Detekce polarity textu s využitím mezijazyčné transformace

Jakub Šmíd[1]

1 Úvod

Detekce polarity textu je úloha zpracování přirozeného jazyka, jejímž cílem je klasifikace vstupního textu dle jeho polarity. Pro tuto úlohu jsou potřeba anotovaná data, kterých může být v některých jazycích nedostatek. Proto se hledají metody, které umožňují použití anotovaných dat z jednoho jazyka v jazyce druhém. Jednou z možností jsou mezijazyčné transformace.

Cílem této práce bylo vytvořit program pro detekci polarity českého textu a pomocí experimentů ověřit možnosti využití dat z jiného jazyka (konkrétně angličtiny) s použitím mezijazyčných transformací a zhodnotit jejich vliv na úspěšnost detekce polarity.

2 Lineární mezijazyčné transformace

Lineární mezijazyčné transformace transformují předtrénované jednojazyčné slovní vektory do společného prostoru pomocí lineárního zobrazení a dvojjazyčných slovníků, jak shrnul např. Ruder et al. (2019). Lineární zobrazení umožňuje transformaci mezi dvěma vektorovými prostory prostřednictvím afinních transformací (např. rotace, posun a zrcadlení). Tyto přístupy vychází z pozorování, která ukazují, že rozmístění vektorů slov zdrojového jazyka je po provedení vhodné lineární transformace geometricky velmi podobné rozmístění vektorů slov jejich překladů. Cílem je nalézt transformační matici $\mathbf{W}$, která umožní transformovat vektorový prostor zdrojového jazyka s do vektorového prostoru cílového jazyka t. Vynásobením vektoru $\mathbf{x}^s$ slova v původním prostoru transformační maticí $\mathbf{W}$ je získán příslušný vektor $\mathbf{x}^t$ v prostoru cílovém (viz rovnice 1). V této práci byla pro zisk matice $\mathbf{W}$ použita metoda nejmenších čtverců (MSE), kterou navrhl Mikolov et al. (2013), a ortogonální metoda, kterou popsal Xing et al. (2015). Obě metody ponechávají vždy jeden prostor nezměněný a druhý na něj mapují.

$$\mathbf{x}^t = \mathbf{W}\mathbf{x}^s \tag{1}$$

3 Experimenty a výsledky

Experimenty byly provedeny s použitím tří datových sad obsahujících filmové recenze – jedné české (ČSFD) a dvou anglických (SST a IMDb). Pro detekci polarity textu byly navrženy a implementovány dva modely neuronových sítí – konvoluční neuronová síť (CNN) a rekurentní neuronová síť (konkrétně obousměrná LSTM) – v kombinaci s natrénovanými slovními vektory *fastText*. Modely byly nejprve natrénovány a vyhodnoceny pouze na českých datech. Dále byly s využitím lineárních mezijazyčných transformací natrénovány na anglických datech a vyhodnoceny na češtině. Transformace proběhla z angličtiny do češtiny i naopak. Pro tvorbu transformační matice bylo vybráno 5 000 nebo 20 000 slov. Tabulka 1 ukazuje dosažené výsledky porovnané s jednojazyčnými experimenty.

[1] student bakalářského studijního programu Inženýrská informatika, obor Informatika, e-mail: biba10@students.zcu.cz

