GloVe: Global Vectors for Word Representation

*EMNLP*2014

Jeffrey Pennington

Richard Socher

Christopher D. Manning

Stanford

2015 年 6 月 17 日

Glo(bal)Ve(ctor)

- 词向量模型
  - global matrix factorization (LSA)
  - local context window (skip-gram)
- word2vec 的实现方式
- GloVe
  - 模型（目标函数）
  - 对比 skip-gram
- 实验
  - Word analogies (A-B=C-?)
  - Word Similarity
  - Named entity recognition

词向量模型

vector models

global matrix factorization

- 可以利用预料中出现频率信息，但是处理 A-B=C-? 问题效果不好

local context window

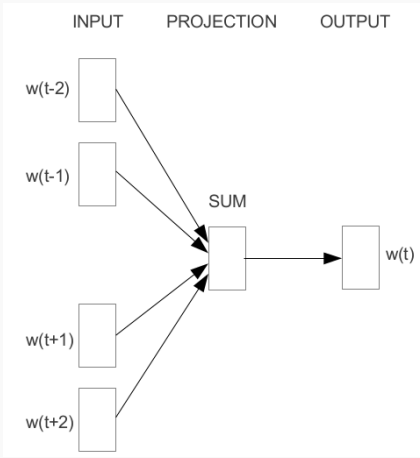
- 处理 A-B=C-? 效果很好，但是利用局部单词的左右窗口内词作为上下文信息，不能利用全局层面的共现信息

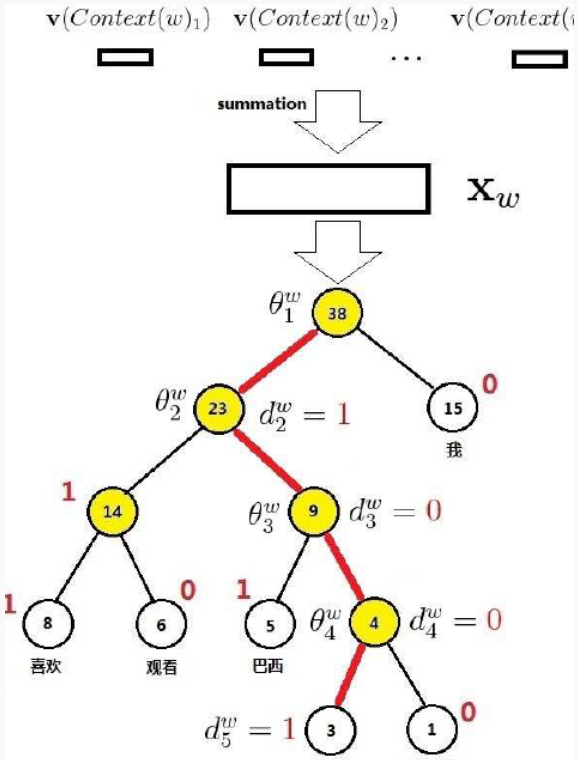
GloVe

- （综合）建立满足 A-B=C-D 的模型，利用共现信息（共现频率）作为权值标准，最优化模型误差

word2vec 的实现方式

word2vec 实现：CBOW





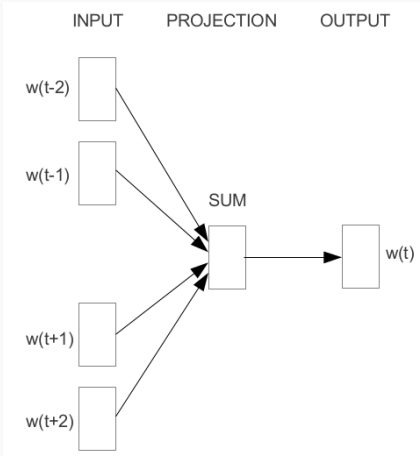
- 前后窗口（宽度 1-4 随机）内所有单词的词向量平均作为隐层词向量
- 所有单词按照词频构建 huffman 树，每个中间结点对应一个判别向量，和隐层词向量、目标词向量维数相同
- 每个中间结点的词向量和隐层词向量的点积做 sigmod 变换后的值，作为选择左分支（1）的概率，否则为走右支（0）的概率
- 每个单词是对应一个 huffman 编码

