



# 中文词汇语义相似度计算研究

---

答辩人：裴家欢  
指导老师：黄德根 教授



# 主要内容

## Main Contents

1 研究背景

2 研究现状

3 研究目标

4 研究过程

5 研究结论

6 发表论文



01

研究背景



# 研究方向

研究背景

研究现状

研究目标

研究过程

研究结论

发表论文

中文词汇语义相似度计算

**Chinese Word Semantic Similarity Computation**

旨在将“词汇相似”这个抽象关系通过特定的计算方法映射成能够表达两个词语对象中蕴含的意义相近程度、计算机可以处理的数值。



# 基本概念

研究背景

- **语义**

自然语言的意义或含义，代表在现实世界中所对应事物的概念

研究现状

- **相似性→相似度**

人类对两个对象之间关系刺激所产生的、定性比较所产生的感知和比较过程。比如，人们面对父与子的外貌信息会产生“很相似”的生理反应。

研究目标

- **词汇语义相似度**

—可以通过两个对象之间在语义内容层面上的距离衡量

—可以通过支持其含义或描述其性质的信息比较获得的数值描述

—可以定量地估计词汇之间语义关系的强度

研究过程

- **相似与相关**

• “公交车” is a “汽车” ——相似

• “汽车”，“道路” ——相关

• “积极”，“消极” ——相关（反义词）

研究结论

发表论文





# 背景意义

研究背景

研究现状

研究目标

研究过程

研究结论

发表论文



## 信息过载

海量信息不断生成，信息过载问题给人们带来的影响日益显著。

## 语义理解

自动理解自然语言中所隐含的语义，迅速而准确地获取有效信息。

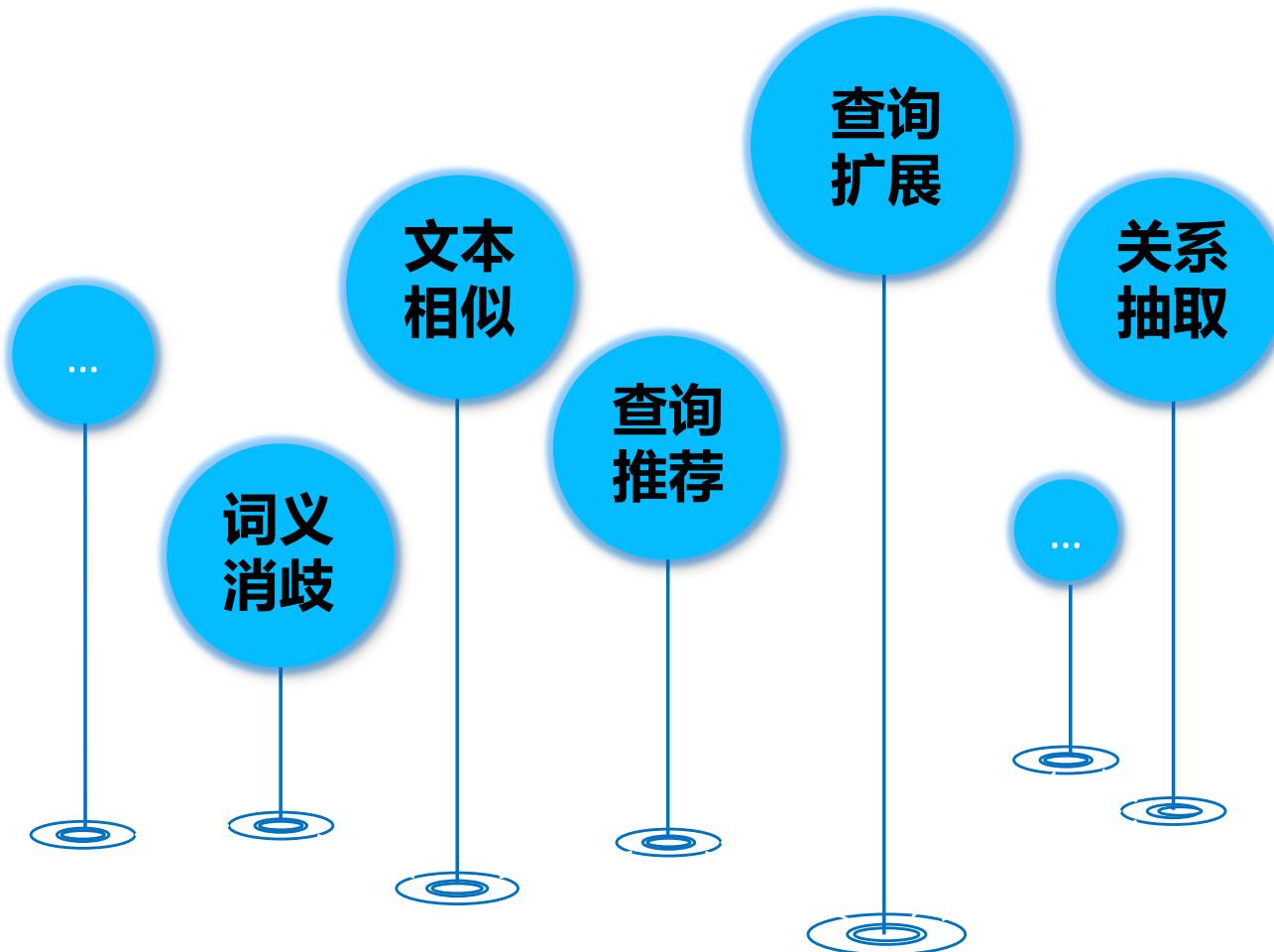
## 词汇语义

词汇是可以表示完整语义的基本单元，是AI利用计算机模仿人类对相似性认知过程的核心步骤。



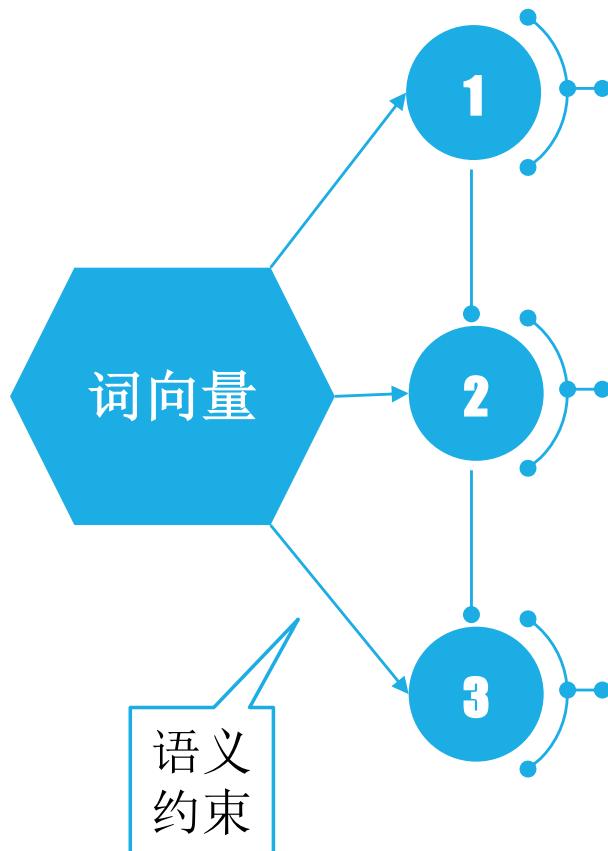
# 背景意义

- 研究背景
- 研究现状
- 研究目标
- 研究过程
- 研究结论
- 发表论文



# 创新之处

- 研究背景**
- 研究现状
- 研究目标
- 研究过程
- 研究结论
- 发表论文



## 机器翻译改进词向量

利用机器翻译构建中英文词向量的关系，并用英文词向量选择性地替换中文词向量，改进性能

## LSTMs改进词向量

学习词对共现句子提升词向量模型的性能

## 融入语义约束的词向量模型

将同义约束、反义约束、相似约束和向量空间存留用于优化后处理。



02

研究现状

# 常用方法

- 研究背景
- 研究现状
- 研究目标
- 研究过程
- 研究结论
- 发表论文

## 语义词典

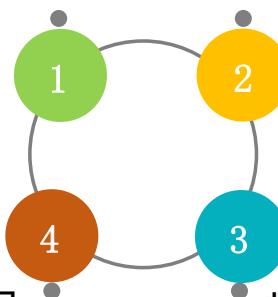
根据语义词典结构特点及词汇之间语义关系，计算词对中词汇之间的语义距离，从而定义相似度计算公式。

