%	0.52%
group artist	4831	5.88%	0.48%	8.88%	39.93%
mythology location	4826	22.71%	6.05%	2.53%	33.73%
person artist	4535	5.8%	1.21%	37.51%	4.5%
location person	4405	5.15%	5.02%	17.8%	3.56%
artwork	3096	50.71%	17.83%	5.91%	1.26%
location person nationality	2643	3.82%	0.3%	1.21%	4.39%
group	2566	54.25%	6.27%	0.0%	0.0%
visual_art_medium location	2504	8.67%	1.6%	4.43%	39.14%
visual_art_form	2455	41.34%	2.85%	0.0%	0.0%
person group	2295	3.97%	0.78%	4.97%	26.97%
location event nationality	2102	81.59%	15.89%	0.0%	0.0%
location artist	1955	14.27%	3.63%	9.77%	4.5%
visual_art_genre	1930	53.94%	0.05%	2.02%	1.5%
museum	1117	40.02%	36.88%	0.0%	0.0%
visual_art_form visual_art_medium	1038	86.51%	0.1%	0.0%	0.1%
artwork location	1028	42.51%	17.51%	0.0%	32.3%

Tabuľka 6.2: Tabuľka analyzuje rozdiel medzi kontextovým a bez kontextovým Semantic Enrichment Component. Stĺpce sú podrobnejšie vysvetlené v texte.

bez kontextu dostane iba odkaz na pomenovanú entitu (Tabuľka 6.2 stĺpec neúspešná extrakcia). Druhý faktor môže byť nesprávne rozpoznanie pomenovanej entity. Dobrý systém na rozpoznávanie sa snaží získať pre zjednoznačňovanie čo najmenší zoznam kandidátnych pomenovaných entít. Pri snahe o čo uľahčenie práce prezjednoznačňovanie môže vypustiť správnu pomenovanú entitu zo zoznamu kandidátnych pomenovaných entít (Tabuľka 6.2 stĺpec neúspešné rozpoznanie). Kontext by mal spresniť zjednoznačnenie pomenovanej entity, ale najpopulárnejšia pomenovaná entita sa bude vyskytovať častejšie ako menej populárne pomenované entity. Kontext môže zlepšiť zlepšiť zjednoznačnenie, ak je menej populárna entita správne vybraná ako (Tabuľka 6.2 stĺpec pozitívne zjednoznačnenie), ale kontext môže zhoršiť zjednoznačnenie, ak nie je najpopulárnejšia entita správne vybraná (Tabuľka 6.2 stĺpec pozitívne zjednoznačnenie).

Na výsledky má najväčší vplyv kontextové zjednoznačňovanie. Kontextové zjednoznačňovanie má horšie výsledky ako zjednoznačňovanie podľa popularity. Druhý najväčší vplyv má extrakcia. Identifikovať odkaz na pomenovanú entitu v texte je ako hľadať ihlu v kope sena, ale podobnosť medzi textom a odkazom je väčšia ako podobnosť medzi ihlou a senom, a preto je to skôr ako hľadať drevenú ihlu v kope sena.

Kategorizácia pomenovaných entít

Semantic Enrichment Component obsahuje aj kategórie pre pomenované entity. Príklad kategórií sú napríklad: osoba, miesto alebo udalosť. Pomenovaná entita môže byť v dvoch kategóriách naraz. Pomenované entity v kategórií umelec sú takmer vždy aj v kategórií osoba. Ak má pomenovaná entita kategóriu, tak sa zvyšuje dôvera v danú pomenovanú entitu. Neprítomnosť kategórie dôveru nezníži, pretože databáza pre Semantic Enrichment Component nie je úplná. Ak pomenovaná entita nemá kategóriu, tak neznamená že k niektorej kategórií nepatrí. Na zlepšenie kategorizovania pomenovaných entít je vhodné použiť viacero zdrojov dát.