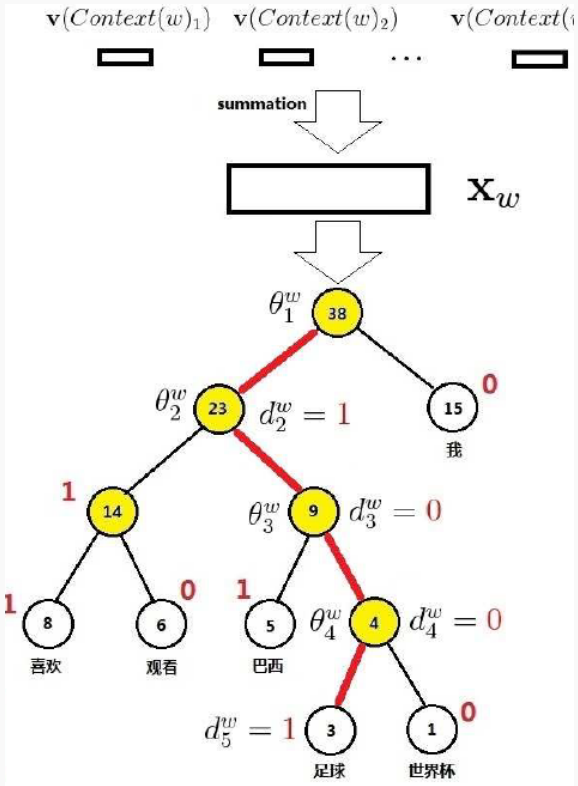
  足球 *fb*（*1001*），观看（*110*）

$$p(fb|context(fb)) = \prod_{j=2..5} p(d_j^w|X_w, \theta_{j-1}^w)$$

$$p(d_j^w|X_w, \theta_{j-1}^w) = \begin{cases} \sigma(x_w^T \theta_{j-1}^w) & d_j^w = 0 \\ 1 - \sigma(x_w^T \theta_{j-1}^w) & d_j^w = 1 \end{cases} \quad (1)$$

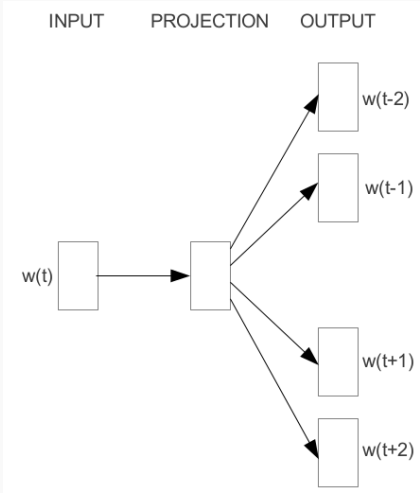
word2vec 实现：CBOW

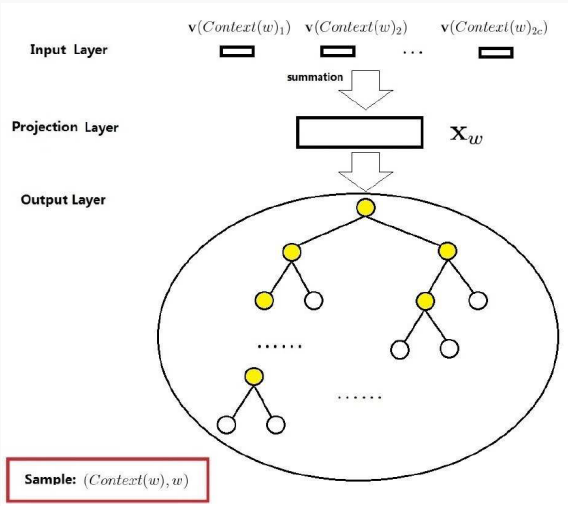




- 前后窗口（宽度 1-4 随机）内所有单词的词向量平均作为隐层词向量
- 所有单词按照词频构建 huffman 树，每个中间结点对应一个判别向量，和隐层词向量、目标词向量维数相同
- 每个中间结点的词向量和隐层词向量的点积做 sigmod 变换后的值，作为选择左分支（1）的概率，否则为走右支（0）的概率
- 每个单词是对应一个 huffman 编码
- 最优化 $\sum_{w \in C} \log p(c|context(c))$

word2vec 实现:Skip-gram





- 类似 CBOW
- 隐层就是输入的当前单词向量
- 不同的是每个隐层向量需要生成多个单词（窗口内的所有单词都预测，不分顺序）
  - 多一个 $\prod$ 循环
- 最优化 $\sum_{w \in C} \prod_{w_j \in window_c} \log p(w_j|c)$

