现代汉语名词歧义度研究

李 安

2012年10月24日星期三

歧义度的提出、研究目标和框架

- 词义消歧实现
- 义项关系和义项的语义距离

结论1：歧义度与义项距离的关系

结论2：三种义项关系及歧义度成因

结论3：人机词典的不同与现汉的不足

词义消歧

- 用计算机为使用中的多义词标注一个确定的义项。
- 词义消歧在自然语言处理中有重要作用。

词义消歧研究概述

- 人工编写规则
- 统计方法
- 词典方法

词义消歧中存在的问题

- 研究集中于人工智能角度关注，评测多以平均正确率和召回率为标准，对词义的个性研究不足。
- 不同多义词消歧正确率不同。体现的更多的不是方法问题，而是词义问题。

歧义度

- 歧义度是多义词体现在计算机词义消歧中的难度和可能达到效果的歧义性的度量。
- (1)歧义度是多义词义项间区别性形式特征有无、多少的数字描述；(2)歧义度是多义词各个义项在分布上差异大小的数字描述；(3)歧义度是多义词在计算机词义消歧中困难程度的数字描写。

歧义度计算

$$P = \left(1 - \frac{M}{N}\right)\% = \left(1 - \frac{\sum m_i}{N}\right)\% \quad (2.8)$$