通过考虑多种计算方法的结果，或同时利用多种资源得到组合模型来完成词汇语义相似度计算。

## 组合策略

## 语料库传统统计方法

根据语料库中的统计信息定义相似度，常见方法有相关熵、互信息、LSA、LDA、PMI-IR等。



基于上下文与目标词之间关系建模得到的一种分布式表示、含有语义信息的实数型数值向量，然后根据向量距离定义相似度。

## 语料库词向量

# 方法缺陷

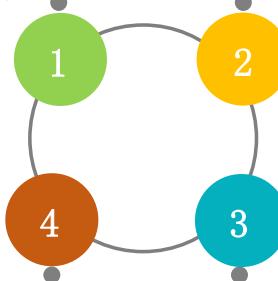
- 研究背景
- 研究现状
- 研究目标
- 研究过程
- 研究结论
- 发表论文

## 语义词典

- ① 依赖人工构建，未登录词计算失效。
- ② 依赖于语义关系较单一
- ③ 词性不同或歧义则计算失效。

- ① 同类别方法组合难以避免一类方法存在的通用问题
- ② 词向量融入语义知识仍在探索阶段。

## 组合策略



## 语料库传统统计方法

- ① 离线语料库，未登录词失效
- ② Web在线语料库，检索返回文档存在噪声和冗余
- ③ 只考虑统计信息，忽略语义关系

- ① 难以区分语义相似与概念相关。
- ② 无法捕获近义词与反义词的不同。
- ③ 无法解决一词多义的问题。

## 语料库词向量



03

研究目标



# 研究目标

- 研究背景
- 研究现状
- 研究目标
- 研究过程
- 研究结论
- 发表论文



在不融入语义约束的前提下，提升词向量模型。



将语义约束融入词向量模型，提升整体计算性能。



# 研究方法

- 研究背景
- 研究现状
- 研究目标
- 研究过程
- 研究结论
- 发表论文

## 无语义约束的词向量模型

- 利用机器翻译构建中、英文词向量关系改进原有中文词向量计算性能。
- 通过LSTMs网络学习词对共现的句子序列，建模词汇关系。

01

## 融入语义约束的词向量模型

- 改进Counter-fitting模型将同义、反义、相似和向量空间存留等语义约束，以后处理优化方式融入已有词向量模型。

02



04

研究过程



# 任务描述

研究背景

研究现状

研究目标

研究过程

研究结论

发表论文

- 任务

给定任意一对词汇，要求设计一种计算方法给出它们语义相似程度的实数值评分（1-10分），如果两个词汇的意义或者语义内容之间的距离越小，则语义相似程度越高、评分越高，反之则分数越低。

- 例子

词对 ( $w_1, w_2$ )	相似度
(没戏, 没辙)	4.9
(只管, 尽管)	4
(GDP, 生产力)	6.5



# 任务描述



## 评价方法

— Spearman

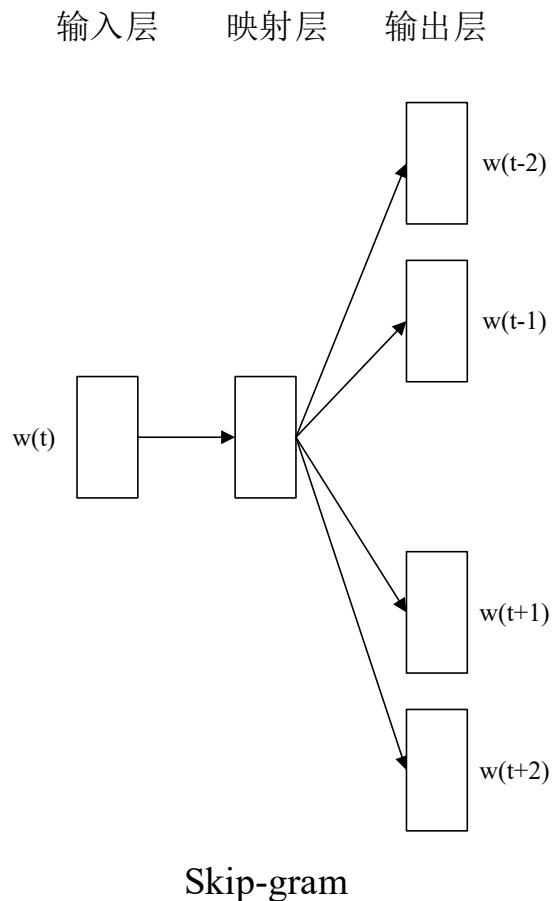
$$\rho = 1 - \frac{6 \sum_{i=1}^n (R_{X_i} - R_{Y_i})^2}{n(n^2 - 1)}$$

— Pearson

$$r = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^n (X_i - \bar{X})^2} \sqrt{\sum_{i=1}^n (Y_i - \bar{Y})^2}}$$

- $X_i$  和  $Y_i$  为自动评分和人工评分的原始数据。
- $\bar{X}$  和  $\bar{Y}$  是两组数据序列中的平均数。
- $R_{X_i}$  和  $R_{Y_i}$  是标准差等级变量。
- n 是序列中样本的数量。

# 标准Skip-gram词向量模型



1. 199M的数据堂新闻语料 (Datatang)
2. 1.1G的维基百科语料 (Wiki)
3. 261M的微博语料 (Weibo)
4. 词作为Query检索百度新闻语料 (News)
5. “写搜”造句网爬取的语料 (Xieso)

$$p(context(w) \mid w) = \prod_{j=1}^k p(d_j^u \mid v(w), \theta_{j-1}^w)$$



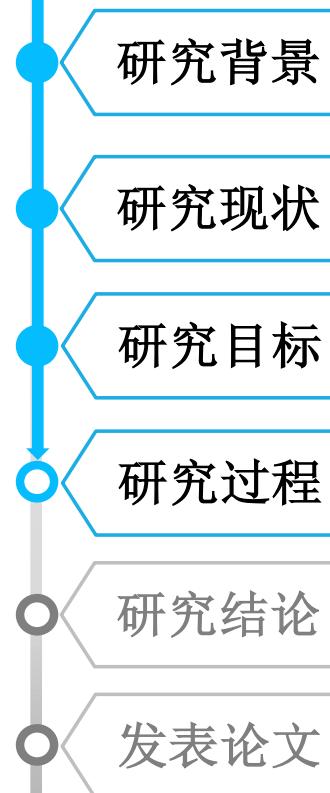
# 标准Skip-gram词向量模型

- 研究背景
- 研究现状
- 研究目标
- 研究过程
- 研究结论
- 发表论文

No.	语料库组合	$\rho$	r
1	Xieso (62M)	0.205	0.203
2	Wiki(1.1G)	0.211	0.213
3	Weibo(261M)	0.227	0.077
4	Datatang(199M)	0.267	0.272
5	News (381M)	0.311	0.305
6	News+Xieso	0.311	0.311
7	News+Xieso+Weibo	0.133	0.123
8	News+Xieso+Wiki	0.178	0.197
9	News+Xieso+Datatang	0.174	0.190
10	News+Xieso+DataTang+Wiki	0.214	0.239
11	News+Xieso+DataTang+Wiki+Weibo	0.214	0.134



