Učenie s prenosom medzi prirodzenými jazykmi

Matúš Pikuliak, Marián Šimko a Mária Bieliková

Ústav informatiky, informačných systémov a softvérového inžinierstva,
Fakulta informatiky a informačných technológií, Slovenská technická univerzita v Bratislave
Ilkovičova 2, 842 16 Bratislava

matus.pikuliak@stuba.sk

Abstrakt. Hlboké učenie sa aktuálne javí ako veľmi perspektívny prístup k mnohým úlohám spracovania prirodzeného jazyka. Tento úspech sa však zatiaľ prejavuje najmä pri jazykoch, ktoré majú dostatočné množstvo zdrojov na natrénovanie komplexných neurónových sietí. Menšie jazyky s menším množstvom zdrojov majú problém tieto nové techniky využiť a priepasť medzi nimi a "bohatými" jazykmi sa tak prehlbuje. V našej práci sa venujeme tomu, ako túto priepasť zmenšiť pomocou prenosu naučenej informácie z jazyka do jazyka. Hlavnou myšlienkou je trénovanie hlbokých neurónových sietí v multilingválnom režime tak, aby sa model naučil využívať znalosti z jedného jazyka aj pre vstupy z iných. Navrhli a vykonali sme experiment s prenosom informácie o sentimente slov pomocou zdieľaného priestoru distribučných vektorov. V experimente sme dosiahli výsledky porovnateľné s nemeckými manuálne vytvorenými lexikónmi sentimentu, pričom sme však nepoužili žiadne nemecké dáta týkajúce sa sentimentu.

Kľúčové slova: spracovanie prirodzeného jazyka, učenie s prenosom, multilingválne učenie, umelé neurónové siete.

1 Úvod

Umelé neurónové siete v posledných rokoch výrazným spôsobom zasiahli do spracovania prirodzeného jazyka. Ukázalo sa, že rozličné architektúry neurónových sietí dokážu veľmi dobre spracovať veľké objemy textových dát a naučiť sa z nich prínosné informácie. V krátkom čase pomocou nich výskumníci dokázali prekonať existujúce riešenia pre veľké množstvo úloh, vrátane tých najpokročilejších, ako napríklad strojový preklad [3], analýza sentimentu [10] či syntaktická analýza [13]. Výhodou týchto neurónových prístupov sú malé nároky na doménovú expertízu. Zdá sa, že strojové učenie sa tu konečne dokáže naučiť všetko, čo potrebuje na úspešné vyriešenie úlohy, bez toho, aby výskumníci museli ručne navrhovať vysoký počet čŕt. Črty sa neurónové siete učia extrahovať samé a môžeme teda viac všeobecne hovoriť o učení sa reprezentácií, resp. učení sa čŕt (angl. *representation learning*) [1].

Tieto zaujímavé pokroky sú podmienené rastúcim množstvom textových dát, ktoré ako ľudstvo vytvárame. Najmä s nástupom Internetu sa zlepšili možnosti tvorby

J. Steinberger, M. Zíma, D. Fiala, M. Dostal, M. Nykl (eds.)
Data a znalosti 2017, Plzeň, 5. - 6. října 2017, pp. 204-208.

a zdieľania rozličných textov, ako napr. recenzií, príspevkov na sociálnych sieťach, blogov či článkov. Aj keď tento rast je globálny, niekoľko svetových jazykov (najmä angličtina) ho pocítilo omnoho viac. Len v týchto jazykoch existujú dostatočne veľké datasety, na ktorých sa danú súčasné modely často postavené na hlbokom učení (angl. *deep learning*) trénovať.

Jedným z možných riešení je pokúsiť sa prenášať informácie naprieč jazykmi. V strojovom učení poznáme koncept učenia s prenosom (angl. *transfer learning*), kedy sa snažíme model natrénovaný na určitej doméne či úlohe použiť v iných doménach či na iných úlohách [8]. Prenos medzi rozličnými jazykmi môžeme chápať ako jeden z druhov takéhoto učenia a v tejto práci sa venujeme práve jemu.