Pretože pomenované entity sú zároveň aj články z Wikipédie, články z Wikipédie môžu byť použité na získanie kategórií. Každý článok z Wikipédie má kategórie, do ktorých je priradený. Problém je v tom že dané kategórie sú vytvorené aby boli použité človekom a nie sú veľmi dobré na strojové použitie. To že je pomenovaná entita v nejakej kategórií neznamená že článok Wikipédie je v tej istej kategórií, ale väčšinou sa nachádza v niektorej podkategórií kategórie pomenovanej entity. Namiesto hľadania jednej kategórie článku Wikipédie sa môže hľadať viac než sto kategórií, ktoré sú podkategórie vyhľadávanej kategórie. Problém nastane keď nie všetky podkategórie niektorej kategórie sa dajú stále považovať za danú kategóriu. Na použitie kategórií článkov Wikipédie je potrebné ručné vetovanie viac ako tisíc kategórií.

Na strojové prácu s dátami z Wikipédie bola vytvorená stránka Wikidata. Stánka Wikidata je tvorená ako Wikipédia priateľská na strojové spracovanie. Zo stránky Wikidata sa kategórie získavajú jednoduchšie ako zo stránky Wikipédie, ale jednoduchšie neznamená jednoduché. Entity Wikidata majú stále problém z hľadaním podkategórií, ale už nie pri všetkých kategóriách. Kategória osoba je naozaj prítomná u všetkých osobách, ale kategória u všetkých miest nie je. U kategórií miesto je stále problém s hľadaním podkategórií, ale oproti kategóriám Wikipédie je stránka Wikidata konzistentnejšia a spojazdniť Wikidata na doplnenie kategorizovania pomocou Semantic Enrichment Component je možné. Aj keď je kategorizácia použitím stránky Wikidata možná nie je súčasťou diplomovej práce a používa sa iba kategorizácia pomocou Semantic Enrichment Component.

Kategorizáciou sa zvyšuje dôvera v pomenovanú entitu, a tým zároveň aj na prvok vyhodnocovacej sady obsahujúci danú pomenovanú entitu. Na zvyšovanie dôvery prvku vyhodnocovacej sady je použitý aj Crowdsourcing. Z pohľadu vylepšovania vyhodnocovacej sady je považujem Crowdsourcing lepší ako kategorizácia pomenovaných entít.

6.2 LingPipe

LingPipe je nástroj na spracovanie prirodzeného jazyka použitím výpočtovej lingvistiky. LingPipe môže byť použitý na úlohy ako:

- Nájsť mená ľudí, organizácií a miest v novinových článkoch.
- Automaticky klasifikovať výsledky vyhľadávania z Twittru do kategórií.
- Overiť pravopis slov v texte a odporučiť správne slová.

Jednou z funkcií systému LingPipe je rozpoznávanie a zjednoznačňovanie pomenovaných entít. LingPipe v skutočnosti poskytuje viacero algoritmov na rozpoznávanie a zjednoznačňovanie pomenovaných entít. Tri algoritmov z LingPipe sú všeobecné a dokážu byť trénované:

