Slovenská technická univerzita v Bratislave
Fakulta informatiky a informačných technológií

FIIT-5212-64362

Ondrej Kaššák

EXTRAKCIA POMENOVANÝCH ENTÍT ZO SLOVENSKÉHO TEXTU

Bakalárska práca

Študijný program: Informatika
Študijný odbor: 9.2.1 Informatika
Miesto vypracovania: Ústav Informatiky a softvérového inžinierstva, FIIT STU
Vedúci práce: Ing. Michal Kompan

máj 2012

Slovenská technická univerzita v Bratislave
Fakulta informatiky a informačných technológií

ZADANIE BAKALÁRSKEHO PROJEKTU

Meno študenta:	**Kaššák Ondrej**
Študijný odbor:	Informatika
Študijný program:	Informatika
Názov projektu:	**Extrakcia pomenovaných entít zo slovenského textu**

Zadanie:
Spracovanie textu hrá významnú úlohu pri metódach odporúčania založeného na obsahu alebo vyhľadávania. Analýza textu a extrakcia pomenovaných entít (identifikácia osôb, organizácií, dátumov a pod.) môže priniesť výrazné zlepšenie a možnosti rozšírenia napr. odporúčacích systémov. Navrhnite metódu extrakcie niektorých pomenovaných entít pre slovenský jazyk. Analyzujte existujúce prístupy aj pre iné jazyky a možnosť aplikovania týchto prístupov pre slovenčinu. Zaoberajte sa možnosťou využitia navrhnutej metódy aj pre iné jazyky. Experimentálne overte navrhnutú metódu na netriviálnej vzorke dát vo vybranej aplikačnej doméne (napr. digitálna knižnica).

Práca musí obsahovať:

Anotáciu v slovenskom a anglickom jazyku
Analýzu problému
Opis riešenia
Zhodnotenie
Technickú dokumentáciu
Zoznam použitej literatúry
Elektronické médium obsahujúce vytvorený produkt spolu s dokumentáciou

Miesto vypracovania: Ústav informatiky a softvérového inžinierstva, FIIT STU, Bratislava
Vedúci projektu: Ing. Michal Kompan

Termín odovzdania práce v zimnom semestri : dňa 13. decembra 2011
Termín odovzdania práce v letnom semestri: dňa 10. mája 2012

Bratislava 19. 9. 2011

prof. Ing. Pavol Návrat, PhD.
riaditeľ ÚISI

Anotácia

Slovenská technická univerzita v Bratislave
FAKULTA INFORMATIKY A INFORMAČNÝCH TECHNOLÓGIÍ
Študijný program: Informatika

Autor: Ondrej Kaššák
Bakalársky projekt: Extrakcia pomenovaných entít zo slovenského textu
Vedúci bakalárskeho projektu: Ing. Michal Kompan

Máj 2012

Spracovanie textu hrá významnú úlohu pri metódach odporúčania založeného na obsahu alebo vyhľadávania. Extrakciou pomenovaných entít dokážeme identifikovať osoby, lokality, organizácie, dátumy a čísla vyskytujúce sa v textoch. Na základe takto zistených údajov vieme odhadnúť presnejšie o čom text pojednáva ako keď vychádzame napríklad len z jeho názvu. Následne vieme vyhľadávať texty na základe entít, ktoré sa v týchto textoch vyskytujú najčastejšie, prípadne extrahované entity využiť pri odporúčaní založenom na obsahu, kde vieme taktiež odporučiť články, v ktorých sa vyskytujú rovnaké entity a teda by mohli používateľa zaujímať.

V práci analyzujeme existujúce prístupy k problematike. Keďže pre slovenský jazyk nevieme o žiadnych existujúcich riešeniach, analyzujeme riešenia pre príbuzné jazyky ako češtinu a poľštinu. Okrem toho porovnávame úspešnosť riešení pre spomenuté jazyky a tiež nemčinu a angličtinu.

Na základe zistenej analýzy sme navrhli vlastnú lingvistickú metódu primárne určenú na extrakciu pomenovaných entít z textov v slovenskom jazyku. Metóda je po výmene slovníkov a stemmera použiteľná aj pre iné flexívne jazyky s podobnými pravidlami tvorby viet, ako má slovenčina.

Annotation

Slovak University of Technology in Bratislava
FACULTY OF INFORMATICS AND INFORMATION TECHNOLOGIES
Degree Course: Informatics

Author: Ondrej Kaššák
Bachelor Theses: Named Entity Recognition for Slovak Language
Supervisor: Ing. Michal Kompan

May 2012

Word processing plays an important role in the methods of recommendations based on content or search. We are able to identify persons, locations, organizations, dates, etc. occurring in the text by extraction of named entities. On the basis of these findings of data we can estimate the text content more precisely than if we assume only by its name. Then we can search texts by the entities which are most abundant in the texts, or use the extracted entities in the recommendations based on content where we can also recommend the articles in which the same entities occur and thus the user might be interested in them.

In this paper we analyzed existing approaches to the problem. We don't know about existing solutions for the Slovak language, thus we analyzed existing solutions for related languages such as Czech and Polish. We analyzed the accuracy of solutions not only for mentioned languages but also for German and English.

Based on this analysis, we proposed our own linguistic method primarily designed to extract entities from the texts in Slovak language. If we exchange the dictionaries and stemmer, the method is usable for other flexible languages with similar rules of sentences construction as in Slovak.

Obsah :

1 Úvod

V dnešnej dobe máme k dispozícii obrovské množstvo informácií. Informuje sa o všetkom, väčšinou sa informácie dajú získať z niekoľkých zdrojov. S rozvojom internetu sa informácie začali šíriť mnohonásobne rýchlejšie ako kedykoľvek predtým a rovnako sa začalo informovať o viacerých témach. V súčasnosti existuje veľa foriem, ktorými je informácie možné šíriť, no veľká časť z nich sa stále vyskytuje v písanej podobe - vo forme textov.

Pre jednotlivca je veľmi namáhavé vyhľadať si o určitej problematike relevantné texty. Z časového hľadiska je pre neho nemožné spracovať všetky dostupné texty. Keby však naopak vychádzal len z názvov alebo nadpisov jednotlivých textov často by niektorý relevantný text prehliadol alebo by sa začal zaoberať niektorým, ktorý s jeho témou nesúvisí.

Pri strojových prístupoch, napríklad pri odporúčacích metódach taktiež vieme používateľovi texty odporúčať efektívnejšie v prípade, že máme aspoň základnú predstavu o obsahu daného textu. Pri vyhľadávaní vieme vďaka takémuto obsahu zvýšiť presnosť vyhľadaných výsledkov.

V našej práci sme navrhli metódu, určenú na rozpoznávanie pomenovaných entít v slovenských textoch. Pojem pomenované entity predstavuje osoby, lokality, organizácie, dátumy, peňažné sumy, percentá a ostatné čísla, ktoré sa v textoch vyskytujú. Cieľom metódy je v prvom rade nájsť v texte výrazy, ktoré predstavujú entity a následne určiť ich rozsah. Potom nájdeným entitám určíme kategóriu, do ktorej patria, čiže zistíme či sa jedná napríklad o organizáciu alebo osobu.

Po identifikácii entít, vieme navrhnuté riešenie využiť napríklad pri vyhľadávaní textov, v ktorých sa nachádzajú rovnaké osoby, prípadne pojednávajú o rovnakom dátume. Taktiež môžeme pre každý text v určitej databáze rozpoznať pomenované entity a najčastejšie sa vyskytujúce entity v texte zobraziť používateľom už pri prezeraní zoznamu textov v databáze. Používatelia tak približne zistia, o čom text pojednáva a budú sa môcť rozhodnúť, či majú záujem si ho prečítať. Prípadne môžeme v textoch zvýrazniť alebo farebne odlíšiť jednotlivé druhy entít a používateľom tak sprehľadniť text.

Odporúčacie systémy založené na analýze obsahu používateľom odporúčajú podobné texty na základe rovnakých alebo podobných pomenovaných entít vyskytujúcich sa v daných textoch. Tento prístup by bolo možné využiť napríklad pri novinových článkoch. Články sú väčšinou rozsahovo krátke, venujú sa len jednej téme. Osoby alebo udalosti, na ktoré sa autor v článku zameriava sú spomenuté viackrát, ostatné entity sa v texte vyskytujú väčšinou jedenkrát.

Vhodným miestom pre nasadenie metódy a jej následné otestovanie sa javia spravodajské portály. V nich je denne uverejňované veľké množstvo článkov a sú využívané množstvom používateľov.

V druhej kapitole práce opisujeme pojem pomenované entity, vysvetľujeme aké typy entít existujú a čo predstavujú. Predstavujeme základné prístupy k rozpoznávaniu entít

a spôsoby akými sa vyhodnocuje úspešnosť jednotlivých metód. Taktiež tu opisujeme slovenský jazyk a jeho špecifiká z hľadiska procesu rozpoznávania pomenovaných entít. V závere kapitoly porovnávame vhodnosť vybraných jazykov na proces rozpoznávania pomenovaných entít.

V tretej kapitole sa venujeme existujúcim metódam rozpoznávania. Rozdeľujeme ho do dvoch fáz. Prvú tvorí predspracovanie textu, druhú samotný proces rozpoznávania entít v predspracovanom texte. V tejto kapitole rozpoznávanie rozdeľujeme na lingvistické a štatistické. V závere kapitoly opisujeme pomocné metódy a nástroje rozpoznávania.

Kapitola štyri obsahuje návrh vlastnej lingvistickej metódy rozšírenej o učenie slovníkov. Metóda je určená pre rozpoznávanie v slovenských textoch. Kapitolu tvoria opisy procesov predspracovania textu, vyhľadávania pomenovaných entít a overenia metódy.

2 Pomenované entity

Pojem rozpoznávanie pomenovaných entít (Named Entity Recognition alebo NER) predstavuje proces identifikácie vybraných typov slovných entít v textoch. Ciele problematiky boli stanovené na prvých šiestich „Message Understand Conferences" (MUC) v rokoch 1987-1995. Typy entít, ktoré sa identifikujú a spôsob ich označovania boli zadefinované na šiestej konferencii (MUC 6). Boli stanovené základné typy rozpoznávaných entít: osoby, lokality, organizácie, dátumy a čísla (Grishman & Sundheim, 1996). Všetky typy entít v textoch označujeme vždy párovými značkami ohraničujúcimi celú entitu (entita môže byť viacslovná) (Príklad 1). Jednotlivé typy označujeme nasledovne. Osoby, lokality a organizácie označujeme značkou <*ENAMEX*> s uvedením typu *PERSON*, *LOCATION* prípadne *ORGANIZATION*. Dátumy a čas označujeme značkou <*TIMEX*>, čísla všeobecne <*NUMEX*> s uvedením typu *NUM* (čísla), *PERC* (percentá) a *MONEY* (peniaze). Pokiaľ identifikujeme entity ale nevieme jej priradiť žiaden zo spomenutých typov označíme ju značkou <*MISC*> (Named Entity Task Definition, 1997).

```
Novým dekanom <ENAMEX TYPE="ORGANIZATION">Fakulty informatiky a informačných technológií Slovenskej technickej univerzity v Bratislave</ENAMEX> sa <TIMEX>24. októbra 2011</TIMEX> stal <ENAMEX TYPE="PERSON">doc. Ing. Pavel Čičák , PhD.</ENAMEX>
```

Príklad 1. Text s označenými pomenovanými entitami

Existujú dva druhy prístupov k problému rozpoznávania entít: lingvistické a štatistické. Lingvistické techniky sú založené na ručne skonštruovaných gramatikách, pravidlách, rozhodovacích stromoch a regulárnych výrazoch. Na ich vytvorenie je potrebné tieto rozhodovacie pravidlá a gramatiku vlastnoručne vytvoriť. Na rozdiel od štatistických metód nie je však nutné žiadne natrénovanie metódy na dátach. Štatistické modely sú založené na rozhodovaní na základe predchádzajúcich podobných prípadoch. Je teda potrebné natrénovať ich na veľkej množine ručne anotovaných dát. Metódy následne dokážu entity identifikovať podľa toho, ako boli v predchádzajúcich prípadoch vyhodnotené podobné entity.

Oba prístupy vedia dosiahnuť výborné výsledky, pričom pri lingvistických metódach sa dá dosiahnuť mierne lepšia presnosť (precision), niekedy však za cenu nižšieho pokrytia (recall). Metriky presnosť a pokrytie sa používajú pri vyhodnocovaní úspešnosti použitej metódy. Presnosť predstavuje počet správne rozpoznaných entít voči počtu nájdených, pokrytie vypočítame ako počet správne rozpoznaných voči všetkým, ktoré sa v texte nachádzajú. Celkovú úspešnosť metódy nazývame F-skóre (F-measure, prípadne F-score). Jeho hodnotu dostaneme ako harmonický priemer vysvetlených metrík (Sasaki, 2007).

Lingvistické metódy sa často využívajú pri úlohách, zameraných na niekoľko vybraných typov entít, kedy je ich vytvorenie značne jednoduchšie a výhodnejšie ako vytvorenie štatistickej metódy. Taktiež je ich ale možné využiť ako nástroj na kompletné rozpoznávanie všetkých pomenovaných entít. Ich výhodou je najmä intuitívnejšie

vytvorenie ako pri štatistických metódach a možnosť prispôsobenia tak, aby presne spĺňali požiadavky, ktoré sú na ne kladené. Taktiež je nespornou výhodou fakt, že na to aby dosahovali výborné výsledky ich nie je nutné trénovať na obrovskom množstve ručne anotovaného textu. Štatistické metódy sa využívajú pri zameraní na konkrétnu doménu jazyka, na ktorej po dostatočnom natrénovaní dokážu dosahovať dobré výsledky. Ich výhodou je, že entitám, ktoré z momentálneho kontextu nevieme klasifikovať, vieme určiť typ podľa predchádzajúcich rozhodnutí. Entita pritom mohla byť identifikovaná aj v niektorom z predchádzajúcich článkov, nie nutne v aktuálne spracovávanom.

2.1 Slovenský jazyk

Slovenský jazyk, na ktorý sa zameriavame, zaraďujeme do rodiny slovanských jazykov, presnejšie do západoslovanskej vetvy. Po slovensky hovorí viac ako 6 miliónov ľudí, z toho takmer 5 miliónov žije na území Slovenskej Republiky, kde je slovenčina úradný jazyk.

Slovenská abeceda sa skladá zo 46 znakov. Jej základom je latinská abeceda, ktorá navyše obsahuje diakritickými znamienkami modifikované písmená (Slovake, 2011). Pre porovnanie nemecká abeceda má 30 písmen, anglická si vystačí s 26. V slovenčine je teda napríklad pri tvorbe regulárnych výrazov nutné ošetriť viac možností, prípadne text najprv predspracovať odstránením diakritiky.

Slovenčina je flektívny jazyk. Jej gramatika pozná pre podstatné mená 12 skloňovacích vzorov rozdelených do 3 rodov (mužský, ženský, stredný) po 4 vzory na každý rod. Navyše rozlišuje 2 čísla (jednotné, možné (vo vybraných slovách existuje 3. číslo – pomnožné)) a každé slovo môžeme vyskloňovať do 6 základných pádov. Prídavné mená sa s podstatnými zhodujú v rode, čísle, páde ale navyše pre ne poznáme 6 vzorov, 2 pre každý rod. Slovenčina má teda bohatú morfológiu a jednotlivé tvary slov sa menia v závislosti od významu a konkrétneho použitia vo vete. Úprava slov na ich základný tvar, prípadne nájdenie koreňa slova algoritmicky je náročné a nedá sa takto dosiahnuť vysoká univerzálnosť, práve kvôli morfológii slov v jazyku a mnohým výnimkám oproti štandardnej tvorbe tvarov slova. Získavanie základného tvaru, prípadne jeho koreňa sa spravidla vykonáva vyhľadaním v špeciálnych slovníkoch.

Zmienené špecifiká spôsobujú, že slovenčina je na extrahovanie informácii nízkej úrovne, kam radíme aj rozpoznávanie pomenovaných entít vysoko náročná a pre počítačové spracovanie nevýhodná (Przepiórkowski, 2007). Vďaka špecifickým tvarom slov je však pomerne vhodná na extrahovanie informácií vysokej úrovne ako zisťovanie vzťahov medzi entitami, identifikáciu gramatických rolí a podobne, keďže význam viet je s veľkou pravdepodobnosťou jasný aj bez znalosti okolitého kontextu, s čím môže byť problém v jazykoch s jediným tvarom slov akým je napríklad angličtina.

2.2 Porovnanie jazykov z hľadiska NER

Problematika NER už bola v minulosti riešená pre väčšinu jazykov sveta. Pre dominantné svetové jazyky je vyriešená na rôzne dobrých úrovniach a taktiež pre tieto jazyky existujú hotové nástroje, pomocou ktorých je možné pomenované entity rozpoznávať. Jednotlivé jazyky sú však značne rôznorodé a preto je pre ne problematika NER riešená rôznymi spôsobmi. Takisto sú niektoré jazyky z pohľadu NER jednoduchšie, iné zložitejšie.

Slovenčina a ostatné slovanské jazyky sú v porovnaní s angličtinou prípadne podobnými germánskymi jazykmi značne náročnejšie z dôvodu ich flexivity a spôsobu, akým sú slová usporiadané vo vetách.

Najlepšie systémy rozpoznávajúce pomenované entity pre angličtinu dosahujú F-skóre na úrovni okolo 90%. Pre nemčinu sú dosiahnuté výsledky približne do 70% úspešnosti (Konkol & Konopík, 2011). Oba tieto jazyky patria medzi germánske jazyky no pri NER medzi nimi existuje značný rozdiel. Je to spôsobné najmä rozdielnymi pravidlami pre tvorbu viet, či rôznym písaním veľkých písmen na začiatku slov.

Slovanské jazyky sú na rozpoznávanie NER pre ich bohatú morfológiu ťažšie ako predchádzajúca skupina jazykov. Pre český jazyk boli doposiaľ prezentované výsledky v rozmedzí od 68% po 72% (Konkol & Konopík, 2011). Poľský jazyk je veľmi podobný ako čeština a dosahujú sa preň obdobné výsledky. Dosiahnuté výsledky sú do značnej miery ovplyvnené výberom domény a rozsahom zamerania medzi entitami. Pri zameraní sa len na rozpoznávanie osôb na malej doméne (množina policajných správ a správ z burzy) boli pre poľský jazyk prezentované výsledky s F-skóre 74%, s použitím metódy Skrytých Markovovských modelov (Hidden Markov Models - HMM). Po zapojení regulárnych výrazov na prefiltrovanie výsledkov tejto metódy dokonca 89% F-skóre (Marcińczuk & Piasecki, 2011). Tieto výsledky však boli dosiahnuté na malej špecifickej množine dát a otázne preto je, či by sa rovnaký výsledok dal preukázať aj na rozsiahlej množine textov, prípadne či by skóre bolo rovnaké pri rozpoznávaní viacerých typov pomenovaných entít.

V porovnaní so zložitosťou rozpoznávania entít v spomenutých jazykoch, sa slovenčina nachádza približne na rovnakom stupni náročnosti ako český jazyk a poľský jazyk, kam patrí aj z pohľadu vývoja jazykov. Pre slovenský jazyk nevieme o žiadnej prezentovanej metóde, ktorá by pomenované entity rozoznávala na úrovni porovnateľnej s ostatnými slovanskými jazykmi. Slovanské jazyky sú však do veľkej miery podobné a preto je porovnanie výsledkov a konfrontácia techník namieste. Pokiaľ dokážeme vytvoriť metódu pre jeden jazyk, dá sa predpokladať, že po drobných úpravách, kde túto metódu prispôsobíme špecifikám gramatiky nového jazyka, bude dosahovať výsledky na podobnej úrovni ako pre pôvodný jazyk. Naopak ťažko budeme porovnávať metódu špecializovanú na slovenčinu s nástrojom navrhnutým pre anglické texty. Napríklad pri výsledku F-skóre 72% pre slovenčinu a 75% pre angličtinu na prvý pohľad vyzerá anglický nástroj lepšie, pokiaľ si však uvedomíme existujúce hranice, ktoré sa pre tieto jazyky zatiaľ podarilo dosiahnuť vidíme, že výsledok porovnania je presne opačný.

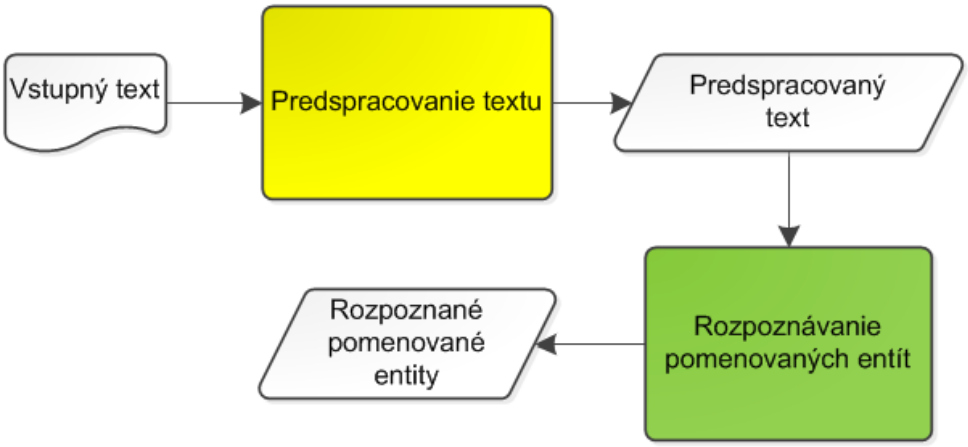
V angličtine majú jednotlivé slová jediný stály tvar. Iba slovesá sa časujú pridaním niekoľkých koncoviek, ktorá je však možné odstrániť jednoduchým odrezaním koncoviek. Výnimku síce tvoria nepravidelné slovesá, ale tak malá množina slov sa dá ošetriť porovnávaním so slovníkom. Slovenský, český či poľský jazyk s komplikovaným skloňovaním podstatných a prídavných mien, časovaním slovies a nepravidelným skloňovaním cudzích slov využívajú príliš veľa tvarov slov a teda množinu, s ktorou sa celé texty v rozumnom čase porovnávať nedajú. Preto je potrebné jednotlivé slová najprv predspracovať. Na to je možné využiť lematizáciu alebo stemming. Po získaní koreňov slov, prípadne slovotvorných základov je možné získané slová porovnať so slovníkmi.