歧义度的研究目标和框架

描写歧义度

- 实现词义消歧
- 标注一定规模语料
- 计算歧义度

解释多义词间歧义度差异的原因

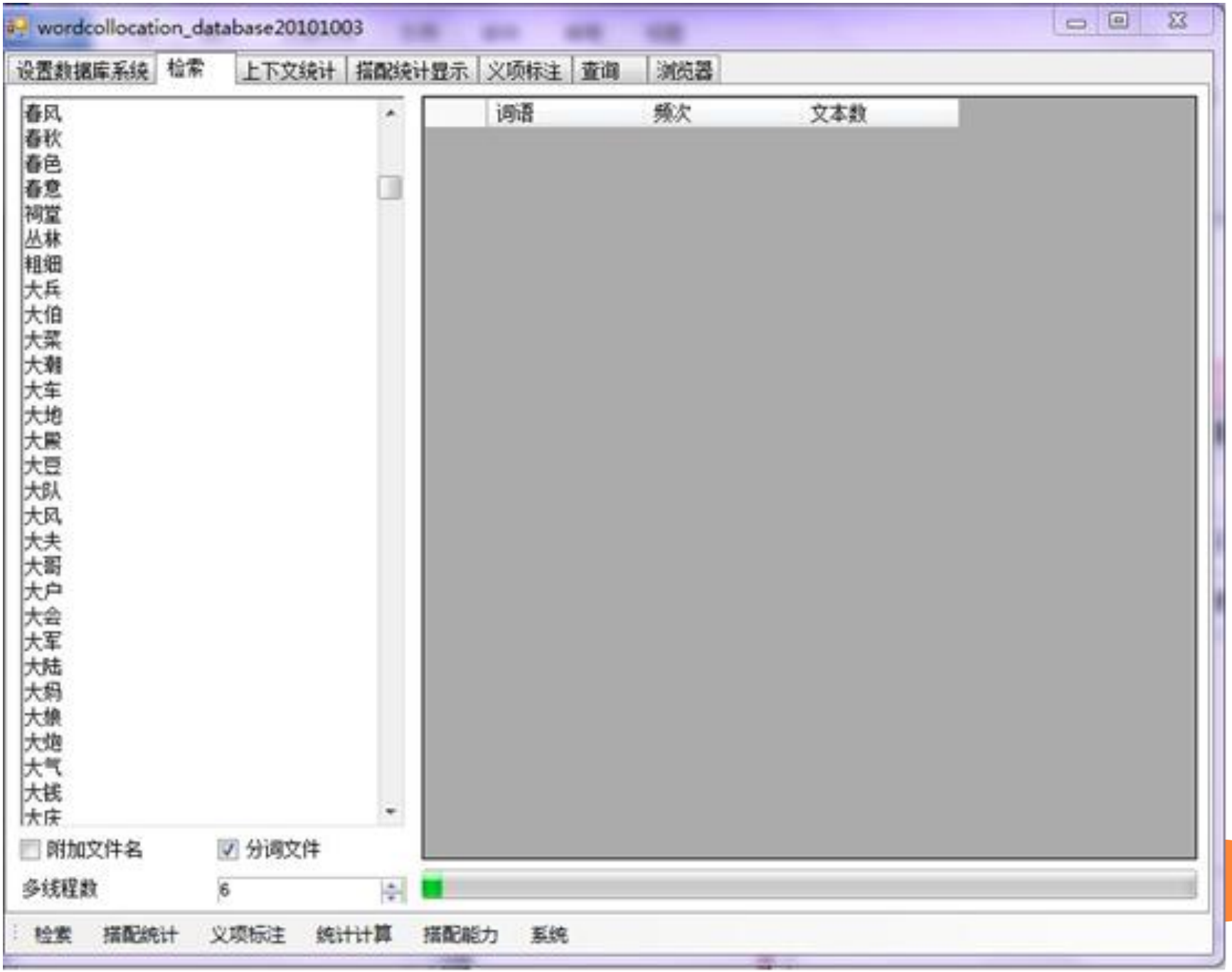
- 义项关系和义项的语义距离