- CharLmHmmChunker: je založený na preformovaní dávkovania ako problém označovania (trochu bohatší a senzitívnejší na kontext ako štandardný interné BIO kódovanie). Model znakového jazyka HMM si potom poradí s anotovaním, použitím modelu znakového jazyka pre každú anotáciu (stav) v HMM s modelom maximálnej pravdepodobnosti bigramu. Dávkovač dokáže dodať prvý najlepší výsledok, n-najlepších výsledkov alebo ohodnotené výsledky. Dávkovač CharLmHmmChunker je najjednoduchší ale najmenej presný s dávkovačov. Je veľmi rýchly ak je zapnutý caching. Zárovaň má dobré vyvolanie kandidátnych entít a dostatočne presné ohodnotenie kandidátnych entít.
- CharLmRescoringChunker: je najpresnejší, ale zároveň najpomalší. Používa CharLmHmmChunker na vygenerovanie hypotéz, ktorým zmení skóre na základe longer-distance modelu znakového jazyka. CharLmRescoringChunker je veľmi pomalý, hlavne keď mení skóre veľkým listom n-najlepších výsledkov. Zároveň dokáže upravovať hodnotenie výsledkov predpokladaním dôvery pomocou zoznamu n-najlepších výsledkov, ale predpokladaná hodnota je závislá na modelu znakového jazyka.
- TokenShapeChunker: je LingPipe 1.0 dávkovať prerobený na LingPipe 2.0 rozhrania. Beží s generatívnym modelom spolu s predpovedaním ďalšieho žetónu založenom na značke predchádzajúcich dvoch žetónoch a predchádzajúcej značke. Neznáme slová sú nahradené s prvkami založenými na inštancii TokenCategorizer. TokenShapeChunker musí byť trénovaný použitím triedy TrainTokenShapeChunker. Je veľmi rýchly, ale poskytuje iba prvý najlepší výstup. Presnosť TokenShapeChunker je väčšinou uprostred medzi CharLmHmmChunker a CharLmRescoringChunker.

Preložené z [1] a [2]

LingPipe má už naučené testovacie príklady na dátových sadách MUC 6 Corpus (noviny), GeneTag Corpus (biomedicína) a Genia Corpus (biomedicína). Žiadna z už naučených

sád nie je vhodná na riešenie mnou vytvorenej vyhodnocovacej sady. Mnou vytvorená vyhodnocovacia sada obsahuje všeobecné dáta, a preto biomedicínske sady GeneTag Corpus a Genia Corpus nie sú vhodné na riešenie mojej vyhodnocovacej sady. Sada MUC 6 Corpus by mohla byť dosť všeobecná, ale iba kategorizuje odkazy na pomenované entity namiesto priradenia odkazu na pomenovanú entitu článku z Wikipédie. Na to aby bol LingPipe použiteľný na mnou vytvorenú vyhodnocovaciu sadu musím dávkovače naučiť.

LingPipe musím najskôr naučiť na nejakých dátach pred tým než sa dá rozpoznávať a zjednoznačňovať. LingPipe môžem učiť na mojej vyhodnocovacej sade pred tým než ho na tej istej sade použijem. LingPipe sa nedá učiť na celej vyhodnocovacej sade, pretože je príliš veľká. Je potrebné získať malú a zároveň reprezentatívnu časť vyhodnocovacej sady, ale to je dosť zložité. Netreba použiť jeden dávkovač na celú vyhodnocovaciu sadu, ale vyhodnocovacia sada sa dá rozdeliť na menšie celky, ktoré budú mať vlastný dávkovač. Celky je najlepšie rozdeliť podľa objektu rozpoznávania. Dávkovač jedného celku bude naučený na rozpoznávanie jedného odkazu, napríklad dávkovač bude naučený na odkazy „John Smith“ a práve tie odkazy bude rozpoznávať a zjednoznačňovať. Vďaka rozdelenie vyhodnocovacej sady podľa odkazov bude rozpoznanie a zjednoznačnenie rýchlejšie a zároveň bude mať veľmi dobrú šancu rozpoznať a zjednoznačniť pomenovanú entitu. Aj keď bude rozpoznanie a zjednoznačnenie rýchlejšie stále môže trvať viac než niekoľko týždňov až mesiac, a preto nie je vyhodnotenie LingPipe uskutočnené na celej vyhodnocovacej sade.

