Toward Incorporation of Relevant Documents in word2vec

Navid Rekabsaz[1], Bhaskar Mitra[2], Mihai Lupu[1], Allan Hanbury[1]
[1] TU Wien, Vienna, Austria
[2] Microsoft, UCL, Cambridge, UK
contact: rekabsaz@ifs.tuwien.ac.at

Problem Definition & Motivation

- Previous studies show the effectiveness of using the word2vec's related terms for document retrieval when applying Generalized and Extended Translation models. (Rekabsaz et al. CIKM 2016, ECIR 2017, SIGIR 2017)
- The translation models however extend each query term independently and don't consider other query terms.
- To address the problem, we introduce a novel *explicit representation* based on the Skip-Gram (SG) model.
- We discuss our ideas of using the explicit representation and local information for query-specific related terms.

Background

Embedding with Negative Sampling

The SG model has two sets of vectors: term and context vectors and aims to optimize the following probability:

$$p(c|w) = \frac{\exp(V_w \widetilde{V}_c)}{\sum_{c' \in W} \exp(V_w \widetilde{V}_{c'})}$$

The Negative Sampling redefines it with the probability that the co-occurrence of terms is genuine:

$$p(y=1|w,c) = \frac{\exp(V_w \widetilde{V}_c)}{\exp(V_w \widetilde{V}_c)+1} = \sigma(V_w \widetilde{V}_c)$$

$$J = -\sum_{\langle w,c\rangle \in X} \left[\log p(y=1|w,c) + k \mathop{\mathbb{E}}_{\check{c}_i \sim N} \log p(y=0|w,\check{c}_i) \right]$$

+ *subsampling* and *context distribution smoothing (cds)*

Explicit Representation

PMI also assesses a genuine co-occurrence:

$$PMI(w,c) = \log \frac{p(w,c)}{p(w)p(c)}$$

$$PPMI(w,c) = max(PMI(w,c),0).$$

$$SPPMI(w,c) = max(PMI(w,c) - \log(k), 0)$$

Levy and Goldberg 2014

SPPMI also considers *subsampling* and *cds* as follows:

$$PMI_\alpha(w,c) = \log \frac{p(w,c)}{p(w)p_\alpha(c)} \qquad p_\alpha(c) = \frac{f(\langle w,.\rangle,X)^\alpha}{\sum_{w' \in W} f(\langle w',.\rangle,X)^\alpha}$$

Explicit Skip-Gram Representation

Theory & Definition

$$ExpSG(w,c) = p(y=1|w,c) = \sigma(V_w \tilde{V}_c)$$

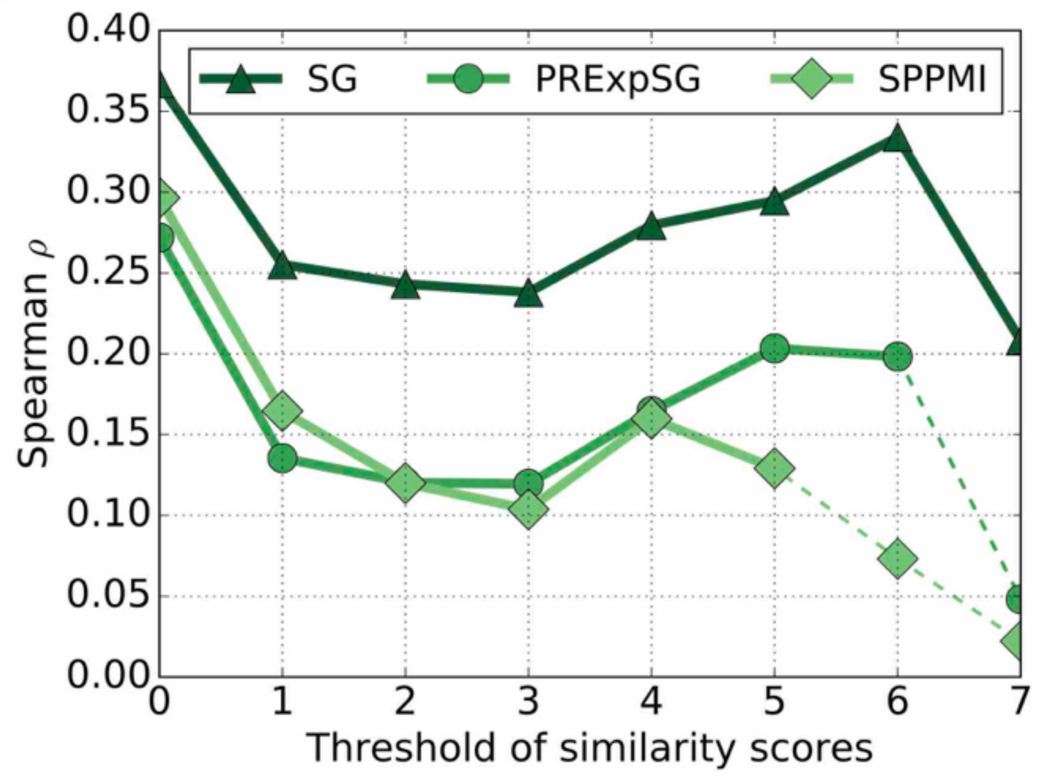
$$RExpSG(w,c) = ExpSG(w,c) - \mathop{\mathbb{E}}_{\check{c} \sim N} p(y=1|w,\check{c}) - \mathop{\mathbb{E}}_{\check{w} \sim N} p(y=1|\check{w},c)$$