# 基于机器翻译的改进方法



- **动机:**

	公开词向量数量	训练难度	研究起步
中文	较少	较大, 需要分词	较晚
英文	较多	较小, 天然分隔	较早

- **一个基本假设:**

当两个词汇从中文到英文翻译完全正确的时候，使用公开的大规模预训练好的英文词向量会比中文词向量语义损失更小、语义相似度计算会更加准确。

- **两条严格约束:**

- (1) 英文译文经拼写检查后无拼写错误。
- (2) 一个中文词汇只翻译成一个英文单词。



# 基于机器翻译的改进方法



- 具体实现的步骤：

- (1) 调用Google Translation API，将每组中文词汇翻译成英文词汇。
- (2) 检查并标记两个译文长度均为1个单词的词对。
- (3) 利用单词拼写检错程序检测步骤(2)中标记的词对，将存在拼写错误的词对剔除。（本文使用了PyEnchant工具包提供的算法进行单词拼写检测。）
- (4) 将剩下的词对取出英文词向量模型中的向量进行相似度计算，其他词对仍然使用中文词向量计算。



# 基于机器翻译的改进方法

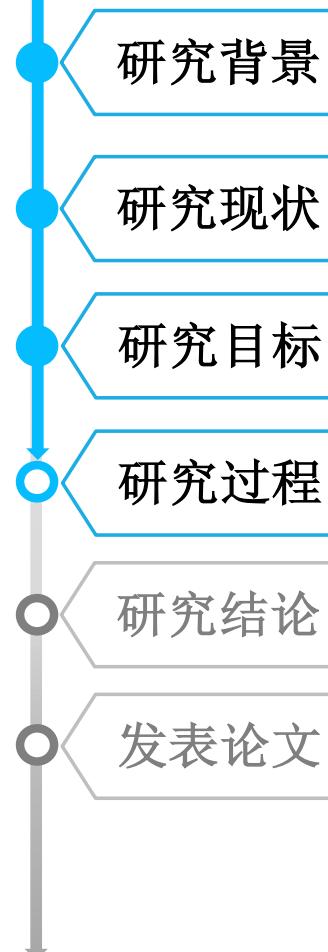


## • 实验结果:

No.	方法	$\rho$	r
1	Baseline	0.311	0.311
2	Baseline +Translation	0.397	0.443



# 基于 LSTMs 的改进方法



- **一个基本假设:**  
除了词语的上下文环境可以反映其语义外，词对中的两个词共现的上下文环境也能够反映它们的某种语义关系。
- **LSTMs 学习句子序列的模型:**  
输入是词对共现的句子序列和将相似度评分取整后的整数标签 $i=1, \dots, 10$ ；测试时，输入是词对共现的句子序列，输出是最终的预测词对的语义相似度评分。



# 基于LSTMs的改进方法



## 具体步骤：

(1) 爬取词对共现句子及预处理。

加拿大对焊接碳钢管违法课徵反倾销税，影响我国钢铁业者，台湾状告世界贸易组织(wto)，经济部今天表示，这不仅是争商机、也是争法理，判决预计年底前出炉，有信心胜诉。**CRLE**  
近日，日本、美国及欧盟以违反世界贸易组织(wto)协议为由针对我国限制稀土出口问题向wto提起诉讼。

- (2) 构造一个 $4n$ 维度的含有距离信息的词向量( $n=150$ )
- (3) LSTMs学习过程词对关系(句子分类)。



$$\begin{aligned} i_t &= \sigma(W_i x_t + U_i h_{t-1} + b_i) & C_t &= i_t * \tilde{C}_t + f_t * C_{t-1} \\ \tilde{C}_t &= \tanh(W_c x_t + U_c h_{t-1} + b_c) & o_t &= \sigma(W_o x_t + U_o h_{t-1} + b_o) \\ f_t &= \sigma(W_f x_t + U_f h_{t-1} + b_f) & h_t &= o_t * \tanh(C_t) \end{aligned}$$