V angličtine taktiež začína veľkým písmenom každé slovo v entite, v slovenčine okrem vlastných podstatných mien len prvé slovo. Je tak náročnejšie identifikovať rozsah a ohraničenie entít. V nemčine začína veľkým písmenom dokonca každé podstatné meno, čiže je z tohto hľadiska znevýhodnená najviac.

Nemecký jazyk spolu s angličtinou majú miernu výhodu v pevnej vetnej skladbe, kde podmet leží bezprostredne vedľa prísudku a podobne. V slovenčine rovnako ako v češtine a poľštine sa vetná skladba využiť nedá.

3 Existujúce metódy

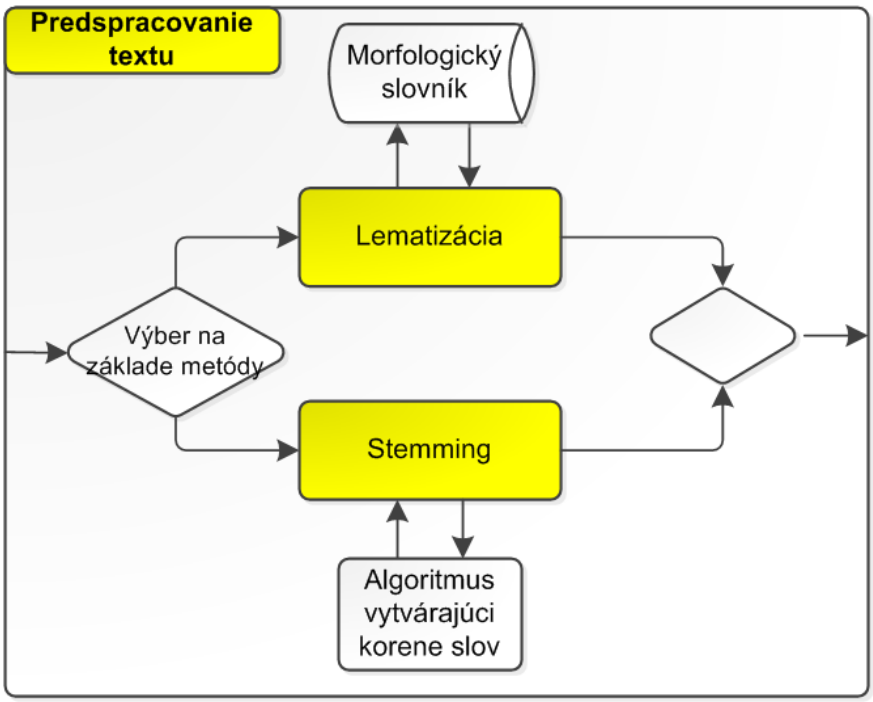
Proces spracovania textu, ktorý vedie k extrakcii pomenovaných entít, môžeme rozdeliť do dvoch hlavných krokov (Obrázok 1). Prvým je predspracovanie textu, ktorého vstupom je text v prirodzenom jazyku vo formáte, s ktorým vie metóda pracovať. Jeho výstupom a zároveň vstupom druhého kroku je predspracovaný text. Druhý krok predstavuje samotné rozpoznávanie pomenovaných entít v texte. Jeho výstupom je množina rozpoznaných pomenovaných entít identifikovaných v zadanom texte.



Obrázok 1. Extrakcia pomenovaných entít s aplikáciou predspracovania textu

3.1 Predspracovanie textu

Predspracovanie textu je proces slúžiaci na úpravu textu do tvaru, s ktorým vieme ďalej pracovať, porovnávať jednotlivé jeho slová so slovníkmi a taktiež v ňom v neskorších fázach identifikovať jednotlivé pomenované entity. Do procesu predspracovania textu teda môžeme zaradiť úvodné načítanie textu, s ktorým bude program pracovať, ďalej jeho rozdelenie na jednotlivé slová. Tieto môžeme rozdeliť napríklad podľa bielych znakov a z nich vytvoriť zoznam slov so zachovaným poradím voči vstupnému textu. Poznáme dva základné spôsoby predspracovania textu, lematizáciu a stemmnig (Obrázok 2). V rôznych metódach sa zvyčajne využíva jeden typ predspracovania textu.



Obrázok 2. Metódy pre predspracovanie textu. V závislosti od typu metódy sa využíva lematizácia, prípadne stemming.

Lematizácia

Proces lematizácie spočíva v nájdení základného tvaru hľadaného slova. Hľadanie prebieha porovnaním výrazu s morfologickým slovníkom naplneným ľudskými expertmi, ktorý pre každé slovo uvádza jeho základný tvar. Pre ilustráciu výraz „Národnej rade Slovenskej republiky“ vyhodnotí ako „národný rad/rada slovenský republika“. Ako vidno, metóda upraví každé slovo nezávisle. Výsledný tvar je v nominatíve jednotného čísla a pokiaľ je to možné, tak mužského rodu. Tvar jednotlivých slov môže byť odvodený od rôznych základov, pričom výhodou lematizácie je, že vráti všetky potenciálne základy. Správny základ následne môžeme zistiť z okolitého kontextu alebo ďalších výskytov slova. Nevýhodou je, že nedokážeme identifikovať slovo ktoré nie je uvedené v slovníku, s ktorým výraz porovnávame, čiže musíme neustále aktualizovať slovník o novovzniknuté slová. Ďalšou nevýhodou je nutnosť pamätať si všetky potenciálne lemy slova.

Stemmnig

Spočíva v určení koreňa slova. Koreňom sa myslí časť slova rovnaká pre všetky tvary, ktoré môže slovo nadobudnúť. Výhodou stemmingu je možnosť algoritmizácie procesu. Stemmer však musí byť vytvorený pre konkrétny jazyk, pretože je silne závislý od spôsobu tvorby slov, ktorá je pre jednotlivé jazyky špecifická. Pre slovenčinu je tento proces pomerne náročný, pre značnú početnosť možných tvarov jednotlivých slov. Výhodou však je, že na rozdiel od lematizácie vyhodnotí novovzniknuté slová rovnako dobre ako ostané, pretože porovnáva na základe predpon a prípon a nie podľa slovníkov. Predchádzajúci príklad „Národnej rade Slovenskej republiky“ stemmer upraví na tvar „národ rad slovensk republik“. Tento tvar slova je možné skombinovaním s regulárnymi výrazmi (napríklad na

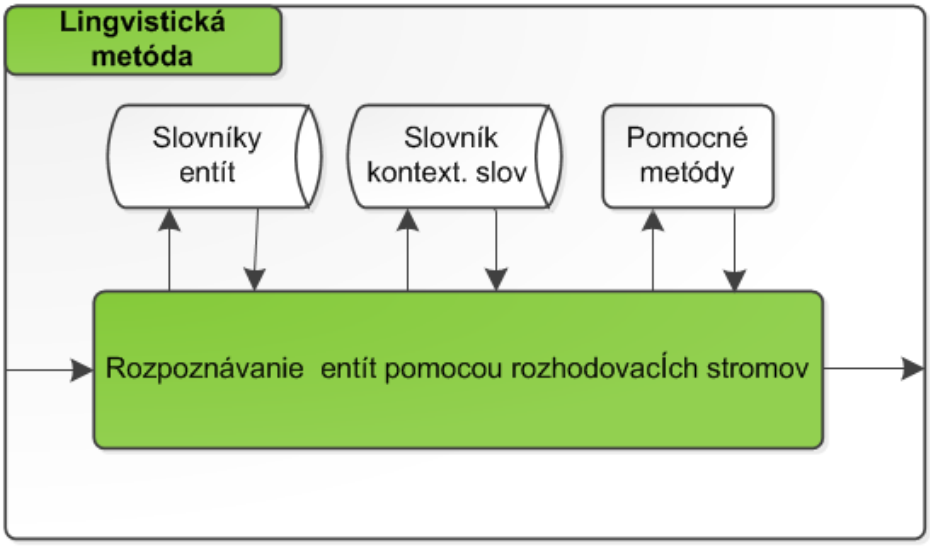
tvar „národ* rad* slovensk* republik*“) a použiť ako dopytovací tvar pri hľadaní v slovníkoch.

3.2 Rozpoznávanie pomenovaných entít

Metódy slúžiace na rozpoznávanie pomenovaných entít radíme do dvoch hlavných skupín, lingvistických a štatistických. Úlohou metód z oboch skupín je nájdenie pomenovaných entít v predspracovanom texte a následná identifikácia typov kam tieto rozpoznané entity patria. Jednotlivé skupiny metód riešia problematiku odlišne, no v praxi sa využívajú oba prístupy.

3.2.1 Lingvistické metódy

Hlavnou črtou lingvistických metód je ručné vytvorenie pravidiel, pomocou ktorých hľadáme v texte entity a následne im priraďujeme typ. Na ich zostrojenie je potrebný doménový expert. Pre tento typ metód je taktiež charakteristické využitie niekoľkých pomocných techník ako rozhodovacie stromy, špecifické slovníky, či regulárne výrazy. Jednotlivé techniky sa navzájom nevylučujú, naopak ich kombinácia často krát prináša najlepšie výsledky (Obrázok 3).



Obrázok 3. Lingvistická metóda. Využíva kombinácie rozhodovacích stromov, slovníkov entít a kontextových slov a ďalšie metódy.

Rozhodovacie stromy

Lingvistické metódy na usporiadanie pravidiel často využívajú rozhodovacie stromy, pomocou ktorých sú tvorené. Jednotlivé pravidlá sú usporiadané v stromovej štruktúre. Proces identifikácie entít v texte zodpovedá prechodu rozhodovacím stromom. Proces vždy začína v tom istom bode – v koreni stromu. Podľa toho v akom liste po prechode stromom výraz skončil je tento identifikovaný ako entita a je jej priradený typ alebo je vyhodnotený ako bežná časť vety. Pre zníženie zložitosti stromov, zjednodušenie a zovšeobecnenie pravidiel tak, aby dosahovali lepšie výsledky je potrebné text predspracovať lematizáciou, stemmingom alebo kombináciou oboch metód.

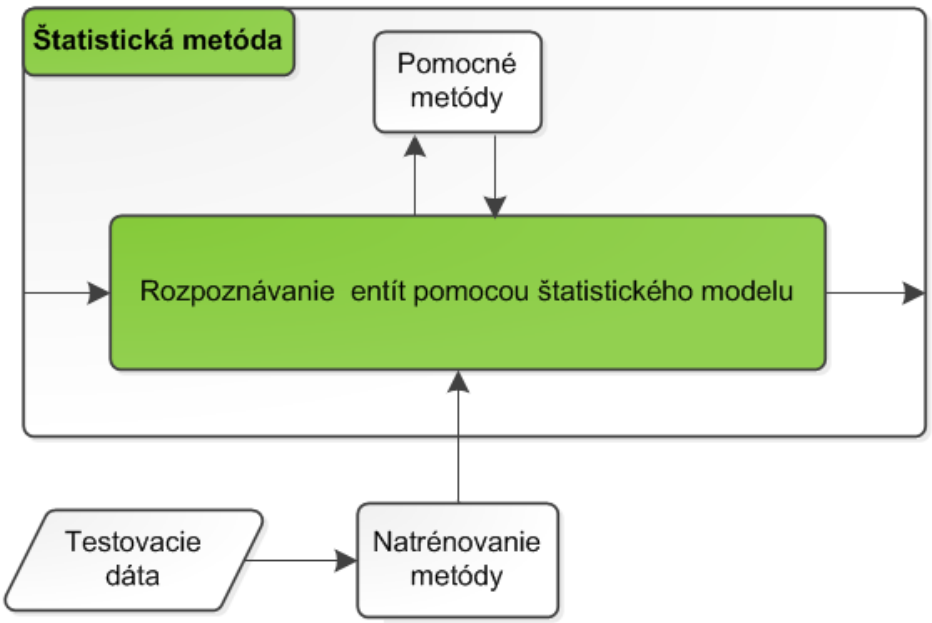
Slovníky
Text, v ktorom rozpoznávame entity je po predspracovaní možné porovnávať s pripraveným slovníkom obsahujúcim pre hľadané slová alebo výrazy priamo priradený typ entity. Prípadne môžeme slová alebo výrazy porovnávať s niekoľkými slovníkmi, kde každý reprezentuje iný typ entít a podľa toho, v ktorom slovníku hľadanú entitu nájdeme, určíme jej typ.

Regulárne výrazy
Lingvistické metódy taktiež často využívajú regulárne výrazy. Jedná sa o špeciálne textové reťazce určené na vyhľadávanie vyhovujúcich reťazcov v texte. Tvar regulárnych výrazov, závisí od programovacieho jazyka, v ktorom sú vytvorené. Skladajú sa z písmen, čísel, špeciálnych znakov a množín znakov. Ich kombináciou vytvoríme filter ktorý po zadaní vstupného textu vyfiltruje vyhovujúce, pripadne odfiltruje nevyhovujúce reťazce. Regulárne výrazy pri extrakcii pomenovaných entít môžeme využiť napríklad na vyhľadanie dátumov čí peňažných súm, rozdelenie textu podľa stop znakov (napríklad podľa interpunkčných znamienok) alebo určenie potenciálnych hraníc entít.

Lingvistické metódy sa v praxi často využívajú pri hľadaní jedného vybraného typu entít, kde sa po predspracovaní textu potenciálne entity porovnávajú s jedným slovníkom. Tento prístup sa využíva z dôvodu, že pre menšiu množinu hľadaných entít (napríklad len jeden typ) dokážeme gramatiku napísať ľahšie. Nasadenie štatistickej metódy je v tomto prípade zrejme zbytočné, pretože sa bez nej dokážeme zaobísť. Nepotrebujeme trénovať na zbytočne veľkej množine dát a pomocou vhodne zvolených rozhodovacích pravidiel dokážeme získať dobré výsledky. Lingvistické metódy sú ale rovnako výhodne použiteľné aj pri vyhľadávaní viacerých typov entít. Pomocou kontextových slov určíme o aký typ entity sa jedná a následne porovnáme výraz s príslušným slovníkom. Keďže rozhodovacie pravidlá sú vytvorené ručne, dokážu metódy tohto typu dosiahnuť pri rozpoznávaní veľkú presnosť, niekedy však za cenu nižšieho pokrytia.

3.2.2 Štatistické metódy

Druhú z hlavných skupín metód určených na rozpoznávanie pomenovaných entít tvoria štatistické metódy. Tento prístup vychádza z predpokladu, že pokiaľ sme nejakú entitu pri trénovaní ručne anotovanými textami niekoľkokrát označili za určitý typ, metóda pri testovaní rovnaké, prípadne určitým spôsobom podobné výrazy označí rovnako. Tento typ metód je pomerne často používaný. Existuje niekoľko typov štatistických metód, medzi najpoužívanejšie patria Skryté Markovovské modely (Hidden Markov models), Podmienené náhodné polia (Conditional random fields), Modely maximálnej entropie (Maximum entropy models) a Podporné vektorové stroje (Support vector machines). Princíp štatistických metód znázorňuje Obrázok 4.



Obrázok 4. Princíp práce štatistických metód - pred používaním metódu treba natrénovať na množine testovacích dát. Na rozpoznávanie využíva najmä štatistický model.

Metóda Maximálnej entropie

Metóda Maximálnej entropie (Maximum entropy method) vychádza z princípu, že pokiaľ nevieme presne rozhodnúť o type rozpoznávanej entity, snažíme sa dosiahnuť pravdepodobnostné rozdelenie s najväčšou možnou entropiou (Nigam, Lafferty, & McCallum, 1999). Slovo entropia môžeme interpretovať ako mieru neinformovanosti. Nadobúda hodnoty v intervale (0 - maximálna informovanosť, 1 - maximálna neinformovanosť). Nulová hodnota entropie nastáva v prípade, že entite vieme s určitosťou priradiť kategóriu. V tomto prípade dosiahneme najinformatívnejšie rozdelenie. Ak ale kategóriu s určitosťou rozlíšiť nevieme, pričom je nám známy konečný počet kategórií, kam rozpoznávané entity radíme, je vhodné všetkým potenciálnym kategóriám priradiť rovnaké pravdepodobnostné rozdelenie, čím dosiahneme maximálnu entropiu *log k*, kde *k* je počet všetkých kategórií.

Metódu maximálnej entropie kombinovanú s metódou Sémantických priestorov (Semantic spaces) využíva existujúci nástroj pre český jazyk zo Západočeskej univerzity v Plzni (Konkol & Konopík, 2011).

Metóda sémantických priestorov definuje každému slovu polohu v priestore identifikovanú vektorom. Podobnosti jednotlivých slov sú určené ich vzájomnou vzdialenosťou. Podobné slová sú zaraďované do rovnakých skupín. V prípade, že objavíme neznáme slovo vieme mu s veľkou pravdepodobnosťou priradiť typ entity podľa toho, v akej kategórii sa nachádzajú najbližšie slová alebo slová v rovnakej skupine.

Skryté Markovovské modely

Skryté Markovovské modely (Hidden Markov models) sú ďalšou často používanou štatistickou metódou. Model si môžeme predstaviť ako konečný stavový automat, ktorého stavy nevidíme (Rabiner & Juang, 2007). Vieme jedine zadať vstupný výraz a model nám

vráti výsledok. Výber nasledujúceho stavu pri prechode medzi jednotlivými stavmi modelu je vykonaný na základe najvyššej pravdepodobnosti spomedzi možných. Každá hrana v modeli má priradenú konkrétnu pravdepodobnosť. Každý výraz pri prechode modelom overíme podľa potrebných parametrov, na základe ktorých sme schopní určiť výsledok. Po prechode do niektorého konečného stavu porovnávanú entitu priradíme do kategórie, ktorú konečný stav určil. Pre každé slovo v slovníku máme vytvorenú vlastnú množinu stavov. Pokiaľ narazíme na nové slovo, ktoré sa v slovníku nenachádza postupne ho podľa sledovaných parametrov posúvame medzi jednotlivými stavmi automatu do momentu, kedy nie sme schopní rozhodnúť, do ktorej kategórie patrí.

Podmienené náhodné polia

Metóda Podmienených náhodných polí (Conditional random fields) je štatistická metóda, do značnej miery podobná metóde skrytých markovovských modelov (HMM). Rovnako ako ona je reprezentovaná konečným stavovým automatom, tento krát ale jediným pre celý model (Lafferty, McCallum, & Pereira, 2001). Na rozdiel od HMM, kde rozhodovanie o type aktuálne rozpoznávanej entite závisí len od nej, v opisovanej metóde sa do úvahy berú rovnako už vyhodnotené entity, prípadne sa výsledok môže zmeniť podľa ešte neidentifikovaných ale v rovnakom texte sa vyskytujúcich entít. Toto je značnou výhodou, pretože entita môže byť na jednom mieste v texte určiteľná jednoznačne, inde typ nemusí byť identifikovateľný jednoznačne.

Metódu využíva napríklad Stanford Named Entity Recognizer (NER)[1] (nazývaný tiež CRF Classifier). Tento nástroj je určený pre anglický jazyk. Bol vyvinutý Standfordskou univerzitou, je voľne dostupný na výskumné účely. Rozpoznáva tri druhy entít: osoby, organizácie a lokality.

Podporné vektorové stroje

Podporné vektorové stroje (Support vector machines) je ďalšia z často používaných štatistických metód. Pôvodne bola navrhnutá iba pre rozhodovanie medzi dvomi možnými skupinami kam môže element patriť. Metóda spočíva v tom, že jednotlivým entitám priradíme vektorovú polohu v priestore (Chen, Lin, & Schölkopf, 2005). V úvode pomocou ručne anotovanej trénovacej množiny zistíme priestorovú polohu určitej množiny dát, aby sme pri využívaní metódy mali už od začiatku nové entity s čím porovnať a podľa toho rozhodnúť. Medzi dvoma skupinami entít zostrojíme plochu, pokiaľ to je možné rovinu tak, aby sme dosiahli najväčšiu možnú vzdialenosť od prvkov týchto skupín. Plocha musí oddeľovať skupiny tak, aby prvky oboch skupín ležali v rozdielnych polpriestoroch vytvorených touto plochou. Vzdialenosť sa meria od najbližších prvkov oboch skupín k deliacej ploche a musí byť rovnaká od oboch týchto prvkov. Podľa toho, v ktorom polpriestore sa zatrieďovaný prvok nachádza mu priradíme rovnaký typ entity ako majú okolité prvky.

Pri n typoch tried ($n > 2$) kam môže entita patriť je riešením postupné rozhodovanie. Najprv rozhodneme, do ktorej z prvých dvoch tried by sme entitu priradili, následne rozhodujeme znovu medzi triedou ktorú sme vybrali z prvých dvoch a treťou.

[1] *http://www-nlp.stanford.edu/ner/*

Potom obdobne rozhodujeme vždy medzi triedou kam sa entita najlepšie hodí a ešte neskúmanou triedou. Po *n-1* porovnaniach dostaneme triedu, do ktorej zaraďovaná entita patrí s najväčšou pravdepodobnosťou.

3.2.3 Pomocné metódy

Ktorákoľvek z vyššie opísaných metód nemôže fungovať samostatne. Existujú pomocné metódy, ktoré pri správnom využití pomôžu výrazne zvýšiť úspešnosť zvolenej metódy. Tieto pomocné metódy je potrebné zaradiť na správne miesta a vhodne ich skombinovať tak, aby pomohli vo vyhľadávaní a následnej identifikácii pomenovaných entít v texte.

