Detekce slov s nepravidelnou výslovností v českém textu

Jan Lehečka[1]

1 Úvod

Český jazyk v současné době podléhá stále více tendenci používat slovní zásobu pocházející z jiných jazyků. Do běžně používaného jazyka se tak postupně zakořeňují cizí slova, jejichž výslovnost se neřídí pravidly české výslovnosti, ale je nutné číst je dle pravidel jazyka, ze kterého pocházejí. Do slovníků lidí se tak dostávají jména cizích firem, produktů, vlastních jmen a zeměpisných názvů, které často ani není možné do češtiny přeložit, a je nutno je vyslovovat v originálním znění („Windows“, „Shakespeare“, „Washington“ atd.). Pravidlům české výslovnosti se ale vymykají i běžně používaná přejatá slova, která mají historický původ v jiném jazyce, zejména v latině (např. slova „medicína“, „kontinent“ nebo „univerzita“ by se dle pravidel české výslovnosti měla číst měkce, tedy s 'ď', 'ť' a 'ň').

To způsobuje problémy zejména v českých TTS (Text To Speech) systémech umělé inteligence, které převádějí text na řeč. Slova s jinou než českou výslovností jsou pak nesrozumitelná. Cílem této práce je detekce všech takových slov v českých textech. Těmto slovům říkáme slova s nepravidelnou výslovností a platí pro ně, že jejich výslovnost není možné odvodit pomocí pravidel české fonetické transkripce.

Pro řešení je použit klasifikátor, který na základě spočtených příznaků každého slova roztřídí všechna slova textu do dvou tříd, a to do třídy slov s pravidelnou výslovností a do třídy slov s nepravidelnou výslovností. Natrénovaný klasifikátor zohledňuje i slovník výjimek zabudovaný v existujícím fonetickém transkriberu.

Zkoumaný problém byl řešen pomocí několika různých klasifikátorů, ze kterých byl na základě vyhodnocení klasifikace technikou křížové validace a ověřením detekce slov v reálných textech nakonec vybrán klasifikátor podle *k*-nejbližšího souseda.

2 Příznaky slov pro klasifikaci

Celkem bylo zvoleno 9 příznaků pro popis každého slova. Pět příznaků popisuje automatickou detekci jazyka, ze kterého slovo pochází. Tyto příznaky vyjadřují odhad pravděpodobnosti výskytu slova v češtině, angličtině, němčině, latině a francouzštině. Dva další příznaky vyjadřují odhad pravděpodobnosti výskytu slova v jazyce pravidelné a nepravidelné výslovnosti na základě jazykových modelů počítaných z trénovacích dat. Poslední dva příznaky představují délku slova a relativní počet znaků mimo českou abecedu.

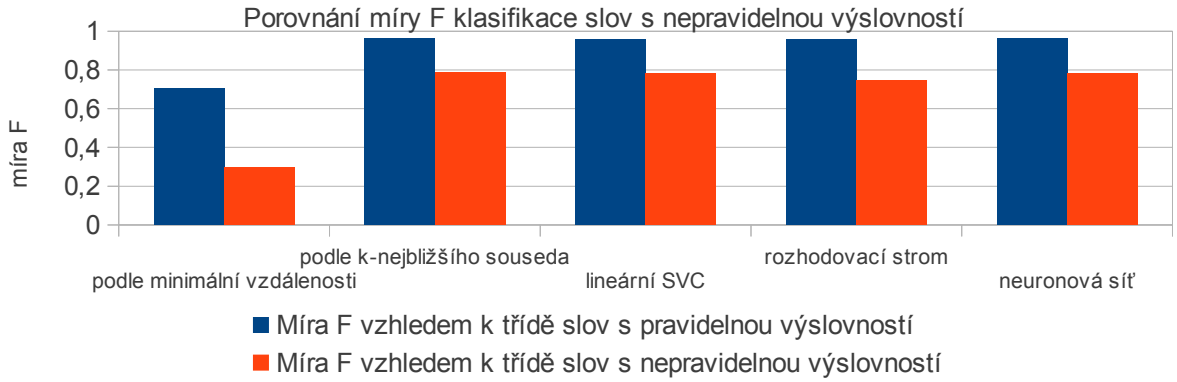
3 Porovnání detekce pomocí různých klasifikátorů

Pro vyhodnocení klasifikace a porovnání kvality různých klasifikátorů byla použita míra F, která je harmonickým průměrem přesnosti a úplnosti klasifikace a počítá se vždy vzhledem k jedné, tzv. pozitivní třídě. Přesnost P udává, jaký podíl ze všech slov klasifikovaných do pozitivní třídy do ní skutečně patří, a úplnost R udává, jaký podíl slov skutečně náležících do pozitivní třídy bylo do této třídy klasifikováno. Míra F vzhledem k pozitivní třídě se pak počítá vztahem pro harmonický průměr (1).

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

$$F = \frac{2\,PR}{P+R} \tag{1}$$

Aby nedošlo k přetrénování klasifikátorů, byla použita technika desetinásobné křížové validace, která zaručuje nestranný odhad míry F. Graf na obr. 1 ukazuje výslednou míru F různých klasifikátorů. Nejlepší míra F vyšla pro klasifikátor podle *k*-nejbližšího souseda, ale i lineární SVC a neuronové sítě vykazují podobnou míru F klasifikace.



Obrázek 1: Vyhodnocení klasifikace pomocí různých klasifikátorů

Otestováním detekce slov s nepravidelnou výslovností na reálných zpravodajských textech pomocí těchto klasifikátorů bylo zjištěno, že klasifikátor podle *k*-nejbližšího souseda skutečně řeší zkoumaný problém nejlépe.

4 Závěr

Klasifikátor podle *k*-nejbližšího souseda byl původně zkoumán jen jako zástupce jednodušších klasifikátorů, aby sloužil pro srovnání, o kolik jsou lepší modernější a složitější klasifikátory (zejména SVC a neuronové sítě). Ukázalo se však, že díky velmi komplikovanému rozložení obrazů obou tříd v příznakovém prostoru nejlépe zadané úloze vyhovuje jednoduchý klasifikátor, který nehledá žádné oddělovače tříd, ale zkoumá pouze vzorové obrazy v bezprostředním okolí klasifikovaných slov v příznakovém prostoru.

Složitost rozložení tříd v příznakovém prostoru je způsobena tím, že fonetický transkriber obsahuje výslovnostní výjimky, které by dle posloupnosti znaků měly patřit do třídy slov s nepravidelnou výslovností, ale díky tomu, že je transkriber umí přepsat správně, je nutno klasifikovat je jako slova s pravidelnou výslovností, aby je program v textech neoznačoval. Takovouto změnou klasifikace několika vybraných slov vznikají v příznakovém prostoru izolované body nebo skupiny bodů zcela obklopené body z druhé třídy. To je pravděpodobně důvod, proč selhaly klasifikátory hledající v tomto příznakovém prostoru nějakou oddělující nadrovinu.

Literatura

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B, Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E., 2011. Scikit-learn: Machine Learning in Python, *Journal of Machine Learning Research*, 12, 2825—2830

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

Stolcke, A., 2002. SRILM - An Extensible Language Modeling Toolkit