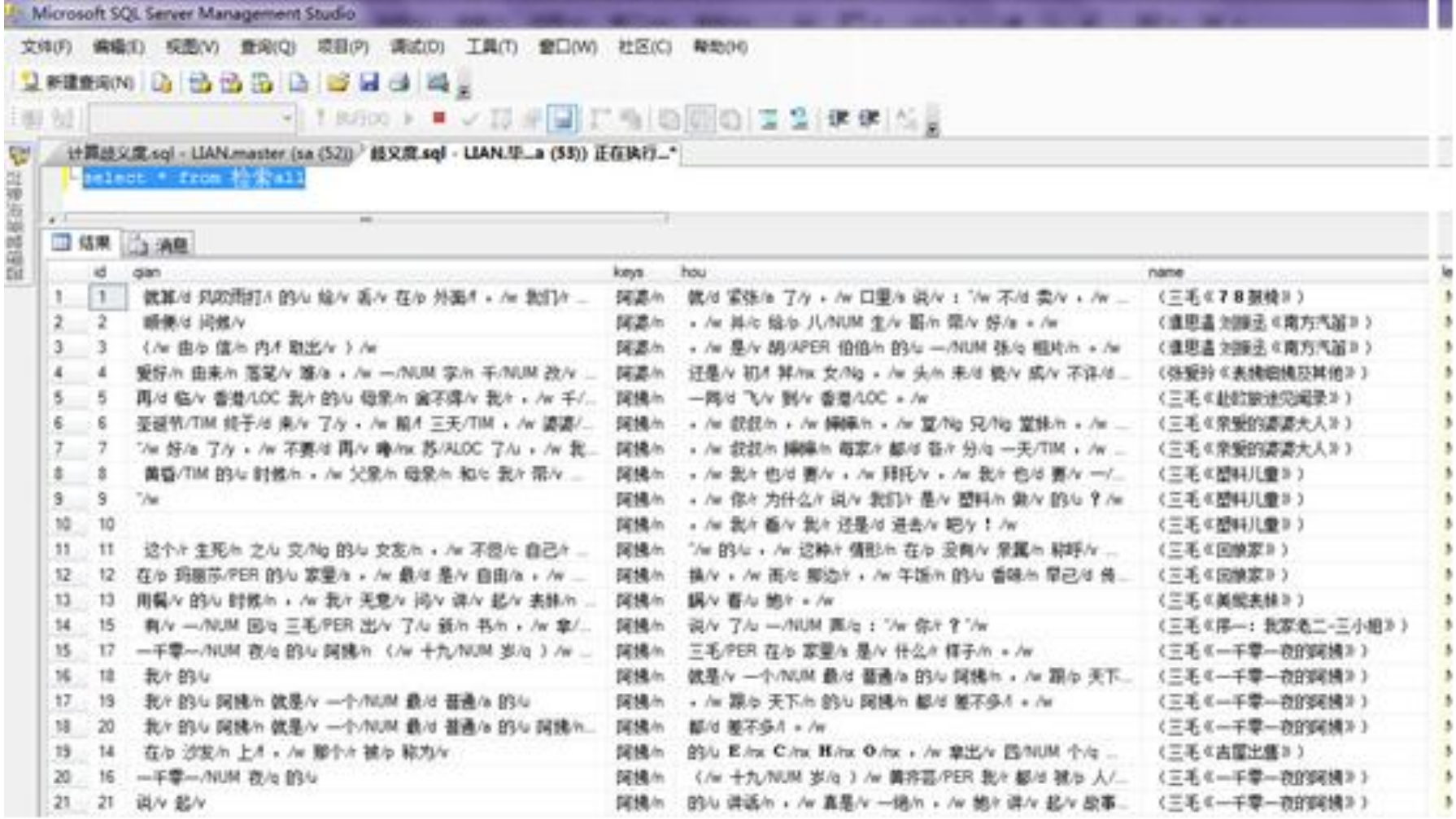
返回

计算歧义度

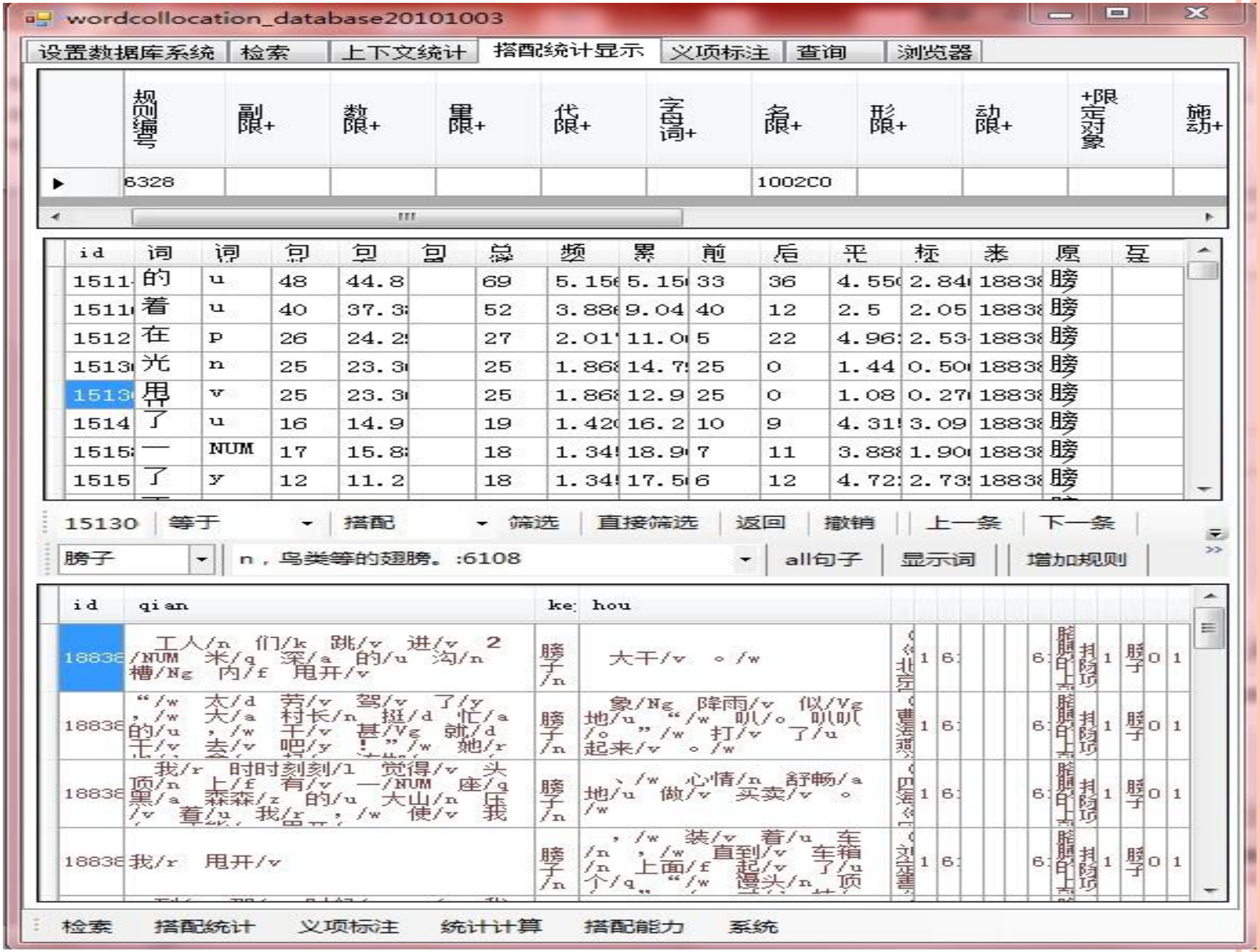
- 采用统计与规则相结合的方式实现。
- 应用到了大规模语料库；搭配库；义类词典。


