DETERMINATION OF THE BASIC SHAPE OF THE WORDS

Pavel Šanda
Master Degree Programme (2), FEEC BUT
E-mail: xsanda01@stud.feec.vutbr.cz

Supervised by: Jan Karásek
E-mail: karasek.jan@phd.feec.vutbr.cz

Abstract: *Lemmatization* is an important preprocessing step for many applications of text mining. *Lemmatization* is similar to word stemming but it does not produce only stem of the word. *Lemmatization* has to replace the suffix of a word by different suffix to get normalized word. The main contribution of this paper is to present methods for impoving *lemmatization* for czech language. The paper also presents new testing set of data which can be freely used in experiments like this.

Keywords: STEMMING, LEMMATIZATION, TEXT PROCESSING, STOP WORDS

1 ÚVOD

Text je spolu s řečí velice důležitou součástí lidské komunikace. V dnešní moderní době je text předáván většinou v elektronické formě a tím pádem je nutné umět tyto elektronické texty počítačově zpracovat. Právě zpracování a následná úprava textu k dalšímu použití je tématem této práce. Hlavním přínosem tohoto článku je představení možností, které by mohly vést ke zdokonalení *lemmatizace* českého jazyka. Dále vytvoření testovací množiny dat, které bude také veřejně dostupná pro další skupiny zabývající se touto problematikou. Tato práce je součástí většího projektu jehož výstupem bude shlukování dokumentů nebo určení emocí z textu. [1] Před samotným lemmatizátorem je zařazena ještě kumulace dat z internetu a následný spell-checking.

2 PŘEDZPRACOVÁNÍ TEXTU, STOP-SLOVA, LEMMATIZACE

Text je k dispozici v různých formách (strukturovaný, semi-strukturovaný a nestrukturovaný). Nejčastěji je však nestrukturovaný. To znamená, že nemá žádnou specifickou formu a bylo by s ním obtížné pracovat. Proto je nutné tento nestrukturovaný text nejprve připravit na další zpracování. [4]

Předzpracování textu neboli *preprocessing* je procedura, která následuje bezprostředně po *tokenizaci* (rozložení vět na jednotlivé důležité části). Funkce předzpracování textu spočívá v odstranění slov s minimální významovou hodnotou. Tyto slova se nazývají *stop-slova*. Řadí se mezi ně např. spojky, předložky apod.. Dále také v eliminaci interpunkčních znamének, která mají také minimální význam. [2] Následným krokem ve zpracování textu může být *stemming*, což je v podstatě výpočetní procedura, která u slov určuje jejich kořen (tzv. *stem*). [3] Druhým způsobem je proces *lemmatizace*, který oproti *stemmingu* neurčí pouze kořen slova, ale také jeho celý základní tvar.

3 IMPLEMENTACE LEMMATIZÁTORU

Tato práce se zabývá vývojem algoritmu pro *lemmatizaci* českého jazyka. Již existují algoritmy pro *lemmatizaci*, ale ani jeden z nich není naprogramován v jazyce Java, v kterém je vyvíjen zde uvedený. Jelikož tedy v Javě nejsou žádné dostupné algoritmy, tak je tento *lemmatizátor* inspirován dostupnými pravidly v jiných jazycích (konkrétně C++), která byla vytvořena ve Slovinsku. A protože tyto pravidla neposkytují stoprocentní úspěšnost je zde prostor k vylepšení.

Jak již bylo řečeno, tak pro vývoj algoritmu jsou použita již známá pravidla pro *lemmatizaci* českého jazyka. Vzhledem k tomu, že v jazyce Java tyto pravidla nebyla k dispozici bylo nutné je přepracovat. Je zde použita tzv. metoda *suffix stripping*, která pracuje na principu oddělování (nahrazování) přípon slov tak, aby vznikl základní tvar. Avšak někdy je nutné odebrat také předponu slova tzv. *prefix*.

3.1 KOMPLIKACE

Jedním z hlavních problémů českého jazyka je mnohoznačnost slov. Jako příklad lze uvést slovo *jí*, které může vyjadřovat sloveso (*jíst*), ale také zájmeno (*ona*). V těchto případech je nutné zjišťovat význam slova z kontextu věty. Dalším z problémů je to, že není možné odebrat známou příponu slova bez ověření (to lze provádět v anglickém jazyce, kde jsou přípony jako *-ing* a jsou vždy odebírány), zda výsledný kořen bude mít nějaký význam.

Na výše uvedené problémy se tato práce zaměřuje a protože český jazyk známe mnohem podrobněji (jako rodilí mluvčí) než ve Slovinsku je možné tyto chyby eliminovat.

3.2 ROZŠÍŘENÍ LEMMATIZÁTORU

Jelikož existují tyto výše uvedené problémy, tak je nutné tento *lemmatizátor* vylepšit, protože pravidla, ze kterých se vychází, toto neošetřují a mohou se tedy vyskytnout chyby. Již byla vytvořena základní tabulka *stop slov* obsahující základní všeobecně známá slova, která budou před samotným zpracováním odebrána (spojky, předložky a další významově bezcenná slova). Další rozšíření spočívá v zařazení algoritmu *Brute Force*, který se provede ještě před samotným *suffix stripping* algoritmem. Jedná se vlastně o tabulku, ve které jsou uloženy kombinace slov: *základní tvar - slovo časované (skloňované)*. V této tabulce mohou být uloženy páry slov jako *být - jsem*, kde není možné pomocí metody *suffix stripping* odebrat příponu a tím zjistit základní tvar. Tato tabulka je zatím ve stádiu vývoje, neboť při implementaci se stále objevují nová a nová slova, která patří do této tabulky. V neposlední řadě také úprava stávajícího *suffix stripping* algoritmu, aby docházelo k větší úspěšnosti určení základního tvaru slova. V některých případech jsou stávající pravidla již takzvaně "přetrénována". To znamená, že jsou zde uvedena zbytečně, jelikož je lze zpracovat dřívějšími jednoduššími pravidly. Tato úprava je také vhodná pro optimalizaci *lemmatizátoru*. Celý vývoj algoritmu je konstruován tak, aby byla zaručena co největší modularita. Není tedy sebemenší problém použít jen jeho jednotlivé části určené pro nějaký úkol a implementovat do jiné aplikace (algoritmu).