GloVe Model

GloVe Model

共现概率

Probability and Ratio	$k = solid$	$k = gas$	$k = water$	$k = fashion$
$P(k\|ice)$	1.9×10^{-4}	6.6×10^{-5}	3.0×10^{-3}	1.7×10^{-5}
$P(k\|steam)$	2.2×10^{-5}	7.8×10^{-4}	2.2×10^{-3}	1.8×10^{-5}
$P(k\|ice)/P(k\|steam)$	8.9	8.5×10^{-2}	1.36	0.96

GloVe Model

- X_{ij} : $word_j$ 作为 $word_i$ 上下文出现次数
- $X_i = \sum_k X_{ik}$
- $P_{ij} = P(j|i) = X_{ij}/X_i$
- $F(w_i, w_j, \widetilde{w}_k) = P_{ik}/P_{jk}$

1. 改用向量差别 $F(w_i - w_j, \tilde{w}_k) = \frac{P_{ik}}{P_{jk}}$
2. 保持线性结构 $F\left((w_i - w_j)^T \tilde{w}_k\right) = \frac{P_{ik}}{P_{jk}}$
3. 我们采用 $F = exp$ 或者 $w_i^T \tilde{w}_k = \log(P_{ik}) = \log(X_{ik}) - \log(X_i)$，使得下面的式子成立

   $F(w_i^T \tilde{w}_k) = P_{ik} = \frac{X_{ik}}{X_i}$ $F\left((w_i - w_j)^T \tilde{w}_k\right) = \frac{F(w_i^T \tilde{w}_k)}{F(w_j^T \tilde{w}_k)}$
4. 如果是采用 $w_i^T \tilde{w}_k = \log(P_{ik}) = \log(X_{ik}) - \log(X_i)$，通过增加偏置项使得模拟 $w_i^T \widetilde{w}_k$ 与 X_{ij} 之间的关系：

   $w_i^T \tilde{w}_k + b_i + \tilde{b}_k = \log(X_{ik})$

GloVe Model

- X_{ij} : $word_j$ 作为 $word_i$ 上下文出现次数
- $X_i = \sum_k X_{ik}$
- $P_{ij} = P(j|i) = X_{ij}/X_i$
- $F(w_i, w_j, \widetilde{w}_k) = P_{ik}/P_{jk}$

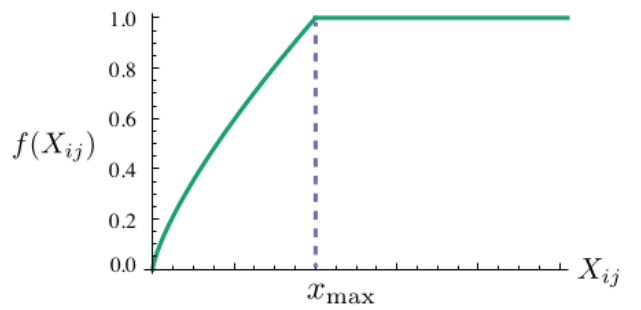
1. 改用向量差别 $F(w_i - w_j, \tilde{w}_k) = \frac{P_{ik}}{P_{jk}}$
2. 保持线性结构 $F\left((w_i - w_j)^T \tilde{w}_k\right) = \frac{P_{ik}}{P_{jk}}$
3. $F(w_i^T \tilde{w}_k) = P_{ik} = \frac{X_{ik}}{X_i}$ $F\left((w_i - w_j)^T \tilde{w}_k\right) = \frac{F(w_i^T \tilde{w}_k)}{F(w_j^T \tilde{w}_k)}$
4. $w_i^T \tilde{w}_k + b_i + \tilde{b}_k = \log(X_{ik})$
5. 目标函数 $J = \sum_{i,j=1}^{V} f\left(X_{ij}\right)\left(w_i^T \tilde{w}_j + b_i + \tilde{b}_j - \log X_{ij}\right)^2$

