KATEDRA INFORMATIKY A VÝPOČETNÍ TECHNIKY

Michal Tušl

Diplomová práce

Inženýrská informatika
Softwarové inženýrství
2018/2019

Vedoucí práce:
Ing. Tomáš Brychcín, Ph.D.

Vícejazyčná sémantická podobnost textů

Abstrakt

Tato práce se zabývá metodami strojového učení bez učitele pro měření sémantické podobnosti textů napříč různými jazyky. Pro monolinguální reprezentaci textu bylo natrénováno několik modelů na korpusu z Wikipedie. Pro vytvoření jazykově nezávislé reprezentace významu jsou monolinguální sémantické prostory transformovány do společného prostoru pomocí lineární transformace. Práce zkoumá lineární transformace za pomocí metody nejmenších čtverců, kanonické korelační analýzy a ortogonální transformace. Kromě standardní transformace na slovech práce představuje dva nové přístupy, a to transformaci na větách a transformaci Paragraph2Vec modelu. Experimenty jsou provedeny na vícejazyčných datasetech SemEval-2017 a GoranGlavas a je měřena Pearsonova a Spearmanova korelace oproti člověku. Zkoumané metody dosahují slibných výsledků na těchto datasetech.

Úvod

Tato práce se zabývá metodami učení bez učitele pro sémantickou podobnost 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.

Distribuční sémantika

Distribuční sémantika nabízí poměrně jednoduchý a praktický způsob, jak reprezentovat sémantiku jednotlivých slov v textu. Modely distribuční sémantiky jsou založeny na předpokladu, že význam slova je dán okolím, ve kterém se slovo vyskytlo. Většina metod pro sémantickou reprezentaci je založena na trénování bez učitele, jediné co potřebují, je velké množství textu, ze kterého se lze naučit souvislosti mezi slovy nebo větnými celky. Metody distribuční sémantiky jsou založeny na distribuční hypotéze, která říká, že dvě slova vyskytující se ve stejných kontextech by měla být sémanticky podobná.

Význam slova je reprezentován jako vektor reálných čísel v mnoharozměrném vektorovém prostoru. 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 a vektorům jednotlivých slov sémantický vektor. Podobnost dvou slov se vypočte kosínem úhlu jejich vektorů.

Sémantická reprezentace textů

Metody pro sémantickou reprezentaci textu se dělí na dvě skupiny, první jsou metody, kde nezáleží na pořadí slov v textu, pro který se sestavuje vektor. Mezi tyto metody patří Bag-of-Words a lineární kombinace. Do druhé skupiny patří metody, které berou v úvahu pozice jednotlivých slov v textu. Sestavování těchto vektorů je již složitější, jedná se tedy o komplexní metody Paragraph2Vec a Skip-thoughts.

Transformace sémantických prostorů

Lineární transformace vektorových prostorů je způsob, jak jeden sémantický prostor transformovat do jiného sémantického prostoru. Lineární transformace mohou být použity pro provedení afinních transformací, jako je posunutí, otáčení, zrcadlení, zkosení, a další transformace. Toho lze využít pro získání jednotné sémantické reprezentace slov a vět napříč různými jazyky.

Transformace jsou trénovány na slovních vektorech, které jsou transformovány do sémantického prostoru jiného jazyka. Kromě tohoto přístupu jsou v této práci popsány nové, dosud nepublikované metody, jejichž myšlenska je transformovat nikoliv vektory slov, ale rovnou vektory celých vět. Tyto způsoby transformací jsou pojmenovány jako transformace na větách a transformace Paragraph2Vec modelu.

Dosažené výsledky

Testování metod a měření experimentů bylo prováděno na datasetech z konferencí SemEval a datasetu GoranGlavas. Celkově byly metody trénovány na porovnávání těchto jazyků: angličtina, španělština, italština, arabština, turečtina a chorvatština. Většina těchto metod dosáhla na všech jazycích a dostupných datasetech velmi dobrých výsledků.

Nejlepší byla metoda Word2Vec se způsobem transformací na větách. 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.

	Způsob transformace					
	na slovech		na větách		Paragraph2Vec	
Trans.	PC	SC	PC	SC	PC	SC
LST	0,254	0,272	0,336	0,352	0,181	0,164
CCA	0,277	0,295	0,354	0,372	0,150	0,147
ORT	0,293	0,308	0,347	0,360	0,155	0,168

Tabulka 1: Shrnutí průměrných výsledků na datasetu SemEval-2017.

	Způsob transformace					
	na slovech		na větách		Paragraph2Vec	
Trans.	PC	SC	PC	SC	PC	SC
LST	0,546	0,546	0,535	0,518	0,296	0,322
CCA	0,568	0,559	0,534	0,531	0,200	0,228
ORT	0,566	0,556	0,547	0,532	0,192	0,222

Tabulka 2: Shrnutí průměrných výsledků na datasetu GoranGlavas.

Závěr

Přínosem této práce jsou dva nové způsoby, jak natrénovat transformační matici. První způsob je transformace na větách. Místo trénování na vektorech slov, se trénuje rovnou na vektorech vět, které po vynásobení transformační maticí vyjadřují význam věty v cílovém jazyce. Tím se získá sémantický reprezentace textu napříč různými jazyky.