BiLSTM							
Počet tříd	**CS**	**Dataset**	**Metoda**	**5 000**		**20 000**	
				EN→CS	**CS→EN**	**EN→CS**	**CS→EN**
3	84,92 ± 0,30	SST	MSE	46,06 ± 0,75	**48,91** ± 2,59	45,86 ± 0,92	**47,53** ± 3,72
			ortog.	**47,25** ± 3,10	**48,14** ± 1,63	**49,15** ± 2,10	**49,71** ± 3,03
		SST (fráze)	MSE	46,83 ± 0,60	**46,52** ± 2,89	46,42 ± 0,98	**47,59** ± 4,29
			ortog.	**49,53** ± 2,91	**48,09** ± 1,39	**50,09** ± 2,35	**49,10** ± 2,27
2	94,29 ± 0,29	SST	MSE	82,82 ± 0,71	84,91 ± 0,71	83,37 ± 2,49	82,54 ± 2,49
			ortog.	79,61 ± 2,69	82,41 ± 2,09	81,05 ± 1,88	82,47 ± 2,49
		SST (fráze)	MSE	82,60 ± 2,69	82,57 ± 2,71	83,33 ± 0,93	84,32 ± 0,93
			ortog.	82,42 ± 1,94	83,57 ± 1,25	82,41 ± 2,46	83,14 ± 2,64
		IMDb	MSE	**86,20** ± 1,84	**87,48** ± 1,84	85,17 ± 0,89	**87,98** ± 0,89
			ortog.	**86,89** ± 0,64	**88,25** ± 1,14	**87,59** ± 0,48	**88,57** ± 0,36

CNN							
Počet tříd	**CS**	**Dataset**	**Metoda**	**5 000**		**20 000**	
				EN→CS	**CS→EN**	**EN→CS**	**CS→EN**
3	83,18 ± 0,14	SST	MSE	46,83 ± 0,09	**58,87** ± 1,10	47,12 ± 0,10	**57,60** ± 1,61
			ortog.	**50,32** ± 1,79	50,58 ± 1,10	**51,53** ± 1,42	50,38 ± 1,08
		SST (fráze)	MSE	46,81 ± 0,10	**57,71** ± 1,21	47,18 ± 0,09	**57,63** ± 0,92
			ortog.	**50,07** ± 1,27	51,50 ± 0,86	**51,46** ± 0,63	50,42 ± 0,75
2	93,86 ± 0,12	SST	MSE	86,46 ± 0,19	86,66 ± 0,46	86,99 ± 0,12	86,19 ± 0,65
			ortog.	84,25 ± 1,44	85,46 ± 0,80	86,18 ± 0,30	86,82 ± 0,31
		SST (fráze)	MSE	86,31 ± 0,37	86,47 ± 0,22	86,86 ± 0,17	86,83 ± 0,40
			ortog.	84,80 ± 1,17	86,36 ± 0,40	85,93 ± 0,44	86,41 ± 0,40
		IMDb	MSE	88,05 ± 0,20	86,64 ± 1,35	**88,36** ± 0,10	**88,11** ± 1,72
			ortog.	87,18 ± 0,34	87,81 ± 0,37	**88,16** ± 0,17	**88,99** ± 0,34

Tabulka 1: Makro F-míra v procentech pro modely BiLSTM a CNN se slovními vektory *fastText* (natrénované na datech z IMDb a SST datasetů pro angličtinu, z ČSFD datasetu pro češtinu) při použití mezijazyčných transformací s transformační maticí získanou z 5 000 či 20 000 slov metodou nejmenších čtverců nebo ortogonálně a transformací z angličtiny do češtiny (**EN→CS**) i naopak (**CS→EN**). Trénováno jen na datech anglických (sloupec **Dataset**) a vyhodnoceno na českých testovacích. Sloupec **CS** slouží k porovnání s jednojazyčnými výsledky. Pro každý model, počet tříd a směr transformace je nejlepší výsledek zvýrazněný **tučně**, při překrývání intervalu spolehlivosti je navíc **podtržený**.

4 Závěr

Porovnání jednojazyčných experimentů s experimenty s využitím mezijazyčné transformace ukázalo, že s dostatečným množstvím výhradně anglických dat lze dosáhnout velmi dobrých výsledků, které jsou jen o 5 až 6 % horší v porovnání s modely trénovanými jen na českých datech.

Literatura

Mikolov, T., Le, Q. V., a Sutskever, I. (2013) Exploiting Similarities among Languages for Machine Translation. *CoRR*. Dostupné z: `http://arxiv.org/abs/1309.4168`.

Ruder, S., Vulić, I., a Søgaard, A. (2019) A survey of cross-lingual word embedding models. *Journal of Artificial Intelligence Research*, s. 569–631.

Xing, C. et al. (2015) Normalized word embedding and orthogonal transform for bilingual word translation. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, s. 1006-1011.