

Bc. Adam Mištera

Diplomová práce

Inženýrská informatika
Softwarové inženýrství
2020/2021Vedoucí práce:
Ing. Tomáš Brychcín Ph.D.

Mezi-jazyčné transformace sémantických prostorů

Abstrakt

Mezi-jazyčná reprezentace významu je aktuální téma v oblasti zpracování přirozeného jazyka. Sémantické prostory z různých jazyků jsou transformovány do jednoho sdíleného univerzálního prostoru. Tento fakt umožňuje přenést znalosti z jazyků, které jsou velmi bohaté na zdroje, do jazyků, které jsou na zdroje velmi chudé. Cílem této diplomové práce bylo analyzovat, porovnat a implementovat dostupné metody pro jazykově nezávislou sémantickou reprezentaci. Současně byly prozkoumány a implementovány vybrané nelineární transformace založené na umělých neuronových sítích a shlukování pro transformaci sémantických prostorů. Kvalita lineárních i nelineárních metod byla následně vyhodnocena na různých úlohách, jako například datové sady slovních podobností, slovních analogií a strojovém překladu. Nově implementované nelineární transformace v mnoha ohledech překonaly lineární transformace ve všech sledovaných kategoriích na rozdílných jazycích s odlišných jazykových rodin.

Úvod

Cílem diplomové práce je průzkum nejmodernějších metod pro jazykově nezávislou reprezentaci významu a následný návrh a implementace nelineárních metod s využitím umělých neuronových sítí a metody shlukování určené pro transformaci sémantických prostorů. Implementované metody pro jazykově nezávislou reprezentaci významu jsou dále otestované na datasetech slovních podobností, slovních analogií a na strojovém překladu slov.

Sémantické prostory

Sémantický prostor lze definovat jako matematickou funkci, která mapuje slovo w do Euklidovského prostoru o dimenzi d . Slovo je následně reprezentováno pomocí vektoru reálných čísel. Získaný vektor posléze v daném sémantickém prostoru reprezentuje samotný význam slova w .

Lineární transformaci mezi sémantickými prostory můžeme posléze definovat jako

vynásobení transformační maticí T . Pomocí uvedeného násobení matic lze provést takzvanou *afinní transformaci*. Mezi nejvýznamnější zástupce lineárních transformací se řadí *metoda nejmenších čtverců*, *ortogonální transformace* a dále také transformace pomocí *kanonické korelace*.

Navržené transformace

- Transformace s **hinge-loss** – Cílem této metody je zařadit správný překlad y_i slova x_i výše než všechny ostatní možné překlady y_j z cílového sémantického prostoru. Stanovením minimální vzdálenosti, kterou má predikovaný vektor y_i od všech ostatních vektorů y_j udržovat, lze následně vylepšit vlastnosti transformace.
- Transformace se **shlukováním** – Klíčovou myšlenkou tohoto přístupu pro vytvoření transformace bylo využití shlukové analýzy pro roztržení slov sémantického prostoru do shluků, které sdílejí stejný či podobný význam. Cílem této metody bylo pokusit se s pomocí lokálních transformací vytvořených pro jednotlivé shluky vylepšit vlastnosti lineární transformace.

Transformace	SemEval	Analogie	MT@1	MT@5	Skewness	Průměr
de – es	0.620	0.450	0.346	0.509	5.5	0.481
de – it	0.613	0.472	0.342	0.503	5.0	0.483
en – de	0.650	0.402	0.590	0.732	2.7	0.594
en – es	0.667	0.453	0.751	0.890	2.9	0.690
en – it	0.660	0.500	0.559	0.717	2.2	0.609
es – it	0.665	0.501	0.467	0.636	3.8	0.567
en – cs	-	0.448	0.350	0.541	3.4	0.446
en – hr	-	0.409	0.303	0.469	4.4	0.394
Průměr	0.646	0.454	0.464	0.625	3.7	0.547

Dosažené výsledky

Z experimentů pro lineární transformace dosáhla nejmenší průměrné hodnoty *metoda nejmenších čtverců*, konkrétně **49.9 %**. Tuto hodnotu přesáhly všechny navržené

nelineární transformace využívající neuronové sítě s cenovou funkcí *hinge-loss* a také shlukování. Nejlepší lineární transformace s pomocí *kanonické korelace* dosáhla průměrné hodnoty **53.6 %**.

Ze všech transformací dosáhla nejlepšího průměrného výsledku **54.7 %** neuronová síť využívající regularizaci a cenovou funkci *hinge-loss*. Kompletní výsledky pro tuto transformaci lze vidět v tabulce.

Závěr

Hlavním přínosem této diplomové práce je především výzkum na poli nelineárních metod určených pro transformaci sémantických prostorů, zejména návrh zcela nových metod řešících tuto úlohu. Velmi slibné výsledky prokázaly především nelineární metody založené na umělých neuronových sítích využívajících cenovou funkci *hinge-loss*. Ve všech sledovaných kategoriích překonaly všechny analyzované lineární transformace. Konkrétně neuronová síť využívající regularizaci dosáhla nejvyšší hodnoty pro slovní podobnost a strojový překlad. Oproti nejlepší lineární transformaci pomocí *kanonické korelace* došlo ke zlepšení o více než jedno procento.