2 Existujúce prístupy

Pri analýze existujúcich prístupov ku učeniu s prenosom medzi jazykmi sme identifikovali 3 lingvistické úrovne, na ktorých môže učenie medzi jazykmi prebiehať. Ide o (1) morfologickú úroveň, (2) úroveň slov a (3) úroveň viet. Postupne tieto úrovne predstavíme.

Morfologická úroveň. Na morfologickej úrovni prebieha učenie najmä medzi lingvisticky podobnými jazykmi, ktoré teda do istej miery zdieľajú podobnú morfológiu a slovnú zásobu (napr. Slovenčina-čeština, španielčina-portugalčina). Takéto učenie sa dá aplikovať pri modeloch, ktoré pracujú na úrovni znakov alebo iných častí slov. Predpokladáme, že znalosti získané o morfológii jedného jazyka sú aplikovateľné aj pre iné jazyky. Takéto učenie bolo použité napr. pri rozpoznávaní pomenovaných entít [2] alebo pri značkovaní viet [12].

Úroveň slov. V spracovaní prirodzeného jazyka sa stali populárne tzv. vektory latentných čŕt slov (angl. *word embeddings*) [6]. Ide o reprezentácie slov vytvorené v skrytých vrstvách neurónových sietí, ktoré zachytávajú sémantiku slov. Takto vznikajú aj multilingválne reprezentácie, kde sú do jedného vektorového priestoru premietané slová z viacerých jazykov [11]. Toto potom môže byť použité pre prenos informácie medzi jazykmi. Náš experiment patrí do tejto kategórie.

Úroveň viet. Posledná a v istom zmysle najnáročnejšia úroveň je úroveň viet. Na tejto úrovni sa výskumníci snažia vytvárať reprezentácie viet tak, aby zachovávali ich sémantickú podobnosť naprieč jazykmi. Táto úroveň je najmenej preskúmaná, riešenia existujú napríklad pre strojový preklad [3] alebo analýzu sentimentu [14].

3 Experiment s analýzou sentimentu

Na základe analýzy tejto oblasti sme navrhli náš vlastný experiment, ktorého úlohou je preniesť informáciu o sentimente slov z jedného jazyka do druhého. Na analýzu sentimentu sa často používajú tzv. lexikóny sentimentu. V podstate ide o zoznam ohodnotených slov, kde hodnota značí či je dané slovo pozitívne alebo negatívne ladené. Zostavovanie takýchto lexikónov je prácna úloha a automatické prístupy ani zďaleka nedosahujú potrebnú kvalitu [9]. Našou základnou myšlienkou je využiť informáciu z kvalitného cudzojazyčného lexikónu. Pre účely experimentu sme sa

zamerali na anglické lexikóny, ktoré chceme zužitkovať pre analýzu sentimentu v nemeckom jazyku.

Pre prenos sentimentu používame tzv. multilingválne vektory latentných čŕt [11]. Ide o vektorový priestor, v ktorom sú reprezentované slová z viacerých jazykov (v našom prípade z angličtiny a nemčiny) tak, že je zachovaná ich sémantická podobnosť. Anglické slova *good* a *fine* sú v takomto priestore blízko pri sebe. Taktiež sa však v ich blízkosti nachádzajú aj nemecké *gut* a *fein*. Nad takýmto multilingválnym priestorom potom trénujeme model sentimentu, ktorý sa naučí klasifikovať slová na pozitívne, negatívne a neutrálne.