Úplné vyhodnotenie prešlo včas na 977996 prvkoch. CharLmHmmChunker úspešne rozpoznal a zjednoznačnil 962639 (98.42%) prvkov, CharLmRescoringChunker úspešne rozpoznal a zjednoznačnil 455501 (46.57%) prvkov a TokenShapeChunker úspešne rozpoznal a zjednoznačnil 455501 (46.57%) prvkov. Dávkovač CharLmHmmChunker má najlepšie skóre, aj napriek tomu že je najjednoduchší. Môže to byť spôsobené tým že je testovaný na rovnakých dátach ako je trénovaný. Ak sa dávkovače CharLmRescoringChunker a TokenShapeChunker snažia o lepšiu generalizáciu, tak budú mať pri situácií, kde sú testované na rovnakých dátach ako sú trénované, majú horšie skóre.

Kapitola 7

Vyhodnocovacie sady

Mnou vytvorená vyhodnocovacia sada obsahuje pomenovanú entitu ako odkaz na článok Wikipédie a text, ktorý je rozdelený na odkaz, kontext pred odkazom a kontext za odkazom. Táto kapitola sa zaoberá ďalšími vyhodnocovacími sadami, ktoré sa dajú použiť na rozpoznanie a zjednoznačnenie pomenovaných entít.

Mnou vytvorená vyhodnocovacia sada má 3612908 textov a 259521 pomenovaných entít, ktoré sa zaujímajú o 145649 odkazov.

MedTag a GeneTag

MedTag obsahuje anotované dáta GeneTag a MedPost. Ide o projekt organizácie National Center for Biotechnology Information. Projekt MedTag je podporovaný vládou Spojených Štátov amerických. Výsledný software a databáza sú volne dostupné na použitie verejnosťou. Štátna Knižnica pre Medicínu a Vláda Spojených Štátov Amerických neustanovili žiadne obmedzenia na použitie a kopírovanie databáz MedTag. Štátna Knižnica pre Medicínu a Vláda Spojených Štátov Amerických neručia za spoľahlivosť a presnosť dát projektu MedTag. [3]

GeneTag je vyhodnocovacia sada na rozpoznávanie a zjednoznačňovanie génov a proteínov. Vyhodnocovacia sada bola vytvorená automatizovaným anotovaním kombinovaným s manuálnym opravovaním anotácií. Vyhodnocovacia sada GeneTag sa skladá z viet. Ku každej vete aspoň jedna anotácia, ktorá označuje hlavné meno génu alebo proteínu. Ak má veta viacero anotácií, tak označujú alternatívne mená génu alebo proteínu. Ak program na rozpoznávanie a zjednoznačňovanie nájde alternatívne meno, tak môže dostať menej nodov než keď nájde celé meno génu alebo proteínu. Vyhodnocovacia sada GenTag obsahuje 20000 rozdelená do štyroch častí. Časť na trénovanie obsahujúca 7500 viet, časť na testovanie obsahujúca 2500 viet, časť prvé kolo obsahujúce 5000 viet a druhé kolo obsahujúce 5000 viet. [29]

- Vyhodnocovacia sada GeneTag je tematicky špecifická a mnou vytvorená vyhodnocovacia sada je všeobecná.
- Vyhodnocovacia sada GeneTag je oveľa menšia ako mnou vytvorená vyhodnocovacia sada.
- Vyhodnocovacia sada GeneTag je podrobnejšia ako mnou vytvorená sada.
- Vyhodnocovacia sada GeneTag je spoľahlivejšia ako mnou vytvorená vyhodnocovacia sada.

CoNLL-2003

CoNLL-2003 je vyhodnocovacia sada rozpoznávanie a zjednoznačňovanie všeobecných pomenovaných entít. Obsahuje prvky vo dvoch jazykoch angličtina a nemčina. Anotácie dát boli získané automatizovaným systémom podporeným manuálnym opravením človekom. Vyhodnocovacia sada je rozdelená na tri časti trénovacia sada obsahujúca 14987 viet, vývojová sada obsahujúca 3068 viet a testovacia sada obsahujúca 3684 viet. Pomenované entity sú rozdelené do štyroch kategórií: miesto, organizácia, osoba a ostatné. Sada je uložená systémom, kde každé slovo alebo špeciálny znak je na samostatnom riadku. Každé slovo má priradené tri anotácie. Jedna určuje typ slova (podstatné meno, sloveso, ...), druhá určuje dávkovú značku a tretia určuje či slovo patrí pomenovanej entite a jej kategóriu. [30]

