Segmentace textu dle tématu

Jan Beneš[1]

1 Úvod

Cílem této práce je návrh, implementace a vyhodnocení různých segmentačních algoritmů. Výstupem segmentačních algoritmů je informace, zda se na dané pozici vyskytuje tématický předěl, či nikoliv. Z této formulace je možné usoudit, že je segmentace úzce spjata s vědní disciplínou klasifikace.

V práci jsou využity převážně tradiční metody typu SVMs. SVMs podávají excelentní výsledky při klasifikaci textu, což je důvod, proč jsou zde tyto algoritmy aplikovány. Výjimku v tradičnosti pak tvoří použití moderních rekurentních neuronových sítí typu LSTM (Hochreiter et al. (1997)).

2 Řešení

Jelikož mi bylo dáno k dispozici poměrně velké množství dat s přidělenými tématy, vytvořil jsem algoritmus, který využívá informaci o tématu článků pro segmentaci textu. Alternativní přístupy (Georgescul et al. (2006)), které též používají v procesu segmentace klasifikátor SVM, většinou pracují pouze s informací o tématických hranicích článků, nikoliv však s informací o samotných tématech.

Vstupními daty jsou dva soubory (trénovací a testovací data) s českými zpravodajskými články s přidělenými tématy. V trénovacím korpusu je 205128 článků, v testovacím 43847 článků.

Nejprve je nutné z trénovacích dat vyjmout tzv *held out* data (také někdy označováno jako *cross validation* data). Tato data se používají pro optimalizaci hyperparametrů (parametr, jenž je nastaven před samotným trénováním). V tomto případě jsem vyjmul 10 % dat, což je v absolutních číslech 20513 článků.

Takto zpracovaná data jsem dále vektorizoval. K tomuto úkonu jsem použil třídu *TfidfVectorizer* z knihovny *scikit-learn*. Na takto zpracovaných datech jsem dále natrénoval klasifikátor SVM.

V této práci jsem vyzkoušel několik různých přístupů.

Nejlepším přístupem se ukázalo být natrénování rekurentní neuronové sítě typu LSTM na tématických predikcích z klasifikátoru SVM. Neuronovou síť jsem implementoval v knihovně *Keras*. Síť sestává ze 2 vrstev. První vrstva obsahuje 32 neuronů typu LSTM. V druhé vrstvě je jeden neuron s aktivační funkcí *sigmoid*.

[1] student bakalářského studijního programu Aplikované vědy a informatika, obor Kybernetika, e-mail: benesjan@students.zcu.cz

3 Výsledky

V tabulce níže jsou seřazeny výsledky různých segmentačních přístupů. *WindowDiff* (Pevzner, Lev, Hearst, Marti (2002)) a P_k (Beeferman, Berger, Lafferty (1999)) metriky udávají míru chybovosti algoritmu, proto mají tyto metriky oproti F míře inverzní charakter.

	F_1	P_k	*WindowDiff*
Triviální algoritmus	0,216	0,479	0,836
Prahování rozdílu od předešlých vzdáleností	0,392	0,339	0,495
K-means	0,400	0,387	0,626
Binární SVM	0,444	0,376	0,660
Prahování rozdílu od okolí	0,491	0,299	0,418
Prahování euklidovské vzdálenosti	0,507	0,337	0,404
Sekvenční shlukovací algoritmus	0,662	0,240	0,325
Prahování kosinové vzdálenosti	0,678	0,234	0,284
LSTM - nezpracované predikce	**0,854**	**0,105**	**0,133**

Tabulka 1: Tabulka výsledků

Vidíme, že nejvyšší dosažená hodnota F míry **0,854** je o 17,6 % lepší než druhý nejlépe fungující přístup. Toto zjištění není překvapující, jelikož data mají sekvenční charakter, čehož je tento typ neuronové sítě schopen využít.

4 Závěr

V této práci jsem zjistil, že rekurentní neuronové sítě podávají dobré výsledky při aplikaci na problém segmentace textu.

Tento typ neuronových sítí by zároveň mohl být i cestou k dosažení ještě lepších výsledků. Nevýhodou modelů hlubokého strojového učení je, že existuje velké množství hyperparametrů, kterými se dá ovlivnit chování modelu. Počítačová optimalizace těchto parametrů vyžaduje obrovský výpočetní výkon, proto jsou pro efektivní nastavení vyžadovány expertní znalosti. Z toho důvodu se dá usoudit, že nastavení modelu pravděpodobně není optimální. Výhodou těchto moderních přístupů oproti tradičním však je, že při volbě dostatečně komplexní struktury zvětšení trénovacího datasetu téměř vždy přinese zlepšení prediktivních schopností modelu. Proto je pravděpodobné, že zvětšení datasetu je cesta k dosažení ještě lepších výsledků.

Literatura

Hochreiter, Sepp and Schmidhuber, Jürgen (1997) Long Short-Term Memory, *Neural Comput.*, Cambridge, MA, USA

Georgescul, Maria and Clark, Alexander and Armstrong, Susan (2006) Word Distributions for Thematic Segmentation in a Support Vector Machine Approach, *Proceedings of the Tenth Conference on Computational Natural Language Learning*, Stroudsburg, PA, USA

Pevzner, Lev and Hearst, Marti A. (2002) A Critique and Improvement of an Evaluation Metric for Text Segmentation, *Comput. Linguist.*, Cambridge, MA, USA

Beeferman, Doug and Berger, Adam and Lafferty, John (1999) Statistical Models for Text Segmentation *Mach. Learn.*, Hingham, MA, USA