$$\mathop{\mathbb{E}}_{\check{w} \sim N} p(y=1|\check{w},c) = \frac{\sum_{i=1}^{|W|} f(\check{w}_i, C) \cdot \sigma(V_{\check{w}_i} \widetilde{V}_c)}{\sum_{i=1}^{|W|} f(\check{w}_i, C)} \qquad \mathop{\mathbb{E}}_{\check{c} \sim N} p(y=1|w,\check{c}) = \frac{\sum_{i=1}^{|W|} f(\check{c}_i, C)^\alpha \cdot \sigma(V_w \widetilde{V}_{\check{c}_i})}{\sum_{i=1}^{|W|} f(\check{c}_i, C)^\alpha}$$

$$PRExpSG(w,c) = max(RExpSG(w,c),0)$$

Evaluation

Method	Sparsity	WS Sim.	WS Rel.	MEN	Rare	SCWS	SimLex
PPMI	98.6%	.681	.603	.702	.309	.601	.284
SPPMI	99.6%	**.722**	**.661**	.704	.394	.571	**.296**
ExpSG	0%	.596	.404	.645	.378	.549	.231
RExpSG	0%	.527	.388	.606	.311	.507	.215
PRExpSG	94.1%	.697	.626	**.711**	**.406**	**.614**	.272
SG	0%	.770	.620	.750	.488	.648	.367



Integration of Local Information

Let us refer to each cell of the matrix of explicit vector representations as $v(w,c)$

Based on the set of local documents *(F)*, we alter the cell values with the following formula:

$$\hat{v}(w,c) = \frac{1}{1+e^{-(a+b \cdot f(w,c,F))}} v(w,c)$$

Here are different suggestions for the function *f*

$$f_1(w,c,F) = f_1(c,F) = \mathbb{1}\left[f(c,F) > 0\right]$$

$$f_2(w,c,F) = f_2(c,F) = \frac{p(c|F)}{p(c|C)} = \frac{f(c,F)/\sum_{d \in F}|d|}{f(c,C)/\sum_{d \in C}|d|}$$

$$f_3(w,c,F) = \frac{p(w,c|X_F)}{p(w,c|X_C)} = \frac{f(\langle w,c\rangle,X_F)/|X_F|}{f(\langle w,c\rangle,X_C)/|X_C|}$$

$$f_4(w,c,F) = f_4(c) = p(c|\Theta_F) = \sum_{\theta_d \in \Theta_F} p(c|\theta_d) \prod_{q \in Q} p(q|\theta_d)$$

$$f_5(w,c,F) = p(w,c|\Theta_F) = \sum_{\theta_d \in \Theta_F} p(w|\theta_d)p(c|\theta_d) \prod_{i \in Q} p(q|\theta_d)$$

This work is funded by: Self-Optimizer (FFG 852624) in the EUROSTARS programme, funded by EU- REKA, the BMWFW and the European Union, ADMIRE (P 25905-N23) by FWF, and the Austrian Ministry for Science, Research and Economy. Thanks to Joni Sayeler and Linus Wretblad for their contributions in SelfOptimizer. Bhaskar Mitra is a part-time PhD student at University College Lond.

FACULTY OF !NFORMATICS