GloVe Model

- X_{ij} : $word_j$ 作为 $word_i$ 上下文出现次数
- $X_i = \sum_k X_{ik}$
- $P_{ij} = P(j|i) = X_{ij}/X_i$
- $F(w_i, w_j, \widetilde{w}_k) = P_{ik}/P_{jk}$

1. 改用向量差别 $F(w_i - w_j, \tilde{w}_k) = \frac{P_{ik}}{P_{jk}}$
2. 保持线性结构 $F\left((w_i - w_j)^T \tilde{w}_k\right) = \frac{P_{ik}}{P_{jk}}$
3. $F(w_i^T \tilde{w}_k) = P_{ik} = \frac{X_{ik}}{X_i}$ $F\left((w_i - w_j)^T \tilde{w}_k\right) = \frac{F(w_i^T \tilde{w}_k)}{F(w_j^T \tilde{w}_k)}$
4. $w_i^T \tilde{w}_k + b_i + \tilde{b}_k = \log(X_{ik})$
5. 目标函数 $J = \sum_{i,j=1}^{V} f\left(X_{ij}\right)\left(w_i^T \tilde{w}_j + b_i + \tilde{b}_j - \log X_{ij}\right)^2$

$$f(x) = \begin{cases} (x/x_{\max})^\alpha & \text{if } x < x_{\max} \\ 1 & \text{otherwise} \end{cases}$$



GloVe Model

GloVe 的目标函数 $J = \sum_{i,j=1}^{V} f\left(X_{ij}\right)\left(w_i^T \tilde{w}_j + b_i + \tilde{b}_j - \log X_{ij}\right)^2$ 对比 window-based methods (skip-gram and ivLBL) 的目标函数：

- 当前词 $word_i$ 预测其上下文单词 $word_j$: $Q_{ij} = \frac{\exp(w_i^T \tilde{w}_j)}{\sum_{k=1}^{V} \exp(w_i^T \tilde{w}_k)}$
- 目标函数为 $J = -\sum_{\substack{i \in \text{corpus} \\ j \in \text{context}(i)}} \log Q_{ij}$
- 合并目标函数中的相同项: $J = -\sum_{i=1}^{V}\sum_{j=1}^{V} X_{ij} \log Q_{ij}$
- 根据 $X_i = \sum_k X_{ik}$， $P_{ij} = P(j|i) = X_{ij}/X_i$，转化目标函数 $J = -\sum_{i=1}^{V} X_i \sum_{j=1}^{V} P_{ij} \log Q_{ij} = \sum_{i=1}^{V} X_i H(P_i, Q_i)$

GloVe Model

GloVe 的目标函数 $J = \sum_{i,j=1}^{V} f\left(X_{ij}\right)\left(w_i^T \tilde{w}_j + b_i + \tilde{b}_j - \log X_{ij}\right)^2$ 对比 window-based methods (skip-gram and ivLBL) 的目标函数：

- 转化目标函数 $J = -\sum_{i=1}^{V} X_i \sum_{j=1}^{V} P_{ij} \log Q_{ij} = \sum_{i=1}^{V} X_i H(P_i, Q_i)$
- 归一化的 P_{ij}, Q_{ij} 计算复杂，不使用交叉熵，改用压缩过的最小均方误差衡量 P, Q 差别大小 $\hat{J} = \sum_{i,j} X_i (\hat{P}_{ij} - \hat{Q}_{ij})^2 = \sum_{i,j} X_i (w_i^T \tilde{w}_j - \log X_{ij})^2$.
- Mikolov 提出限制权重来消除高频词的过度影响 $\hat{J} = \sum_{i,j} f(X_{ij})(w_i^T \tilde{w}_j - \log X_{ij})^2$
  - GloVe 的目标函数是合理的 (?)

