

Retrofitting Word Vectors to Semantic Lexicons

NAAACL2015

**Manaal Faruqui, Jesse Dodge
Sujay K. Jauhar, Chris Dyer
Eduard Hovy, Noah A. Smith**

Carnegie Mellon University

2015 年 6 月 16 日

Outline

- 论文方法概述
- 传统词向量生成工具
 - Glove, Skip-Gram, Global Context, Multilingual Vectors
- 语义词典 (Semantic Lexicon)
 - PPDB, WordNet, FrameNet
- 实验
 - 数据集
 - Word Similarity: WS-353, RG-65, MEN
 - Syntactic Relations, Synonym Selection, Sentiment Analysis
 - 顺序模型: Retrofitting with Semantic Lexicons
 - 联合模型: Semantic Lexicons during Learning
 - lazy method, periodic method
 - 实验分析

论文方法概述

- 把语义信息加入词向量；语义相关的单词词向量有更大的相似性
- 两类方法
 - 顺序模型：先使用传统词向量生成工具，再使用语义信息修正词向量
 - 联合模型：修正传统词向量训练时的目标函数，在其中加入语义信息

传统词向量生成工具

传统词向量生成工具 tools

Glove

- stanford:Jeffrey Pennington, Richard Socher
- 收集单词对的共现情况

word2vec

Skip-Gram Vectors

Global Context Vector

tree-RNN + local and global (document) context features

Multilingual Vector

SVD + CCA

语义词典

PPDB

- 复述 (paraphrase) 预料集
- 220 million paraphrase pairs
- 6 个版本: S,M,L,XL,XXL,XXXL, 容量依次增大, 质量依次降低
- [VBN] ||| pruned ||| cropped ||| $p(e|f)=4.33$ $p(f|e)=4.88$... ||| 0-0

FrameNet

- frame: *Cause_change_of_position_on_a_scale*
 \leftrightarrow push, raise, ..., growth

WordNet

WN_{syn} : 只对同义词连边

WN_{all} : 同义词、上位词、下位词都连边

实验时 $\alpha_i = 1, \beta_{ij} = degree(i)^{-1}$

实验数据集

Word Similarity

WS-353

- 353 个英语单词对 (200 个 13 个人标, 153 个 16 个人标)
- 相似度: 0-10 (可 0.5)

RG-65

- 65 对英语名词

MEN

- 3000 对单词 (共现 700 次)

Word Similarity

WS-353

- 353 个英语单词对 (200 个 13 个人标, 153 个 16 个人标)
- 相似度: 0-10 (可 0.5)

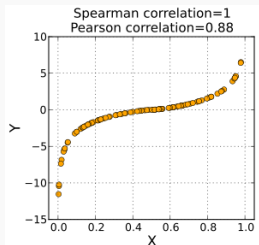
RG-65

- 65 对英语名词

MEN

- 3000 对单词 (共现 700 次)

评价结果好坏采用斯皮尔曼等级相关系数



$$\rho = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2 \sum_i (y_i - \bar{y})^2}} \in [-1, 1]$$

Syntactic Relation(SYN-REL)

- Mikolov 给出 (word2vec)
- 给定 a, b, c , 找到最合适的 d , 满足: a is to b as c is to d
- 实验时找和 $(q_a - q_b + q_c)$ 余弦相似度最大的单词作为 q_d

Synonym Selection(TOEFL)

- 80 个问题, 找出候选中与目标最相近的单词
 $rug \rightarrow \{sofa, ottoman, \textbf{carpet}, hallway\}$

Sentiment Analysis (SA)

- Socher 给出 (Glove)
- 6920(train)+872(dev)+1821(text) 个句子, 正负情感极性

实验

顺序模型

Retrofitting with Semantic Lexicons

顺序模型: Retrofitting with Semantic Lexicons

Notation

$V = \{w_1, w_2, \dots, w_n\}$: vocabulary

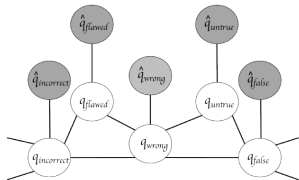
$E = \{(w_i, w_j), \dots\} \subseteq V \times V$: edges

$\Omega = (V, E)$: ontology

$\hat{q}_i, q_i \in \mathbb{R}^d$: word vector

$\hat{Q} = (\hat{q}_1, \dots, \hat{q}_n)$: original matrix

$Q = (q_1, \dots, q_n)$: target matrix;



顺序模型: Retrofitting with Semantic Lexicons

Notation

$V = \{w_1, w_2, \dots, w_n\}$: vocabulary

$E = \{(w_i, w_j), \dots\} \subseteq V \times V$: edges

$\Omega = (V, E)$: ontology

$\hat{q}_i, q_i \in \mathbb{R}^d$: word vector

$\hat{Q} = (\hat{q}_1, \dots, \hat{q}_n)$: original matrix

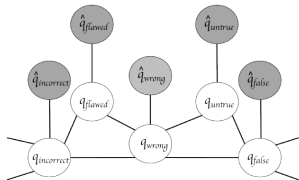
$Q = (q_1, \dots, q_n)$: target matrix;

