FAKULTA APLIKOVANÝCH VĚD ZÁPADOČESKÉ UNIVERZITY V PLZNI

STUDENTSKÁ VĚDECKÁ KONFERENCE 2017

Explicitní sémantická analýza

Michal Tušl[1]

1 Úvod

Zpracování přirozeného jazyka – NLP (*Natural Language Processing*) je rozvíjejícím se oborem, který je čím dál víc využíván. Používají ho velké společnosti jako jsou Google nebo Facebook ve svých vyhledávačích, zobrazování příspěvků a dalších funkcích. Jádrem tohoto zpracování je sémantika, která se zabývá významem výrazu z různých úrovní jazyka. V tomto případě se tedy jedná hlavně o zpracování a částečné porozumění textu počítačem. V současné době existuje již mnoho metod, které dokážou určit význam textu. Metody trénování bez učitele se učí samy na velkém korpusu dat. Tyto metody jsou založené na distribuční hypotéze, předpokladu, že význam slova lze odvodit z jeho použití (distribuce v textu).

2 Popis metod

Jednou z těchto metod je i metoda ESA (*Explicit Semantic Analysis*) (Gabrilovich a Markovitch , 2009). ESA vytváří vícerozměrné vektory, kde každý prvek vektoru je jeden kontext. Tyto vektory reprezentují sémantiku jednotlivých slov. Na základě vypočtených vektorů pak lze porovnávat podobnost významu slov i souvislých textů. K učení významu potřebuje metoda velké množství dat. K tomu je použita Wikipedie, která obsahuje velké množství článků z různých oborů (kontexty). Výsledek této reprezentace lze využít pro další aplikace jako je například vyhledávání informací, strojový překlad, opravy pravopisu a gramatiky, jazykové modelování pro rozpoznávání řeči, dialogové systémy a další. Metoda ESA používá TF-IDF pro výpočet spojistosti mezi slovem t a kontextem d. Tyto hodnoty ukládá do matice $\boldsymbol{M}$, kde řádky *i* představují slova a sloupce *j* články (kontexty), v kterých se slova vyskytla:

$$\boldsymbol{M}_{i,j} = \mathrm{tf}(t_i, d_j) \cdot \log \frac{n}{df_i}. \tag{1}$$

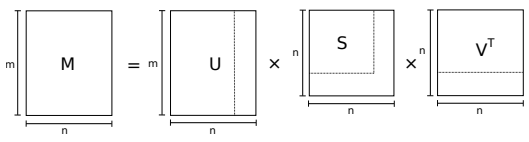
Hodnota *tf* je definována jako: $1 + \log \mathrm{N}(t_i, d_j)$, kde $\mathrm{N}(t_i, d_j)$ je počet výskytů slova t_i v dokumentu d_j. Hodnota $df_i = |\{d_k : t_i \in d_k\}|$ představuje počet dokumentů (článků), ve kterých se vyskytuje výraz (slovo) t_i. Sémantická interpretace slova t_i je řádek i z matice $\boldsymbol{M}$. Význam slova je dán vektorem kontextů spojeným s jejich TF-IDF hodnotami, které odráží spojitost mezi každým tématem a vybraným slovem.

Tuto metodu jsem následně rozšířil o použití kategorií ve Wikipedii jako kontextů. Výpočet míry asociace mezi slovem a kategorií se vypočte pomocí PMI (*Pointwise mutual information*). PMI určuje do jaké míry jsou dva jevy na sobě závislé. V tomto případě tedy do jaké míry je slovo t_i závislé na určité kategorii c_j:

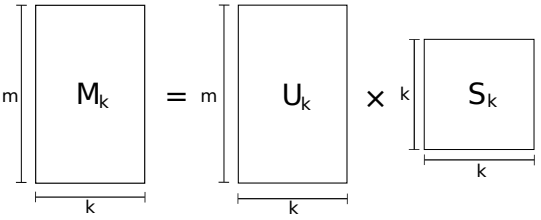
$$\mathrm{pmi}(t, c) = \log \frac{\mathrm{p}(t, c)}{\mathrm{p}(t)\mathrm{p}(c)} = \log \frac{\mathrm{p}(t|c)}{\mathrm{p}(c)}. \tag{2}$$

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

Na výslednou matici této metody byl dále aplikován singulární rozklad (SVD), který redukuje dimezi matice, čímž znižuje paměťové nároky a zároveň zlepšuje výsledky metody. SVD je rozklad matice na tři matice $\boldsymbol{U}, \boldsymbol{S}, \boldsymbol{V}$, viz Obrázek 1. Pro redukci dimenze je potřeba nejdříve redukovat matice $\boldsymbol{U}$ a $\boldsymbol{S}$ na požadovanou dimenzi k. Aproximace původní matice $\boldsymbol{M}$ s redukovanou dimenzí na k se získá vynásobením matic $\boldsymbol{U}_k$ s maticí $\boldsymbol{S}_k$, součinem matic pak vznikne matice $\boldsymbol{M}_k$, viz Obrázek 2. Tato metoda nebyla ještě nikde představena, jedná se tedy o zcela novou metodu nazvanou LSC (*Latent Semantic Categories*).



Obrázek 1: SVD – rozklad matice



Obrázek 2: SVD – redukce dimenze

Singulární rozklad byl též aplikován na metodu ESA, z které takto vznikla metoda LSA (*Latent Semantic Analysis*) (Landauer et al. , 1998). U metod LSC a LSA lze měnit velikost cílové dimenze, proto jsem u nich provedl měření pro různé velikosti dimenze. Nejlépe vyšly výsledky pro dimenzi 2500, které jsou uvedeny v tabulce 1.

3 Výsledky

Všechny prezentované metody byly testovány na datasetech *RG-65* a *WS-353* obsahující lidmi definovanou významovou podobnost párů slov. Shodnost výsledků implementace s datasety byla určována pomocí Spearmanovy (SC) a Pearsonovy korelace (PC). Výsledky metod jsou uvedeny v tabulce 1. Všechny metody dosáhly dobrých výsledků pro český i anglický jazyk.

	angličtina				čeština				čeština – stemming			
	RG-65		**WS353**		**RG-65**		**WS353**		**RG-65**		**WS353**	
Model	**PC**	**SC**	**PC**	**SC**	**PC**	**SC**	**PC**	**SC**	**PC**	**SC**	**PC**	**SC**
ESA	0,57	0,76	0,41	0,51	0,70	0,75	0,32	0,51	0,38	0,83	0,38	0,49
PMI	0,51	0,58	0,49	0,48	0,58	0,62	0,41	0,45	0,47	0,70	0,43	0,44
LSA	0,68	0,72	0,52	0,54	0,65	0,67	0,46	0,44	0,66	0,70	0,49	0,51
LSC	0,63	0,62	0,55	0,56	0,57	0,62	0,43	0,42	0,59	0,59	0,48	0,48

Tabulka 1: Výsledky metod pro anglický a český jazyk.

Literatura

Gabrilovich, E., Markovitch, S. (2009) Wikipedia-based Semantic Interpretation for Natural Language Processing. *J. Artif. Int. Res*, pp. 443—498.

Landauer, T. K., Foltz, P. W., Laham, D. (1998) An Introduction to Latent Semantic Analysis. *Discourse Processes*, pp. 259—284.