# 基于LSTMs的改进方法

研究背景

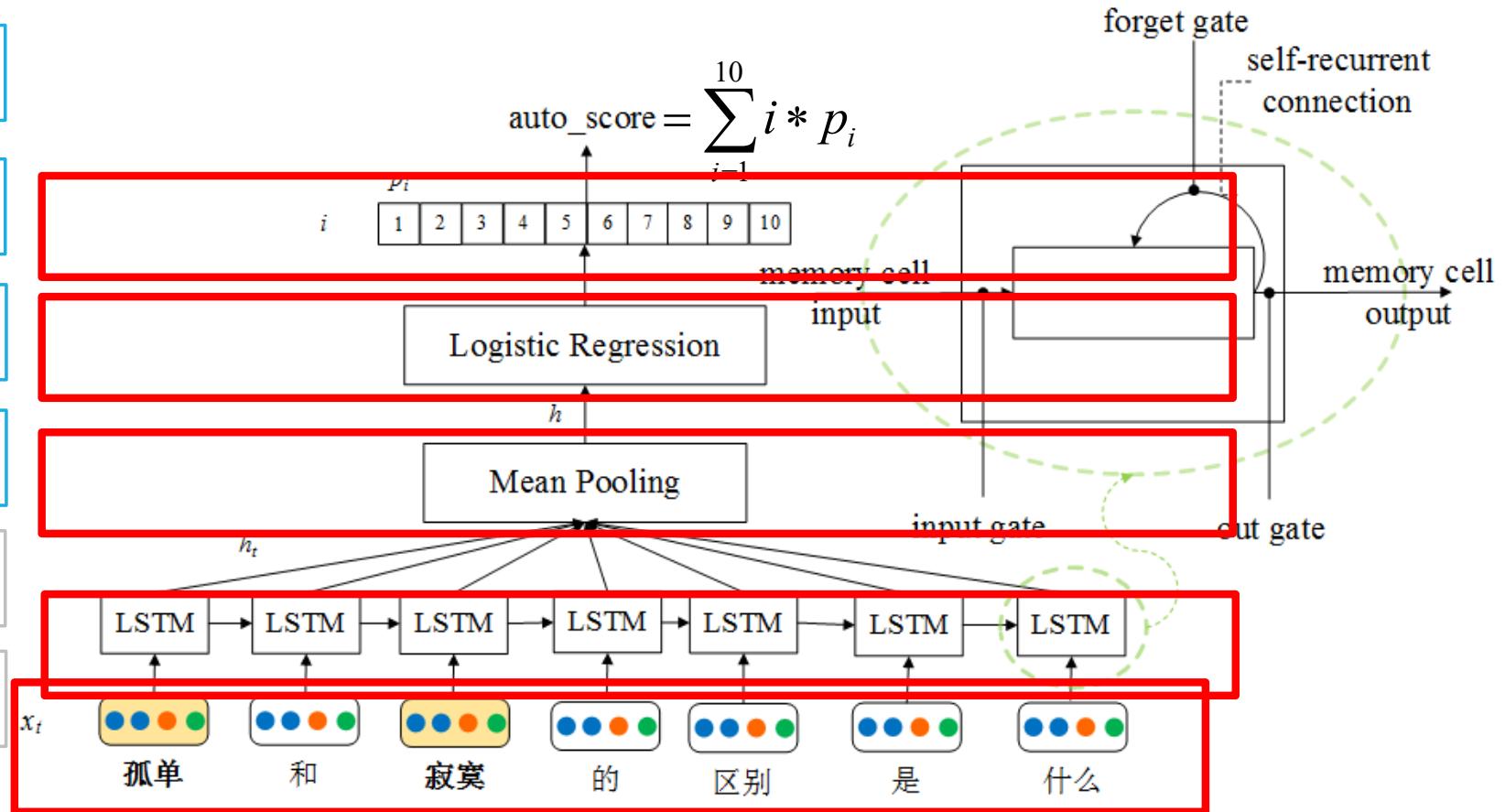
研究现状

研究目标

研究过程

研究结论

发表论文





# 基于 LSTMs 的改进方法



- 实验结果:

No.	方法	$\rho$	r
1	Baseline	0.311	0.311
2	Baseline+ LSTMs	0.351	0.364



# 融入语义约束的词向量模型



- **动机:**

由于词向量的训练是基于上下文环境来进行的，所以词向量模型很难学习到某些语言学上的约束条件，例如同义词或者反义词。而**counter-fitting**的主要想法是根据同义和反义关系对来对训练得到的原始词向量进行微调、优化，使其在词汇相似度计算任务中能够表现出更好的性能。

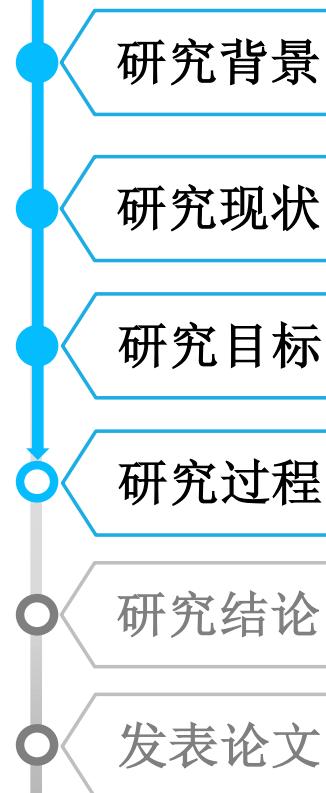
- **基本思想:**

词向量+语义知识→组合策略

已知一个原始词向量空间 $V$ , 找出一组语义约束条件 $C$ , 使得关于新词向量空间 $V'$ 和语义约束条件 $C$ 的目标函数经数次迭代达到最值，在此过程中，便可以得到一个融入了语义约束的新词向量空间 $V'$ 。



# 融入语义约束的词向量模型



- 模型构建: 目标函数

$$\tau(x) \triangleq \max(0, x)$$

$$C(V, V') = k_1 AR(V') + k_2 SA(V') + k_3 RSI(V') + k_4 VSP(V, V')$$

- (1) 反义排斥 (Antonym Repel, AR)

$$AR(V') = \sum_{(u,w) \in A} \tau(\delta - d(v'_u, v'_w))$$

- (2) 同义词吸引 (Synonym Attract, SA)

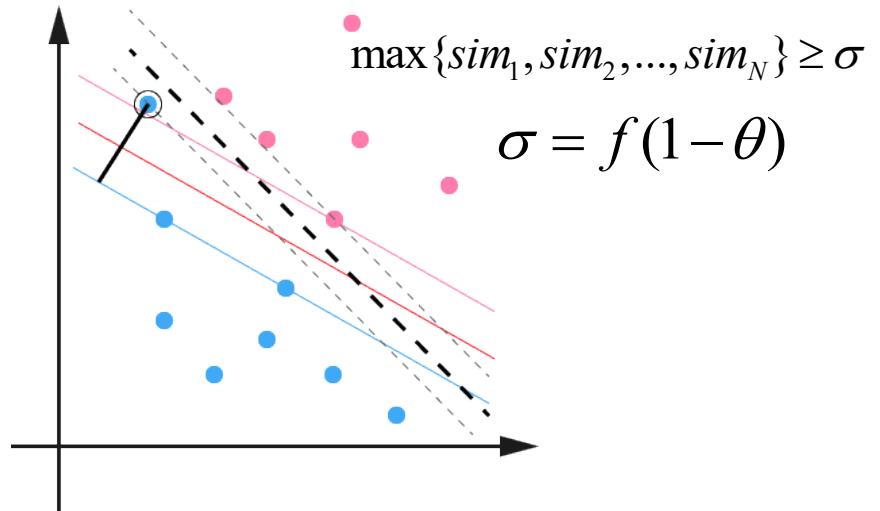
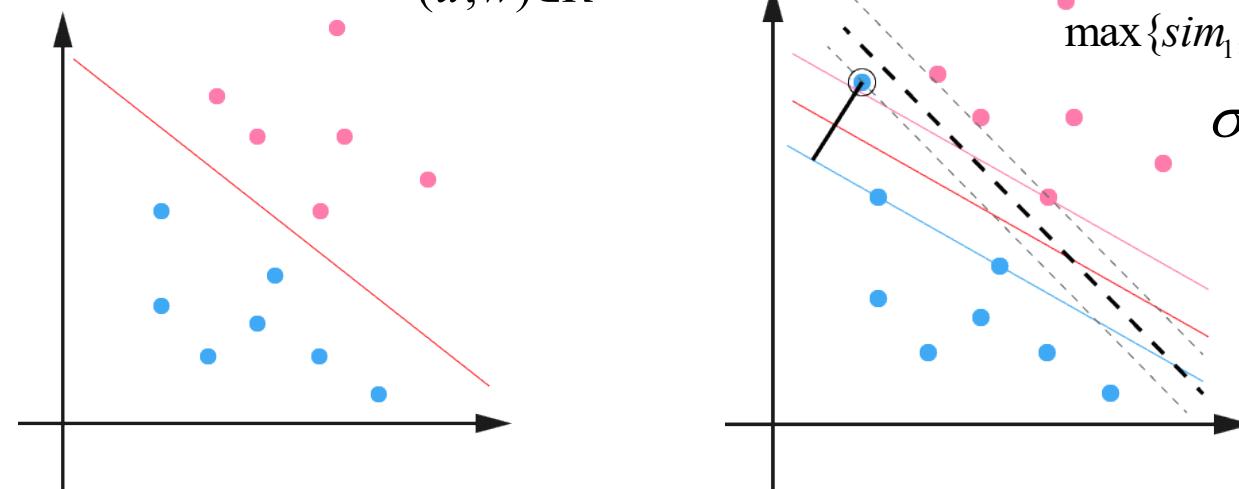
$$SA(V') = \sum_{(u,w) \in S} \tau(d(v'_u, v'_w) - \gamma)$$

# 融入语义约束的词向量模型

- 研究背景
- 研究现状
- 研究目标
- 研究过程
- 研究结论
- 发表论文

- 模型构建:
- (3) 剩余相似指数 (Residual Similarity Index, RSI)