Kontextové slová

Jedná sa o slová pomocou ktorých vieme priamo alebo nepriamo určiť, prípadne vylúčiť typ kam overovaná entita spadá. Môže sa jednať buď o časti entít alebo slová nachádzajúce sa v okolí entít, ktoré však nie sú ich súčasťou. Príkladmi slov, ktoré sú súčasťou entity sú napríklad slová „univerzita, ulica“. Slová, ktorých výskyt vo vete znamená, že okolitá entita zrejme bude konkrétneho typu sú napríklad „profesor, pán/pani, povedal, minister“. Slovami vylučujúcimi niektoré typy entít môžu byť napríklad určité predložky. Povedzme „v“ znamená že entita bezprostredne za ním nie je osoba a ak začína veľkým písmenom tak ani časová miera („v Stredu“ sa v slovenčine nepoužíva).

Predchádzajúce identifikované entity

Pred rozhodnutím o type určitej entity je vhodné prejsť už rozpoznané entity, ktoré boli v texte identifikované skôr. Je napríklad pomerne bežným javom, že v článku sa pri prvej zmienke o osobe uvedie celé jej meno, niekedy dokonca s titulom, či funkciou a neskôr sa už uvádza len priezvisko. To by sme mohli identifikovať ako entitu zmiešaného typu alebo pri priezvisku typu „Kováč, Malý, Líška“ dokonca neidentifikovať ako entitu. Po prezretí skôr rozpoznaných entít ju však zaradíme správne.

Slovníky

Slovníky boli opísané pri lingvistických metódach. Je ale potrebné poznamenať, že slovník v tomto význame obsahuje slová v ich základnom tvare. Slová sú abecedne radené pre rýchlu orientáciu v slovníku. Pre každý výraz je tu uvedený tep entity, ktorú predstavuje. Alternatívou môžu byt slovníky v zmysle abecedne usporiadaných zoznamov slov. Každý typ entity má vlastný zoznam a entitu identifikujeme podľa toho, v ktorom zozname sme našli zhodu s porovnávanou entitou. Ako príklady uveďme zoznam slovenských a svetových mien, priezvisk, organizácií, zemepisných lokalít alebo konkrétnejšie štátov, miest, riek a podobne.

Hranice entity

Určiť hranice entity nie je v Slovenčine vždy najľahšie. S istotou vieme povedať len to, že prvé písmeno bude veľké. Aj toto tvrdenie pozná výnimku v prípade, že identifikujeme časové entity ako napríklad „v januári, rok 2012“. Všeobecne ale môžeme vychádzať z predpokladu, že keď slovníkovým overením nájdeme spomenuté časové miery, môžeme

sa v texte orientovať podľa veľkých začiatočných písmen. Ďalšou výnimkou, ktorú je potrebné odfiltrovať sú začiatočné slová viet, ktoré zároveň nie sú začiatkom entity.

Číselné entity

Identifikácia číselných entít prebieha jednak nájdením numerických znakov v texte a taktiež overením slovne zadaných čísel pomocou slovníka. Podľa špeciálnych znakov „%, $, €, dd.mm.yy“ vyskytujúcich sa bezprostredne pred alebo za číslom dokážeme navyše identifikovať percentá, peňažné sumy, prípadne dátumy.

3.2.4 Existujúce pomocné nástroje

Tieto nástroje predstavujú existujúce služby dostupné online na internete. Poskytujú rôznu funkcionalitu, ktorá môže byť nápomocná pri overení, či je identifikovaný výraz skutočne entita, prípadne či sme správne identifikovali hranice entity. Taktiež môžu pomôcť napríklad pri spracovaní textu (lematizácií).

Slovenský národný korpus[1]

„Slovenský národný korpus (SNK) je elektronická databáza obsahujúca slovenské texty z rôznych štýlov, žánrov, vecných oblastí, regiónov a pod., vybavená výkonným vyhľadávacím systémom a prídavnými jazykovými informáciami. Slúži ako referenčný zdroj poznatkov o slovenčine a jej reálnom používaní, nenahrádza však kodifikačné príručky.“ (Jazykovedný ústav Ľ. Štúra SAV, 2011).

Každý z textov v SNK je ručne anotovaný ľudským expertom. Po zadaní výrazu, ktorý hľadáme do vyhľadávača, dostaneme množinu všetkých výskytov tohto výrazu v textoch SNK. Výraz môžeme zadať v ľubovoľne vyskloňovanom tvare, na rozdiel od väčšiny slovníkov a pokiaľ sa v textoch nachádza, dostaneme všetky výskyty v tomto tvare. Pred vyhľadaním máme možnosť vybrať, že chceme pre každé slovo zobraziť aj jeho lemu. Napríklad po zadaní výrazu „Slovenskú technickú univerzitu“ nám SNK vráti 57 výskytov výrazu v tvare „Slovenskú /slovenský technickú /technický univerzitu /univerzita“. Okrem toho máme možnosť nastaviť si koľko znakov z pôvodného textu, kde sa výraz vyskytuje, chceme okolo výrazu zobraziť. SNK poskytuje aj ďalšiu funkcionalitu, tá však pre účely rozpoznávania pomenovaných entít nemá väčší význam.

Online slovníky slovenských priezvisk a obcí[2]

Na webovom portáli môžeme okrem iných slovníkov nájsť databázu slovenských priezvisk a databázu obcí Slovenska. Obe sú síce pomerne staré, údaje v databáze priezvisk sú z roku 1995, v databáze obcí z roku 1997. Napriek dátumu ich vzniku sa však stále dajú využiť, pretože predpokladáme, že od vzniku databáz nedošlo k v výrazným zmenám alebo vzniku nových obcí či priezvisk. Údaje teda stále pokrývajú značné percento z momentálne existujúcich množín a môžu tak byť veľmi nápomocné pri identifikácii týchto typov entít v textoch.

[1] *http://korpus.juls.savba.sk/*
[2] *http://slovniky.korpus.sk/*

Do slovníkov je potrebné zadať výraz v základnom tvare, prípadne po zvolení niektorej z možností stačí jeho prefix, sufix alebo časť slova. V prípade, že zadaný výraz sa nachádza v databáze obcí, dostaneme zoznam názvov tejto obce od roku 1773 do roku 1997. Pre nás majú význam aktuálne názvy. Ak sa zadaný výraz nachádza v databáze priezvisk, dostaneme zoznam výskytov tohto priezviska na Slovensku v roku 1995. Pre každý výskyt je uvedená lokalita a počet osôb s týmto priezviskom.

Využitie slovníkov spočíva v overení existencie slov, ktoré boli identifikované v textoch. V prípade, že sa určité slovo v slovníku nachádza znamená, že to že existuje a predstavuje entitu typu osoba.

Wikipédia

Wikipédia[1] je najväčšia webová encyklopédia dostupná online. Verzia Wikipédie existuje takmer pre každý jazyk sveta. Najviac obsahu existuje v angličtine, kde môžeme nájsť viac ako 3 907 000 článkov[2]. Jej obsah tvoria samotní používatelia a preto nie sú informácie, ktoré obsahuje zaručeným zdrojom. Napriek tomu sa z hľadiska NER dá využiť na overenie existencie entít. Potom čo je v texte identifikovaná určitá entita, zadá sa jej názov do Wikipédie a v prípade že entita existuje, je veľká šanca, že tu o nej existuje článok. V prípade, že po zadaní identifikovaného názvu entity zistíme, že článok neexistuje, je značná pravdepodobnosť, že sme hranice entity neurčili správne. Pre vyhľadávanie entít v slovenských textoch je ich výhodné primárne hľadať v slovenskej verzii a v prípade neúspechu overiť ich existenciu v anglickej verzii. Slovenská verzia[3] obsahuje oproti anglickej verzii viacero lokálnych názvov, ktoré anglická nemôže obsiahnuť. Anglická verzia zase obsahuje prehľad svetových pojmov, o ktorých sa v slovenských článkoch často píše, ale slovenská verzia Wikipédie nie je natoľko rozvinutá, aby sa v nej nachádzali všetky.

Okrem priameho pristupovania do Wikipédie, je možné získať offline obrazy jednotlivých jej častí, ako napríklad články, diskusie, revízie článkov alebo zoznamy kategórií. Túto službu poskytuje samotná Wikipédia, ktorá svoj obsah dodáva v pomerne ľahko spracovateľných XML súboroch, stiahnuteľných priamo z oficiálneho zdroja[4]. Pre naše účely sme využili obraz obsahujúci iba aktuálne verzie všetkých slovenských a anglických článkov.

[1] *http://en.wikipedia.org/*
[2] *http://www.wikipedia.org/*
[3] *http://sk.wikipedia.org/*
[4] *http://dumps.wikimedia.org/*

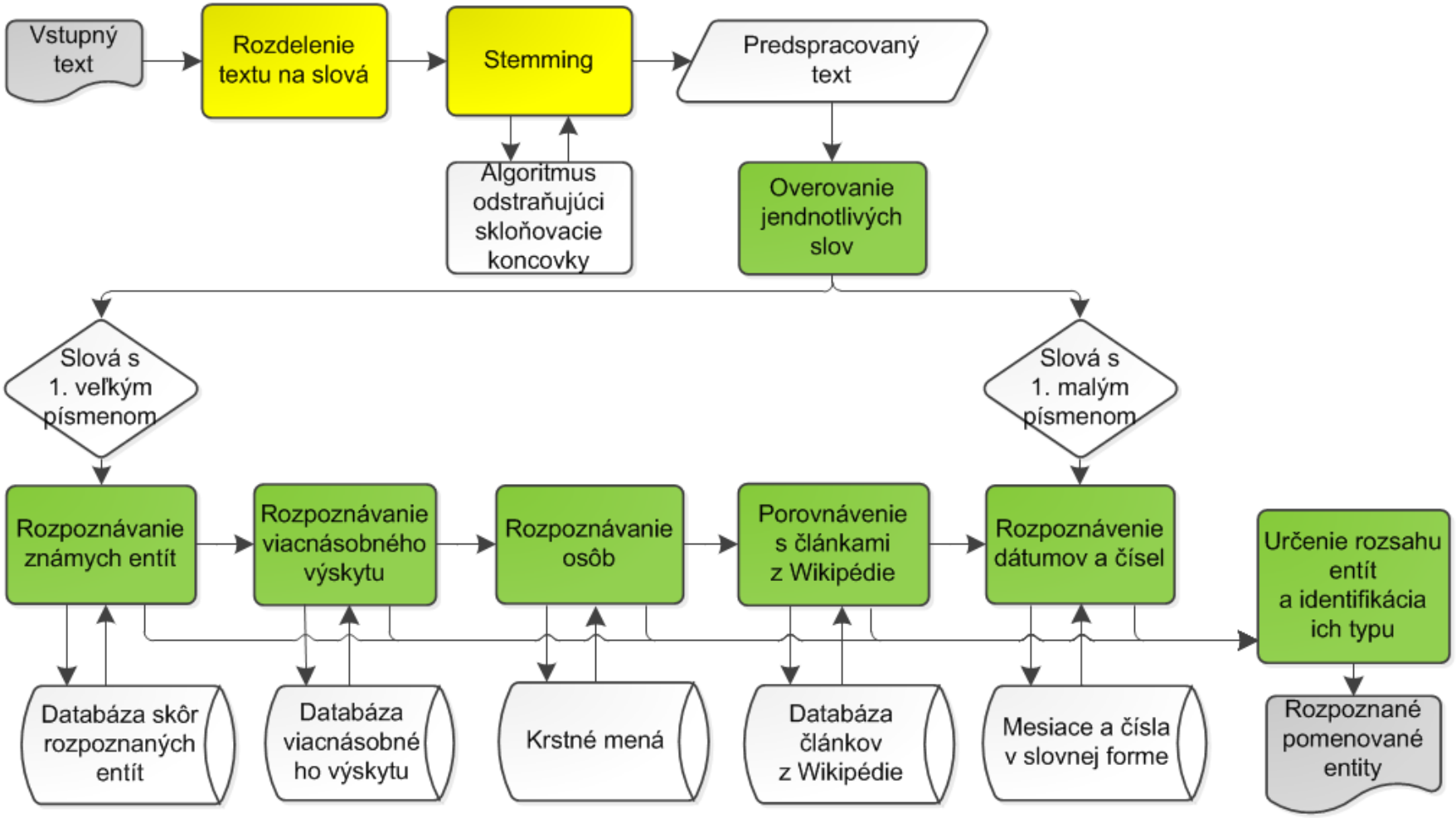
4 Navrhovaná metóda

V predchádzajúcej kapitole sme opísali základné členenie prístupov k rozpoznávaniu pomenovaných entít a načrtli špecifiká jednotlivých metód. Nástroje využívajúce tieto metódy sú však väčšinou určené na rozpoznávanie textov písaných vo svetových jazykoch. Pre tieto jazyky boli dosiahnuté najvyššie hodnoty úspešnosti rozpoznávania a taktiež sa im venuje najviac pozornosti. Existuje však aj niekoľko metód riešiacich problematiku pre slovanské jazyky. Tieto nástroje pracujú s rozdielnou úspešnosťou a rozpoznávajú rôzny počet typov entít. Pre slovenský jazyk ale nevieme o žiadnom konkrétnom nástroji, ktorý by rozpoznával všetky definované typy entít.

Metóda, ktorú v tejto práci navrhujeme rozpoznáva všetky typy pomenovaných entít definovaných na MUC 6. Jej základ tvorí nástroj na predspracovanie textu do podoby, v ktorej sme schopní pomenované entity rozpoznávať. Medzi kľúčové prvky metódy ďalej patria funkcie rozpoznávajúce entity pomocou databázy skôr rozpoznaných entít, databázy článkov získaných z obrazov slovenskej a anglickej Wikipédie či zoznamu mien osôb. Metóda rovnako využíva funkcie identifikujúce entity na základe slov určujúcich slovenské čísla (zoznam slovných zápisov čísiel), dátumy (zoznam kalendárnych mesiacov), peňažné meny, či špecifikujúcich typy rozpoznaných entít na základe okolitých kontextových slov.

Metóda pomocou vlastnoručne vytvorených pravidiel vyhľadá začiatky potenciálnych entít, z ktorých odfiltruje napríklad začiatky viet, ktoré nie sú entity. Ostávajúcim určí ich rozsah a následne pomocou viacerých navrhnutých techník určí príslušný typ. Z tohto pohľadu navrhovaná metóda vyzerá ako lingvistická, no v jednom aspekte sa radí aj do štatistickej vetvy. Z tohto prístupu sme prevzali myšlienku natrénovania metódy. Natrénovanie slúži na rozšírenie databázy skôr rozpoznaných entít, kedy sa pre zníženie výpočtovej náročnosti a zvýšenie presnosti metódy, metóda učí - entity rozpoznané na základe dopytovania sa do Wikipédie a typické začiatky viet, ktoré nie sú entity získané zo SNK sú lokálne zaznamenané.

Navrhovaná metóda je primárne určená pre slovenský jazyk. Po výmene stemmera, databáz a množiny kontextových slov, s ktorým metóda pracuje a miernej modifikácii niektorých pravidiel, je však metódu možné použiť aj pre iné jazyky s podobnou štruktúrou a spôsobom tvorby viet. Príkladom sú jazyky patriace do skupiny slovanských jazykov.



Obrázok 5. Schéma navrhovanej metódy. Procesy patriace do predspracovania sú označené žltou farbou, zelené procesy predstavujú rozpoznávanie pomenovaných entít, kde postupne prechádzame každé slovo vstupného textu. V prípade, že aktuálne overované slovo začína veľkým písmenom začíname identifikáciu Rozpoznávaním známych entít, pre slovo začínajúce malým písmenom začíname Rozpoznávaním dátumov čísel. V prípade, že pri ktoromkoľvek procese rozpoznávania nájdeme entitu prejdeme k procesu Určenia rozsahu entít a identifikácii typu, inak prejdeme na proces napravo od aktuálneho, alebo na konci zahodíme slovo (po neúspešnom Rozpoznávaní dátumov a čísel).

4.1 Predspracovanie textu

Vstupom pre metódu je text, ktorý môže byť vložený vo viacerých formátoch. Prvá možnosť je zadanie url, ukazujúcej na text umiestnený na internete, druhá je spracovanie textu, ktorý používateľ zadal priamo. Potom, čo metóda získa text, môže s ním začať pracovať a pripraviť ho na samotné rozpoznávanie pomenovaných entít.

Proces predspracovania textu sa skladá z niekoľkých častí. V prvej fáze pridáme okolo interpunkčných znamienok textu z oboch strán medzery. Pre lepšiu predstavu napríklad text „sused, ktorý“ bude upravený na tvar „sused , ktorý“, ale text „3,14“ alebo „s.r.o“ po tejto úprave ostane nezmenený. Po tejto izolácii bielych znakov v druhej časti rozdelíme text na jednotlivé slová.

V rámci predspracovania textu všetkým slovám pomocou vlastnoručného stemmera odstránime ich koncovky, ktoré vznikli skloňovaním. Získané tvary slov síce netvoria slovotvorné základy jednotlivých slov, čo je štandardný výstup stemmerov, ale pre naše účely sú ešte vhodnejšie, keďže zachovávajú významy jednotlivých slov. Stemmer napríklad zo slova „Ministerstvu“ vyrobí výraz „Ministerstv“, ale pri slovách „prezident“ a „viceprezident“ zachová rozdiel. Opísaný proces nie je príliš časovo náročný a umožní nám vysporiadať sa s bohatým skloňovaním slovenského jazyka. Jednotlivé slová potom môžeme porovnávať úplne nezávisle od toho, v akom tvare sú uvedené v texte.

4.2 Vyhľadávanie jednotlivých typov entít

Po spracovaní vstupného textu do požadovanej formy sa začne prechádzať pole s rozdeleným vstupným textom. Pre jednotlivé jeho prvky postupne overujeme, či začínajú veľkým písmenom, číslom prípadne sa jedná o slovne zapísané číslo.

V prípade, že slovo vyhodnotíme ako začiatok entity označíme celú entity a začíname proces identifikácie od začiatku pre slovo nasledujúce za koncom entity. Ak naopak slovo neidentifikujeme ako entitu prejdeme len na nasledujúci krok. Kroky sú opísané v príslušných podkapitolách. Posledný krok je však vždy tvorený posunom na nasledujúce slovo a opätovným spustením procesu pre nasledujúce slovo.

4.2.1 Identifikácia slovných entít

Identifikácia slova začínajúceho veľkým písmenom sa skladá z niekoľkých krokov. V každom kroku sa však pokúšame určiť, či toto slovo predstavuje začiatok entity, následne určiť jej rozsah a typ. Ak v konkrétnom kroku určíme entitu, ukončíme pre ňu proces identifikácie a pokračujeme znovu od začiatku pre slovo nasledujúce za koncom tejto entity. Ak v kroku nerozpoznáme entitu, pokračujeme nasledujúcim slovom textu.

Databáza skôr rozpoznaných entít

Prvý krok identifikácie spočíva v porovnaní overovaného slova s databázou skôr rozpoznaných entít. Jedná sa o entity, ktoré naša metóda správne identifikovala v predchádzajúcich textoch. Cieľom tohto kroku je jednoduchá identifikácia entít, pre ktoré bol celý proces už v minulosti vykonaný.

V nej teda vyhľadáme všetky zhody začínajúce overovaným slovom. Následne v rámci výsledkov nájdeme najdlhšiu celkovú zhodu medzi množinou výrazov z databázy a textom na sledujúcim za overovaným slovom. Kladný výsledok si zaznamenáme a skončíme, v opačnom prípade prejdeme na nasledujúci krok.

Databáza entít rozpoznaných skôr v danom texte

Tento krok spočíva v porovnaní overovaného slova s množinou entít, ktorú sme rozpoznali v aktuálnom texte. Tieto entity majú veľkú šancu, že sa neskôr dostanú do databázy skôr rozpoznaných entít, najskôr však musia byť potvrdené ľudským expertom. V rámci tohto textu sú však brané ako plnohodnotne rozpoznané entity, pretože s danými vedomosťami metóda dokáže entitu identifikovať prinajlepšom rovnako a teda opäť preskočíme zvyšné kroky a entitu určíme podľa predchádzajúcich riešení.

V textoch je totiž pomerne bežné, že napríklad pri prvom výskyte osoby sa spomenie celé jej meno, prípadne tituly. Na základe nich sme schopní určiť, že sa jedná o entitu typu osoba. Následne sa však v textoch spomína len priezvisko, pretože je z pohľadu človeka zbytočné písať opätovne celé meno. Pri nájdení takéhoto priezviska by sme však nemuseli vždy určiť, že sa jedná o osobu. Porovnávanie s rozsiahlym slovníkom priezvisk by zase zbytočne spomaľovalo vykonávanie a nikdy by sme úplne nepokryli všetky možnosti. Preto si pri prvom nájdení osoby zapamätáme jej celé meno a rovnako samostatne priezvisko, čo nám umožní identifikovať prípadné následné výskyty niektorej formy zápisu.

Identifikácia osôb

Samotné rozpoznávanie osôb, ktoré v texte predtým neboli spomenuté, je ďalší krok rozpoznávania. K nemu pristupujeme, pokiaľ sme pre overované slovo nenašli zhodu medzi skôr rozpoznanými entitami. Proces začína zistením, či je overované slovo titul, iniciál prípadne meno. Ak áno zväčšujeme rozsah entity, dovtedy, kým nachádzame súvislý rad mien alebo iniciál. Ak nenarazíme na slovo začínajúce veľkým písmenom, ktoré nedokážeme identifikovať vyhlásime ho za priezvisko a entitu označíme. Ak v úvode nerozpoznáme zástupcu niektorého zo spomenutých typov, prejdeme na nasledujúci krok.

Databáza článkov z Wikipédie

Tento krok spočíva vo vyhľadaní overovaného slova v databázach obsahujúcich články zo slovenskej a anglickej Wikipédie. Články totiž obsahujú aj rôzne číselné názvy, ktoré by nám mohli znížiť úspešnosť rozpoznania týchto entít.