- Vyhodnocovacia sada CoNLL-2003 je všeobecná rovnako ako mnou vytvorená vyhodnocovacia sada.
- Vyhodnocovacia sada CoNLL-2003 je oveľa menšia ako mnou vytvorená vyhodnocovacia sada.
- Vyhodnocovacia sada CoNLL-2003 je podrobnejšia ako mnou vytvorená sada.
- Vyhodnocovacia sada CoNLL-2003 je na rozdiel od mnou vytvorenej vyhodnocovacej sady neposkytuje odkazy na články Wikipédie.

WNUT2017

Vyhodnocovacia sada WNUT2017 sa zaoberá nezvyčajnými pomenovanými entitami. Ako zdroje dát boli použité Twitter, Reddit, YouTube a StackExchange. Prvky vyhodnocovacej sady sú v angličtine. Pomenované entity sú rozdelené do kategórií osoba, miesto, organizácia, produkt, tvorivá práca (pieseň, kniha, ...) a skupina. Vyhodnocovacia sada WNUT2017 je rozdelená na dve časti vývojová s 1008 dokumentmi a 835 pomenovanými entitami a testovacia s 1287 dokumentmi a 1070 pomenovanými entitami. Vo vyhodnocovacej sade je každé slovo je na samostatnom riadku a pre každé slovo je jedna anotácia, ktorá určuje či slovo patrí pomenovanej entite a jej kategóriu. [14]

- Vyhodnocovacia sada WNUT2017 je všeobecná rovnako ako mnou vytvorená vyhodnocovacia sada.
- Vyhodnocovacia sada WNUT2017 je oveľa menšia ako mnou vytvorená vyhodnocovacia sada.
- Vyhodnocovacia sada WNUT2017 je zložitá na rozpoznávanie a zjednoznačňovanie pomenovaných entít.
- Vyhodnocovacia sada WNUT2017 je na rozdiel od mnou vytvorenej vyhodnocovacej sady neposkytuje odkazy na články Wikipédie.

Kapitola 8

Záver

Práca sa zaoberá vytvorením vyhodnocovacej sady pre pokročilé problémy rozpoznania a zjednoznačňovania pomenovaných entít. Ako základ na získanie dát pre vytvorenie vyhodnocovacej sady som použil projekt Wikilinks výskumnej skupiny KNOT[5] a komprimovaný archív internetu Common Crawl[1].

Wikilinks som najskôr aplikoval na Common Crawl a tým som získal odkazy na Wikipédiu z reálnych webových stránok. Z odkazov je najskôr potrebné získať iba reprezentatívne odkazy. Za reprezentatívny odkaz považujem odkaz, ktorý obsahuje aspoň časť mena pomenovanej entity. Z reprezentatívnych odkazoch nasledovne získam relevantné odkazy. Relevantné odkazy sú vhodné do pokročilej vyhodnocovacej sady na rozpoznávanie a vyhodnocovanie pomenovaných entít. Relevantné odkazy majú vhodný kontext na zjednoznačňovanie a o odkaz sa musia zaoberať aspoň dve pomenované entity. Vďaka to že sa o odkaz zaujímajú aspoň dve pomenované entity, algoritmy naučené na vyhodnocovacej sade nemajú úplnú úspešnosť.

Vyhodnocovacia sada má 3612908 textov a 259521 pomenovaných entít, ktoré sa zaujímajú o 145649 odkazov. Viac ako tri a pol milióna prvkov robí vyhodnocovaciu sadu extrémne veľkú, ale nie každý prvok vyhodnocovacej sady je kvalitný. Crowdsourcing je potrebný na vyčistenie vyhodnocovacej sady od nekvalitných prvkov, a preto som vytvoril hru, ktorá vyhodnocovaciu sadu požíva. Pri hraní hry sú získavané informácie o prvkoch vyhodnocovacej sady. Žiaľ nemal som dostatok užívateľov na úplné vyhodnotenie crowdsourcingu, ale z vlastného testovania predpokladám jeho správnu funkčnosť.