Experiments

Experiments

Word analogies

Model	Dim.	Size	Sem.	Syn.	Tot.
ivLBL	100	1.5B	55.9	50.1	53.2
HPCA	100	1.6B	4.2	16.4	10.8
GloVe	100	1.6B	67.5	54.3	60.3
SG	300	1B	61	61	61
CBOW	300	1.6B	16.1	52.6	36.1
vLBL	300	1.5B	54.2	64.8	60.0
ivLBL	300	1.5B	65.2	63.0	64.0
GloVe	300	1.6B	80.8	61.5	70.3
SVD	300	6B	6.3	8.1	7.3
SVD-S	300	6B	36.7	46.6	42.1
SVD-L	300	6B	56.6	63.0	60.1
CBOW†	300	6B	63.6	67.4	65.7
SG†	300	6B	73.0	66.0	69.1
GloVe	300	6B	77.4	67.0	71.7
CBOW	1000	6B	57.3	68.9	63.7
SG	1000	6B	66.1	65.1	65.6
SVD-L	300	42B	38.4	58.2	49.2
GloVe	300	42B	**81.9**	**69.3**	**75.0**

- 19544 组问题 a *is to* b *as* c *is to* ___?
  - SVD-S: 矩阵元素值压缩 $\sqrt{X_{trunc}}$
  - SVD-L: 压缩 $\log(1 + X_{trunc})$
  - SVD-$f(X)$?
- CBOW 效果提升不明显

Experiments

Word Similarity

Model	Size	WS353	MC	RG	SCWS	RW
SVD	6B	35.3	35.1	42.5	38.3	25.6
SVD-S	6B	56.5	71.5	71.0	53.6	34.7
SVD-L	6B	65.7	72.7	75.1	56.5	37.0
CBOW†	6B	57.2	65.6	68.2	57.0	32.5
SG†	6B	62.8	65.2	69.7	58.1	37.2
GloVe	6B	65.8	72.7	77.8	53.9	38.1
SVD-L	42B	74.0	76.4	74.1	58.3	39.9
GloVe	42B	**75.9**	**83.6**	**82.9**	**59.6**	**47.8**
CBOW*	100B	68.4	79.6	75.4	59.4	45.5

Experiments

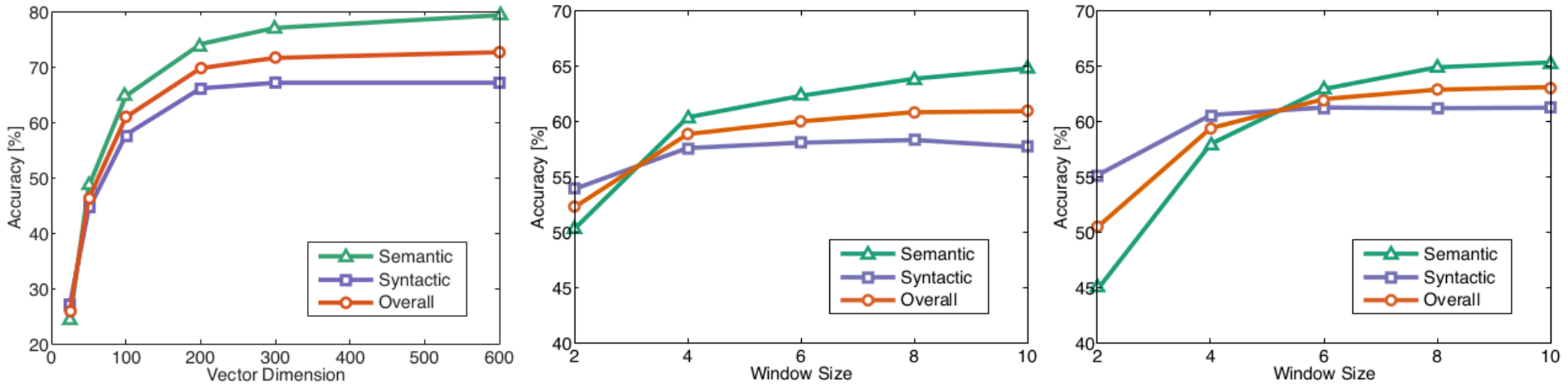
NER

Model	Dev	Test	ACE	MUC7
Discrete	91.0	85.4	77.4	73.4
SVD	90.8	85.7	77.3	73.7
SVD-S	91.0	85.5	77.6	74.3
SVD-L	90.5	84.8	73.6	71.5
HPCA	92.6	**88.7**	81.7	80.7
HSMN	90.5	85.7	78.7	74.7
CW	92.2	87.4	81.7	80.2
CBOW	93.1	88.2	82.2	81.1
GloVe	**93.2**	88.3	**82.9**	**82.2**