Posledným krokom rozpoznania týchto entít je teda porovnanie overovaného slova s článkami z Wikipédie. Ich názvy sme spolu s kategóriami, do ktorých patria získali parsovaním offline obrazov. Primárne sa pokúšame nájsť zhodu v tabuľke so slovenskými článkami, v prípade neúspechu hľadáme v anglických. Princíp je podobný ako pri porovnávaní s databázou skôr rozpoznaných entít, opísanom v prvom kroku. Podstatný rozdiel je však v spôsobe akým určujeme typ entít. V porovnávaní s databázou skôr rozpoznaných entít a s entitami rozpoznanými skôr v texte sme typ určili podľa entity pri ktorej sme našli zhodu. V aktuálnom kroku máme množinu kategórií, kam konkrétny

článok patrí. Typ určujeme na základe ich porovnania s množinou kľúčových slov, kde vyberáme prvú zhodu.

4.2.2 Identifikácia číselných a dátumových entít

Tento krok predstavuje rozpoznanie dátumov a čísiel, prípadne ich ekvivalentov zapísaných slovne. Po zistení, že prvok poľa, ktorý práve overujeme je číslo, pomocou vlastnoručne vytvorených pravidiel identifikujeme rozsah tejto entity a na základe jej častí určíme o aký typ číselnej entity sa jedná. V tomto type entít majú kontextové slová veľký význam. Vďaka nim vieme rozpoznať samotné čísla od percent, peňažných súm, či časových jednotiek. Tieto slová sú špecifické pre konkrétny jazyk.

Okrem čísel rozpoznávame aj dátumy. Tieto identifikujeme po nájdení kalendárneho mesiaca na začiatku entity, prípadne za číslom. Ako dátumy označujeme postupnosti, ktoré spĺňajú niektorý z predpokladaných formátov. Tie sú definované prevažne regulárnymi výrazmi. Mesiace sú uvedené v samostatnom zozname.

Do opisovanej časti metódy patria taktiež čísla, ktoré sú zapísané slovne – s veľkým alebo malým prvým písmenom. Tento krok teda vykonávame pre každé slovo z vstupného textu, ktoré sme skôr neoznačili ako súčasť inej entity.

V metóde máme uvedenú množinu slov, z ktorých sú tieto čísla zložené. Z overovaného slova postupne odoberáme časti, ktoré sa nachádzajú v spomenutej množine. Pokiaľ nám na konci ostane prázdne slovo vyhodnotíme overované slovo ako slovne zapísané číslo. Pre lepšiu predstavu uveďme príklad. Slovo „dvanásťtisíc“ dokážeme rozdeliť na „dva – násť - tisíc“, čo sú slová tvoriace číslovku. Naopak to nedokážeme pre slová „tribúna“, kde vieme oddeliť len „tri-“ alebo slovo „sova“, kde nedokážeme oddeliť nič.

4.2.3 Charakteristika jednotlivých typov entít

Identifikácia typu, ktorý rozpoznanému výrazu prislúcha prebieha pre každý typ špecificky. Pri určitých typoch sú primárne dôležité kontextové slová nachádzajúce sa v okolí entity, pri iných rozhodujeme najmä na základe určitých častí samotnej entity.

Identifikovanie osôb

Osoby môžu mať svoje mená uvedené rôznym spôsobom. Vieme však povedať, že všetky časti mena začínajú veľkým písmenom. Väčšinou sa stretneme s dvoma typmi zápisov slov. Prvý predstavuje zápis „Meno Priezvisko“ alebo „Priezvisko Meno“. Druhým typom je forma „Priezvisko“. Tento kratší zápis sa vyskytuje väčšinou po tom, čo už osoba bola v texte spomenutá celým menom. Preto je vhodné pamätať si aké osoby už boli v texte spomenuté a keď na priezvisko narazíme priradíme ho k danej osobe. Preto si do zoznamu skôr rozpoznaných entít ukladáme okrem celého mena osoby samostatne aj jej priezvisko. Pri ďalších výskytoch ho potom identifikujeme správne už pri porovnávaní so skôr rozpoznanými entitami. Keby totiž samotné priezvisko identifikujeme nezávisle, mohol by (najmä pri zriedkavejších alebo zahraničných priezviskách) nastať prípad kedy nedokážeme entitu správne identifikovať a priradíme jej nesprávny typ alebo ju označíme ako entitu zmiešaného typu.

V určitých prípadoch môže byť osoba pomenovaná tak, že meno alebo dokonca priezvisko, môže byť nahradené jeho prvým písmenom s bodkou za ním (iniciály). Osoba môže taktiež mať okrem prvého mena jedno alebo niekoľko stredných mien, ktoré môžu byť uvedené v plnom znení alebo nahradené iniciálami. Osoby identifikujeme v záverečnom výpise párovými značkami <*ENAMEX TYPE=“PERSON“*> (Príklad 2).

```
pán <ENAMEX TYPE=“PERSON“> Ing. Peter Buday</ENAMEX>
```

Príklad 2. Ukážka označenia pomenovanej entity typu osoba

Ako je uvedené aj v príklade, osoby môžu mať pred alebo za menom titul. Ten sa na rozdiel od oslovenia do entity zaraďuje spolu s samotným menom.

Overenie, či je výraz pomenovaná entita typu osoba sa uskutočňuje v prípade, že sme pomocou kontextových slov typu „pán, pani“ zistili že výraz by mal byť tohto typu. Výraz teda najprv porovnáme so zoznamom krstných mien. V ňom môžeme nájsť všetky oficiálne slovenské a najpoužívanejšie svetové mená. V prípade, že sme pri porovnávaní výrazu našli zhodu s niektorým prvkom slovníka, zistili sme, že výraz predstavuje meno osoby. V takomto prípade začneme overovať nasledujúce slová pokiaľ nenarazíme na slovo začínajúce malým písmenom alebo interpunkčné znamienko. Pre slovo s veľkým písmenom overíme či sa nejedná o titul, ak áno nepriradíme ho k entite a ukončíme hľadanie rozsahu entity. Ak sa jedná o slovo (alebo iniciál) s veľkým písmenom, ktoré nie je titul priradíme ho k entite a pokračujeme v určovaní rozsahu.

Po dokončení určovania rozsahu entity, označíme posledné označené slovo za priezvisko v prípade, že entita nie je jednoslovná alebo posledné slovo nie je len iniciála. Do zoznamu identifikovaných entít zaradíme celé identifikované meno a taktiež samostatne priezvisko, kvôli ľahšej prípadnej identifikácii ďalších výskytov tejto osoby v texte.

V článkoch slovenskej Wikipédie vieme osoby veľmi dobre identifikovať na základe kategórií, „narodenia v ...“, prípadne aj „úmrtia v ...“. V anglickej verzii potom „ ... births“ a „ ... deaths“. Bodky predstavujú rok kedy sa daná osoba narodila a prípadne umrela.

Identifikovanie lokalít

Pojem lokality označuje vo všeobecnosti geografické názvy. Radíme sem názvy štátov, miest, obcí, ale napríklad aj názvy vrchov a riek. Môžeme o nich s určitosťou povedať, že názov v slovenčine začína veľkým písmenom. Názvy sú väčšinou uvádzané v plnom znení, ale môžeme sa stretnúť aj so skráteným tvarom názvov, kde skracujeme najmä časti slov, ktoré nie sú podstatné mená. Napríklad „Pov. Bystrica“ alebo „Dubnica n. Váhom“. Musíme preto dokázať identifikovať aj tieto skrátené verzie. Okrem toho sa v textoch a najmä v novinových článkoch stretávame s tvarom kedy je lokalita uvedená so všetkými písmenami veľkými. Tento tvar sa vyskytuje väčšinou keď sa uvádza miesto kde sa opisované udalosti odohrali. Metóda dokáže identifikovať aj tento tvar, napriek tomu, že sa nevyskytuje v článkoch z Wikipédie.

Lokality identifikujeme v záverečnom výpise párovými značkami <*ENAMEX TYPE=“LOCATION“*> (Príklad 3).

```
<ENAMEX TYPE=“LOCATION“>Vysoké Tatry</ENAMEX>
```

Príklad 3. Ukážka označenia pomenovanej entity typu lokalita

V článkoch slovenskej i anglickej Wikipédie vieme lokality pomerne dobre identifikovať. Kategórie síce nie sú tak jednoznačné ako u osôb, ale až na niekoľko výnimiek sú označené na úrovni „štáty“, „mestá“, „moria“ a podobne. V anglickej verzii potom sú potom uvedené príslušné anglické slová.

Identifikovanie organizácií

Organizácie sú pomerne dynamicky sa rozvíjajúca skupina pomenovaných entít. Nové spoločnosti vznikajú denne. Spomedzi entít sú zrejme najpočetnejšia skupina, ktorú je prakticky nemožné zapísať do jediného zoznamu. Väčšina štátov má síce svoju formu obchodného registra, no tieto nie sú nikde centralizované. Pomôckou preto môžu byť *kontextové slová*. Napríklad slová „spoločnosť“, „firma,“ alebo koncovky „a.s.“, „n.o., „s.r.o.“. Koncovky indikujú výskyt entity typu organizácia s najväčšou pravdepodobnosťou. Miernou komplikáciou pri entitách tohto typu je však fakt, že môžu mať vo svojom názve interpunkčné znamienka – čiarky a bodky.

Organizácie identifikujeme v záverečnom výpise párovými značkami <*ENAMEX TYPE=“ORGANIZATION“*> (Príklad 4).

```
<ENAMEX TYPE=“ORGANIZATION“>Halali, n.o.</ENAMEX>
```

Príklad 4. Ukážka označenia pomenovanej entity typu organizácia

V článkoch Wikipédie sú organizácie spomedzi hľadaných entít najhoršie rozpoznateľné. Kľúčové slová sú najmä v slovenskej verzii často úplne nepostačujúce a preto niekedy dochádza k zámene s entitami zmiešaného typu. V prípade, že sú však dobre označené, vieme ich vyhľadať podľa slov ako „organizácie“, „založené v ...“, „univerzity“, „strany“.

Identifikovanie dátumu a času

Dátumy sa v texte môžu vyskytovať v niekoľkých formách zápisu. Teoreticky ich síce môže byť mnoho, ale bežne sa dá identifikovať niekoľko prípadov, pomocou ktorých vieme dátumy úspešne rozpoznávať. Dátum začneme identifikovať po nájdení slovne zapísaného mesiaca v texte, prípadne po nájdení reťazca „čč. čč. ččč“ kde *č* predstavuje číslicu. Okrem prípadu, kedy sú jednotlivé časti oddelené bodkami, sa často stretávame s lomítkom v úlohe oddeľovača.

Čas v texte identifikujeme pomocou číslice, za ktorou sa nachádza nejaká forma zápisu časových jednotiek (4. hod, piateho januára).

Dátumy a čas identifikujeme v záverečnom výpise párovými značkami <*TIMEX*> (Príklad 5).

```
<TIMEX>11. apríl 2010</TIMEX>
```

Príklad 5. Ukážka označenia pomenovanej entity typu dátum a čas

Overenie, či je výraz pomenovaná entita typu dátum prípadne čas sa uskutočňuje po tom, čo v texte odhalíme slovo predstavujúce kalendárny mesiac, alebo po načítaní čísla, za ktorým sa nachádza mesiac, prípade určitá forma časovej jednotky. Po nájdení slova zo slovníka dátumov a času skúsime či sa pred ním nachádza číslo, pri dátume oddelené bodkou. Ak bolo nájdené slovo mesiac, overíme ešte či sa za ním nenachádza číslo predstavujúce rok. Takto identifikovaný reťazec označíme ako entitu typu dátum alebo čas.

Identifikovanie čísel

Čísla môžu v textoch predstavovať viacero rôznych entít. Môže sa jednať o peňažné sumy, percentá, vyššie spomenuté dátumy, čas alebo bežné číslo. Medzi číslami, percentami a sumami rozlišujeme pomerne jednoducho. Stačí zistiť či sa v okolí overovaného čísla vyskytuje slovo alebo symbol predstavujúci percentá (%), prípadne niektorú z peňažných mien ($, €).

Čísla všeobecne identifikujeme v záverečnom výpise párovými značkami <*NUMEX*> s uvedením typu *NUM* pre všeobecné čísla, *PERC* pre percentá (Príklad 6) a *MONEY* pre peniaze.

```
<NUMEX TYPE="PERC">15 %</NUMEX>
```

Príklad 6. Ukážka označenia pomenovanej entity typu percento

Po identifikácii čísla vo vstupnom texte overíme prítomnosť kontextových slov alebo znakov (percent, %, $, Eur, €). Ak sa tu niektoré z nich nachádza, identifikujeme entitu podľa tohto typu. Ak nenájdeme ani prítomnosť slov tohto typu identifikujeme výraz ako bežnú číselnú entitu.

Identifikovanie entít zmiešaného typu

V prípade, že sme rozpoznali entitu, ktorá nepatrí do žiadneho z vyššie uvedených typov, zaradíme ju do tohto, všeobecného, typu. Definované typy totiž rozčleňujú entity len do základných skupín. Zvyšné entity sú označované jednotne (Príklad) ako pomenované entity zmiešaného typu.

Typ entít primárne identifikujeme podľa kľúčových slov. Kontextové slová tu totiž majú podstatne nižší vplyv. Do opisovanej skupiny entít zaraďujeme entity, ktoré sa nám podarilo identifikovať, určiť im rozsah, ale nedokážeme im na základe kľúčových slov určiť kategóriu. Keďže nevieme aký typ entite prislúcha, predpokladáme, že sa jedná o entitu zmiešaného typu.

V záverečnom výpise daný typ entít označujeme párovými značkami <*MISC*> (Príklad 7).

```
<MISC>Slovenský pohár</MISC>
```

Príklad 7. Ukážka označenia pomenovanej entity zmiešaného typu

4.2.4 Vplyv kontextových slov

Po správnej identifikácii rozsahu entity je dôležité určiť jej prislúchajúci typ. Táto akcia nastáva vždy po identifikácii rozsahu entity v ktoromkoľvek z vyššie opísaných krokov. Navrhovaný typ určíme v prípade rozpoznania na základe zhody s niektorou databázou entít podľa tejto zhody. V prípade rozpoznania na základe článku z Wikipédie podľa kategórií kam článok patrí. Kategórie porovnáme s množinou charakteristických kľúčových slov, podľa ktorých sme schopní jednotlivé typy entít rozlíšiť. Jedná sa napríklad o slová ako „mestá“, „štáty“, „organizácie“, „založené“, „narodenia“ a podobne.

Tento typ však nemusí byť vo všetkých prípadoch správny. Niekedy totiž môže závisieť od kontextu v ktorom bola entita použitá. Preto pred finálnym rozhodnutím o type entity overíme, či sa v jej okolí, alebo pri číselných entitách priamo v nej nenachádza niektoré zo slov, ktoré pomôžu identifikovať konkrétny prípad správne.

Navrhovaná metóda pokladá pri identifikácii jednotlivých entít za kontextové slová len tie, ktoré konkrétnu entitu bezprostredne obklopujú, prípadne sú jej súčasťou. Kontextové slová nám môžu pomôcť identifikovať typ, do ktorého príslušnú entitu radíme, prípadne ukázať, že rozsah entity nemožno zväčšiť kvôli výskytu stop slov na okolitých pozíciách entity.

Kontextové slová, ležiace mimo entity, ktoré napomáhajú správne identifikovať entity typu osoba, sú napríklad slová „pán“, „pani“, pre lokality sa napríklad jedná o slová ako „mesto“. Slová odhaľujúce organizácie mávajú pred sebou označenia ako „agentúra“ alebo „spoločnosť“, entity neznámeho typu napríklad slovo „portál“.

Kontextové slová sú v niektorých prípadoch aj priamo súčasť aktuálne identifikovanej entity. Príkladom môžu byť akademické tituly osôb alebo označenia spoločností ako „s.r.o.“, či „a.s.“. Kalendárne mesiace sú taktiež množina kontextových slov, ktoré sú priamo súčasťou entít. Veľmi dobre napomáhajú identifikácii dátumov. Okrem nich dátumy identifikujú aj okolité slová ako napríklad slovo „dňa“.

Kontextové slová sú veľmi dôležité z dôvodu zvýšenia úspešnosti rozpoznávania pomenovaných entít. Keby sme neoverovali typ entity získaný pomocou kontextových slov, mohol by nastať prípad, kedy by sme dve entity s rozdielnym významom identifikovali ako dva výskyty jednej entity na základe ich na prvý pohľad rovnakého názvu. Príkladom môže byť osoba, ktorá vlastní firmu nazvanú svojim priezviskom (Príklad 8), prípadne osobu s priezviskom pripomínajúcom názov mesta, žijúcu v meste pomenovanom po osobe (Príklad 9). S využitím kontextových slov nemáme problém identifikovať tieto, inak náročné vety.

```
Pán <ENAMEX TYPE="PERSON">Vojtek</ENAMEX> zakladateľ spoločnosti <ENAMEX TYPE="ORGANIZATION">Vojtek s.r.o.</ENAMEX>
```

Príklad 8. Ukážka dvoch podobne vyzerajúcich entít rozdielneho typu

V texte sme najskôr označili osobu (napríklad podľa slova „Pán“ ležiaceho pred entitou) a neskôr sme narazili na ďalší výskyt rovnomennej entity. Kontextové slovo „s.r.o.“ na jej konci však ukazuje, že sa jedná o entitu typu organizácia.

```
Pani <ENAMEX TYPE="PERSON">Trnavská</ENAMEX> sa narodila v meste <ENAMEX
TYPE="LOCATION">Martin</ENAMEX>.
```

Príklad 9. Ukážka osoby s menom pripomínajúcom lokalitu a lokality s názvom podobným ľudskému krstnému menu

V tomto príklade sme taktiež identifikovali osobu podľa oslovenia na začiatku, mesto má pred svojim názvom kontextové slovo „mesto“.

5 Realizácia navrhnutej metódy

Navrhnutú metódou sme implementovali ako webovú službu, prípadne sa dá využiť aj ako samostatná knižnica. Takto máme možnosť teoretický návrh overiť v praxi a zároveň sme vytvorili riešenie, ktoré umožňuje extrakciu pomenovaných entít z textov pre širokú komunitu ľudí, ktorí ju môžu využiť pri rôznom type odporúčaní prípadne vyhľadávaní. Implementácia metódy je zložená z niekoľkých časti.

5.1 Štruktúra metódy

Metóda sa skladá z predspracovania textu do podoby, s ktorou vieme ďalej efektívne pracovať a zo samotného jadra metódy. Jadro metódy je zložené z niekoľkých celkov. Jeho cieľom je samotná identifikácia potenciálnych entít, určenie ich rozsahu v texte a následné rozpoznanie typov, kam jednotlivé entity patria.

5.1.1 Predspracovanie textu

V úvode predspracovania načítame vstupný text. Ten môže byť zadaný priamo, alebo formou odkazu naň. V prípade odkazu, je nutné aby používateľ zadal link, kde na internete je text možné nájsť. Takto zadaný text následne zo stránky potrebujeme vydolovať. Na to využívame externú službu Readability[1], ktorá po zadaní adresy vráti čistý hlavný text očistený od okolitých metadát, reklám a ostatných, pre nás nepodstatných častí.

Po získaní očisteného textu nasleduje séria ďalších krokov, ktoré text upravia do podoby s ktorou vieme efektívne pracovať. Jedná sa o tieto kroky:

- Obkolesenie nealfanumerických znakov - okrem bodiek a čiarok ležiacich v strede slov a čísel (napr. desatinná čiarka) a bielych znakov – medzerami
- Výmena spojovníkov za pomlčky (články z Wikipédie sú v tomto nejednotné, pretože znaky pre ľudí vyzerajú rovnako, preto výmenu spravíme v texte aj v porovnávaných článkoch Wikipédie a sme schopní porovnávať bez zbytočných chýb)
- Výmena znakov nového riadku za znak „*“ (inak by sme neskôr nevedeli určiť kde končili nadpisy a teda ktoré slová za nimi sú začiatky viet)
- Rozdelenie textu podľa bielych znakov na pole slov
- Stemmovanie každého prvku poľa (slová porovnávame so zoznamom skloňovacích koncoviek radených od najdlhších ku kratším, pre každé slovo postupne porovnávane s jednotlivými koncovkami, kým nenájdeme zhodu alebo neprejdeme všetky koncovky. Ak nájdeme zhodu odrežeme koncovku a opakujeme proces so vzniknutým slovom. Koncovku odstraňujeme len ak bude mať vzniknuté slovo aspoň tri znaky. Toto pravidlo sme zaviedli z dôvodu zachovania predložiek v texte. Ak nenájdeme zhodu prejdeme na nasledujúce slovo.)

[1] *http://peweproxy.fiit.stuba.sk/metall/*

5.1.2 Rozpoznávanie entít

Druhá časť procesu je v porovnaní s prvou mierne obšírnejšia, pretože sa jedná o samotné jadro metódy. To sa skladá sa z niekoľkých funkcií opísaných v tejto kapitole.

Postupným prechádzaním poľa slov získaných zo vstupného textu hľadáme slová písané s prvým veľkým písmenom, čísla alebo slovne zapísané čísla. Pre slová s veľkým písmenom vykonáme všetky nižšie opísané kroky. Proces prerušíme len ak v niektorom kroku identifikujeme slovo ako pomenovanú entitu alebo jej začiatok. Čísla sa pokúsime identifikovať v príslušnom kroku, pre slová s malým písmenom overíme, či sa nejedná o slovný zápis dátumu alebo čísla. Priebeh metódy vhodne dokumentuje Obrázok 5 uvedený v návrhu metódy.