Vyhodnocovaciu sadu som testoval na dvoch systémoch a piatich algoritmoch. Dva systémy sú Semantic Enrichment Component a LingPipe. Zo Semantic Enrichment Component som použil kontextový a bez kontextový algoritmus. Z LingPipe som použil tri algoritmy, ktoré sa dokážu učiť na príkladoch. Z môjho testovania majú jednoduchšia algoritmy lepšiu úspešnosť ako komplikovanejšie algoritmy.

[1] https://commoncrawl.org/

Literatúra

[1] *LingPipe Home*. [Online; naštívené 16.5.2019].
URL http://alias-i.com/lingpipe/

[2] *LingPipe: Named Entity Tutorial*. [Online; naštívené 16.5.2019].
URL http://alias-i.com/lingpipe/demos/tutorial/ne/read-me.html

[3] *MEDTAG: README*. [Online; naštívené 18.5.2019].
URL ftp://ftp.ncbi.nlm.nih.gov/pub/lsmith/MedTag/README

[4] *Semantic Enrichment Component*. [Online; naštívené 14.5.2019].
URL http://sec.fit.vutbr.cz/sec_cs.html

[5] *Wikilinks*. [Online; naštívené 10.1.2018].
URL http://knot.fit.vutbr.cz/wikilinks/wikilinks_cs.html

[6] von Ahn, L.: *Massive-scale online collaboration*. [Online; naštívené 7.1.2018].
URL https://www.ted.com/talks/luis_von_ahn_massive_scale_online_collaboration

[7] Bontcheva, K.; Derczynski, L.; Roberts, I.: Crowdsourcing named entity recognition and entity linking corpora. In *Handbook of Linguistic Annotation*, Springer, 2017, s. 875–892.

[8] Brabham, D. C.: Crowdsourcing as a model for problem solving: An introduction and cases. *Convergence*, ročník 14, č. 1, 2008: s. 75–90.

[9] Brito, J.; Vieira, V.; Duran, A.: Towards a framework for gamification design on crowdsourcing systems: the GAME approach. In *2015 12th International Conference on Information Technology-New Generations*, IEEE, 2015, s. 445–450.

[10] Bunescu, R.; Paşca, M.: Using encyclopedic knowledge for named entity disambiguation. In *11th conference of the European Chapter of the Association for Computational Linguistics*, 2006.

[11] 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.

[12] Demartini, G.; Difallah, D. E.; Cudré-Mauroux, P.: ZenCrowd: leveraging probabilistic reasoning and crowdsourcing techniques for large-scale entity linking. In *Proceedings of the 21st international conference on World Wide Web*, ACM, 2012, s. 469–478.

[13] Derczynski, L.; Maynard, D.; Rizzo, G.; aj.: Analysis of named entity recognition and linking for tweets. *Information Processing & Management*, ročník 51, č. 2, 2015: s. 32–49.

[14] Derczynski, L.; Nichols, E.; van Erp, M.; aj.: *Results of the WNUT2017 Shared Task onNovel and Emerging Entity Recognition. Proceedings of the Workshop on Noisy, User-generated Text, at EMNLP*, [Online; naštívené 19.5.2019].
URL http://noisy-text.github.io/2017/pdf/WNUT18.pdf

[15] Goncalves, J.; Hosio, S.; Ferreira, D.; aj.: Game of words: tagging places through crowdsourcing on public displays. In *Proceedings of the 2014 conference on Designing interactive systems*, ACM, 2014, s. 705–714.

