JMZW: DETEKCE VÝZNAMNÝCH SLOV MIMO SLOVNÍK

Jan LEHEČKA[1]

1 ÚVOD

Jazykové modelování z webu (JMZW) je systém vyvíjený na Katedře kybernetiky, který vytváří a automaticky doplňuje rozsáhlou databázi textů (především internetových článků) tříděných dle témat a klíčových slov. Na základě statistik počítaných z těchto dat mohou být sestavovány tématické jazykové modely, které se využívají například při automatickém titulkování živě vysílaných televizních pořadů nebo při diktování novinových článků.

V úloze automatického titulkování živě vysílaných pořadů převádí řeč na titulky automatický rozpoznávač řeči. Rozpoznávač řeči využívá slovník, který by měl obsahovat co nejvíce řečených slov. Vzhledem k bohaté slovní zásobě českého jazyka a vzhledem k tomu, že většina českých slov se může vyskytovat v mnoha různých tvarech (každý odvozený tvar slova odpovídá ve slovníku jedné položce), je nutné slovník omezit pouze na nejčastěji používaná slova dle apriorní znalosti tématu nebo oblasti řečníkovy promluvy.

Slovní zásoba řečníků se každým dnem obohacuje o nová aktuální slova z oblasti, o které řečník hovoří (nová jména, zeměpisné údaje, názvy firem atd.). Tato slova jsou často klíčová k porozumění řečníkova sdělení, proto by měla být ve slovníku rozpoznávače. Automatickou detekcí těchto významných slov v nových textech se zabývá tato práce.

2 FILTROVÁNÍ OOV SLOV

Všechna slova, která nejsou ve slovníku rozpoznávače řeči, se zkráceně označují OOV (out of vocabulary). Tato slova jsou vyhledávána v nových článcích databáze systému JMZW. Cílem je vybrat ze všech nalezených OOV slov pouze ta významná, tj. oddělit běžně používaná slova v málo používaných tvarech od slov, která jsou díky určité významné události v současné době aktuální a začnou být používána řečníky.

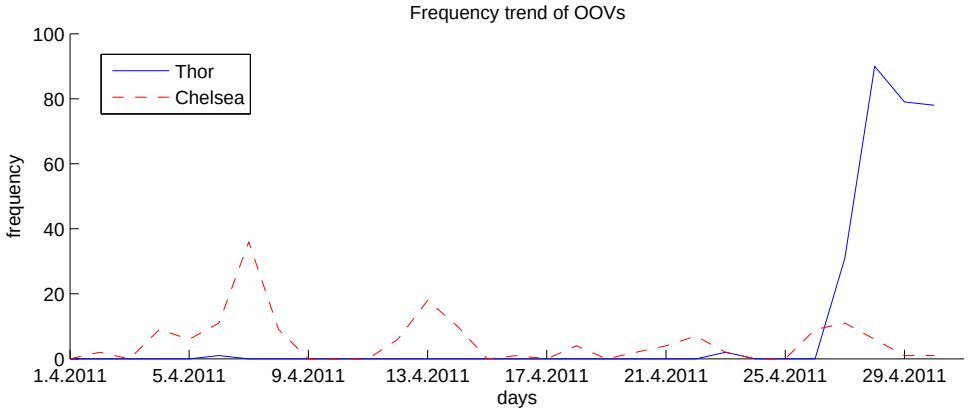
Texty, ve kterých se OOV slova vyhledávají, jsou již zpracované modulem dekapitalizace, který převádí velké písmeno na začátku věty na malé, pokud je jisté, že se nejedná o slovo, kde by se mělo správně psát velké písmeno na začátku. Všechna nalezená OOV slova jsou proto nejprve rozdělena na dvě skupiny: slova, která začínají malým písmenem, a slova začínající velkým písmenem. Většina nových významných slov bude začínat velkým písmenem (jména, místa, firmy, ...), první filtr tedy odebere všechna OOV slova, která začínají malým písmenem (tato slova by měla být po odebrání podrobena dalšímu zkoumání, ale to není předmětem této práce).

Pro další filtr je nutné spočítat četnost výskytu slov v textech za určité časové období. Z množiny OOV slov (po první filtraci, tedy jen ty s velkým písmenem na začátku) odebereme slova, jejichž celková četnost ve zkoumaném období je nižší, než zvolený práh. Práh volíme buď absolutně (přiměřeně k rozsahu zkoumaných textů) nebo relativně k celkovému počtu slov v textech. Vhodně zvolený práh od OOV slov oddělí málo používaná slova.

[1] Jan Lehečka, student navazujícího studijního programu Aplikované vědy a informatika, obor Kybernetika a řídicí technika, specializace Umělá inteligence, e-mail: jlehecka@students.zcu.cz

V dalším kroku porovnáme množinu zbylých OOV slov se slovníkem rozpoznávače, ale tentokrát ignorujeme velikost písmen. Tím z množiny OOV slov oddělíme běžně používaná slova, kterým dekapitalizace ponechala velké písmeno na začátku nebo která byla v textu z nějakého důvodu napsána velkými písmeny.

Poslední důležitý filtr zkoumá, zda se OOV slovo v nedávné historii začalo používat s větší frekvencí, což je vlastnost, kterou hledaná významná slova nepochybně mají. Pro tento účel je porovnávána průměrná denní četnost slova v textech za delší časové období a průměrná denní četnost v nedávné historii (například průměrná denní četnost za poslední dva měsíce ku průměrné denní četnosti za posledních čtrnáct dnů). Zvýšila-li se na konci zkoumaného období průměrná denní četnost výrazně (nutné zvolit mezní poměr), slovo je pravděpodobně významné a mělo by být přidáno do slovníku rozpoznávače. Obr. 1 zobrazuje rozdíl mezi časovým průběhem četnosti slova, které není nové (Chelsea - aktuální pouze když fotbalový klub Chelsea hraje zápas), a slova, které je nové a na konci zkoumaného období aktuální (Thor - nový americký film).



Obr. 1: Ukázka časového průběhu četnosti některých OOV slov

3 ZÁVĚR

Tato práce navrhuje postup, jak vybrat z velkého množství OOV slov (řádově sta tisíce) pouze několik set pravděpodobně významných, která mohou být podrobena dalšímu zkoumání a následně přidána do slovníku rozpoznávače. Budoucí práce spočívá v experimentálním nalezení optimálních prahů a mezí navržených filtrů. Dále pak navržení detekce významných OOV slov mezi slovy začínajícími malými písmeny a propojení detekce významných OOV slov s algoritmem pro detekci nestandardní výslovnosti.

Poděkování: Příspěvek byl podpořen grantovým projektem Západočeské univerzity v Plzni, projekt č. SGS-2010-054, a za podpory Ministerstva průmyslu a obchodu, projekt č. MPO FR-TI1/486.

LITERATURA

Psutka, J., Müller, L., Matoušek, J., Radová, V., 2006. *Mluvíme s počítačem česky*. Praha.

Švec, J., Skorkovská, L., Vavruška, J., Ircing, P., Lehečka, J., Pražák, A., Kanis, J., Hoidekr, J., Pressl, D., Stanislav, P., Soutner, D., 2010. *Výzkumná zpráva projektu jazykové modelování z webu*, Výzkumná zpráva interního grantu Západočeské univerzity v Plzni č. SGS-2010-054.

Venkataraman, A., Wang, W., 2003. *Techniques for effective vocabulary selection*. Geneva.