Po úspešnej identifikácii entity, prípadne neúspešnom prejdení všetkých krokov pre dané slovo, sa ukazovateľ inkrementuje na aktuálne overované slovo a posunieme sa na nasledujúce (Fragment kódu 1). Funkcie vykonávajúce sa pri identifikácii potenciálnych entít sú porovnanie overovaných slov s databázou skôr rozpoznaných entít, porovnanie s entitami, ktoré boli rozpoznané skôr v danom texte, identifikácie osôb, porovnanie s databázami obsahujúcimi články Wikipédie a identifikácia dátumov a čísiel.

```
def recognize_NER #funkcia sluziaca na identif. potenc. entit v poli @edited_text
  i = 0               # atribut reprezentujuci index aktualne spracov. prvku pola

  while i < @edited_text.length                  # postupne prechadzame cele pole
   result = false
   if @full_text[i].length >= 1 && start_capital(i)   # ak overovane slovo zacina
                                                 # velkym pismenom a sklada s z aspon 1 znaku
    result, i = is_in_database?(i)      # skusime potencialnu entitu identifikovat
                                         # podla databazy skor rozpoznanych entit
    if ! result                          # ak sa slovo nenachadza v danej databaze
     result, i = find_before_recognized(i)       # overime, ci sme rovnaku entitu
                                                 # nerozpoznali skor v texte
     if !result   # ak overovane slovo nie je nasobny vyskyt skor identif. entity
      result , i = find_persons(i) # overime, ci rozpozn. slovo nie zaciatok mena
      if !result                                             # ak sa nejedna o osobu
       result, i = find_dates_and_nums(i,i)   # zaciname identif. datumov a cisel
       if !result                       # ak overovane slovo nie je ciselna entita
        result, i = find_text_date(i) # overime, ci je slovo textovo zapis. datum
        if !result                                    # ak sme stale nenasli zhodu
         result, i = find_in_svk_dump(i)        #skusime potencialnu entitu najst
                                         # v databaze clankov slovenskej wikipedie
         if !result                                   # ak sme stale nenasli zhodu
          result, i = find_in_eng_dump(i)          # skusime potenc. entitu najst
                                          # v databaze clankov anglickej wikipedie
          if !result               # ak sme overovane slovo neidentif. ako entitu
           i +=1                      # posunieme ukazovatel na nasledujuce slovo
          end
         end
        end
       end
      end
     end
    end
   end
  end
```

```
else                                    # inak(ak slovo zacina s malym pismenom)
 result, i = find_dates_and_nums(i,i)   # zaciname identifikaciu datumov a cisel
 if !result                            # ak overovane slovo nie je cislena entita
  result, i = find_text_date(i)    # overime, ci je slovo textovo zapisany datum
  if !result                # ak sme overovane slovo nedentifikovali ako entitu
   i +=1                            # posunieme ukazovatel na nasledujuce slovo
  end
 end
end
end
...
```

Fragment kódu 1. Priebeh jednotlivých krokov identifikácie potenciálnych entít v texte

Databáza skôr rozpoznaných entít a entity rozpoznané v danom texte

Metóda v prvom kroku využíva dve rôzne funkcie, porovnávajúce začiatok potenciálnej entity so skôr identifikovanými entitami. Najskôr sa pokúšame nájsť zhodu s databázou, kde máme uložené pomenované entity, ktoré metóda rozpoznala v minulosti. V prípade neúspechu sa pokúsime nájsť zhodu overovaného slova so začiatkami entít v zozname entít, ktoré už boli identifikované v danom texte.

Hľadanie v oboch metódach vráti množinu výsledkov. Tá sa následne pre obe spracováva identicky, preto je tento proces reprezentovaný vlastnou funkciou. Táto funkcia nájde najdlhšiu zhodu medzi množinou získanou zo skôr rozpoznaných entít a textom začínajúcim overovaným slovom. Ak je výsledná zhoda nenulová, pristúpime k označeniu entity. Novej entite navrhneme typ podľa zhodnej entity. Samotný proces označovania ešte kontroluje, či navrhovaná entita nie je bežné slovo ležiace na začiatku vety a overuje, či kontextové slová neukážu iný ako navrhnutý typ entity. Proces označenia je opísaný nižšie.

Identifikácia osôb

Tento krok sa vykonáva v prípade, že sme v predchádzajúcom nenašli zhodu. Overované slovo sa porovná so zoznamami krstných mien, akademických titulov a regulárnym výrazom identifikujúcim iniciály. Zoznam mien bol vytvorený spojením slovenských mien uvádzaných v kalendároch získaným z webovej stránky kalendar.azet.sk[1] a výberom najpoužívanejších svetových mien získaným zo stránky behindthename.com[2].

Po identifikovaní niektorej z uvedených možností overuje funkcia nasledujúce slová textu. Pokiaľ opätovne nachádza súvislý rad mien, iniciál alebo titulov zväčšuje rozsah entity. V prípade, že slovo nasledujúce za posledným označeným začína veľkým písmenom, priradíme ho k menu ako priezvisko. Pri označovaní entity predstavujúcej osobu označíme celé meno, do zoznamu identifikovaných entít vložíme celé meno a samostatne aj priezvisko, do databázy skôr identifikovaných entít vložíme len samotné priezvisko.

Databáza článkov z Wikipédie

V prípade, že sme stále neidentifikovali entitu, porovnáme v tomto kroku overované slovo pomocou vyhľadania v databázach tvorených článkami získanými z Wikipédie. Každá databáza sa skladá z článkov v jednom jazyku. Rozpoznávanie slovenských textov využíva

[1] http://kalendar.azet.sk/meniny/
[2] http://www.behindthename.com/top/

slovenskú a anglickú databázu, pre iné jazyky je slovenská databáza nahradená databázou článkov v príslušnom jazyku.

Každá položka v databáze sa skladá z mena, podľa ktorého sú položky usporiadané a z kategórií. Menu zodpovedajú názvy jednotlivých článkov, v kategóriách sú vypísané slová tvoriace kategórie kam v štruktúre Wikipédie príslušné články patria.

Proces identifikácie, či je overované slovo entita a určenia jeho rozsahu je rovnaký ako pri identifikácii v Databáze skôr rozpoznaných entít. Aktuálne opisovaný proces je však schopný identifikovať aj skrátené zápisy entít ako „Pov. Bystrica“ alebo „Nové mesto n. Váhom“, narozdiel od predchádzajúceho procesu, ktorý identifikuje len presnú zhodu.

Podstatný rozdiel, je však v spôsobe určenia typu entity. Pri predchádzajúcom spôsobe máme v databáze uvedený typ akým sme identifikovali zhodnú entitu v minulosti a na základe toho navrhneme typ, ktorému aktuálne overovaná entita prislúcha. Tento typ však priradíme overovanej entite, len v prípade, že nenájdeme rozpor s okolitými kontextovými slovami. V tomto procese (porovnanie s článkami z Wikipédie) prechádzame kategórie rovnomenného článku ako identifikovaná entita a porovnávame ich so zoznamom kľúčových slov, ktorým vieme určiť typ. Po nájdení takéhoto slova v kategóriách článku, navrhneme podľa neho typ rozpoznávanej entity. Tento potom prípadne ešte overíme podľa kontextových slov.

Proces vyhľadávania začína najskôr v slovenskej databáze alebo v príslušnej jazykovej verzii, pre ktorú je metóda použitá. Následne v prípade neúspechu pristupujeme k anglickej verzii.

Parsovanie Wikipédie

Skôr ako sme v práci pristúpili k overovaniu existencie entít pomocou databázy článkov z Wikipédie, prebiehalo overenie priamo dopytovaním do online verzie Wikipédie. Proces bol však z dôvodu optimalizácie času vykonávania metódy nahradený súčasnou podobou, kedy hľadáme v databáze vytvorenej z obrazu portálu.

Podobne ako v aktuálne vykonávanom procese sme najprv skúšali zadať overované slovo do slovenskej verzie a následne v prípade neúspechu sme prešli na anglickú verziu.

Po nájdení zhodného článku, prípadne plnej formy entity, ktorá bola v texte zadaná len skratkou, sme rovnako ako v aktuálnom systéme pomocou kategórií článku navrhli entite typ.

Identifikácia číselných a dátumových entít

Metóda slúžiaca na identifikáciu týchto typov entít sa aktivuje ak je overované slovo číslo, prípade metóda slúžiaca na nájdenie slovne zapísaného čísla vyhodnotí overované slovo kladne.

Po nájdení určitej formy čísla identifikujeme nasledujúce slová z vstupného textu pokiaľ nachádzame zhodu s niektorou vetvou rozhodovacieho stromu slúžiaceho na identifikáciu číselných a dátumových entít. Tento strom (Fragment kódu 6 – uvedený v prílohe A.) je zložený z možností ako môžu jednotlivé druhy entít vyzerať a v každej vetve je uvedené aký typ entite priradiť ak navštívila daný list. Metóda sa rozhoduje najmä podľa charakteristických slov ako sú kalendárne mesiace, peňažné hodnoty („€“,

„dolárov“, „Sk“, ...), časové obdobia („rok“, „hodín“, „sekundy“, „storočie“...), percentá a podobne. Časť stromu uvádzame pre ilustráciu v prílohe A.

Pri identifikácii dátumových entít vychádzame z poznatku, že dátumy sa zapisujú niekoľkými formátmi, ktoré je možné ošetriť regulárnymi výrazmi.

V prípade, že identifikujeme len jednoslovné slovne zapísané číslo, vyhodnotíme ho ako bežné číslo nepredstavujúce entitu a neoznačíme ho.

5.1.3 Označenie entít

Proces označenia pomenovaných entít sa skladá z niekoľkých krokov, ktorých záverom je samotné označenie každej entity príslušnými značkami. Jednotlivé procesy označovania sú podrobne opísané v kapitole Navrhovaná metóda.

Okrem označenia, sú niektoré entity naviac zaradené do zoznamu entít rozpoznaných v danom texte. Do tohto zoznamu zaraďujeme iba entity predstavujúce osobu, lokalitu, organizáciu alebo entity zmiešaného typu. Zoznam slúži na porovnávanie s neskôr nájdenými entitami, aby sme predchádzali opätovnému zdĺhavému identifikovaniu rovnakých entít.

Proces sa aktivuje vždy keď v niektorej z vyššie opísaných metód rozpoznáme entitu a navrhneme jej typ. V úvode tohto procesu overíme, či potenciálna entita leží na začiatku vety a zároveň je tvorené jediným slovom. V prípade že áno, musíme skôr ako ju označíme vylúčiť, že sa jedná o bežné slovo ležiace na začiatku vety, ktoré bolo nesprávne označené ako entita. Slovo teda porovnáme s databázou typických začiatkov viet a ak v nej slovo nájdeme, ukončíme preň proces rozpoznávania entity a vyhodnotíme ho ako bežné slovo. Ak slovo v databáze nenájdeme, otestujeme ho ešte pomocou SNK. Tu zadáme slovo s prvým písmenom malým a v prípade, že dostaneme viac ako polovicu výsledkov začínajúcich s malým písmenom zrušíme identifikáciu entity a toto slovo pridáme do databázy typických začiatkov viet.

Pri overovaní potenciálnych entít v SNK sme dospeli k poznatku, že SNK pomerne často nedokáže vyhovieť rýchlemu sledu požiadaviek, ktorým ho metóda počas rozpoznávania entít dokáže zahrnúť. Z tohto dôvodu sme boli nútení v prípade, že sa hľadané slovo nenachádza v databáze typických začiatkov viet a overenie pomocou SNK nie je možné, automaticky rozhodnúť ako požiadavky vyhodnotíme. Rozhodli sme sa pre negatívny výsledok testov, teda že slová nie sú entita. Zistili sme totiž, že vety častejšie začínajú bežnými slovami ako jednoslovnými entitami, ktoré sme nedokázali identifikovať pomocou databázy skôr rozpoznaných entít.

V prípade, že SNK vrátil odpoveď a zistili sme, že začiatočné slovo vety vo väčšine prípadov začína veľkým písmenom prípadne sa jedná o viacero slov predstavujúcich jeden výraz, rozhodneme že sa jedná o entitu. Tu nastupuje proces overenia správnosti navrhnutého typu. Proces spočíva v overení okolitých slov entity. Tieto porovnáme so zoznamami slov potenciálne ležiacich pred a za entitou. Typy slov boli detailnejšie rozpísané v návrhu metódy. Z pohľadu implementácie je podstatné, že každé zo slov má uvedený typ, ktorý je entite potrebné priradiť.

Po overení kontextových slov metóda entitu v texte označí príslušnými značkami, pridá ju do množiny entít, ktorú po spracovaní celého textu vrátime používateľovi

a v prípade slovnej (nečíselnej a nedátumovej) entity ju taktiež pridá do zoznamu entít rozpoznaných v danom texte, slúžiaceho na porovnávanie s nasledujúcimi entitami..

V prípade, že sme počas identifikácie potenciálnych entít rozhodli, že slovo začínajúce vetu nie je entita, pridáme ho do databázy typických začiatkov viet. V prípade, že sa tu už nachádza, iba zvýšime počet jeho výskytov. Slová, ktoré sme viac ako 10 krát vyhodnotili ako bežný začiatok vety, takto pri následných výskytoch označujeme automaticky, aby sme mohli vynechať časovo náročný prístup do SNK a vyhli sa problémom s častou nedostupnosťou tejto služby.

5.2 Možnosti práce s implementovanou metódou

Metódu sme v rámci implementácie nasadili ako webovú službu. Služba poskytuje možnosť spracovania textu zadaného formou linky priamo naň, prípadne dokáže spracovať priamo zadaný text. Taktiež v úvode musíme zašpecifikovať, aký typ výstupu chceme vrátiť. Možnosti sú dve a to buď celý spracovávaný text s entitami označenými podľa štandardov MUC (Named Entity Task Definition, 1997), alebo kompletný zoznam nájdených entít vo formáte JSON.

Okrem využitia metódy ako spomenutej webovej služby je možné využívať ju priamo ako knižnicu. Po jej nainštalovaní podľa inštalačnej a používateľskej príručky uvedenej v prílohe tejto práce, je možné využívať jej jednotlivé funkcie.

Podrobný manuál k používaniu oboch variantov je uvedený v prílohe, v časti inštalačná a používateľská príručka.

6 Overenie metódy

Úspešnosť navrhnutej metódy pre extrakciu pomenovaných entít sme overili na reálnych spravodajských textoch. Na kvantitatívne vyjadrenie úspešnosti sme zvolili štandardne využívané metriky. V experimentoch sme overili dva rôzne jazyky. Vďaka tomu sme mohli zistiť skutočnú úroveň jazykovej závislosti metódy v rámci skupiny slovanských jazykov, ktorú sme predtým predpokladali.

6.1 Vyhodnocovanie úspešnosti

Úspešnosť metód rozpoznávajúcich pomenované entity v texte sa zvyčajne meria pomocou F-skóre (F-measure, prípadne F-score). Tento výraz označuje mieru správnosti overovanej NER metódy (metódy určenej na rozpoznávanie pomenovaných entít) (Sasaki, 2007). Vyjadrujeme ním pomer dvoch zisťovaných metrík.

Prvá metrika sa nazýva presnosť (precision). Vypočítame ju ako počet správne identifikovaných entít delený počtom všetkých nájdených entít v zvolenom texte. Ako vyplýva už z názvu jedná sa o parameter, ktorý vypovedá o tom s akou presnosťou dokáže metóda správne identifikovať entitu po tom čo ju v texte objaví. Vzorec na výpočet presnosti p

$$p = \frac{i}{n} \tag{1}$$

kde i predstavuje počet správne identifikovaných entít v texte a n predstavuje počet všetkých nájdených entít v zvolenom texte.

Druhú metriku nazývame pokrytie (recall). Jej hodnotu dostaneme po vydelení počtu správne identifikovaných entít v texte počtom všetkých entít, ktoré sa v texte nachádzajú. Vyjadruje teda nakoľko metóda dokáže odhaliť a následne správne identifikovať entitu v texte. Pokrytie označujeme r, vypočítame ho

$$r = \frac{i}{v} \tag{2}$$

pričom i predstavuje počet správne identifikovaných entít v texte a v predstavuje počet všetkých entít, ktoré sa v zvolenom texte nachádzajú.

Presnosť a pokrytie sú navzájom veľmi úzko prepojené. Zvyšovanie jednej metriky zvyčajne zapríčiní pokles druhej. Sú teda vo vzťahu nepriamej úmernosti. Ani jedna z nich osamotene nevypovedá o úspešnosti overovanej metódy (Laclavík, 2007). Tá sa vyjadruje pomocou F-skóre, ktoré vypočítame ako harmonický priemer oboch opísaných metrík. Slúži teda na vyjadrenie, ako úspešne sme dokázali skombinovať obe metriky (Sasaki, 2007). Výpočet F-skóre (označíme ho ako F) formálne zapíšeme

$$F = 2 \cdot \frac{p \cdot r}{p + r} \tag{3}$$

Pri výpočte F-skóre sa môžeme zamerať na niektorú zložku viac a to tak, že jej prisúdime väčšiu váhu ako druhej. Štandardne sa však využíva opísaná technika, ktorú sme si pri overení vybrali aj my.

6.2 Priebeh experimentu

Prvý experiment sme vykonali na sérii novinových článkov dostupných na internete. Overovali sme výsledky F-skóre, ktoré metóda dosahuje na článkoch v slovenskom jazyku. Náhodne sme vyberali články uverejňované na portáloch SME.sk[1], HNonline.sk[2] a topky.sk[3], pre ktoré sme potom vyhodnocovali dosiahnuté výsledky. Zoznam identifikovaných entít sme porovnali voči štandardu, ktorý bol vytvorený doménovým expertom.

Toto prvé komplexnejšie textovanie bolo vykonané s využitím predbežnej verzie metódy, kvôli prezentácii na študentskej vedeckej konferencii IIT.SRC 2012[4] (Kaššák, 2012).

Dosiahnuté výsledky (Tabuľka 1) však boli získané na pomerne malom datasete, 60 článkov. Z celkových 1620 entít sa metóde podarilo správne rozpoznať 1204, u ďalších 229 entít síce identifikovala, že sa jedná o entitu ale neurčila im správny typ. Dosiahnuté F-skóre 79% pokladáme za veľmi dobrý výsledok.

Najväčší počet neidentifikovaných entít zapríčinil fakt, že vo Wikipédii sa nenachádzajú články o pomerne veľkom množstve entít. V prípade slovenskej verzie sú kategórie článkov niekedy uvedené príliš nejednoznačne na to aby sme dokázali správne určiť typ entity. Metóda tiež nedokáže identifikovať neúplne názvy jednotlivých entít, ktoré sa v textoch niekedy uvádzajú. Výsledok najviac degradovali entity zmiešaného typu. Ak by sme ich nerozpoznávali, dosiahli by sme úspešnosť na úrovni F-skóre 84%.

Tabuľka 1. *Výsledky dosiahnuté v prvom experimente – Slovenský jazyk*

Typ	Presnosť	Pokrytie	F-skóre
Osoby	0.97	0.80	0.88
Organizácie	0.94	0.67	0.78
Lokality	0.83	0.73	0.78
Dátumy	0.97	0.76	0.85
Čísla	0.90	0.87	0.88
Percentá	0.83	0.68	0.75
Peňažné sumy	1.00	0.76	0.86
Zmiešané	0.50	0.66	0.57
Celkovo	**0.84**	**0.74**	**0.79**
Celkovo bez zmiešaných entít	**0.92**	**0.77**	**0.84**

Druhé experimentálne overenie na slovenských článkoch bolo vykonané na rovnakých 3 webových portáloch ako pri prvom overení, navyše sme metódu testovali aj na článkoch portálu TERAZ.sk[5]. Portál má totiž pri každom článku uvedených niekoľko kľúčových slov, na základe ktorých sme si dokázali overiť, či sa nám podarilo nájsť najpodstatnejšie entity z pohľadu autora textu, ktorý ich tu zadal. Po vykonaní experimentov sme zistili, že metóda v texte vždy našla explicitne vyjadrené entity. Tento fakt sme overovali

[1] http://www.sme.sk/
[2] http://hnonline.sk/
[3] http://www.topky.sk/
[4] http://www.fiit.stuba.sk/generate_page.php?page_id=3529
[5] http://www.teraz.sk/

zhodnotením doménového experta. Výsledky experimentu zameraného na rozpoznanie entít v samotnom texte sme rovnako ako v prvom experimente overili zhodnotením ľudského experta. Dosiahnuté hodnoty sú zobrazené v Tabuľke 2.

Tabuľka 2. *Výsledky dosiahnuté v druhom experimente – Slovenský jazyk*

Typ	Presnosť	Pokrytie	F-skóre
Osoby	0.97	0.83	0.89
Organizácie	0.92	0.75	0.83
Lokality	0.92	0.83	0.87
Dátumy	0.99	0.95	0.97
Čísla	0.97	0.96	0.97
Percentá	0.98	0.87	0.92
Peňažné sumy	0.99	0.99	0.99
Zmiešané	0.53	0.71	0.61
Celkovo	**0.88**	**0.83**	**0.85**
Celkovo bez zmiešaných entít	**0.95**	**0.84**	**0.89**

V tomto experimente sme metódu nechali vyhodnotiť 200 článkov. Z každého portálu sme náhodne vybrali 50 textov. Texty sme vyberali z aktuálne publikovaných článkov. V textoch sa metóde podarilo rozpoznať 4212 pomenovaných entít z 5083. U ďalších 574 entít síce metóda rozpoznala, že sa jedná o entitu, ale neurčila jej správne typ. Výsledok 85% F-skóre predstavuje výrazné zlepšenie oproti prvému experimentu. Tento rozdiel je podľa nás spôsobený kombináciou 2 faktorov. V prvom rade sme zoptimalizovali metódu tak, aby sme zamedzili pravidelne sa vyskytujúcim prípadom, kedy metóda v prvom overení metódy nerozhodla správne a v druhom rade tým, že prípadné chyby mali pri rozsiahlejšej množine menšiu váhu oproti prvému testu. V druhom experimente sme totiž spracovali viac ako trojnásobný počet textov oproti prvému testovaniu. Výsledok experimentu opäť najviac znižovali entity zmiešaného typu. Entitu totiž označíme ako entitu zmiešaného typu akonáhle nevieme rozhodnúť, že patrí inému typu. Výsledok bez zarátania tohto typu entít je uvedený v spodnej časti Tabuľky 2.