Tento model je jednoduchá neurónová sieť (viacvrstvový perceptrón) s jednou skrytou vrstvou, ktorá sa snaží z vektorovej reprezentácie slova vyrátať jeho hodnotenie. Vstupom do tejto neurónovej siete sú predpripravené vektory slov, zatiaľ čo výstupom je vektor s 3 položkami, pričom každá zodpovedá jednej triede sentimentu. Ako trénovaciu množinu pre tento model použijeme kvalitný anglický lexikón sentimentu. Na výsledný vektor sa aplikuje softmaxová funkcia. Ako stratovú funkciu používame cross entropy funkciu, trénovanie prebieha s použitím bežného algoritmu Stochastic Gradient Descent. Natrénovaný model potom môžeme použiť pre klasifikáciu sentimentu nemeckých slov, keďže ich máme k dispozícii ich reprezentácie zo zdieľaného anglicko-nemeckého vektorového priestoru slov.

Pri experimentoch sme vyskúšali použiť viacero alternatívnych architektúr – testovali sme vplyv počtu skrytých vrstiev aj ich šírky. Napokon sa ako najlepšie riešenie ukázala sieť s jednou skrytou vrstvou s dĺžkou 10. Takáto neurónová sieť dokázala prekonať lineárnu aj logistickú regresiu. Myslíme si, že je spôsobené nelinearitou neurónových sietí. Sentiment vo vektorovom priestore totižto nemusí byť lineárne separovateľný. Správanie slov a ich vlastností v takomto priestore môže byť jedným zo smerov ďalšieho výskumu.

Takýto model sa naučí na anglických slovách odhadovať ich sentiment. Keďže je však zachovaná sémantická príbuznosť, dokážeme tento model rovno použiť aj na ohodnocovanie nemeckých slov. My sme model použili na ohodnotenie každého nemeckého slova v priestore. Získaný lexikón sentimentu pre nemecký jazyk sme potom použili v štandardnej klasifikácii na analýzu sentimentu nemeckých viet. Z každej vety sme vyextrahovali určité črty (prítomnosť slov, emotikonov, štylistických prvkov) a navyše sme pridali niekoľko čŕt extrahovaných pomocou lexikónu (počet pozitívnych/negatívnych slov, priemerná hodnota sentimentu). Nad výsledným vektorom čŕt sme natrénovali SVM klasifikátor.

Vo výsledku náš plne automatizovaný prístup presvedčivo predbehol iné automatizované prístupy. Ako nemecké lexikóny sme použili tie vygenerované v [9]. Automaticky vytvorený lexikón TKM napríklad dosiahol súhrnné makro F1 skóre 0,75 pre binárnu, resp. 0,59 pre ternárnu klasifikáciu. Náš prístup dosiahol najlepšie skóre s anglickým lexikónom sentimentu NRC [7]: 0,79, resp. 0,61. Tento náš výsledok je porovnateľný s manuálne vytvorenými state-of-the-art lexikónmi pre nemecký jazyk. Najlepší z nich, GPC, dosahuje výsledky 0,80, resp. 0,62. Vidíme, že sa s pomerne jednoduchou metódou dokážeme priblížiť kvalitným, manuálne vytvoreným lexikónom. Potrebovali sme pritom len paralelný anglicko-nemecký korpus a anglický lexikón sentimentu. Namiesto rátania reprezentácii slov v multilingválnom priestore

z paralelného korpusu sme v našich experimentoch použili už hotový anglicko-nemecký vektorový priestor [5], ktorý obsahuje 95,195 nemeckých a 40,054 anglických slov. Dĺžka vektorov je 512.

4 Záver

Učenie s prenosom medzi jazykmi má potenciál priniesť kvalitné riešenia pre jazyky, ktoré nemajú dostatok zdrojov. Tento smer je obzvlášť aktuálny teraz, keď v spracovaní prirodzeného jazyka sa do popredia dostáva hlboké učenie. Reprezentácie, ktoré sa pri hlbokom učení vytvárajú, sú totižto obzvlášť vhodné na učenie s prenosom. V tejto práci sme predstavili problematiku učenia s prenosom medzi jazykmi a uviedli sme aj tri úrovne, na ktorých takéto učenie môže prebiehať. Na úrovni slov sme ho aplikovali aj pri našom experimente, kedy sa nám podarilo automaticky zostrojiť kvalitný lexikón sentimentu pre nemecký jazyk.