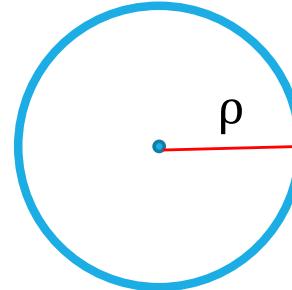
$$RSI(V') = \sum_{(u,w) \in R} \tau(d(v'_u, v'_w) - \theta)$$



基本SVM和引入松弛因子的软间隔SVM示意图



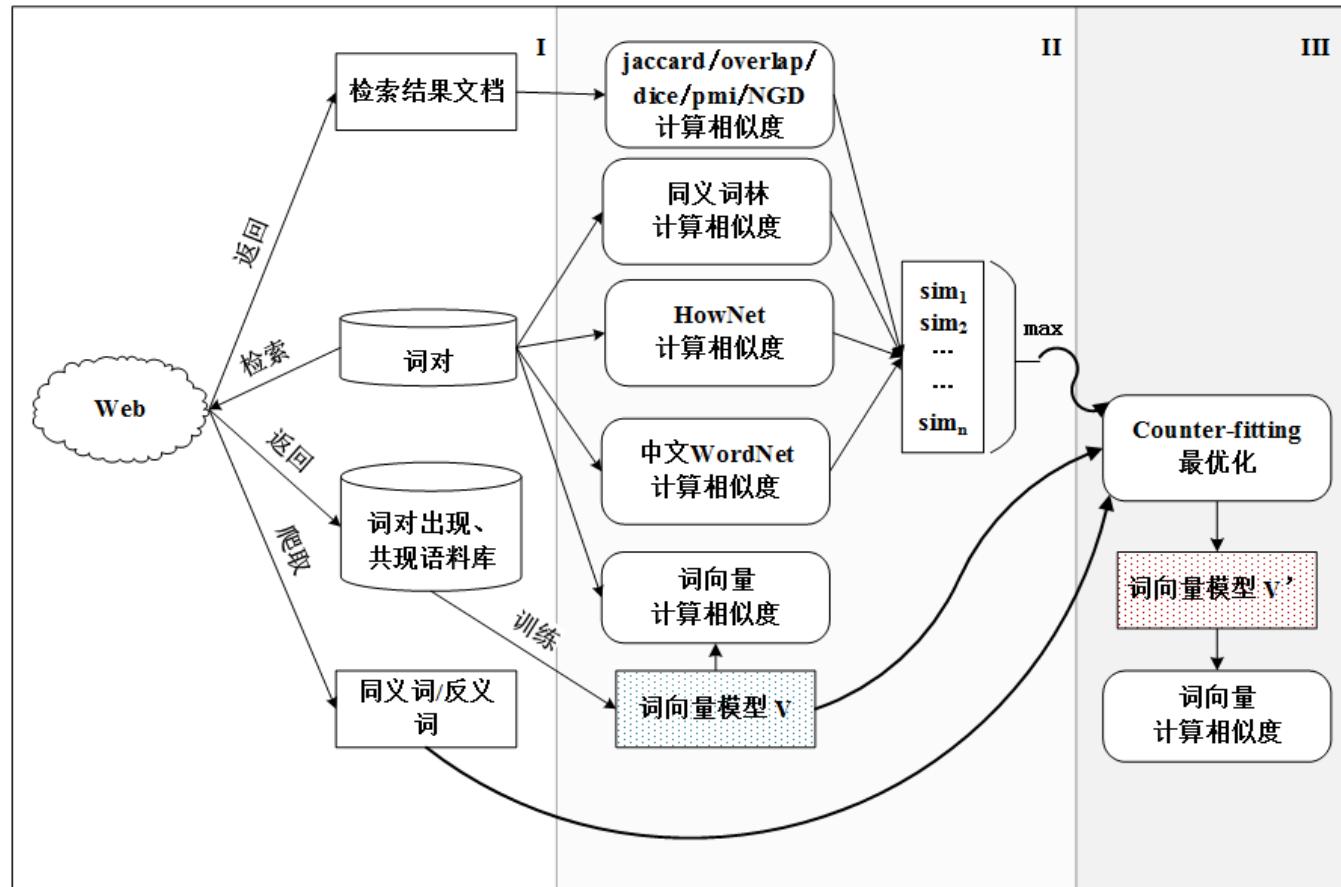
# 融入语义约束的词向量模型

- 研究背景
  - 研究现状
  - 研究目标
  - 研究过程
  - 研究结论
  - 发表论文
- 模型构建:
  - (4) 向量空间存留 (Vector Space Preservation, VSP)
$$VSP(V') = \sum_{i=1}^N \sum_{j \in N(i)} \tau(d(v_u, v_w) - d(v'_u, v'_w))$$
  - (5) 最小化目标函数

# 融入语义约束的词向量模型

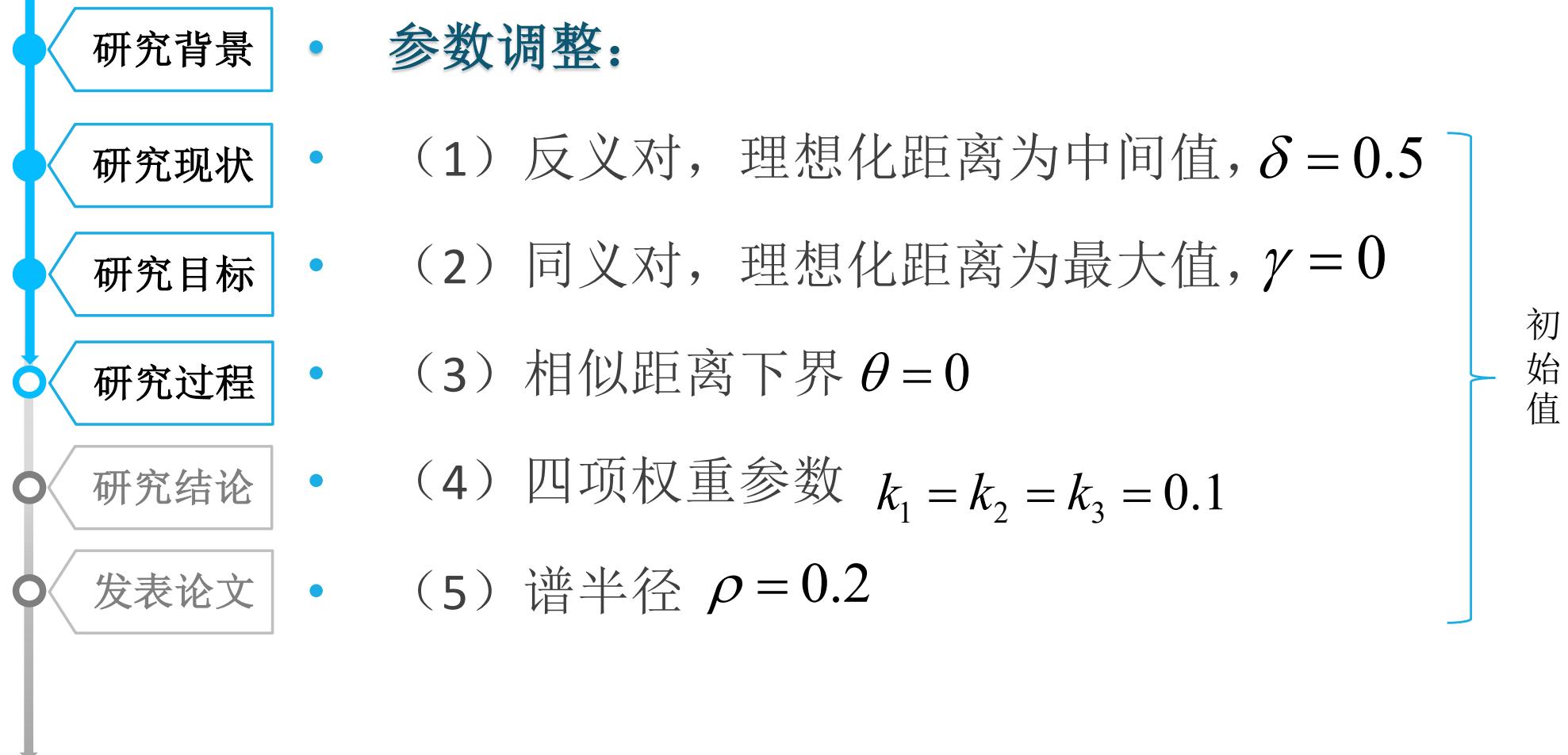
- 研究背景
- 研究现状
- 研究目标
- 研究过程
- 研究结论
- 发表论文