❶ 传统工具 (word2vec) 生成初始向量空间 \hat{Q}

❷ 根据语义字典生成 Ω

❸ 最优 (小) 化 $\Psi(Q) =$

$$\sum_{i=1}^n \left[\alpha_i \|q_i - \hat{q}_i\|^2 + \sum_{(i,j) \in E} \beta_{ij} \|q_i - q_j\|^2 \right]$$



顺序模型: Retrofitting with Semantic Lexicons

Notation

$V = \{w_1, w_2, \dots, w_n\}$: vocabulary

$E = \{(w_i, w_j), \dots\} \subseteq V \times V$: edges

$\Omega = (V, E)$: ontology

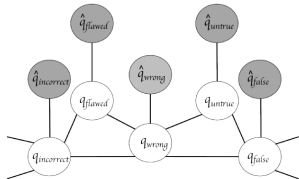
$\hat{q}_i, q_i \in \mathbb{R}^d$: word vector

$\hat{Q} = (\hat{q}_1, \dots, \hat{q}_n)$: original matrix

$Q = (q_1, \dots, q_n)$: target matrix;

- 1 传统工具 (word2vec) 生成初始向量空间 \hat{Q}
- 2 根据语义字典生成 Ω
- 3 最优 (小) 化 $\Psi(Q) =$

$$\sum_{i=1}^n \left[\alpha_i \|q_i - \hat{q}_i\|^2 + \sum_{(i,j) \in E} \beta_{ij} \|q_i - q_j\|^2 \right]$$



- (?) $\Psi(Q)$ 是凸函数 \Rightarrow 沿切线更新 $\Rightarrow q_i = \frac{\sum_{j: (i,j) \in E} \beta_{ij} q_j + \alpha_i \hat{q}_i}{\sum_{j: (i,j) \in E} \beta_{ij} + \alpha_i}$

顺序模型: Retrofitting with Semantic Lexicons

Notation

$V = \{w_1, w_2, \dots, w_n\}$: vocabulary

$E = \{(w_i, w_j), \dots\} \subseteq V \times V$: edges

$\Omega = (V, E)$: ontology

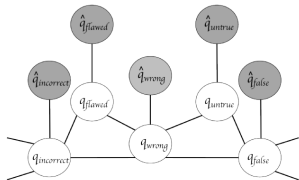
$\hat{q}_i, q_i \in \mathbb{R}^d$: word vector

$\hat{Q} = (\hat{q}_1, \dots, \hat{q}_n)$: original matrix

$Q = (q_1, \dots, q_n)$: target matrix;

- 1 传统工具 (word2vec) 生成初始向量空间 \hat{Q}
- 2 根据语义字典生成 Ω
- 3 最优 (小) 化 $\Psi(Q) =$

$$\sum_{i=1}^n \left[\alpha_i \|q_i - \hat{q}_i\|^2 + \sum_{(i,j) \in E} \beta_{ij} \|q_i - q_j\|^2 \right]$$



- $(?)\Psi(Q)$ 是凸函数 \Rightarrow 沿切线更新 $\Rightarrow q_i = \frac{\sum_{j:(i,j) \in E} \beta_{ij} q_j + \alpha_i \hat{q}_i}{\sum_{j:(i,j) \in E} \beta_{ij} + \alpha_i}$
- 10 次迭代收敛 ($(?)$ 邻接矩阵距离小于 10^{-2})

Experiment

顺序模型 (Retrofitting with Semantic Lexicons)

Lexicon	MEN-3k	RG-65	WS-353	TOEFL	SYN-REL	SA
Glove	73.7	76.7	60.5	89.7	67.0	79.6
+PPDB	1.4	2.9	-1.2	5.1	-0.4	1.6
+WN _{syn}	0.0	2.7	0.5	5.1	-12.4	0.7
+WN _{all}	2.2	7.5	0.7	2.6	-8.4	0.5
+FN	-3.6	-1.0	-5.3	2.6	-7.0	0.0
SG	67.8	72.8	65.6	85.3	73.9	81.2
+PPDB	5.4	3.5	4.4	10.7	-2.3	0.9
+WN _{syn}	0.7	3.9	0.0	9.3	-13.6	0.7
+WN _{all}	2.5	5.0	1.9	9.3	-10.7	-0.3
+FN	-3.2	2.6	-4.9	1.3	-7.3	0.5
GC	31.3	62.8	62.3	60.8	10.9	67.8
+PPDB	7.0	6.1	2.0	13.1	5.3	1.1
+WN _{syn}	3.6	6.4	0.6	7.3	-1.7	0.0
+WN _{all}	6.7	10.2	2.3	4.4	-0.6	0.2
+FN	1.8	4.0	0.0	4.4	-0.6	0.2
Multi	75.8	75.5	68.1	84.0	45.5	81.0
+PPDB	3.8	4.0	6.0	12.0	4.3	0.6
+WN _{syn}	1.2	0.2	2.2	6.6	-12.3	1.4
+WN _{all}	2.9	8.5	4.3	6.6	-10.6	1.4
+FN	1.8	4.0	0.0	4.4	-0.6	0.2