V ďalšej práci chceme nadviazať na tento experiment a vylepšiť metódu prenosu informácie medzi slovami. Taktiež sa chceme venovať prenosu s učením na úrovni viet, ktoré predpokladá vytváranie reprezentácií s pokročilými metódami hlbokého učenia. Pri tomto učení sa chceme zamerať aj na interpretáciu natrénovaných modelov a lepšie pochopenie toho, čo sa vlastne umelé neurónové siete pri spracovaní prirodzeného jazyka učia.

Literatúra

1. Bengio, Y., Courville, A., Vincent, P.: Representation Learning: A Review and New Perspectives. IEEE Trans. on Pattern Analysis & Machine Intelligence 35.8 (2013), 1798 1828.
2. Gillick, D., et al.: Multilingual Language Processing From Bytes. In: Proc. of the 2016 Conf. of the North American Chapter of the ACL. ACL, San Diego (2016), 1296 1306.
3. Johnson, M., et al.: Google's Multilingual Neural Machine Translation System: Enabling Zero-Shot Translation. arXiv preprint arXiv:1611.04558 (2016).
4. Levy, O., Sogaard, A., Goldberg, Y.: A Strong Baseline for Learning Cross-Lingual Word Embeddings from Sentence Alignments. In: Proc. of the 15th Conf. of the European Chapter of the ACL 1. ACL, Valencia (2017), 765 774.
5. Luong, T., Pham, H., Manning, C.D.: Bilingual word representations with monolingual quality in mind. In Proc. of the 1st Workshop on Vector Space Modeling for NLP. ACL, Berlin (2015), 151 159.
6. Mikolov, T., et al.: Distributed representations of words and phrases and their compositionality. In: Advances in neural information processing systems 26, (2013), 3111-3119.
7. Mohammad, S.M., Turney, P.D.: Crowdsourcing a word–emotion association lexicon. Computational Intelligence 29.3 (2013), 436 465.
8. Pan, S.J., Yang, Q.: A survey on transfer learning. IEEE Transactions on knowledge and data engineering 22.10 (2010), 1345 1359.
9. Sidarenka, U., Stede, M.: Generating Sentiment Lexicons for German Twitter. In: Proc. of the Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media. The COLING 2016 Organizing Committee, Osaka (2016), 80 90.

10. Socher, R., et al.: Recursive deep models for semantic compositionality over a sentiment treebank. In: Proc. of the Conf. on Empirical Methods in Natural Language Processing.ACL, Seattle (2013), 1631 1642.
11. Upadhyay, S., et al.: Cross-lingual Models of Word Embeddings: An Empirical Comparison. In: Proc. of the 54th Annual Meeting of the ACL 1. ACL, Berlin (2016), 1661 1670.
12. Yang, Z., Salakhutdinov, Z., Cohen,WW.: Transfer Learning for Sequence Tagging with Hierarchical Recurrent Networks. In: 5th Int. Conf. on Learning Representations. (2017).
13. Zhou, H., et al.: A Neural Probabilistic Structured-Prediction Model for Transition-Based Dependency Parsing. In: Proc. of the 53rd Annual Meeting of the ACL 1. ACL, Beijing (2015), 1213 1222.
14. Zhou, X., Wan, X., Xiao, J.: Cross-Lingual Sentiment Classification with Bilingual Document Representation Learning. In: Proc. of the 54th Annual Meeting of the ACL 1. ACL, Berlin (2016), 1403 1412.

Poďakovanie: Tento článok vznikol vďaka čiastočnej podpore projektov VEGA 1/0646/15 a KEGA 009STU-4/2014.

Annotation:

Transfer Learning between Natural Languages

Deep learning is currently a popular approach to natural language processing. It is however very data-demanding and languages without proper resources can not utilize it fully. Transfer learning between languages tackles this problem as it transfer trained models from one language to another. We discuss such learning here and we also present our own experiment concerned with word-level induction of sentiment from English to German.