Detekce přízvuků na datech z Russian National Corpus

Anastasiia Chizhova [1]

1 Úvod

V rámci práce pro převod textu do syntetizované řeči v ruském jazyce se zabývám slovy, která mají volný nebo pohyblivý přízvuk, tzv. homografy. Přízvuk takových slov je závislý na kontextu a je ho třeba různými způsoby syntetizovat. Např.:

- зáмок (*čes. hrad)*
- замóк *(čes. zámek)*

- в следующем годý (*čes. v příštím roce*)
- к 2010 гóду (*čes. do roku 2010*)

Pro většinu slov lze pro určení pozice přízvuků používat slovník, ale pro homografy existuje několik případů, protože umístění přízvuku ovlivňuje pád nebo celkový význam slova. Cílem mého experimentu je predikovat přízvuk pouze na základě textu. Při zpracování daného problému používám metodu "Učení s učitelem" (Psutka (2016)).

2 Trénovací a testovací data

Pro experiment bylo vybráno několik nejčastěji se vyskytujících slov v novinách s nejednoznačným přízvukem. Z novinových textů jsem pro každé slovo našla 200 vět s různými případy kontextu a přízvuku, které představovaly trénovací data. Dále jsem ručně provedla klasifikaci do dvou tříd:

- je přízvuk (1)
- není přízvuk (0).

Loni (Chizhova (2017)) jsem stejná data používala i pro testování, s tím že všechna data byla rozdělena na testovací a trénovací v poměru 1 ku 99. V dané práci jsem pokročila tak, že klasifikátor bude trénován na mnou připravených datech a pak testován na mnohem větším objemu dat z Ruského Národního korpusu (http://www.ruscorpora.ru/en/index.html), poskytnutého Ruskou Akademií věd, což by mělo dát reálnou představu o úspěšnosti metody klasifikace a klasifikátorů.

3 Klasifikátory a klasifikace

Řetězce znaků jsem převedla na číselné vektory pomocí metody DictVectorizer z balíčku scikit-learn (Pedregosa *et al.* (2011)). Každá položka vektoru tak odpovídala výskytu konkrétního znaku na konkrétní pozici v řetězci. Zkoušela jsem 2 klasifikátory: Logistic Regression (LogReg) a Support Vector Machine (SVM), kde pro SVM jsem zkoušela 2 různá nastavení: rbf a linear. Testovala jsem různě dlouhé levé a pravé kontexty. Výsledky klasifikace byly porovnány se správnými odpověďmi pomocí funkce f1_score.

V Tab. 1 a Tab. 2 je uvedena část výsledků pro testování na původních datech a testování na datech z RNK pro vybrané pravé a levé kontexty a vybrané homografy:

[1] studentka bakalářského studijního programu Inženýrská informatika, obor Inteligentní komunikace člověk - stroj, e-mail: chizhova@students.zcu.cz

Typ klasif.	LogReg			SVM		
Kontext	L-20 P-5	L-10 P-10	L-5 P-8	L-20 P-5	L-10 P-10	L-5 P-8
Году	93,97 %	93,50 %	92,96 %	94,74 %	93,97 %	92,57 %
Города	89,72 %	89,45 %	85,57 %	86,58 %	86,74 %	83,82 %
Самом	67,8 %	71,72 %	72,04 %	66,05 %	70,56 %	70,3 %

Tabulka 1 Výsledky klasifikátoru LogReg a SVM(linear) pro původní data (f1_score)

Typ klasif.	Počet výskytů slova	LogReg			SVM		
Kontext		L-20 P-5	L-10 P-10	L-5 P-8	L-20 P-5	L-10 P-10	L-5 P-8
Году	416	96,60 %	97,09 %	97,61 %	95,53 %	95,7 %	97,83 %
Города	126	92,80 %	94,49 %	91,36 %	88,71 %	90,84 %	90,69 %
Самом	337	90,60 %	89,38 %	93,02 %	80,24 %	70,22 %	85,04 %

Tabulka 2 Výsledky klasifikátoru LogReg a SVM (linear) pro RNK (f1_score)

4 Závěr

Z výsledků experimentu je vidět, že je možné pouze z textového okolí slova s nejednoznačným přízvukem s relativně vysokou úspěšností určit pozici přízvuků ve slově. Průměrná úspěšnost na všech slovech možná neočekávaně prokazuje, že neznámá data lze klasifikovat lépe (viz. Tab. 3). Ale je třeba zmínit, že v daném případě byly výsledky testování na RNK také ovlivněny počtem výskytů obou variant přízvuků v textu, ve všech uvedených případech je to v poměru 5-10% jedné varianty a 90-95% druhé.

Manuální data	LogReg	SVM (linear)	SVM (rbf)
Min	56,06 %	56, 08 %	56,08 %
Max	83,46 %	82,66 %	73,44 %
RNK			
Min	61,38 %	64,48 %	59,86 %
Max	81,60 %	78,37 %	75,02 %

Tabulka 3 Porovnani uspesnosti 2 skupin testovacich dat (puvodni a RNK)

Lepším z používaných klasifikátorů pro oba případy stálé zůstává LogisticRegression.

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.,a Duchesnay, E. (2011) Scikit-learn: Machine Learning in Python, Journal of Machine Learning Research, 12, pp. 2825-2830.

Psutka, J. (2016) Učební texty z předmětu Základy strojového učeni a rozpoznávaní, ZČU v Plzni.

Chizhova A. (2017) Detekce přízvuků v ruštině s použitím klaifikátoru, Studentská vědecká konverence FAV 2017.