## • 具体实现流程:





# 融入语义约束的词向量模型





$\theta = 0$

$k_1 = k_2 = k_3 = 0.1$

$\rho = 0.2$

# 融入语义约束的词向量模型

研究背景

研究现状

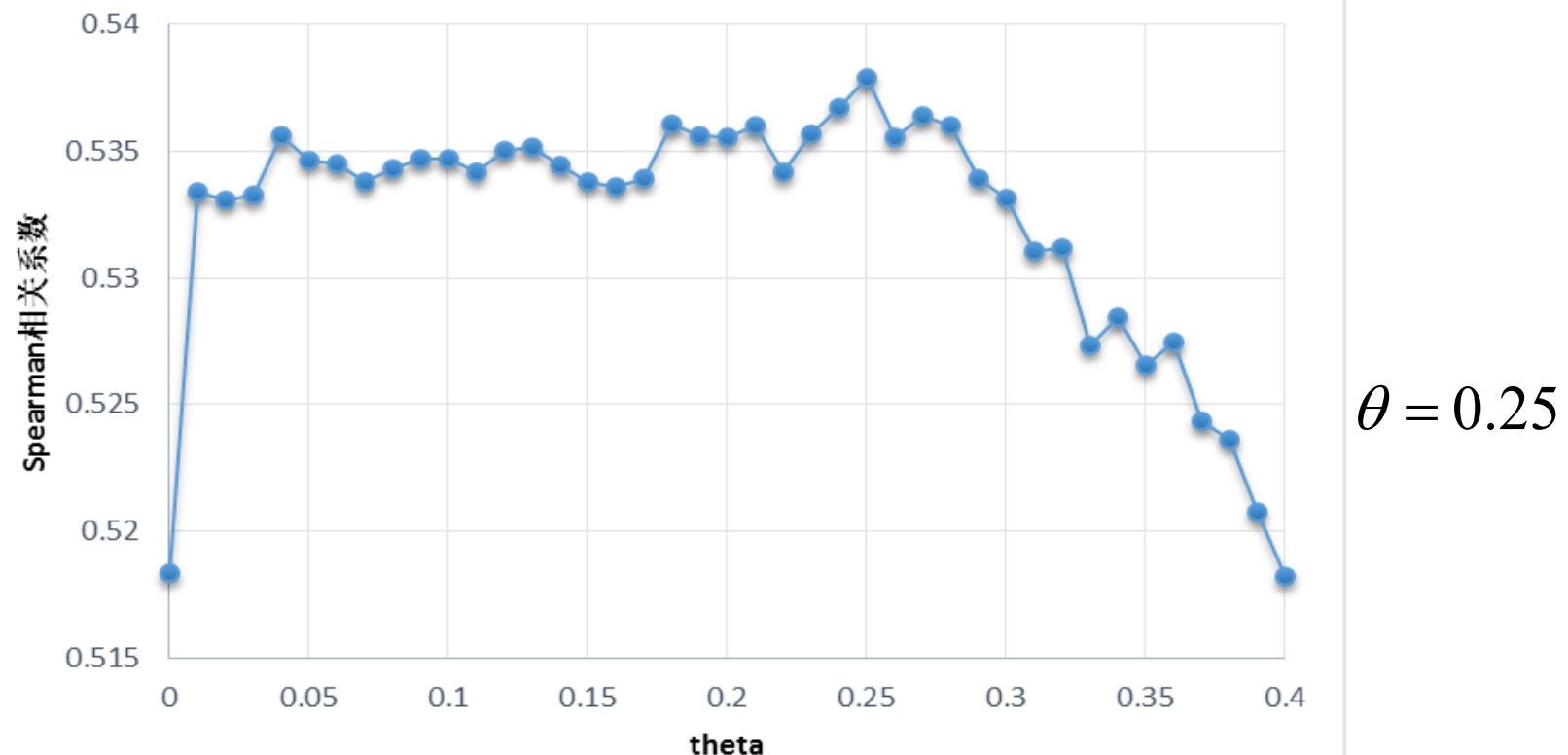
研究目标

研究过程

研究结论

发表论文

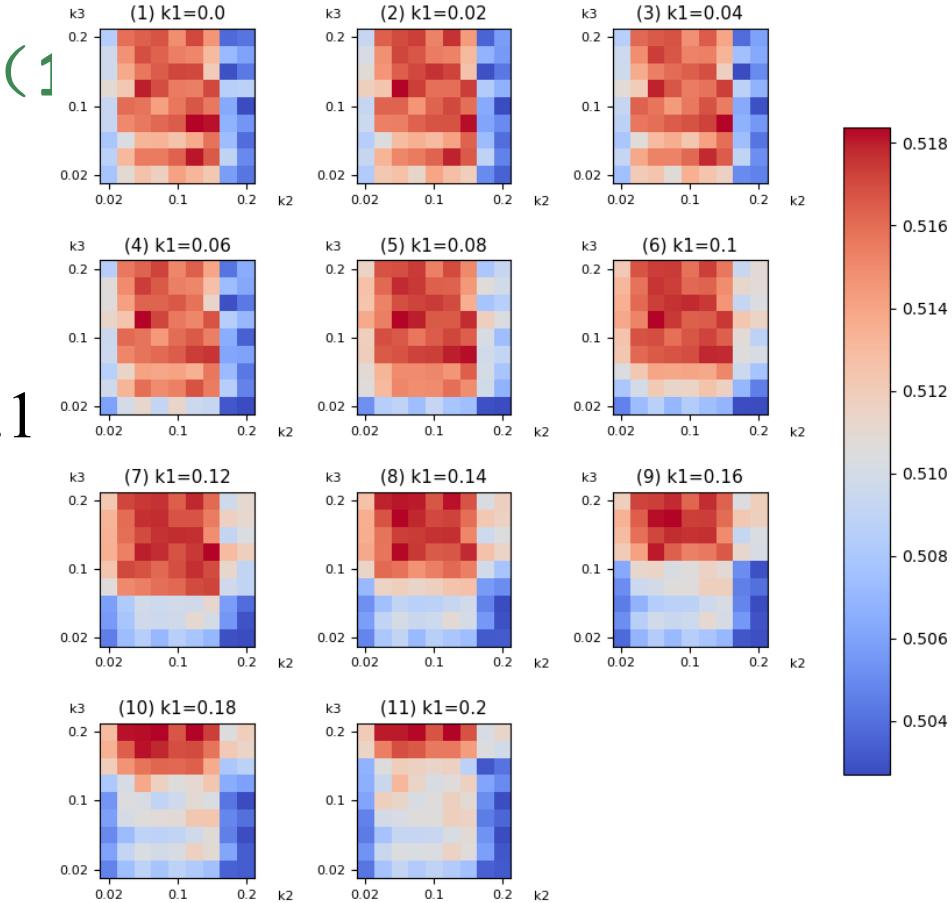
- 参数调整：（1）相似距离下界 theta



# 融入语义约束的词向量模型

- 研究背景
- 研究现状
- 研究目标
- 研究过程
- 研究结论
- 发表论文