wordcollocation_database20101003
设置数据库系统
检索
上下文统计
搭配统计显示
义项标注
查询
浏览器
数据库服务器
LN-PC
不使用用户名和密码
使用用户名密码
sa
选择一个数据库
毕业论文
刷新
停止
状态
连接到Sql Server数据库
打开ACCESS数据库
词表_n
词列表
验证语料_n
检索语料存放表
统计_n
上下文统计结果表
文本类型表
wuji
义类词典
总字频表
规则分析表
词典校对版
现代汉语词典
名词
名词规则
动词
动词规则
形容词
形容词规则
词性
词性规则
create table 词表 (id autoincrement not null,word char (80),pinci int, pinlv single,wenbenshu single)
create table 检索 (id autoincrement not null,qian ntext,keys
执行命令
原始语料文件夹
统计结果存放目录
保存当前命令
检索
搭配统计
义项标注
统计计算
搭配能力
系统


wordcollocation_database20101003
设置数据库系统
检索
上下文统计
搭配统计显示
义项标注
查询
浏览器
春风
春秋
春色
春意
祠堂
丛林
粗细
大兵
大伯
大菜
大褂
大车
大地
大厦
大豆
大队
大风
大夫
大哥
大户
大会
大军
大陆
大妈
大娘
大炮
大气
大钱
大庆
词语
频次
文本数
附加文件名
分词文件
多线程数
6
检索
搭配统计
义项标注
统计计算
搭配能力
系统