[16] Good, B. M.; Loguercio, S.; Griffith, O. L.; aj.: The cure: design and evaluation of a crowdsourcing game for gene selection for breast cancer survival prediction. *JMIR Serious Games*, ročník 2, č. 2, 2014: str. e7.

[17] Hachey, B.; Radford, W.; Curran, J. R.: Graph-based named entity linking with wikipedia. In *International conference on web information systems engineering*, Springer, 2011, s. 213–226.

[18] Hachey, B.; Radford, W.; Nothman, J.; aj.: *Evaluating Entity Linking with Wikipedia.* [Online; naštívené 30.12.2018].
URL http://benhachey.info/pubs/hachey-aij12-evaluating.pdf

[19] Hachey, B.; Radford, W.; Nothman, J.; aj.: Evaluating entity linking with Wikipedia. *Artificial intelligence*, ročník 194, 2013: s. 130–150.

[20] Howe, J.: The rise of crowdsourcing. *Wired magazine*, ročník 14, č. 6, 2006: s. 1–4.

[21] Kittur, A.; Chi, E. H.; Suh, B.: Crowdsourcing user studies with Mechanical Turk. In *Proceedings of the SIGCHI conference on human factors in computing systems*, ACM, 2008, s. 453–456.

[22] Luo, G.; Huang, X.; Lin, C.-Y.; aj.: Joint entity recognition and disambiguation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, 2015, s. 879–888.

[23] Pastorek, P.: *Příprava vyhodnocovací sady pro složité problémy rozpoznávání a zjednoznačňování pojmenovaných entit pomocí crowdsourcingu. 2019, semestrální projekt. Vysoké učení technické v Brně, Fakulta informačních technologií. Vedoucí práce doc. RNDr. Pavel Smrž, Ph.D.*

[24] Sabou, M.; Bontcheva, K.; Derczynski, L.; aj.: Corpus Annotation through Crowdsourcing: Towards Best Practice Guidelines. In LREC*, 2014, s. 859–866.*

[25] Salk, C. F.; Sturn, T.; See, L.; aj.: Assessing quality of volunteer crowdsourcing contributions: lessons from the Cropland Capture game. International Journal of Digital Earth*, ročník 9, č. 4, 2016: s. 410–426.*

[26] Sayeed, A. B.; Meyer, T. J.; Nguyen, H. C.; aj.: Crowdsourcing the evaluation of a domain-adapted named entity recognition system. In Human language technologies: the 2010 annual conference of the North American chapter of the association for

computational linguistics, *Association for Computational Linguistics, 2010, s. 345–348.*

[27] Shen, W.; Wang, J.; Han, J.: Entity Linking with a Knowledge Base: Issues, Techniques, and Solutions. [Online; naštívené 30.12.2018].
URL http://dbgroup.cs.tsinghua.edu.cn/wangjy/papers/TKDE14-entitylinking.pdf

[28] Stern, R.; Sagot, B.; Béchet, F.: A joint named entity recognition and entity linking system. In Proceedings of the Workshop on Innovative Hybrid Approaches to the Processing of Textual Data, *Association for Computational Linguistics, 2012, s. 52–60.*

[29] Tanabe, L.; Xie, N.; Thom, L. H.; aj.: GENETAG: a tagged corpus for gene/protein named entity recognition. BMC Bioinformatics, *ročník 6, č. 1, May 2005: str. S3, ISSN 1471-2105, doi:10.1186/1471-2105-6-S1-S3.*
URL https://doi.org/10.1186/1471-2105-6-S1-S3

[30] Tjong, E. F.; Sang, K.; Meulder, F. D.: Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition. [Online; naštívené 30.12.2018].
URL http://www.aclweb.org/anthology/W03-0419

[31] Zhang, W.; Su, J.; Tan, C. L.; aj.: Entity linking leveraging: automatically generated annotation. In Proceedings of the 23rd International Conference on Computational Linguistics, *Association for Computational Linguistics, 2010, s. 1290–1298.*