- frameNet 数据少, 效果差
- [Append:A] 语义:
Glove 好, 句法:
word2vec 好
- PPDB, WN_{all} 好
- retrofitting 对于句法
信息没有提升效果

联合模型：Semantic Lexicons during Learning

修改传统模型的训练过程，加入语义信息

lazy mode

核心思想：在传统模型的目标函数中加入体现语义信息的正则项

$$Q \text{ 的先验: } p(Q) \propto \exp \left(-\gamma \sum_{i=1}^n \sum_{j:(i,j) \in E} \beta_{ij} \|q_i - q_j\|^2 \right)$$

- ① $p(Q)$ 加入目标函数中
- ② 每次更新 k 个单词的向量 (lazy update)

periodic mode

核心思想：递归的过程中每更新 k 个词后使用下式更新所有的单词向量

$$q_i = \frac{\sum_{j:(i,j) \in E} \beta_{ij} q_j + \alpha_i \hat{q}_i}{\sum_{j:(i,j) \in E} \beta_{ij} + \alpha_i}$$

Experiment

联合模型效果测试

- log-bilinear (LBL) vectors 为基准 (Mnih and Teh, 2012)
- lazy** Mode: $k=100,000$

Method	k, γ	MEN-3k	RG-65	WS-353	TOEFL	SYN-REL	SA
LBL (Baseline)	$k = \infty, \gamma = 0$	58.0	42.7	53.6	66.7	31.5	72.5
LBL + Lazy	$\gamma = 1$	-0.4	4.2	0.6	-0.1	0.6	1.2
	$\gamma = 0.1$	0.7	8.1	0.4	-1.4	0.7	0.8
	$\gamma = 0.01$	0.7	9.5	1.7	2.6	1.9	0.4
	$k = 100M$	3.8	18.4	3.6	12.0	4.8	1.3
LBL + Periodic	$k = 50M$	3.4	19.5	4.4	18.6	0.6	1.9
	$k = 25M$	0.5	18.1	2.7	21.3	-3.7	0.8
LBL + Retrofitting	-	5.7	15.6	5.5	18.6	14.7	0.9

Experiment

对比实验

Yu and Dredze (2014)

- word2vec(CBOW)+retrofitting(PPDB)

Corpus	Vector Training	MEN-3k	RG-65	WS-353	TOEFL	SYN-REL	SA
WMT-11	CBOW	55.2	44.8	54.7	73.3	40.8	74.1
	Yu and Dredze (2014)	50.1	47.1	53.7	61.3	29.9	71.5
	CBOW + Retrofitting	60.5	57.7	58.4	81.3	52.5	75.7

Xu et al. (2014)

- word2vec(CBOW)+retrofitting(PPDB)

Wikipedia	SG	76.1	66.7	68.6	72.0	40.3	73.1
	Xu et al. (2014)	–	–	68.3	–	44.4	–
	SG + Retrofitting	65.7	73.9	67.5	86.0	49.9	74.6

多语言效果实验（每种语言
独立测试）

- retrofitting(WN_{all})

Language	Task	SG	Retrofitted SG
German	RG-65	53.4	60.3
French	RG-65	46.7	60.6
Spanish	MC-30	54.0	59.1

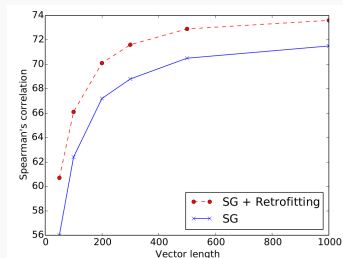
Experiment

多语言效果实验（每种语言
独立测试）

● retrofitting(WN_{all})

Language	Task	SG	Retrofitted SG
German	RG-65	53.4	60.3
French	RG-65	46.7	60.6
Spanish	MC-30	54.0	59.1

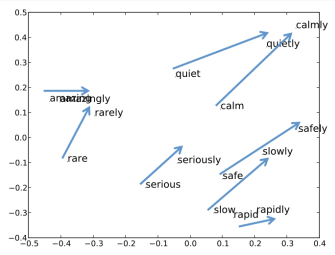
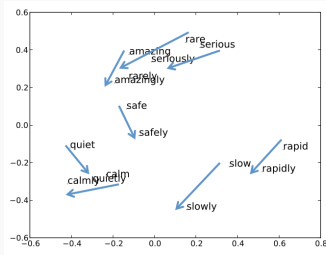
retrofitting 和向量长度效用实验



Experiment

可视化

利用 PCA 从 SG 训练得到的 100 维向量压缩为 2 维。左右图分别为使用 retrofitting 前后的向量位置



谢谢

大多数数据集可以在上面找到: <http://www.cs.cmu.edu/~mfaruqui/suite.html>

Append:A

GloVe vs word2vec

Model	Semantic	Syntactic	Total
GloVe (W+C)	79.6	61.0	69.4
word2vec (W)	72.7	65.8	68.9

<https://docs.google.com/document/d/1ydlujJ7ETSZ688RGfU5IMJJsbxAi-kRI8czSwpti15s/mobilebasic>