id	qian	keys	hou	name	le
1	1 就算/d 风吹雨打/i 的/u 仙/v 丢/v 在/p 外面/f 。/w 我们/r …	阿姨/n	就/d 紧张/a 了/y 。/w 口里/s 说/v ：“/w 不/d 卖/v 。/w …	《三毛《7 8 狼棒》》	
2	2 顺便/d 问候/v	阿姨/n	。/w 并/c 给/p 八/NUM 生/v 哥/n 带/v 好/a 。/w	《谭思嘉 刘丽丞《南方汽笛》》	
3	3 （/w 由/p 信/n 内/f 取出/v ）/w	阿姨/n	。/w 是/v 胡/APER 伯伯/n 的/u 一/NUM 张/q 相片/n 。/w	《谭思嘉 刘丽丞《南方汽笛》》	
4	4 爱好/n 由来/n 落笔/v 难/a 。/w 一/NUM 字/n 千/NUM 改/v …	阿姨/n	还是/v 初/f 算/nx 女/Ng 。/w 头/n 未/d 梳/v 成/v 不许/d …	《张爱玲《表情细胞及其他》》	
5	5 再/d 临/v 香港/LOC 我/r 的/u 母亲/n 舍不得/v 我/r 。/w 千/…	阿姨/n	一阵/d 飞/v 到/v 香港/LOC 。/w	《三毛《赴拉萨途见闻录》》	
6	6 圣诞节/TIM 前/f 夕/v 了/y 。/w 前/f 三天/TIM 。/w 婆婆/…	阿姨/n	。/w 叔叔/n 。/w 婶婶/n 。/w 宜/Ng 兄/Ng 堂妹/n 。/w …	《三毛《亲爱的婆婆大人》》	
7	7 “/w 好/a 了/y 。/w 不要/d 再/v 难/nx 苏/ALOC 了/u 。/w 我…	阿姨/n	。/w 叔叔/n 婶婶/n 每家/r 都/d 各/r 分/q 一天/TIM 。/w …	《三毛《亲爱的婆婆大人》》	
8	8 黄昏/TIM 的/u 时候/n 。/w 父亲/n 母亲/n 和/c 我/r 带/v …	阿姨/n	。/w 我/r 也/d 要/v 。/w 拜托/v 。/w 我/r 也/d 要/v 一/…	《三毛《塑料儿童》》	
9	9 “/w	阿姨/n	。/w 你/r 为什么/r 说/v 我们/r 是/v 塑料/n 做/v 的/u ？/w	《三毛《塑料儿童》》	
10	10	阿姨/n	。/w 我/r 看/v 我/r 还是/d 退去/v 吧/y ！/w	《三毛《塑料儿童》》	
11	11 这个/r 生死/n 之/u 交/Ng 的/u 女友/n 。/w 不但/c 自己/r …	阿姨/n	“/w 的/u 。/w 这种/r 情形/n 在/p 亲戚/v 朋属/n 聊呼/v …	《三毛《回娘家》》	
12	12 在/p 玛丽莎/PER 的/u 家里/s 。/w 最/d 是/v 自由/a 。/w …	阿姨/n	换/v 。/w 舌/c 那边/r 。/w 午饭/n 的/u 香味/n 早已/d 勾…	《三毛《回娘家》》	
13	13 用餐/v 的/u 时候/n 。/w 我/r 无意/v 问/v 讲/v 起/v 表舅/n …	阿姨/n	锅/v 看/v 她/r 。/w	《三毛《美丽表舅》》	
14	15 有/v 一/NUM 回/q 三毛/PER 出/v 了/u 领/n 书/n 。/w 拿/…	阿姨/n	说/v 了/u 一/NUM 声/q ：“/w 你/r ？”/w	《三毛《序一：我家老二-三小姐》》	
15	17 一千零一/NUM 夜/q 的/u 阿姨/n （/w 十九/NUM 岁/q ）/w …	阿姨/n	三毛/PER 在/p 家里/s 是/v 什么/r 样子/n 。/w	《三毛《一千零一夜的阿姨》》	
16	18 我/r 的/u	阿姨/n	就是/v 一个/NUM 最/d 普通/a 的/u 阿姨/n 。/w 跟/p 天下…	《三毛《一千零一夜的阿姨》》	
17	19 我/r 的/u 阿姨/n 就是/v 一个/NUM 最/d 普通/a 的/u	阿姨/n	。/w 跟/p 天下/n 的/u 阿姨/n 都/d 差不多/f 。/w	《三毛《一千零一夜的阿姨》》	
18	20 我/r 的/u 阿姨/n 就是/v 一个/NUM 最/d 普通/a 的/u 阿姨/n…	阿姨/n	都/d 差不多/f 。/w	《三毛《一千零一夜的阿姨》》	
19	14 在/p 沙发/n 上/f 。/w 那个/r 被/p 称为/v	阿姨/n	的/u E/nx C/nx H/nx O/nx 。/w 拿出/v 四/NUM 个/q …	《三毛《吉屋出售》》	
20	16 一千零一/NUM 夜/q 的/u	阿姨/n	（/w 十九/NUM 岁/q ）/w 黄容芸/PER 我/r 都/d 被/p 人/…	《三毛《一千零一夜的阿姨》》	
21	21 说/v 起/v	阿姨/n	的/u 讲话/n 。/w 真是/v 一绝/n 。/w 她/r 讲/v 起/v 故事…	《三毛《一千零一夜的阿姨》》	


wordcollocation_database20101003
设置数据库系统
检索
上下文统计
搭配统计显示
义项标注
查询
浏览器
规则编号
副限+
数限+
量限+
代限+
字母词+
名限+
形限+
动限+
+限定对象
施动+
6328
1002C0
id
词
词
句
句
句
总
频
累
前
后
平
标
来
原
互
的
着
在
光
用
了
一
了
15130
等于
搭配
筛选
直接筛选
返回
撤销
上一条
下一条
膀子
n，鸟类等的翅膀。:6108
all句子
显示词
增加规则
id
qian
ke
hou
工人/n 们/k 跳/v 进/v 2/NUM 米/q 深/a 的/u 沟/n 槽/Ng 内/f 甩开/v
膀子/n
大干/v 。/w
我/r 甩开/v
膀子/n
检索
搭配统计
义项标注
统计计算
搭配能力
系统


wordcollocation_database20101003
设置数据库系统
检索
上下文统计
搭配统计显示
义项标注
查询
浏览器
仅概率
规则与概率
标注句子表
标注搭配表
前半句长
9
后半句长
10
逗号断开
线程数
开始
暂停
继续
停止
刷新数据列
导出规则
导出规则词
进度词
进度句子
获取词表
select word from @cibiao group by word order by word asc
获取词典
select 义项编号,词性,义频顺序,sense,义项号 from @cidian where word ='@str_ci'
操作
选择导出到位置默认C
执行
检索
搭配统计
义项标注
统计计算
搭配能力
系统