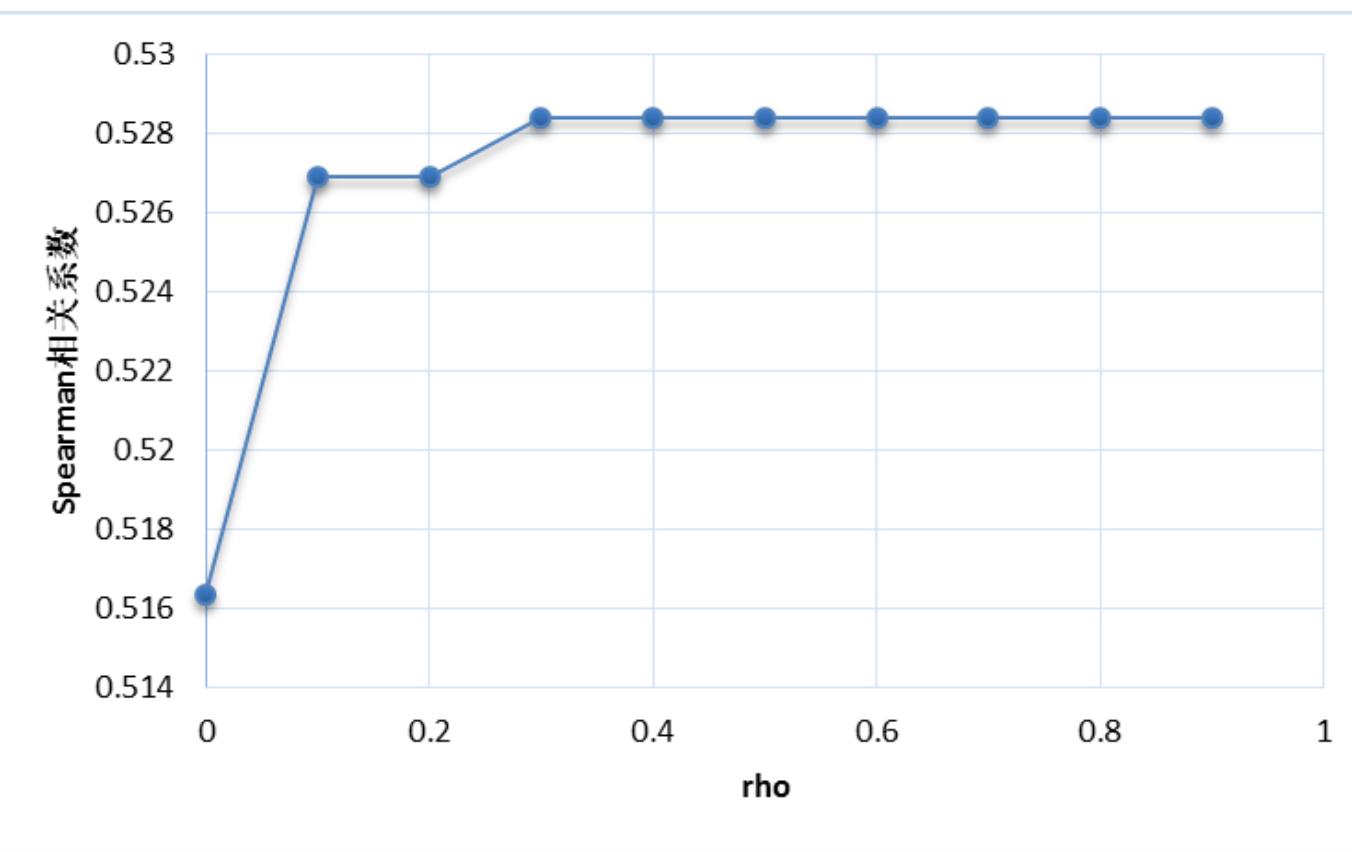
## • 参数调整:



# 融入语义约束的词向量模型

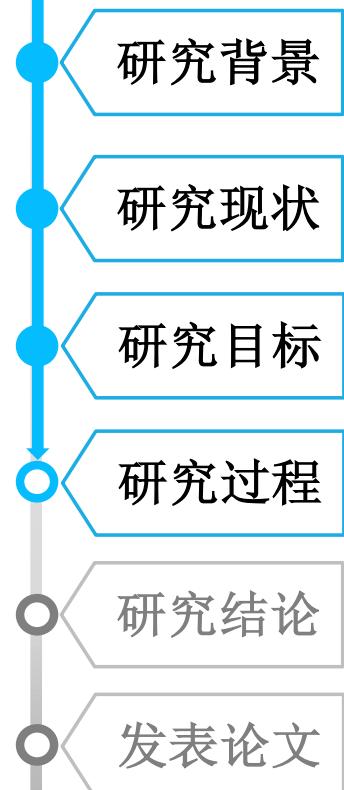


- 参数调整: (3) 谱半径rho





# 融入语义约束的词向量模型



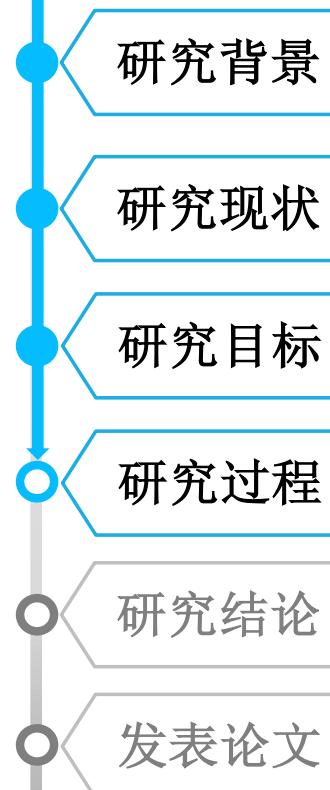
- 实验结果:

- 五组随机采样实验结果

Group	Spearman $\rho$	Pearson r
1	0.551	0.525
2	0.527	0.511
3	0.528	0.491
4	0.526	0.467
5	0.506	0.453
All	0.552	0.513



