FAKULTA APLIKOVANÝCH VĚD ZÁPADOČESKÉ UNIVERZITY V PLZNI

STUDENTSKÁ VĚDECKÁ KONFERENCE 2016

Neuronové sítě v úloze identifikace tématu z textu

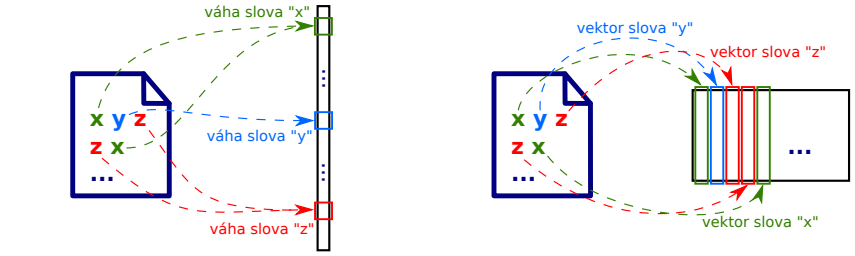
Jan Lehečka[1]

1 Úvod

Se stále narůstajícím počtem online textových dokumentů roste i potřeba tato data efektivně filtrovat a automaticky z nich extrahovat užitečné znalosti. Jednou z populárních úloh, které získávají znalosti z textu, je i *úloha automatické identifikace tématu*, která má za úkol přiřadit ke každému textovému dokumentu jeden nebo více tzv. *štítků*, které reprezentují určitá témata či kategorie. Tato znalost může být dále využívána např. pro efektivní filtrování rozsáhlých textových korpusů, adaptaci systémů na dané téma atd.

Tradiční přístup k řešení této úlohy je reprezentovat každý dokument jako vektor vysoké dimenze, tzv. Bag-of-Words (BOW, viz obr. 1a), a z těchto vektorů natrénovat klasifikátor, typicky SVM (Support Vector Machine). Nevýhodou BOW je ignorování pořadí slov v textu, vysoká dimenze vektorů, a s tím související velký počet trénovaných parametrů.

Velmi populárními se v nedávné době staly také slovní vektory (*word2vec*) publikované v Mikolov et al. (2013), které mapují slova do prostoru nízké dimenze. Z těchto vektorů je možné poskládat maticovou reprezentaci dokumentu vhodnou pro sekvenční zpracování (viz obr. 1b).



(a) Bag-of-Words reprezentace
výsledkem je vektor *(velikost slovníku x 1)*

(b) word2vec reprezentace
výsledkem je matice (*dim x počet slov)*

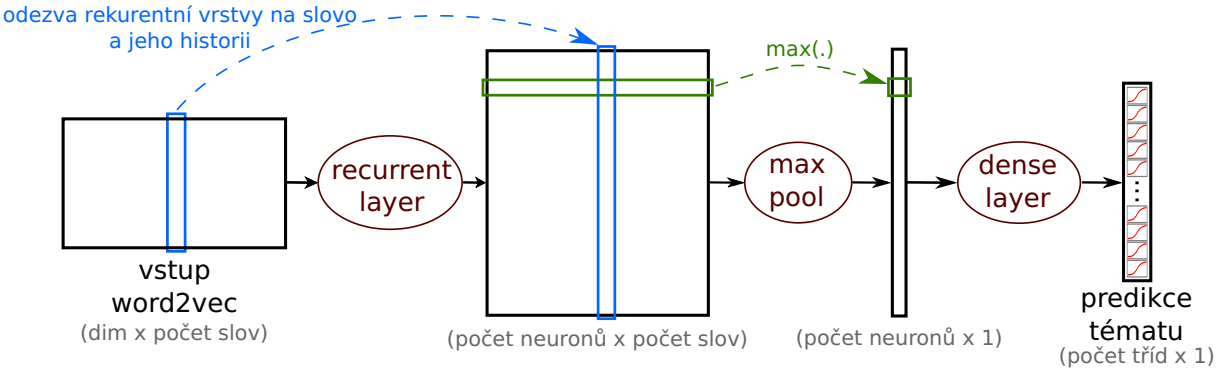
Obrázek 1: Dvě různé reprezentace textového dokumentu

V této práci byly experimentálně porovnány obě reprezentace textových dokumentů v úloze automatické identifikace tématu. Testován byl klasický model (SVM) a dva typy neuronových sítí (NN): dopředné (FFNN) a rekurentní (RNN, konkrétně typy LSTM a GRU). RNN byly testovány také v kombinaci s tzv. podvzorkováním (*poolingem*, viz obr. 2).

2 Experiment

Experiment byl proveden na českých zpravodajských článcích s ručně přiřazenými tématy. Pro trénování bylo použito $195k$ článků a pro test $44k$. Celkem bylo v datech 577 různých témat, průměrně 3 na jeden článek. Dimenze BOW byla $370k$ a dimenze *word2vec* 300. Všechny skryté vrstvy v použitých NN modelech měly 512 neuronů. Pro vyhodnocení byla použita F-míra,

[1] student doktorského studijního programu Aplikované vědy a informatika, obor Kybernetika, e-mail: jlehecka@kky.zcu.cz



Obrázek 2: Schéma RNN s poolingem

která vyžaduje binární rozhodnutí klasifikátoru pro každou dvojici článek & téma. Toho bylo dosaženo dvěma různými strategiemi prahování: (1) *RCut(3)*, která přiřadí 3 témata s nejvyšším skóre ke každému článku, (2) *MCut* publikovaná v Largeron er al. (2012).

Výsledky jsou shrnuty v tabulce 1. Přestože žádná NN nepřekonala SVM při strategii prahování *RCut(3)*, je zřejmé, že pro výstupy NN je vhodnější strategií *MCut*. Již obyčejná 2vrstvá FFNN předčila tradiční SVM. Použitím *word2vec* a RNN bylo dosaženo srovnatelných výsledků při mnohonásobně nižším počtu parametrů. Přidáním poolingu za rekurentní vrstvu bylo dosaženo významného zlepšení, což je zřejmě dáno tím, že příznaky jednotlivých témat v textu není nutné přesně lokalizovat, ale stačí pouze detekovat, zda jsou přítomny.

repr.	*model*	*# param [mil.]*	$F_{RCut(3)}$	F_{MCut}
BOW	SVM (baseline)	213.6	0.711	0.677
BOW	FFNN (1 vrstva)	213.6	0.696	0.703
	FFNN (2 vrstvy)	189.8	0.701	0.728
word2vec	LSTM	1.9	0.660	0.697
	LSTM + pooling	1.9	0.698	0.740
	GRU	1.5	0.675	0.719
	GRU + pooling	1.5	0.697	**0.741**

Tabulka 1: Tabulka výsledků a počtu trénovaných parametrů.

Poděkování

Tento příspěvek byl podpořen grantovým projektem SGS-2016-039.

Literatura

Largeron, C., Moulin, C. and Géry, M., 2012. MCut: a thresholding strategy for multi-label classification. *Advances in Intelligent Data Analysis XI*. Springer Berlin Heidelberg. pp. 172-183.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S. and Dean, J., 2013. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*. pp. 3111-3119.