返回

如何解释多义词歧义度的差别

- 词汇语义理论一直分为分析主义和功能主义两个派别。
- 歧义度是对义项功能差异的描写，我们希望从词义内涵本身寻找原因，把两种主义结合起来。
- 结构主义语言学认为语言要素都存在于系统当中，组合、聚合是这个系统的主要关系。
- 描写义项的关系，就需要在系统中找到其相对位置。

义类词典

- 义类词典是词义系统的体现。
- 义类的相对位置关系体现了意义异同。
- 以义类词典为描写义项关系的工具。

义项关系

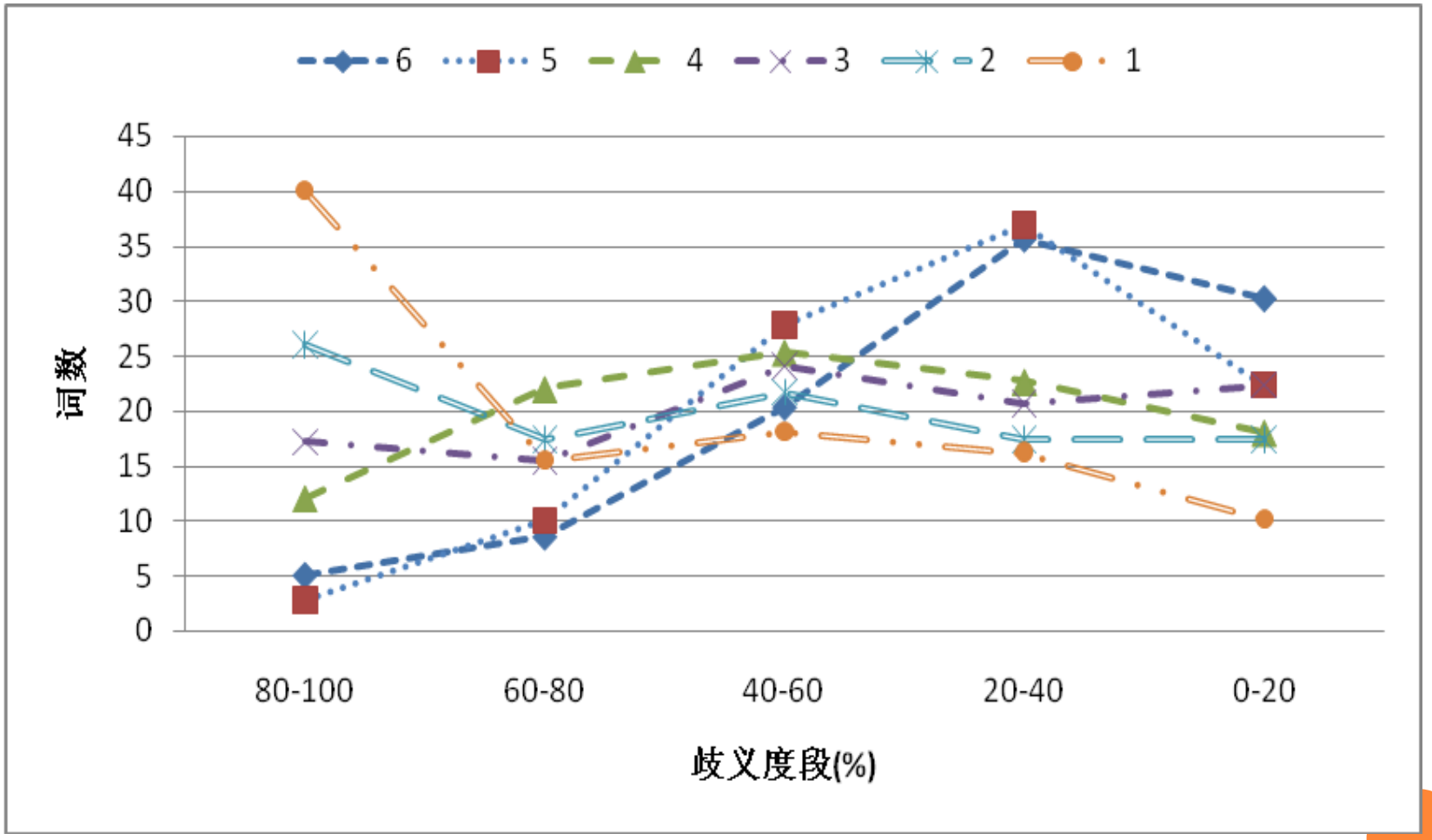
- 指多义词义项间的概念关系，具体地用义项在义类词典中所对应的义类间对立关系表示，如“便衣”：
- 【便衣】①平常人的服装(区别于军警制服)。②身着便衣执行任务的军人、警察等。
- “便衣”①为：“具体物-生活用品-服饰-衣服-便装”类
- “便衣”②为：“生物-人-职业-军人-侦察兵”类，这两个义类的对立就构成了“便衣”的义项间语义关系。