Výsledky experimentu dokazujú, že sa nám pomerne dobre podarilo vysporiadať so špecifikami slovenského jazyka. Vyplýva z nich, že spôsob odstránenia skloňovacích koncoviek, ktorý sme použili, funguje pre dostatočne veľké percento slov a je na jeho základe možné upraviť text do strojovo spracovateľnej formy. Nedokázali sme však pokryť všetky prípady. Problémy stále spôsobujú prípady ako napríklad slovo „Tatry“, ktoré má v určitej forme skloňovania, tvar „Tatier“. V databáze sa vyskytuje základný tvar slova. Stemmer, ktorý využívame upraví slovo „Tatry“ na tvar „Tatr“ a tento tvar očakáva aj v texte. Slovo „Tatier“ však stemmer na základný tvar nedokáže upraviť a pri porovnávaní so slovníkom sa teda nenájde zhoda.

Rovnako sa stretávame s prípadom, kedy chybne identifikujeme typ, kam entitu zaraďujeme. To spôsobuje fakt, že databáza entít, získaných z Wikipédie nemá u časti entít vhodne definované kategórie (viac u slovenských ako anglických) a teda nedokážeme správne rozlíšiť o aký typ entity sa jedná. Pokiaľ entitu nevieme zaradiť do žiadnej z rozpoznávaných kategórií, označíme ju ako entitu zmiešaného typu. Z toho vyplýva, že značná časť chýb spočíva v priradení zmiešaného typu entitám, ktoré sú síce iného typu,

ale z kategórií získaných z Wikipédie nie sme schopní identifikovať správny typ. Zníženiu výskytov chýb tohto typu sme venovali značné úsilie a to najmä identifikáciou množstva kľúčových slov typických pre jednotlivé typy.

Wikipédia okrem nie vždy vhodných kategórií tiež nepokrýva časť organizácií, ktoré sa vyskytujú v textoch. Metóda je síce schopná učiť sa a každú zo správne identifikovaných entít (a opravených nesprávne identifikovaných) si zapamätá a pri ďalšom výskyte v inom texte ju už určí správne, avšak proces je zdĺhavý. Aj po množstve spracovaných textov nachádzame stále nové entity, ktoré metóda nepozná. Postupne sa však jej výsledky zlepšujú, pretože každým spracovaním textu rozširuje svoju databázu entít. Entity, ktoré metóda raz rozpoznala správne si zapamätá a v nasledujúcich textoch ich vyhodnotí rovnako (pokiaľ z kontextových slov nevyplynie nová situácia). Tento fakt okrem správnej identifikácie nepriamo zvyšuje aj rýchlosť spracovania textov keďže entity sú identifikované v prvom možnom kroku a nevyžadujú vykonanie ostatných krokov identifikácie.

6.3 Jazyková závislosť

Metódu sme sa snažili navrhnúť tak, aby bola použiteľná pre viacero jazykov. Jazyková nezávislosť je totiž prostriedok, ako možno zaistiť širšie využite navrhnutého princípu.

Slovanské jazyky majú podobnú štruktúru viet, vzájomné postavenie vetných členov a princíp ohýbania slov (Przepiórkowski, 2007). Každý jazyk má, pravdaže, svoje konkrétne mechanizmy ako vytvárať slovné tvary v jednotlivých rodoch, či časoch, no dá sa povedať, že spôsob spočívajúci v obmieňaní koncoviek so zachovaním slovného základu je jednotný.

Metódu sme teda navrhli tak, aby bola po výmene databáz a zoznamov špecifických pre konkrétne jazyky v čo najvyššej miere univerzálne použiteľná pre čo najviac jazykov. Spôsob akým sa predspracúva vstupný text a následne sa rozpoznávajú entity ju však obmedzuje len na rodinu slovanských, prípadne podobných flexívnych jazykov.

Primárne sme metódu navrhli pre slovenský jazyk. Pre tento jazyk sme taktiež vykonali väčšinu experimentov. Na potvrdenie našich predpokladov o jazykovej nezávislosti sme však chceli metódu overiť pre aj pre iný jazyk. Rozhodli sme sa pre češtinu. Tento jazyk je spolu so slovenským zástupcom západoslovanskej vetvy jazykov, čo značí vysokú podobnosť oboch jazykov. Jedná sa však o samostatné jazyky s vlastnými tvarmi slov a spôsobom skloňovania. Česká abeceda má dokonca oproti slovenskej naviac tri písmená a naopak šesť slovenských vôbec nepoužíva.

Keďže sa teda jedná o samostatný jazyk, bolo možné overiť jazykovú závislosť metódy. Bolo však nutné vymeniť určité jazykovo závislé časti metódy, ako zoznam kalendárnych mesiacov, čísel alebo časti rozhodovacích stromov, za príslušné ekvivalenty. Výsledky experimentov pre cudzí jazyk sú uvedené v Tabuľke 3.

Tabuľka 3. *Výsledky dosiahnuté v treťom experimente – Český jazyk*

Typ	Presnosť	Pokrytie	F-skóre
Osoby	0.96	0.66	0.78
Organizácie	0.98	0.50	0.66
Lokality	0.83	0.62	0.71
Dátumy	0.97	0.71	0.82
Čísla	0.91	0.90	0.90
Percentá	0.98	0.98	0.98
Peňažné sumy	0.99	0.96	0.97
Zmiešané	0.33	0.77	0.46
Celkovo	**0.72**	**0.70**	**0.71**
Celkovo bez zmiešaných entít	**0.93**	**0.68**	**0.78**

V tomto experimente sme spracovali 60 novinových článkov. Po 30 textov sme získali z portálov iDNES.cz[1] a blesk.cz[2]. V textoch sa spolu nachádzalo 1564. Z nich metóda dokázala správne identifikovať 1090. V 424 prípadoch síce entitu našla, no nesprávne určila typ. To zodpovedá úspešnosti na úrovni 71%, bez zarátania entít zmiešaného typu dokonca 78% (F-skóre). Dosiahnutý výsledok ukázal nižšie hodnoty úspešnosti metódy v porovnaní s úspešnosťou, ktorú metóda dosiahla na slovenských textoch. Myslíme si ale, že sa stále jedná o pomerne dobrý výsledok. Metóda bola z dôvodu testovania síce upravená pre potreby českého jazyka (nahradenie databázy s článkami zo slovenskej Wikipédie za českú databázu, výmena niektorých kľúčových slov v rozhodovacích stromoch, výmena slovníkov so špecifickými slovami ako sú kalendárne mesiace a podobne), no netreba zabúdať, že slovenskú verziu sme pred jednotlivými testovaniami počas vývoja natrénovali na oveľa väčšom množstve dát.

V rámci testovania úspešnosti metódy sme taktiež vykonali porovnanie našej metódy s existujúcim riešením pre český jazyk. Riešenie pre slovenčinu sa nám, ako sme spomínali skôr, nepodarilo získať. Porovnávanú českú metódu vytvoril tím vedcov na Karlovej univerzite v Prahe ako niekoľkoročný projekt zameraný na extrakciu pomenovaných entít z českých textov (Ševčíková, Žabokrtský & Krůza, 2007). Táto metóda je založená na metóde SVM a využíva princíp strojového učenia (Kravalová, Žabokrtský, 2009).

Pri rozpoznávaní sa zameriava na vyššiu úroveň práce ako naša metóda. Okrem základných kategórií entít, identifikuje aj ich jednotlivé časti. Napríklad pri osobách rozlišuje tituly, krstné mená, druhé mená, priezviská, mená mýtických postáv a podobne.

Pre účel porovnania metód sme brali do úvahy len označenia prvej úrovne typu entít, teda či sa jedná o osobu, organizáciu a podobne. Z množiny cca 2000 viet, pre ktoré sme mali k dispozícii označkované entity, ktoré našla porovnávaná metóda sme náhodne vybrali časť tak, aby sa v tejto časti nachádzal približne rovnaký počet entít ako v treťom experimente.

Text sme v tomto experimente najprv nechali spracovať českej verzii našej metódy, následne sme ho preložili pomocou služby Google translate[3] do slovenčiny. Takto

[1] http://www.idnes.cz/
[2] http://www.blesk.cz/
[3] http://translate.google.com/

preložený text sme nechali spracovať slovenskej verzii metódy. Úspešnosť porovnávanej metódy sme získali vyhodnotením textu, v ktorom bol zaznamenaný výstup rozpoznávania pomenovaných entít porovnávanou metódou. Dosiahnuté výsledky ukazuje Tabuľka 4.

Tabuľka 4. *Výsledky dosiahnuté v treťom experimente – porovnanie metód (P - presnosť, R- pokrytie, F- F-skóre)*

	Naša metóda SK			Naša metóda CZ			Porovnávaná metóda		
Typ	P	R	F	P	R	F	P	R	F
Osoby	0.93	0.68	0.79	0.94	0.70	0.80	0.98	1.00	0.99
Organizácie	0.84	0.74	0.79	0.85	0.68	0.76	0.96	0.97	0.96
Lokality	0.91	0.51	0.65	0.87	0.52	0.65	0.93	0.96	0.93
Dátumy	1.00	0.99	0.99	0.98	0.98	0.98	1.00	1.00	1.00
Zmiešané	0.40	0.61	0.48	0.31	0.64	0.41	0.96	0.92	0.94
Celkovo	**0.81**	**0.67**	**0.73**	**0.78**	**0.68**	**0.73**	**0.96**	**0.95**	**0.95**
Celkovo bez zmiešaných entít	**0.89**	**0.69**	**0.78**	**0.88**	**0.69**	**0.77**	**0.96**	**0.97**	**0.96**
Čísla	0.96	0.97	0.97	0.96	0.95	0.96	1.00	0.58	0.73
Percentá	1.00	1.00	1.00	1.00	1.00	1.00	-	-	-
Peňažné sumy	1.00	1.00	1.00	1.00	1.00	1.00	-	-	-

V experimente pre český jazyk sme vybrali text, v ktorom sa nachádzalo 1216 entít (okrem čísel, percent a peňažných súm). Porovnávaná metóda z Karlovej univerzity správne rozpoznala 1180 entít, v 49 prípadoch určila nesprávne typ.

Naša metóda v českom texte správne identifikovala 839 entít, v 114 prípadoch určila nesprávny typ. Po preložení textu do slovenčiny sa niektoré entity znehodnotili a nedali sa následne vyhodnotiť. Napríklad niektoré priezviská prekladač zmenil na začínajúce malým písmenom a podobne. V texte sa spolu nachádzalo 1198 entít, ktoré bolo možné identifikovať. Metóda správne rozpoznala 827 entít, v 102 prípadoch určila nesprávny typ.

Porovnávaná metóda nerozpoznáva percentá a peňažné sumy, tie sú preto v tabuľke oddelené a nezapočítavajú sa do výsledkov. Rovnako sú oddelené čísla, pretože porovnávaná metóda ich označovala výrazne menej ako iné typy entít a bolo zbytočné kaziť jej výslednú úspešnosť.

Hlavné príčiny nižšej úspešnosti našej metódy voči porovnávanej spočíva v tom, že porovnávaný text obsahoval mnoho entít, ktoré sa nenachádzali v nami využívaných databázach. Jednalo sa najmä o lokálne názvy organizácií a lokalít, kde neboli uvedené celé názvy, ale len bežne používané varianty. často sa tiež vyskytovali prezývky osôb, prípadne len priezviská bez predchádzajúceho uvedenia mena, rovnako sa tu nachádzalo mnoho cudzojazyčných slov, či názvov umeleckých diel. Ďalším aspektom bolo, že text, na ktorom sme vykonali tento experiment bol tvorený zoskupením nesúvisiacich viet. Jednotlivé vety síce dávali zmysel ale vzájomne spolu nesúviseli. Prevažná väčšina entít bola teda spomenutá len jedenkrát,

Na dosiahnutie vyššej úspešnosti našej metódy by bolo potrebné využívať komplexnejšie databázy osôb, organizácií a lokalít obsahujúce väčší počet entít, rozlíšiteľných z hľadiska príslušnosti do príslušného typu.

Na druhej strane úspešnosť našej metódy sa zvyšuje každým vyhľadávaním, keďže metóda je schopná učiť sa. Preto veríme, že rozsiahlejšie natrénovanie metódy umožní zvýšenie úspešnosti F-skóre.

Dôležitým aspektom vplývajúcim na úspešnosť je podľa nášho názoru aj druh textov, ktoré metóda spracováva. Pri práci so súvislými textami (akými sú napríklad novinové články, na ktorých sme vykonali prvé tri experimenty), kde neboli informácie vytrhnuté z kontextu, dosiahla naša metóda vyššiu úspešnosť F-skóre. Entity sú v týchto textoch zväčša najprv krátko predstavené, napríklad plným názvom entity spolu s jednoduchým okolitým kontextom, ktorý slúži na objasnenie informácií prípadne vzťahov medzi entitami pre čitateľa. Pomocou týchto informácií dokáže naša metóda entity rozlišovať spoľahlivejšie v porovnaní s prípadom, kedy text tvoria náhodne pospájané vety.

7 Záver

V rámci projektu sme navrhli metódu určenú na rozpoznávanie pomenovaných entít pre texty v slovenskom a jazyku. Metódu sme navrhovali tak aby bola využiteľná aj pre iné slovanské jazyky. Metóda má za úlohu rozpoznať v zadanom texte pomenované entity. Okrem ich rozpoznania musí v rámci identifikácie taktiež priradiť entitám typ, do ktorého ich zaraďujeme. Na identifikáciu entít využíva metóda porovnávanie predspracovaného textu s vlastným, postupne rozširovaným slovníkom entít, databázou názvov článkov z Wikipédie a niekoľko menších zoznamov (zoznam krstných mien, kalendárnych mesiacov a pod.). Okrem toho využíva, najmä na identifikáciu čísel a dátumov, rozhodovacie stromy. Kvôli ohybnosti slovanských slov metóda vstupný text najprv upraví pomocou odstránenia koncoviek slov vzniknutých skloňovaním prípadne časovaním.

Metódu sme overili pomocou spracovania množiny novinových článkov z rôznych webových portálov. Takto sme získali veľmi rôznorodé texty, na ktorých sme mohli vyhodnotiť aké výsledky metóda dosahuje. Texty boli po spracovaní vyhodnotené doménovým expertom, ktorý určil správnosť výsledkov.

Zistili sme, že metóda pracuje pre slovenčinu na pomerne dobrej úrovni. Metódu je preto podľa nás možné využívať ako zdroj relevantných entít pre rôzne metódy odporúčania na základe obsahu, prípadne na vyhľadávanie textov. Okrem slovenského jazyka sme metódu overili pre český jazyk. Dosiahnuté výsledky ukázali, že metóda je z veľkej miery jazykovo nezávislá minimálne v rámci skupiny slovanských jazykov. Dosiahnuté výsledky boli mierne horšie ako v prípade slovenského jazyka. Primárne sme sa totiž snažili dosiahnuť najlepšie možné výsledky pre slovenčinu, pre ktorú sme metódu nasadili aj na reálne používanie[1]. Z výsledkov experimentov ale vyplynuli potenciálne smery ďalšej práce, ktorými môže byť skvalitnenie jazykovej nezávislosti metódy, prípadne zameranie sa na konkrétne príbuzné jazyky a špecifikácia metódy tak, aby pre tieto jazyky dosahovala lepšie výsledky.

Metódu sme implementovali ako webovú službu, aby mohla byť využívaná ako podporný nástroj pre spomínané odporúčania, prípadne vyhľadávanie, pričom pracujeme na jej integrácii do služby peweproxy/metall[2].

Metóda sa postupne učí, raz rozpoznané entity ďalej identifikuje bez nutnosti vykonávania opätovného procesu identifikácie, čím postupne skracuje čas svojho vykonania. Naučené entity však nevyužíva, kým ich správnosť nepotvrdí ľudský expert. Na druhej strane nieje potrebné pravidelne potvrdzovať správne entity, pretože metóda ich po prekročení definovaného limitu schváli sama. Takto totiž môže využívať získané poznatky na kvalitnejšie a rýchlejšie rozpoznávanie pomenovaných entít.

[1] http://kassak.visnovsky.sk/ner/
[2] http://peweproxy.fiit.stuba.sk/metall/

Literatúra

CHEN, P. H., LIN, C. J., & SCH**Ö**LKOPF, B. (2005). A tutorial on v-support vector machines. *Applied Stochastic Models in Business and Industry*, *21*(2), (pp. 111–136). Wiley Online Library.

GRISHMAN, R., & SUNDHEIM, B. (1996). Message understanding conference-6: A brief history. *Proceedings of COLING*, *96*, (pp. 466–471).

KAŠŠÁK, O. (2012). Named Entity Recognition for Slovak and Related Languages. S*tudent Research Conference 2012. Vol. 1. 8th Student Research Conference in Informatics and Information Technologies*, Bratislava (Slovakia): STU v Bratislave FIIT, (Vol. 1, pp. 15-20)

KONKOL, M., & KONOPÍK, M. (2011). Maximum entropy named entity recognition for Czech language. *Text, Speech and Dialogue,* (pp. 203–210). Springer.

KRAVALOVÁ, J., & ŽABOKRTSKÝ, Z. (2009). Czech Named Entity Corpus and SVM-based Recognizer. *NEWS '09 Proceedings of the 2009 Named Entities Workshop: Shared Task on Transliteration* (August), (pp. 194-201).

LACLAVÍK, M., ŠELENG, M., & CIGLAN, M. (2007) Vyhľadávanie informácií, (pp. 17-20) Dostupný Z WWW: http://ikt.ui.sav.sk/vi/vi_laclavik.pdf [2011-04-15].

LAFFERTY, J., McCALLUM, A., & PEREIRA, F. C. N. (2001). Conditional random fields: *Probabilistic models for segmenting and labeling sequence data.* Computer.

MARCIŃCZUK, M., & PIASECKI, M. (2011). Study on named entity recognition for Polish based on hidden Markov models. *Text, Speech and Dialogue,* (pp. 142–149). Springer.

NAMED ENTITY TASK DEFINOTION (1997). Dostupný z WWW: http://www-nlpir.nist.gov/related_projects/muc/proceedings/ne_task.html [2012-04-15]

NIGAM, K., LAFFERTY, J., & McCALLUM, A. (1999). Using maximum entropy for text classification. *IJCAI-99 workshop on machine learning for information filtering,* (Vol. 1, pp. 61–67).

PRZEPIÓROWSKI, A. (2007). Slavonic Information Extraction and Partial Parsing. *Computational Linguistics*, (June), (pp. 1-10).

RABINER L. R., & JUANG, B. H. (2007). An introduction to hidden Markov models. *Current protocols in bioinformatics*, (January).

SASAKI, Y. (2007). The truth of the F-measure. (October), (pp. 1-5).

SLOVAKE. O jazyku - Podoba a štruktúra (2011). Dostupný z WWW: http://slovake.eu/sk/intro/language/form [2011-11-24]

SLOVENSKÝ NÁRODNÝ KORPUS. Bratislava: Jazykovedný ústav Ľ. Štúra SAV (2011). Dostupný z WWW: http://korpus.juls.savba.sk [2011-11-24]

ŠEVČÍKOVÁ, M., ŽABOKRTSKÝ, Z. & KRUZA, O. (2007). Zpracování pojmenovaných entit v českých textech. Dostupný z WWW: http://ufal.mff.cuni.cz/~zabokrtsky/reports/techrep-ne-2007.pdf [2011-04-24]

Prílohy

A. Technická dokumentácia
B. Inštalačná a používateľská príručka
C. Príspevok publikovaný na konferencii IIT.SRC 2012
D. Obsah elektronického média

A. Technická dokumentácia

Metódu sme implementovali v jazyku Ruby. Keďže sme výsledky vytvorenej aplikácie chceli poskytnúť najmä ako api, vložili sme aplikáciu do Rails projektu, ktorý sme mohli nasadiť na server. Projekt sme vytvorili vo vývojovom prostredí Aptana Studio 3. Aplikácia bola vytvorená v Ruby 1.9.3 a Rails 3.1.3., využíva MySQL 5.0.8 databázu.