3.3 TESTOVACÍ MNOŽINA DAT

Pro účely ověření funkčnosti vyvíjeného *lemmatizátoru* začala být vytvářena trénovací množina dat. Jedná se zatím o sto souborů s obsahem přibližně deseti vět originálního textu[1] a dalších sto souborů obsahující manuálně zpracovaná data pro porovnání. Ty byly vytvořeny na základě definovaných *stop slov* a *Brute Force* metody. Tato množina je stále ve stádiu vývoje a není tedy ještě zcela kompletní. Avšak alespoň na přibližné určení správnosti výsledků poskytovaných *lemmatizátorem* je dostačující.

4 VÝSLEDKY

Originální text
V Lysé nad Labem v pátek ohrožoval muž svoji družku střelnou zbraní v jejich společném bytě. Policista muže během vyjednávání zastřelil. Násilník se uvnitř zabarikádoval, vyhrožoval vlastní smrtí a pak začal střílet, řekla mluvčí nymburské policie Petra Potočná.

Manuálně upravený text
Lysý Labe pátek ohrožovat muž druh střelný zbraň společný byt. Policista muž vyjednávání zastřelit.

[1] Texty byly získávány z webových stránek www.idnes.cz a www.novinky.cz

Násilník uvnitř zabarikádovat, vyhrožovat vlastní smrt začít střílet, říct mluvčí nymburský policie Petra Potočná.

Výstupní text z existujícího lemmatizátoru[2]
V Lysý nad Lab v pátek ohrožovat muž svůj druh střelný zbraň v jeho společný byt. Policista muž běh vyjednávání zastřelit. Násilník s uvnitř zabarikádovat, vyhrožovat vlastní smrt a pak začít střílet, řeknout mluvčí nymburský policie Petr Potočný.

úspěšnost: 67%

Výstupní text z vyvíjeného lemmatizátoru
Lysý Labe pátek ohrožovat muž svůj druh střelný zbraň jeho společný byt Policista muž běh vyjednávání zastřelit Násilník uvnitř zabarikádovat vyhrožovat vlastní smrt začít střílet řeknout mluvčí nymburský policie Petr Potočný

úspěšnost: 74%

Z výše uvedeného procentuálního porovnání je zřejmé, že oproti již existujícímu *lemmatizátoru* je ten v současné době vyvíjený o několik procent úspěšnější. Hlavní faktor, který ovlivnil větší úspešnost je ten, že byl vytvořen seznam *stop slov*, která mají minimální významovou hodnotu a tyto slova byla odebrána.

5 ZÁVĚR

Současné výsledky *lemmatizace* jsou o několik procent lepší, ale zdaleka ne ideální. Stále není optimální seznam *stop slov* a je třeba ho doplnit. Dále je zde názorně ukázán případ mnohoznačnosti slova, který není ošetřen. Jedná se o slovo "během", které by mělo být v tomto případě odebráno, neboť z kontextu vyplývá, že se jedná o příslovce. Ale jelikož má toto slovo ještě druhý význam (podstatné jméno), tak stávající *lemmatizátor* ho převedl právě na základní tvar tohoto podstatného jména, což je v tomto případě chybné. V neposlední řadě *lemmatizátor* odebírá všechna interpunkční znaménka. Avšak je požadavek, aby jednotlivé věty zůstaly odděleny tak jak byly. Nyní tomu tak není a tím pádem se jedná o další chyby v *lemmatizaci*. V závěru ukázkového textu je také nekorektně převedeno jméno. Mělo by zůstat základním tvaru, ale ženském, tak jak je předvedeno v manuálně upraveném textu.

REFERENCE

[1] Burget, R., Karásek, J., Smékal, Z.: Classification and Detection of Emotions in Czech News Headlines. In The 33rd International Conference on Telecommunication and Signal Processing, TSP 2010. 2010. s. 64-69. ISBN: 978-963-88981-0-4.

[2] Francis, L., Flynn, M.: Text Mining Handbook [online]. [s.l.] : Casualty Actuarial Society E-Forum, 2010. Dostupné z WWW: <http://www.casact.org/pubs/forum/10spforum/Francis_Flynn.pdf>.

[3] Lovins, J. B.: Development of a Stemming Algorithm [online]. [s.l.] : Mechanical Translation and Computational Linguistics, 1968. Dostupné z WWW: <http://journal.mercubuana.ac.id/data/MT-1968-Lovins.pdf>.

[4] Sedláček, P.: Text mining a jeho možnosti (aplikace) [online]. 30.1.2004. Text mining a jeho možnosti (aplikace). Dostupné z WWW: <http://www.fi.muni.cz/usr/jkucera/pv109/2003p/xsedlac5.htm>.

[2] Lemmatizátor je dostupný na webových stránkách http://lemmatise.ijs.si/