语义距离

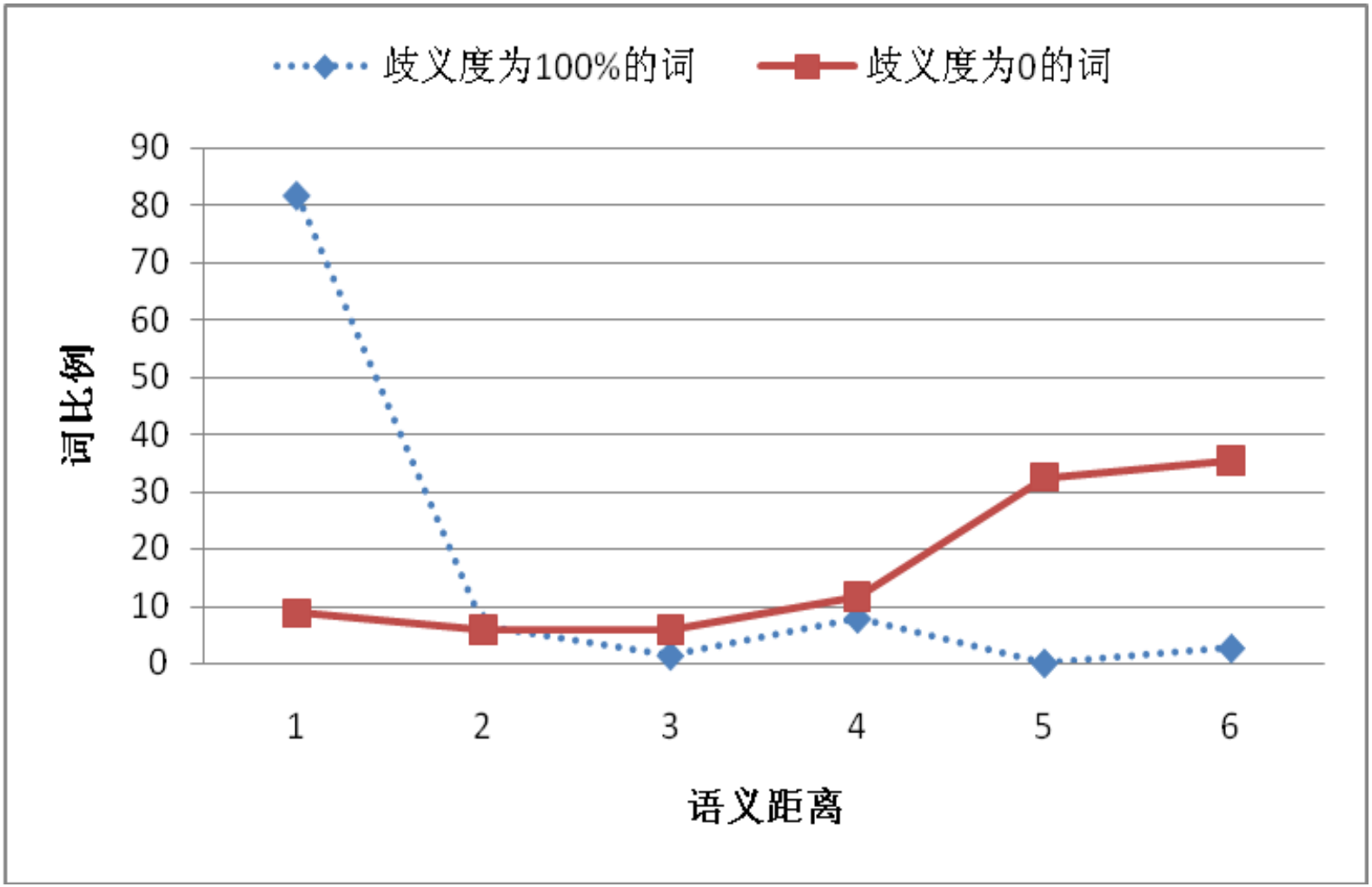
- 语义距离为义项在义类词典中相对位置的远近，是义项间语义关系的数字化表示。分1-6级
- 第一级：两个义项属于同一个五级类，如“大雨”：
- 【大雨】①指24小时内雨量达25－50毫米的雨。②指下得很大的雨。
- 第六级：一级类开始不同，如“心脏”：
- 【心脏】①心①。②比喻中心首都。北京是祖国的～。