A.1. *Príklady použitých funkcií*

Vo fáze predspracovania je tvorí metódu niekoľko menších funkcií. Funkcia *preprocess* Fragment kódu 2) slúžiaca na predspracovanie textu využíva v prípade zadania textu cez url funkciu *readability* (Fragment kódu 3). Tá získa textový obsah zadanej stránky dostupnej cez url využitím externej webovej služby Readability. Predspracovanie využíva funkciu *split_text* (Fragment kódu 4) na rozdelenie textu na pole slov. Na orezanie skloňovacích koncoviek využíva funkciu *chop_array* (Fragment kódu 5).

```
def preprocess(input,link)           # funkcia sluziaca na predsprac vstup textu
 if link
  @full_text = split_text(NER_recognizer.readability(input)
      .force_encoding('UTF-8'))
 else
  @full_text = split_text(input.force_encoding('UTF-8'))
 end
 @edited_text = @full_text.dup             #vytvor. kopie textu, kt. ostemmujeme
 chop_array(@edited_text)       # orezeme sklon. koncovky v slovach vstup. textu
end
```

Fragment kódu 2. Funkcia slúžiaca na predspracovanie textu

```
 def self.readability(link)
  base_uri 'http://peweproxy.fiit.stuba.sk/metall/readability/?url=' + link
                                  # uri sluzby Readability, ktorou ziskavame text z clanku
  post(               #ziskanie textu zo zvoleneho clanku pomocou externej sluzby
   base_uri,
   :headers => {
    'Content-Type' => 'application/x-www-form-urlencoded; charset=utf-8',
    'Accept-Charset' => 'utf-8'
   }
  )
 end
```

Fragment kódu 3. Funkcia slúžiaca na získanie textového obsahu zo stránky zadanej cez url pomocou externej služby Readability

```
def split_text(text)            # funkcia sluziaca na rozdel. textu na pole slov
 rgxp = /([a-zääíýóúäônšžctldl]\.)|0xC2|0xA0|\ /
 text = text.gsub(rgxp){|rgxp|"#{rgxp} "}

 regexp = /([\.\,\:]([\s\„\"\"\,\)]|\z))|[\…\?\!\'\"\„\"\(\)\;\€\$\%\-\-\+\´\/]|\s+/
      # regexp identif. znaky, kt. treba pred rozdel. textu obklopit medzerami
 return (text.gsub(/\n/," * ").gsub(regexp){|regexp|" #{regexp} "}).gsub("–","-")
        .split(" ")
  # v texte nahradime entery znakoom '*', pridame biele znaky okolo interpunkcie
      # a vymenime spojovniky za pomlcky. text rozdelime podla medzier na pole
end
```

Fragment kódu 4. Funkcia slúžiaca na rozdelenie vstupného textu na pole slov

```
def chop_array(text) # funkcia sluziaca na odstran. sklonov. koncoviek zo vstup.
                                                                  # textu
 word_ends = ["atami","atách","atom","iach","ach","ami","ati","atá","ata","atu",
              "eho","emu","iam","ich","imi","och","ych","ymi","ách","ého","ému",
              "ích","ími","ých","ými","am","ej","ho","ia","ie","im","in","iu",
              "mi","mu","om","ou","ov","ym","ám","ím","ín","ým","a","e","i","o",
              "u","y","á","é","í","ú","ý"]
            # zoznam slovenskych koncoviek, kt. potrebujeme stemmovanim odstranit
 for i in 0 ... text.length                  # pre kazdne slovo zo vstupneho pola
  result = true                                        # nastavime result na true
  while result             # ak sme danemu slovu v minulom vykonani orezali konc.
   result = false
   word_ends.each do |aktual|               # pre kazdu koncovku z pola word_ends
    if text[i].end_with?(aktual)            # ak over. slovo konci akt. koncovkou
     if(text[i].length - aktual.length > 2)  # a po orezani koncovky ostanu viac
                                                        # ako 2 pismena
      text[i] = text[i][0... (text[i].length - aktual.length)]
                                              # orezeme koncovku aktualneho slova
      result = true
      break                                 # ukoncime predspracovnaie daneho slova
     end
    end
   end
  end
 end
 return text                                                # vratime upraveny text
end
```

Fragment kódu 5. Funkcia slúžiaca na orezanie skloňovacích koncoviek v poli zadaných slov

Časť metódy riešiaca rozpoznávanie pomenovaných entít je v základe tvorená hlavnou funkciou (Fragment kódu 1 – uvedený v implementácii), ktorá postupne pre každé slovo zo vstupného textu vykonáva sériu overení, ktorými sa pokúša identifikovať slová a rozpoznať prípadné pomenované entity. Overenia vykonáva vždy v samostatných funkciách. Ich konkrétna funkcionalita bola opísaná v návrhu a v implementácii.

Príkladom takejto funkcie je rozhodovací strom určený na identifikáciu dátumov a čísel. Podstatnú časť stromu obsahuje funkcia *find_dates_and_nums* , ktorej začiatok zobrazuje Fragment kódu 6.

```
def find_dates_and_nums (first,i)     # funkcia na rozpoznavanie datumov a cisel

 months = Set["apríl", "august", "december", "decembr", "február", "január",
              "júl", "jún", "kvartál","marc", "marec", "máj", "november",
              "novembr", "október", "októbr", "september", "septembr"]
                                                          # slovnik mesiacov
 number = /\A\#?\d+[\,\.\:]?\d*\z/      # regexp sluziaci na identifikaciu cisel
 date = /\A\d{1,4}[\.\/]{1}\d{1,2}[\.\/]{1}\d{0,4}\z/#regexp na identif. datumov
 j = i + 1               # j = index slova 1. nasled. za over. slovom(indexom i)

 if (@edited_text[i].match(number) || text_number(UnicodeUtils.downcase
     (@edited_text[i]),i,i))                        # ak je overovane slovo cislo
  if ((@edited_text[j] != nil )&&(@edited_text[j].match(number)
     || text_number(UnicodeUtils.downcase(@edited_text[j]),i,j)))
                                                    # ak je 2. nasled. slovo cislo
   return find_dates_and_nums(first,j)      # overujeme rekurzivne nasled. slovo
                             # (oznacime obe najdene cisla do 1 vstupnej entity)
  else                                      # inak (ak nie je nasl. slovo cislo)
   if @edited_text[j] != nil
    case UnicodeUtils.downcase(@edited_text[j])       # ak je 1.nasled slovo ...
    when "."                                                  #... bodka tak:
     if months.include?(UnicodeUtils.downcase(@edited_text[j+1]))
                                              # ak je 2. nasledujuce slovo mesiac
      if @edited_text[j+2].match(/\A\d{1,4}\z/)# ak je 3.nasled slovo cislo(rok)
       j +=2                                  # koniec entity je 3. nasledujuce slovo
       result = mark_entity(first,j,"","T",false,true,false)
                             # oznacime najdenu entitu tagmi pre datum (TIMEX)
      else                              # ak nie je 3. nasledujuce slovo cislo(rok)
       j +=1                                  # koniec entity je 2. nasledujuce slovo
       result = mark_entity(first,j,"","T",false,true,false)
                             # oznacime najdenu entitu tagmi pre datum (TIMEX)
      end
     ...
```

Fragment kódu 6. Ukážka časti rozhodovacieho stromu slúžiaceho na identifikáciu čísel a dátumov

Rozpoznané entity sú reprezentované ako objekty triedy *Entity*. V tejto podobe s entitami pracujeme počas celého priebehu identifikácie. Triedu zobrazuje Fragment kódu 7. Pre každú entitu si pamätáme atribúty *real_name* – názov entity v základnom tvare, *name* – tvar entity s ktorým pracujeme, *type* – identifikovaný typ entity. Nepovinné atribúty sú index začiatku a konca entity v texte (*start_index*, *end_index*).

```ruby
# -*- encoding : utf-8 -*-

class Entity
 attr_accessor :real_name, :name, :type, :start_index , :end_index

 def initialize(*args)                              # inicializacna funkcia objektu
  if args.length == 3                                  # ak boli zadané 3 argumenty
   @real_name = args[0]
   @name = args[1]
   @type = args[2]
  elsif args.length == 5                                        # alebo 5 argumentov
   @real_name = args[0]
   @name = args[1]
   @type = args[2]
   @start_index = args[3]
   @end_index = args[4]
                                                  # inicializujeme object typu Entity
  else                                                  # inak vypiseme chybovu hlasku
   raise ArgumentError, "wrong number of arguments (#{args.length} for (3 or 5))"
  end
 end
end
```

Fragment kódu 7. Reprezentácia triedy *Entity*. Pri vytvorení objektu musí byť zadané real_name, name a type. Indexy ohraničujúce entitu v texte zadané byť nemusia.

A.2. *Model prípadov použitia*

Opis modelu prípadov použitia podlieha Jacobsonovej notácii. Model slúži na objasnenie možných akcií používateľa. Model na Obrázok 6 opisuje využitie webovej služby, model na Obrázok 7 platí pre prácu s metódou dostupnou formou knižnice, kedy si sám používateľ volí kedy vykoná jednotlivé časti poskytovanej metódy.

Obrázok 6. Model prípadov použitia webovej služby

UC1 Získanie pomenovaných entít zadaného textu
Používateľ si zvolí formu zadania textu, v ktorom chce vyhľadať pomenované entity a taktiež si zvolí formu výstupu. Služba spracuje text a vráti výstup v požadovanej forme.

Hlavný tok Získanie zoznamu pomenovaných entít z textu dostupného cez zadanú url
Vstupné podmienky: Text dostupný pomocou zadanej url
Výstupné podmienky: Získanie zoznamu pomenovaných entít
Účastníci: Používateľ

1. Používateľ zadá požiadavku na extrakciu pomenovaných entít z textu, ktorého url zadal. Používateľ rovnako zvolí typ požadovaného výstupu, konkrétne zvolí možnosť ‘words’, čím definuje, že chce vrátiť zoznam nájdených entít.
2. Služba predspracuje text a rozdelí ho na pole slov.
3. Služba postupne prejde celé pole slov a identifikuje pomenované entity.
4. Služba vráti požadovaný zoznam pomenovaných entít.
5. Používateľ môže pracovať so zoznamom pomenovaných entít identifikovaných v zadanom texte.
6. Prípad použitia končí.

Alternatívny tok Získanie označkovaného textu z priamo zadaného textu
Vstupné podmienky: Text dostupný v elektronickej forme
Výstupné podmienky: Získanie označkovaného textu
Účastníci: Používateľ

1. Používateľ zadá požiadavku na extrakciu pomenovaných entít z textu, ktorý zadal službe . Používateľ rovnako zvolí typ požadovaného výstupu, konkrétne zvolí možnosť ‘text’, čím definuje, že chce vrátiť zoznam nájdených entít.
2. Tok prebieha rovnako ako body 2 – 3 hlavného toku.
3. Služba vráti označkovaný text.
4. Používateľ môže pracovať s označkovaným textom.
5. Prípad použitia končí.

Obrázok 7. Model prípadov použitia metódy

UC1 Predspracovanie zadaného textu

Predspracovanie textu je nutný proces, ktorý upravuje texty písané v flexívnom jazyku do podoby, ktorá je ľahšie strojovo spracovateľná

Hlavný tok Predspracovanie priamo zadaného textu
Vstupné podmienky: Text dostupný v elektronickej forme
Výstupné podmienky: Text rozdelený na pole slov, slová neobsahujú skloňovacie koncovky
Účastníci: Používateľ

1. Používateľ zadá požiadavku na predspracovanie textu, ktorý zadal ako vstupný parameter funkcii *preprocess*.
2. Metóda vymedzí interpunkčné znamienka a rozdelí text na pole slov.
3. Metóda odstráni skloňovacie koncovky jednotlivých slov.
4. Používateľ môže pracovať s textom rozdeleným na pole slov a rovnako s jeho variantom s odstránenými koncovkami.
5. Prípad použitia končí.

Alternatívny tok Predspracovanie textu dostupného cez zadanú url
Vstupné podmienky: Text dostupný pomocou zadanej url
Výstupné podmienky: Text rozdelený na pole slov, slová neobsahujú skloňovacie koncovky
Účastníci: používateľ

1. Používateľ zadá požiadavku na predspracovanie textu, ktorého url zadal ako vstupný parameter funkcii *preprocess*.
2. Prípad použitia pokračuje podľa hlavného toku od bodu 2.

UC2 Rozpoznávanie pomenovaných entít

Proces vykonávajúci rozpoznávanie pomenovaných entít pracuje s poľom slov pripraveným pomocou UC1. Jeho výsledkom je zoznam identifikovaných entít a rovnako aj označkovaný text.

Hlavný tok Rozpoznávanie pomenovaných entít v predspracovanom texte
Vstupné podmienky: Predspracovaný text
Výstupné podmienky: Dostupný zoznam identifikovaných entít, označkovaný text
Účastníci: Používateľ

1. Používateľ definuje databázu, ktorú bude metóda využívať priradením databázového klienta do premennej *@client*.
2. Používateľ zadá požiadavku na rozpoznanie pomenovaných entít. Požiadavku zavolá funkciou *recognize_NER*.
3. Metóda postupne prejde celé pole slov a identifikuje pomenované entity.
4. Prípad použitia končí.

Alternatívny tok Rozpoznávanie pomenovaných entít v predspracovanom texte
Vstupné podmienky: žiadne
Výstupné podmienky: žiadne
Účastníci: Používateľ

1. Tok prebieha rovnako ako body 1 – 2 hlavného toku.
2. Metóda zistí, že pole vstupných slov je prázdne, nevykoná nič.
3. Prípad použitia končí.

UC3 Získanie pomenovaných entít

Po spracovaní textu sa v premenných *@full_text* a *@entities* nachádza označkovaný text a zoznam identifikovaných entít. Je s nimi možné pracovať.

Hlavný tok Získanie zoznamu pomenovaných entít
Vstupné podmienky: V texte boli rozpoznané pomenované entity
Výstupné podmienky: Získanie zoznamu pomenovaných entít
Účastníci: Používateľ

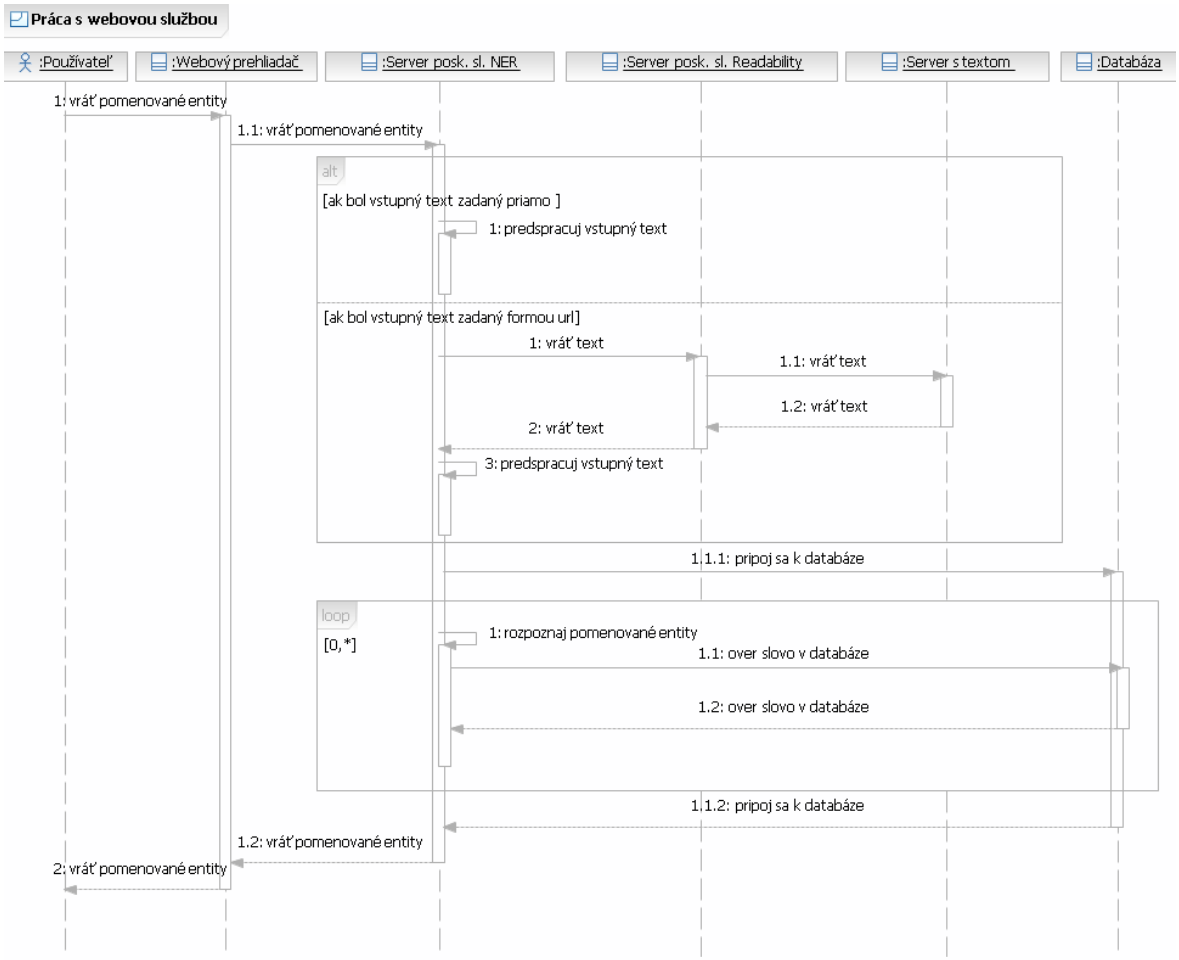
1. Používateľ zadá požiadavku na sprístupnenie zoznamu identifikovaných pomenovaných entít.
2. Metóda vráti obsah premennej *@entities*, obsahujúcej požadovaný zoznam.
3. Používateľ pracuje so získaným zoznamom.
4. Prípad použitia končí.

Alternatívny tok Získanie označkovaného textu
Vstupné podmienky: V texte boli rozpoznané pomenované entity
Výstupné podmienky: Získanie označkovaného textu
Účastníci: Používateľ

1. Používateľ zadá požiadavku na sprístupnenie označkovaného textu.
2. Metóda vráti obsah premennej *@full_text*, obsahujúcej požadovaný text.
3. Používateľ pracuje so získaným textom.
4. Prípad použitia končí.

A.3. *Sekvenčný diagram webovej služby*

Pre lepšiu predstavu o postupnosti jednotlivých krokov, ktoré sa vykonajú po zavolaní opisovanej webovej služby, sme jej priebeh znázornili do jednoduchého sekvenčného diagramu. Tento diagram znázorňuje výmenu správ medzi objektmi usporiadaných podľa ich vzájomnej časovej postupnosti. Diagram na Obrázok 8 znázorňuje časť webovej služby vykonávajúcu rozpoznávanie entít, nie však časť slúžiacu na správu identifikovaných entít.



Obrázok 8. Sekvenčný diagram webovej služby

A.4. *Databáza*

Metóda v práci vyžaduje databázu obsahujúcu niekoľko tabuliek. Tabuľky navzájom nie sú prepojené, pretože slúžia na uloženie odlišných informácií. Neuvádzame preto logický a fyzický model databázy. Pre prehľadnosť sme vzhľadom na situáciu zvolili opis databázy formou tabuliek.

Názov	entities		
Popis	Tabuľka obsahujúca entity, ktoré metóda v minulosti rozpoznala		
Názov stĺpca	Údajový typ	PK	Popis
id	INTEGER(11)	áno	Identifikačné číslo entity, (auto_increment)
real_name	VARCHAR(80)	nie	Názov entity v zákl. tvare, kt. zobrazujeme
name	VARCHAR(80)	nie	Názov entity, s ktorým pracujeme, (unique)
typ	CHAR(1)	nie	Typ, kam entitu radíme (pers, loc, num)
checked	TINYINT(1)	nie	Príznak, či je entita skontrolovaná

Názov	sk_wikipedia		
Popis	Tabuľka obsahujúca obrazy článkov zo slovenskej Wikipédie		
Názov stĺpca	Údajový typ	PK	Popis
name	VARCHAR(80)	áno	Názov článku z Wikipédie
type	TEXT	nie	Zoznam kategórií, kam článok zaraďuje Wikipédia

Názov	en_wikipedia		
Popis	Tabuľka obsahujúca obrazy článkov z anglickej Wikipédie		
Názov stĺpca	Údajový typ	PK	Popis
name	VARCHAR(80)	áno	Názov článku z Wikipédie
type	TEXT	nie	Zoznam kategórií, kam článok zaraďuje Wikipédia

Názov	sentence_begins		
Popis	Tabuľka obsahujúca typické začiatky viet, ktoré nie sú entity		
Názov stĺpca	Údajový typ	PK	Popis
begin_word	VARCHAR(50)	áno	Slovo, ktoré často začína vetu a nie je entita
count	INTEGER(5)	nie	Počet výskytov slova

Názov	cs_entities		
Popis	Tabuľka obsahujúca entity, ktoré metóda správne rozpoznala		
Názov stĺpca	Údajový typ	PK	Popis
id	INTEGER(11)	áno	Identifikačné číslo entity, (auto_increment)
real_name	VARCHAR(80)	nie	Názov entity v zákl. tvare, kt. zobrazujeme
name	VARCHAR(80)	nie	Názov entity, s ktorým pracujeme, (unique)
typ	CHAR(1)	nie	Typ, kam entitu radíme (pers, loc, num)
Poznámka: tabuľka pri rozpoznávaní českých textov nahrádza tabuľku entities			

Názov	cs_wikipedia		
Popis	Tabuľka obsahujúca obrazy článkov z českej Wikipédie		
Názov stĺpca	Údajový typ	PK	Popis
name	VARCHAR(80)	áno	Názov článku z Wikipédie
type	TEXT	nie	Zoznam kategórií, kam článok zaraďuje Wikipédia
Poznámka: tabuľka pri rozpoznávaní českých textov nahrádza tabuľku sk_wikipedia			

B. Inštalačná a používateľská príručka

Navrhnutú metódu je, možné využívať ako webovú službu alebo je možné pracovať s ňou vo forme knižnice. Implementácia metódy je kompatibilná s Ruby 1.9.3. a Rails 3.1.3. Databáza bola navrhnutá pre MySQL 5.0.8.

B.1. *Inštalácia a práca s knižnicou*

Pred použitím knižnice je potrebné vytvoriť MySQL databázu s názvom *ner_recognizer* a importovať do nej priložený súbor *ner_recoqnizer.sql*. To je možné vykonať cez príkazový riadok zadaním príkazu:

```
mysql –user root –p nazov_databazy < ner_recognizer.sql
```

Následne je nutné v projekte kde sa bude knižnica využívať nainštalovať nasledujúce gemy:

```
gem install 'rubygems'
gem install 'httparty'
gem install 'open-uri'
gem install 'nokogiri'
gem install 'unicode_utils/downcase'
gem install 'mysql2'
gem install 'timeout'
gem install 'Set'
```

Po úspešnej inštalácii gemov je knižnica pripravená na používanie. Používateľ má po vytvorení objektu prikazom

```
recognizer = NER_recognizer.new
```

možnosť predspracovať text zadaný formou url na webovú stránku, kde ho môžeme nájsť. Predspracovanie zavoláme príkazom

```
recognizer.preprocess(url,true)
```

Každý text je pred samotným rozpoznávaním nutné predspracovať. Druhou alternatívou predspracovanie je priame zadanie textu príkazom

```
recognizer.preprocess("text",false)
```

Pred rozpoznávaním entít je taktiež nutné inicializovať prístup do databázy zadaním nasledujúceho príkazu. Uvedené hodnoty atribútov zodpovedajú pripojeniu na lokálnu databázu.

```
recognizer.client = Mysql2::Client.new(:adapter => "mysql2",
:host => "127.0.0.1", :username => "root", :database =>
"ner_recognizer")
```

Do databázy pred použitím nutné importovať tabuľky entities, sk_wikipedia, en_wikipedia a sentecne_begins, ktorých obrazy sa nachádzajú na priloženom dátovom nosiči.