# 融入语义约束的词向量模型

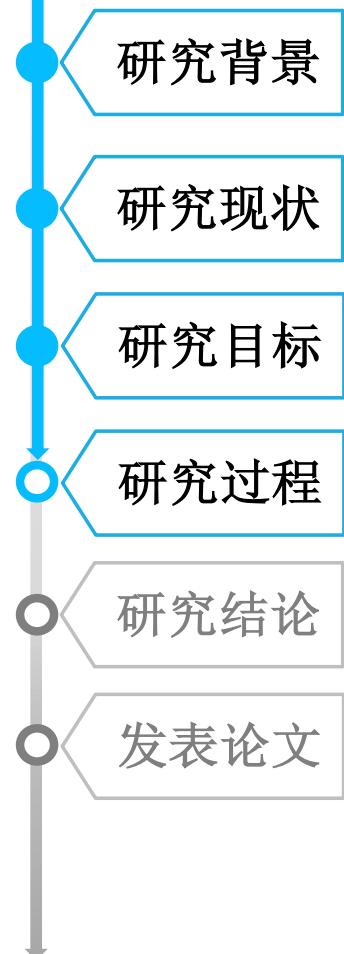


## • 实验结果:

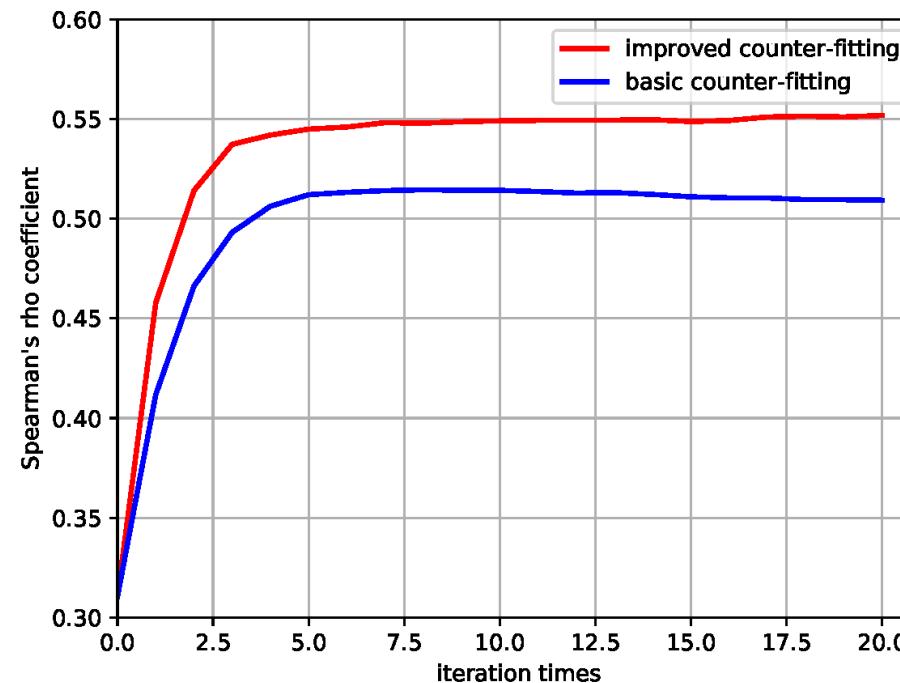
- 基线方法、基本及改进**counter-fitting**实验结果对比

No.	方法	$\rho$	r
1	Baseline: skip-gram	0.311	0.311
2	Baseline +Basic Counter-fitting	0.520	0.496
3	Baseline +Improved Counter-fitting	0.552	0.513

# 融入语义约束的词向量模型



- 实验结果:
  - 基本Counter-fitting和改进Counter-fitting的学习曲线





# 融入语义约束的词向量模型



## • 实验结果:

### • 不同类方法的实验结果对比

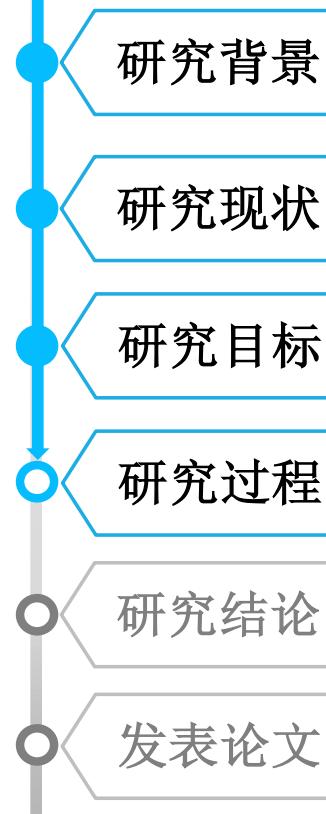
No	方法	方法类别	$\rho$	r
1	同义词林	语义词典模型	0.422	0.472
2	Skip-gram	词向量模型	0.311	0.311
3	Web-pmi	Web语料库模型	0.275	0.311
4	改进counter-fitting	组合模型 (语义+词向量)	<b>0.552</b>	<b>0.513</b>

- No.1-3→语义词典方法在登录词覆盖率较高时有优势。
- No.4 VS No.1-3→融入语义约束的词向量模型可以整合语义词典、统计规律及词向量等多元化资源，具有优势互补的特性。



- 语义词典方法在登录词覆盖率较高时有优势。
- 融入语义约束的词向量模型可以整合语义词典、统计规律及词向量等多元化资源，具有优势互补的特性。

# 融入语义约束的词向量模型



## 实验结果：

### 其它语料上的对比实验（MC-30）

No	方法	方法类别	$\rho$
1	同义词林	语义词典模型	0.520
2	HowNet	语义词典模型	0.682
3	Skip-gram	词向量模型	0.697
4	web-pmi	Web语料库模型	0.371
5	同义词林+ HowNet	组合模型 (词典A+词典B)	0.847
6	改进Counter-fitting	组合模型 (语义+词向量)	0.731

高登录率的小数据集上的天然优势，应用广泛性较差。



05

研究结论

# 研究结论



## (1) 无语义约束的词向量模型。

分别利用机器翻译选择性替换中文向量和LSTMs学习词对共现句子提升词向量模型的性能。实验结果证明，基于机器翻译和LSTMs的改进模型相对于标准词向量模型效果略有提高。

## (2) 融入语义约束的词向量模型。

考虑同义、反义、相似、向量空间存留，对已有词向量进行后处理优化，将强、弱语义关系融入到词向量的改进过程中。实验结果证明，融入语义约束可有效提升词向量模型的性能。

## (3) 方法适用性分析。

基于语义词典的方法在登录词覆盖率较高时有优势，否则，基于词向量方法和基于Web语料库统计的方法更具有实用性。



06

发表论文



# 发表论文

研究背景

研究现状

研究目标

研究过程

研究结论

发表论文

1. Pei, J.\* , Zhang, C., Huang, D. & Ma, J. (2016). Combining Word Embedding and Semantic Lexicon for Chinese Word Similarity Computation. In International Conference on Computer Processing of Oriental Languages (pp. 766-777). Springer International Publishing. (EI检索会议)
2. Pei, J.\* , Huang, D., Ma, J., et al. (2016). DUT-NLP-CH @ NTCIR-12 Temporalia Temporal Intent Disambiguation Subtask, In Proceedings of the 12th NTCIR Conference on Evaluation of Information Access Technologies, pp.253-257. Tokyo, Japan, Jun 7-10. (国际会议)
3. Ma, J., Pei, J.\* , Huang, D & Song, D. (2016). Syntactic Parsing of Clause Constituents for Statistical Machine Translation. International Journal of Computational Science and Engineering (EI检索期刊, 录用待发表)
4. 马建军, 裴家欢\*, 黄德根. (2016). CRFs融合语义信息的英语功能名词短语识别. 中文信息学报, 30(6):59-66 (国内核心期刊)
5. 张聪, 裴家欢\*, et al. (2016). 基于语义图优化算法的中文微博观点摘要研究. 山东大学学报 (理学版) (国内核心期刊, 录用待发表)



Thanks for attention!  
恳请各位老师批评指正