FAKULTA APLIKOVANÝCH VĚD ZÁPADOČESKÉ UNIVERZITY V PLZNI

STUDENTSKÁ VĚDECKÁ KONFERENCE 2019

Vícejazyčná sémantická podobnost textů

Michal Tušl[1]

1 Úvod

Tato práce se zabývá metodami učení bez učitele pro sémantickou podobnost textu. Tyto metody jsou založené na distribuční hypotéze, předpokladu, že význam slova lze odvodit z jeho použití (distribuce v textu). Navíc se v této práci metody rozšiřují, aby bylo možné vyjádřit význam textu napříč různými jazyky. Cílem úlohy je tedy určit, jak moc jsou dvě věty v odlišných jazycích významově podobné.

Význam slov $w \in W$, kde W je slovník všech slov, je reprezentován jako vektor reálných čísel v mnoharozměrném vektorovém prostoru s dimenzí d, $w \in \mathbb{R}^d$. Slova, která se vyskytla ve stejných kontextech, jsou si blízko ve vektorovém prostoru a předpokládá se tedy, že mají podobný význam. Tomuto vektorovému prostoru se říká sémantický prostor (*semantic space*) a vektorům jednotlivých slov sémantický vektor (*semantic vector*). Pro získání sémantické reprezentace slov byly natrénovány modely *GloVe*, *Word2Vec* a *FastText*.

2 Transformace sémantických prostorů

Lineární transformace vektorových prostorů je způsob, jak jeden sémantický prostor transformovat do jiného sémantického prostoru (Brychcín (2018)). Toho lze využít pro získání jednotné sémantické reprezentace slov a vět napříč různými jazyky. V této práci se pro naučení transformací používala metoda nejmenších čtverců (LST), kanonická korelační analýza (CCA) a ortogonální transformace (ORT).

Předpokládejme dva sémantické prostory $\boldsymbol{X}$ a $\boldsymbol{Y}$, Cílem je najít transformační matici $\boldsymbol{T}$ s rozměry $d \times d$ takovou, že platí:

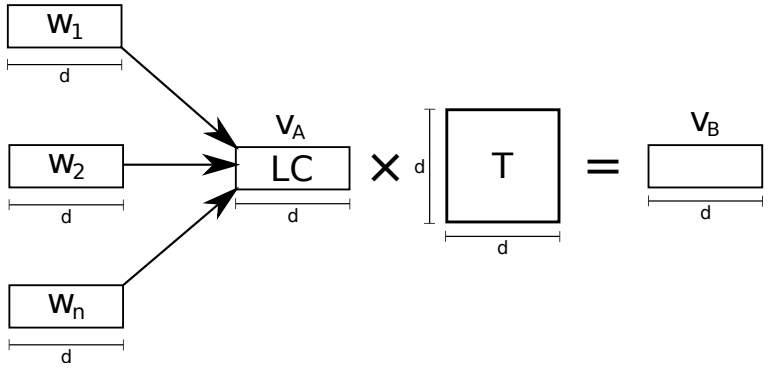
$$\boldsymbol{X} \times \boldsymbol{T} \approx \boldsymbol{Y}. \tag{1}$$

Transformace jsou trénovány na slovních vektorech, které jsou transformovány do sémantického prostoru jiného jazyka. Kromě transformací na slovech se otestovaly nové, dosud nepublikované metody, jejichž myšlenka je transformovat nikoliv vektory slov, ale rovnou vektory celých vět. Tyto způsoby transformací jsou pojmenovány jako transformace na větách a transformace *Paragraph2Vec* (Le - Mikolov (2014)) modelu.

K trénování transformační matice na větách je potřeba paralelní korpus, například takový, který se používá pro strojový překlad. Tento korpus by měl obsahovat větu ve zdrojovém jazyce a její ekvivalent v cílovém jazyce. Vektory pro tyto věty jsou získány metodami *Paragraph2Vec*, *Skip-thoughts* nebo lineární kombinací slovních vektorů. Tyto vektory vytvoří matice $\boldsymbol{X}$ a $\boldsymbol{Y}$, kde každý řádek je vektor jedné věty. Nyní se stejně jako pro transformace slov hledá transformační matice $\boldsymbol{T}$. Po vypočtení transformační matice lze transformovat vektory vět, vytvořené modelem pro zdrojový jazyk, do sémantického prostoru cílového jazyka, viz

[1] student navazujícího studijního programu Inženýrská informatika, obor Softwarové inženýrství, specializace Zpracování přirozeného jazyka, e-mail: tuslm@students.zcu.cz

Obrázek 1, který ilustruje transformaci lineární kombinace slovních vektorů.



Obrázek 1: Transformace lineární kombinace slov ze zdrojového do cílového jazyka.

Dalším novým způsobem transformace vektorů mezi jazyky je transformace na slovech, které jsou uloženy v *Paragraph2Vec* modelu.

3 Výsledky

Testování metod a měření experimentů bylo prováděno na datasetech z konferencí *SemEval* a datasetu *GoranGlavas*. Meřena byla Pearsonova (PC) a Spearmanova (SC) korelace mezi kosínovou podobností dvou vektorů a hodnocením z datasetu. Celkově byly metody trénovány na porovnávání těchto jazyků: angličtina, španělština, italština, arabština, turečtina a chorvatština. Většina těchto metod dosáhla na testovaných datasetech velmi slibných výsledků. Pro dataset *SemEval-2017* bylo nejlepší trénovat transformační matici ortogonální transformací nebo kanonickou korelační analýzou, neboť oba přístupy jsou srovnatelné. Pro dataset *GoranGlavas* byla nejlepší ortogonální transformace. Shrnutí průměrných výsledků metod na obou datasetech je v tabulce 1.

	SemEval-2017						GoranGlavas					
Trans.	na slovech		na větách		Paragraph2Vec		na slovech		na větách		Paragraph2Vec	
	PC	**SC**	**PC**	**SC**	**PC**	**SC**	**PC**	**SC**	**PC**	**SC**	**PC**	**SC**
LST	0,254	0,272	0,336	0,352	**0,181**	0,164	0,546	0,546	0,535	0,518	**0,296**	**0,322**
CCA	0,277	0,295	**0,354**	**0,372**	0,150	0,147	**0,568**	**0,559**	0,534	0,531	0,200	0,228
ORT	**0,293**	**0,308**	0,347	0,360	0,155	**0,168**	0,566	0,556	**0,547**	**0,532**	0,192	0,222

Tabulka 1: Srovnání dosažených průměrných výsledků na vícejazyčných datasetech z konference *SemEval-2017* a *GoranGlavas*.

Literatura

Brychcín, T. (2018) Linear Transformations for Cross-lingual Semantic Textual Similarity. *CoRR*. Dostupné z: `http://arxiv.org/abs/1807.04172`.

Le, Q., Mikolov, T. (2014) Distributed Representations of Sentences and Documents. *In Proceedings of the 31st International Conference on International Conference on Machine Learning - Volume 32*, ICML'14, s. II–1188–II–1196. Dostupné z: `http://dl.acm.org/citation.cfm?id=3044805.3045025`.