Samotné rozpoznávanie pomenovaných entít zavoláme príkazom

```
recognizer.recognize_NER
```

Funkcia v texte rozpozná pomenované entity a naplní jednotlivé premenné (*@entities* *@full_text* a *@edited_text*) obsahom, s ktorým môžeme ďalej pracovať. Premennú *@full_text* (pole reťazcov) tvorí vstupný text s označkovanými entitami, v premennej *@edited_text* (pole reťazcov) sa nachádza predspracovaný vstupný text. *@entites* predstavuje zoznam všetkých nájdených entít (pole *Entít*). *Entita* je objekt s atribútmi *real_name* (názov entity v základnom tvare), *name* (názov entity, s ktorým pracujeme), *type* (identifikovaný typ), *start_index* a *end_index* (indexy rozsahu entít v poli so vstupným textom). Ukážka príkazov na prácu s jednotlivými premennými metódy.

```
recognizer.entities.each do |act|
     puts act.real_name
end
puts recognizer.full_text
puts recognizer.edited_text
```

B.2. *Práca s webovou službou*

Metódu sme nasadili taktiež ako webovú službu. Táto služba poskytuje možnosť spracovania textu zadaného formou url priamo naň, prípadne dokáže spracovať priamo zadaný text. Ako vstupný parameter teda zvolíme buď *url* alebo *content*. Okrem toho musíme zašpecifikovať, aký typ (*type*) výstupu chceme dostať. Možnosti sú dve, prvou je '*text*', ktorý vráti celý spracovávaný text s entitami označenými podľa štandardov MUC 6 (Named Entity Task Definition, 1997). Druhá možnosť je zvoliť *type* ako '*words*'. Táto možnosť nám vráti kompletný zoznam získaných entít vo formáte JSON. Nasledujúce dva príkazy ukazujú príklad zavolania služby pomocou GET metódy. Prvým príkazom zadávame požiadavku na vrátenie zoznamu entít z textu zadaného pomocou *url* s obsahom *some_url*, v druhom príkaz na zobrazenie označkovaného textu, ktorý sme zadali - s*ome_text*.

```
http://kassak.visnovsky.sk/ner/?type=words&url=some_url
```

```
http://kassak.visnovsky.sk/ner/?type=text&content=some_text
```

GET metóda odosiela všetky údaje v HTTP hlavičke, ktoré je obmedzená na 2048 znakov. Nie je preto vhodná na zadávanie dlhých textov. Pre tento prípad je možné zavolať POST metódu, ktorá dokáže vo svojom tele odoslať aj veľmi dlhé texty, ktorých veľkosť limituje až minimum z dvojice maximálny limit prehliadača a maximálny limit serveru. POST metódu je možné adresovať na nižšie uvedenú adresu, pričom je potrebné rovnako ako pri GET metóde zašpecifikovať typ výstupu (parameter type).

```
http://kassak.visnovsky.sk/ner/long_text
```

Základné inštrukcie o funkcionalite metódy a povinných vstupných parametroch sa používateľovi zobrazia po zaslaní GET metódu na adresu

```
http://kassak.visnovsky.sk/ner/
```

V prvok kroku vykonania metódy sa vytvorí objekt typu *NER_recognizer*, pre ktorý sa zavolá funkcia *preproces*. Jej vstupné parametre závisia od zadaného vstupu. Funkcia slúži na predspracovanie textu. Následne po jej vykonaní metóda inicializuje premennú *clienet*, ktorá nastaví parametre prístupu k databáze. Po tejto inicializácii sa vykoná metóda *recognize_NER*, slúžiaca na rozpoznanie entít v texte. Po jej vykonaní zobrazí metóda na výstup podľa zadaného *type*, označkovaný text alebo zoznam entít. Telo metódy zobrazuje (Fragment kódu 8).

```
class NerController < ApplicationController
 def index
  recognizer = NER_recognizer.new                          # vytvorenie objektu
  if !params[:url].nil?                  # predspracovnaie textu po zadani linky
   recognizer.preprocess(params[:url],true)
  else                                  # predspracovanie textu po priamom zadani
   recognizer.preprocess(params[:content],false)
  end
  recognizer.recognize_NER                                 # rozpoznavanie entit

  if params[:type].eql?("text")              # ak type = text, vratime cely text
   @result = String.new                                  # s otagovanymi entitami
   recognizer.full_text.each do |act|
     @result << act + " "
   end
  elsif params[:type].eql?("words")            # ak type = words, vratime zoznam
   @result = Array.new                            # najdenych pomenovanych entit
   recognizer.entities.each do |act|
     @result.push act
   end
  else
   @result = "Error, you don't selected 'type' ('text' OR 'words')"
  end
  respond_to do |format|
   format.html
   format.json { render :json => @result }
  end
 end
...
```

Fragment kódu 8. Priebeh metódy vykonávaný po zavolaní webovej služby

Okrem možnosti samotného rozpoznávania entít poskytuje webová služba aj možnosť pohodlnej práce s neschválenými entitami, ktoré metóda identifikovala v textoch. Táto časť po zadaní príkazu

`http://kassak.visnovsky.sk/ner/entities`

zobrazí zoznam všetkých neschválených entít s ich atribútmi. Pri každej z entít sa nachádzajú tlačidlá na priame akceptovanie alebo odstránenie entity a tlačidlo na zobrazenie editačného formulára, v ktorom je možné meniť jednotlivé atribúty entity a po ich úprave túto akciu potvrdiť alebo zrušiť.
Táto časť však nie je určená pre bežných používateľov, pretože by teoreticky mohli úmyselne meniť entity za nesprávne a následne ich schvaľovať.

C. Príspevok publikovaný na konferencii IIT.SRC 2012

Táto príloha obsahuje článok, ktorý bol v plnom znení prijatý na konferenciu IIT-SRC 2012. Článok je možné nájsť aj v zborníku konferencie.

Named Entity Recognition for Slovak and Related Languages

Ondrej KAŠŠÁK*

Slovak University of Technology in Bratislava
Faculty of Informatics and Information Technologies
Ilkovičova 3, 842 16 Bratislava, Slovakia
ondrej.kassak@gmail.com

Abstract. In this paper we describe a proposal of the method for recognizing and extraction of named entities in texts. We primarily deal with methods designed for the Slovak language, but we also describe possibilities of application for other flective languages. The goal of the described method is to identify potential entities occurring in the processed text, determine its scope and consequently identify the category to which they belong. Proposed approach can be used in various tasks, where the text pre-processing is required as information search or content-based recommendation etc.

Introduction

Nowadays we are literally overwhelmed with information. It is served to user from all sides and we are also acquiring the necessary knowledge from many sources. It is impossible for users to process all the information they find and that is the reason for the huge research interest for the information overload. Although information is now available in many various forms, much of it is stored in a text form. This is a very traditional way of storing knowledge and in the digital age it acquires an unprecedented dimension because of the simplicity with which the text can be created, distributed and stored.

Due to these facts, an area for many methods used for finding information arises there. Various approaches for information gaining have been proposed as personalized recommendations [10] or search methods. The search methods provide results in varying levels depending on the manner in which information is gained. In addition of searching methods, there are methods used for text recommendation. Recommendations based on the content of the text or searching methods using key entities from the texts assume the named entities appearing in the text for their working as an input. It can either directly identify entities in the text or create some form of metadata, data bearing on these entities. Based on them, recommendation algorithms can search and work then more efficiently in comparison with methods working only with the text titles or with the most frequent words in the text.

The aim of our proposed method is to recognize entities in the text and assign them the correct type. We primarily focus on texts in Slovak language because a comprehensive tool for this

* Bachelor degree study programme in field: Informatics
Supervisor: Ing. Michal Kompan, Institute of Informatics and Software Engineering, Faculty of Informatics and Information Technologies STU in Bratislava

language that would identify all entities classified according to the MUC-6 (6th Message understand conference) [1] is missing. The paper also describes the options for deployment of the method in related Slavic languages that have free sentence structure and words inflection. Similar methods for the majority of world languages already exist. Depending on the language, the issue is managed with a variety of high success rate. For German up to 70%, for English about 90%.The best tools for Slavic languages achieve success just around 70% [2].

Related Work

We have no information about an existing tool recognizing all types of entities for the Slovak language, but several approaches have been proposed for the related Slavic languages. The best presented solution for the Czech language achieves the success rate between 68% and 72% [2]. Most of methods use the statistical approach as the Semantic spaces or the Maximum entropy method [2], but we can find the linguistic solutions respectively.

Affinity between Czech and Polish is, similarly, proved by the similarity of the presented results. These are very close to the Czech. In specializing of recognizing only persons in a specific domain of stock exchange messages and police reports was achieved the success rate of 72% using Hidden Markov Models method and even 89% after linguistic optimization [4]. The question is how good results can be reached by this method when used on bigger scale of texts or for recognition of other types of entities.

As the English is one of the most used languages plenty of approaches have been proposed. This language belongs to the group of languages with constant form of words, where dominate statistical methods before linguistic. Three standard methods are the most used for English language: Maximum entropy method [5], Hidden Markov models [7] and Conditional random fields [3]. The actual state of art approaches in the field of named entity extraction for English do not outperform 90%, for German are achieved results aprox. 70% [2]. Both of these languages belong to the group of Germanic languages, but there are different rules for making sentences and different capitalization, thus there is quiet big difference in results for these languages.

Named Entities

The concept of named entities was gradually defined in MUC 1 to 6. It means word entities that are important for us in some way. After recognition of the entity it is marked by well-defined brands. We distinguish between these types of entities:

- Persons – marked <ENAMEX TYPE="PERSON"> and </ENAMEX>
- Locations - marked <ENAMEX TYPE="LOCATION"> and </ENAMEX>
- Organizations - marked <ENAMEX TYPE="ORGANIZATION"> and </ENAMEX>
- Dates - marked <TIMEX> and </TIMEX>
- Numbers - marked <NUMEX TYPE="NUM"> and </NUMEX>
- Percents - marked <NUMEX TYPE="PERC"> and </NUMEX>
- Sums of money - marked <NUMEX TYPE="MONEY"> and </NUMEX>
- Miscellaneous entity types - marked <MISC> and </MISC>

The process of identifying entities in the text consists of two parts - the initial part of pre-processing of the text and then the recognition of the named entities.

Text Pre-processing

As mentioned above, along with other Slavic languages, the Slovak belongs to the group of flective languages which means that most of its words are inflected according to certain grammatical rules. Because of that nouns, adjectives and verbs acquire several different forms and make it impossible to identify the entities directly form the specified text [9].

The words should therefore be first processed by a certain way. Our method uses a form of stemming. The regular stemming method consists of obtaining the word formation root of each word. This is done by removing prefixes and suffixes of words. For our purposes, we decided to remove only the word suffix caused by inflection. The result is an original form of words which is not the word formation root but, with only a few exceptions, we get the uniform forms of words with which can be used in further computation. Larger text editing process would unnecessarily slow down the algorithm. In some cases it would make exactly the same roots of similar words. That would be even counterproductive for named entity recognition.

Suffixes are identified by comparison each word with the set of Slovak word endings. We proceed from the longest to shorter ones to verify the suffixes, so we always remove the entire suffix. We do not change abbreviations made only form capitals, because they are not inflected.
The text pre-processing also divides the text into individual words and we separate the punctuation marks like full stops, quotation marks, commas, etc. from the words. Some of the punctuation marks are used as full-stop symbols. In this case it is clear that they cannot belong to the entity and the entity scope can be identified easily.

Named Entity Recognition

Slavic languages have a rich morphology and transform words to their basic form or find the root is algorithmically difficult and can not be provided universally. This makes low level extracting information, where we advise the named entity recognition, highly demanding and for computer processing disadvantageous [6]. We propose the named entity extraction (Figure 1), which consists of several steps.

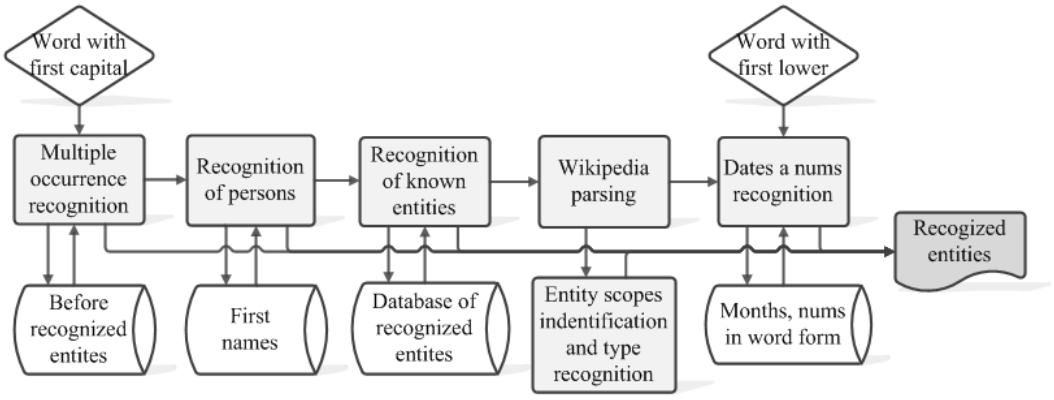


Figure 1. Sequence of steps describing proposed method.

The process of recognizing entities in the pre-processed text written in Slovak is slightly easier because of the fact that the entities always start with a capital letter. Due to this fact, for most types of entities we have just to search for the capital letters in the text. If we then sort out the beginnings of sentences or quotations that are not entities at the same time, we get a set of entity

beginnings. Thus we are able to identify persons, organizations, locations and miscellaneous entities.

When found the beginning of the entity we recognize its scope through comparing the word and next words from the text with the list of before recognized entities, dictionaries or finally web parsing. If found the scope we try to recognize entity type through this compared expression.

After recognizing and marking the entities of these types, we store them into a special list. At the beginning, every new entity is compared with all the stored entities on the list, so we can simply identify entities that have already been recognized before without the long process of standard recognition of a new entity. In case of persons, we also store a surname separately because it is very common to mention the full name of the person at first and then, for simplicity, write only the surname. There is a chance that we do not recognize the surname in the text properly if we identify it by the standard way. For example, let us name a person with the surname "Pekný" standing at the beginning of the sentence.

Peoples' names may appear in different forms in the text. The basic form is to write a name and a surname. But the order of them can be changed or one of the parts can be replaced by initials. There can by two names or surnames in the one person's name. They can be written separately or linked by a hyphen. Person can by directly identified by titles or contextual words. Persons with a Slovak name can by identified by comparing their name with a dictionary of names. In recognition of context words we can avoid incorrect marking if we labelled a company named by its founder incorrectly by ignoring that after the person's name there are words such as "s.r.o." and the like.

We identify locations and organizations the most through dictionaries but if the entity is not in the dictionary, it should be found by parsing the Wikipedia[1] page. We search for scope of entity and the category it belongs to by the keywords gained from Wikipedia article. It is also helpful to determine the type entity by the contextual words such as the already mentioned "s.r.o." or "n.o."

It is necessary to find scopes for the miscellaneous entities, too. That means how many words from the text actually make the entity. We have to avoid a situation when recognize the word "Bratislava" in the text of "Dni Bratislavy" as a location and then wrongly ignore the rest.

In addition to entities beginning with capital letters, our method identifies also date and numeric entities. The numeric ones distinguish between money amount, percentage and a number itself. In identifying the type of numerical entities the context words are the most significant. They are also the most dependent on the language.

Dictionaries

Our method uses several forms of dictionaries. There are the alphabetically arranged arrays of words, word arrays primarily arranged according to length and then alphabetically arranged ones and bigger database of entities consisting of more words. Elements stored in these dictionaries are always prepared by stemming to be directly comparable with a modified processed text.

The only specific dictionary is the one of word endings. We use it to pre-process texts. This list is sorted by suffixes form the longest to the shortest ones and only then by the alphabet. When editing the input text we sequentially compare each word with the all elements of the dictionary until we get the match or go through the whole dictionary.

Similarly, in the form of an alphabetically sorted array other smaller dictionaries with relatively unchanging content are made such as a list of calendar months, or a list of Slovak names. These dictionaries are for time reasons searched by our quicksort implementation which is the reason why elements must be sorted alphabetically.

Large vocabulary of recognized entities is implemented as a database. It provides faster identification of entities which seek to reduce the need for time-consuming verification of the existence of entities using the Wikipedia.

[1] http://sk.Wikipedia.org/

Web Parsing

If we come across the beginning of an entity we could not identify through dictionaries, we try to identify it using querying Wikipedia. There we enter an increasing number of words following the first word in text and look for the article with identical name. We try to find the longest match, so for example if found matches “Slovakia” and “Slovakia Hockey Hall of Fame”, we choose second one. In case we find an article, we get the keywords from it and try to find the entity category from them.

Besides working with the Wikipedia we use the Slovak National Corpus [8]. In this manner we verify if the beginnings of sentences or quotations are common words or entities. We put the name of the potential entity into the corpus. Entity scopes have been identified earlier. We are looking for the term written with a small starting letter. If we get an empty set as a result (e.g., "slovenská technická univerzita" you will look for it uselessly), or the majority of results shows the first capital letter (e.g. "trenčiansky hrad"), it is an entity. Otherwise it is a common word.

Evaluation

To evaluate proposed approach we processed 20 articles from SME.sk[1], 20 from HNonline.sk[2] and 20 from topky.sk[3]. For achieved results (see Table 1) we identify Precision (number of correct results divided by the number of all returned results), Recall (number of correct results divided by the number of results that should have been returned), and F-measure (harmonic mean of precision and recall).

The current version of method reads input in the form of directly entered text or a link to the article available on the Internet. This link is processed through the Readability service (part of the Metall[4] service), which obtain the text of the article.

Using text pre-processing method we divide input text into the array of modified words, in which we recognize named entities. For recognition we use database of before recognized entities and dictionaries implemented directly as part of a method and can not be changed (e.g. vocabulary of calendar months or first names). We use them to compare and recognize potential entities.

Table 1. Achieved results of proposed method

Type	Precision	Recall	F-measure
Persons	0.97	0.80	0.88
Organizations	0.94	0.67	0.78
Locations	0.83	0.73	0.78
Dates	0.97	0.76	0.85
Numbers	0.90	0.87	0.88
Percents	0.83	0.68	0.75
Sums of Money	1.00	0.76	0.86
Miscellaneous	0.50	0.66	0.57
Total	**0.84**	**0.74**	**0.79**

Results were obtained based on very small dataset, which was created and annotated manually by expert. Total we correctly recognized 1204 entities of 1620. We wrong identified 232. These results are used for the first concept evaluation and after finishing proposed method; it will be

[1] http://www.sme.sk/
[2] http://hnonline.sk/
[3] http://topky.sk/
[4] http://peweproxy.fiit.stuba.sk/metall/

tested on bigger and more complex dataset, in order to obtain the more reliable results. Similarly, the comparison to existing approaches for several languages will be performed, because of method language dependency exploration.

Discussion and Conclusions

In this paper we described a method for recognizing of named entities in the texts. We focused on Slovak language, where obtained result of 79% F-measure seems to be a promising result, but we expect slight improvement after completion of method and it's tuning after more complex testing.

The method can be used, after slight modifications, also for other languages with a similar style of making sentences and the type of word inflection. Especially Slavic languages fulfil these conditions. After replacing dictionaries by a new dictionary containing data in the language with which we are going to work, we expect to obtain similar results like for the Slovak language. The method implementation is not complete yet. For example we don't parse the English version of Wikipedia if we didn't find some entity in the Slovak version. After completing the implementation and testing method more complexly we would like to put it as part of the Metall service, to be freely available to users.

The method was developed as a support tool for recommendation or search methods which can benefit in their work from our acquired named entities. The method can be used directly as a standalone tool to graphical highlight important entities in the text which will be then clearer for the reader. Other option is use it as a tool describing the important terms of texts which can serve as a very brief summary or to get keywords and then show them to the user so he can decide whether he will deal with the text.

Acknowledgement. This work was partially supported by the Scientific Grant Agency of Slovak Republic, grant No. VG1/0971/11.

References

[1] Grishman, R., Sundheim, B.: Message understanding conference-6: A brief history. In: Proceedings of COLING, 96, (1996), pp. 466–471.

[2] Konkol, M., Konopík, M.: Maximum entropy named entity recognition for Czech language. In: Text, Speech and Dialogue, Springer, (2011), pp. 203–210.

[3] Lafferty, J., McCallum, A., & Pereira, F. C. N.: Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In: Computer, (2001).

[4] Marcińczuk, M., Piasecki, M.: Study on named entity recognition for Polish based on hidden Markov models. In: Text, Speech and Dialogue, Springer, (2011), pp. 142–149.

[5] Nigam, K., Lafferty, J., McCallum, A.: Using maximum entropy for text classification. In: IJCAI-99 workshop on machine learning for information filtering, (1999), pp. 61–67.

[6] Przepiórkowski, A.: Slavonic Information Extraction and Partial Parsing. In: Computational Linguistics, (June 2007), pp. 1-10.

[7] Rabiner L. R., Juang, B. H.: An introduction to hidden Markov models. In: Current protocols in bioinformatics, (January 2007).

[8] Slovak National Corpus – prim-5.0-public-all. Bratislava: Ľ. Štúr Institute of Linguistics SAV 2011. Available at WWW: http://korpus.juls.savba.sk/index_en.html.

[9] Slovake. About the language - Form and structure. [Online; accessed November 24, 2011] Available at: http://slovake.eu/en/intro/language/form.

[10] Suchal, J., Návrat, P. Full text search engine as scalable k-nearest neighbor recommendation system. In: AI 2010, IFIP AICT 331, Springer, (2010), pp. 165-173.

D. Obsah elektronického média

- Bakalárska práca - elektronická verzia dokumentu
- Implementácia
  - zdrojový kód webovej služby
  - zdrojový kód knižnice pre slovenský jazyk
  - zdrojový kód knižnice pre český jazyk
  - dokumentácia k zdrojovému kódu
  - obraz tabuliek databázy pre jednotlivé metódy
- Konferencia IIT.SRC - článok a poster na konferenciu IIT.SRC 2012