- Discrete: 直接使用 Stanford NER 的输出结果
- 50 维，窗口大小为 5

Experiments

vector length and window size



(a) Symmetric context (b) Symmetric context (c) Asymmetric context

- 200 维就够了
- symmetric：左右窗口，asymmetric：只有左边窗口
- 句法信息从左到右传递
- 语义信息随着窗口大小变化明显，说明是非局部的信息

Experiments

Corpus Size

- Wikipedia 好于 Gigaword，虽然规模小
  - Wikipedia 实体多
  - gigaword 新闻语料可能有错误信息

Run-time

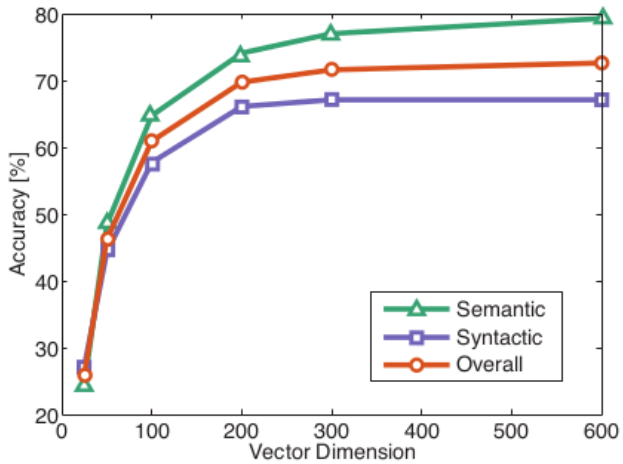
a dual 2.1GHz Intel Xeon E5-2658

- populateing X (single thread): 85min: winSize:10, vocabulary:400,000, token:6 billion
- train Model (32 cores): 14min: 300-dim

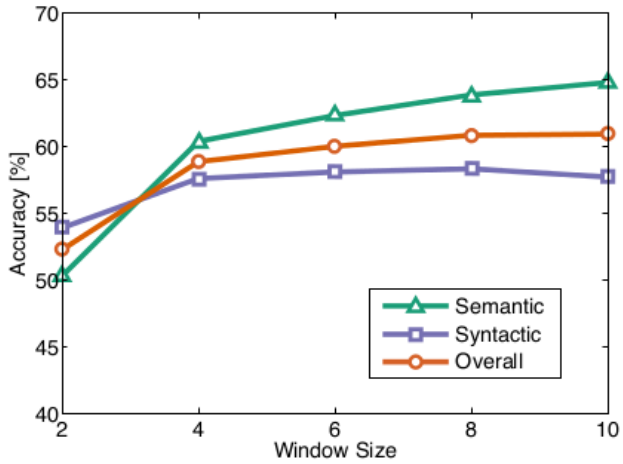
VS word2vec

vsWord2vec

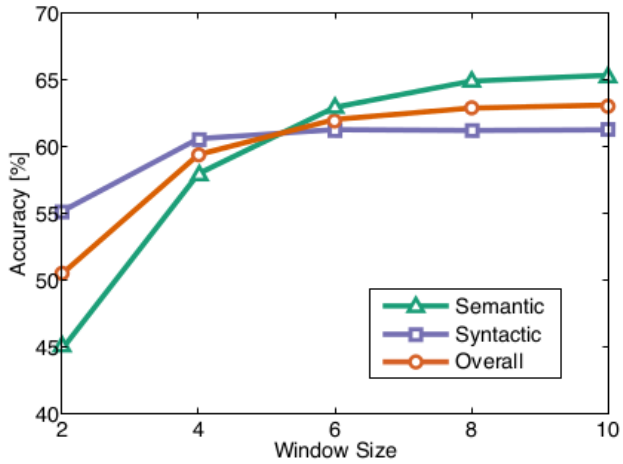
- word2vec 顺序遍历数据，不支持多进程？
  - the code is currently designed for only a single epoch



(a) Symmetric context



(b) Symmetric context



(c) Asymmetric context

- 负采样个数不超过 10 个
- GloVe 收敛更快