返回

不同语义距离词的歧义度分布差异明显

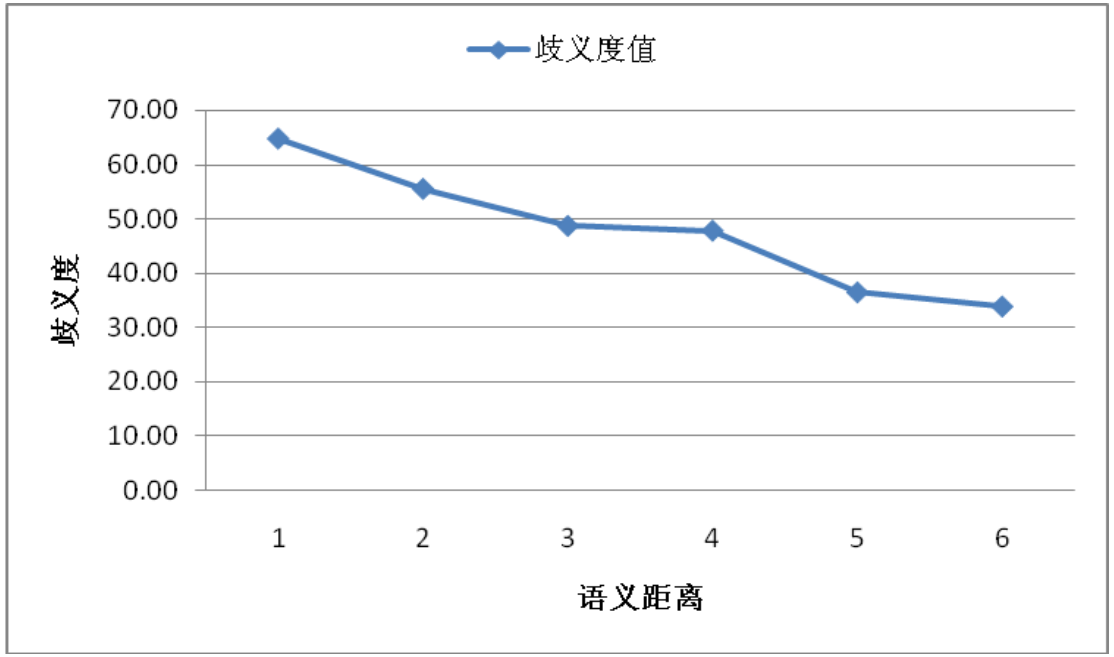


两种极端歧义度词在语义距离的分布迥异



语义距离与歧义度间存在高度负相关关系

语义距离	词数	平均歧义度(%)	歧义度均差(%)
1	265	64.76	18.95
2	69	55.54	9.73
3	58	48.75	2.94
4	150	47.85	2.04
5	219	36.55	-9.26
6	314	33.95	-11.86



使用EXCEL统计得到语义距离与歧义度均值间的相关系数为r=-0.98

返回

三种重要义项关系

同义关系

- 意义基本相同，用法部分或完全相同的关系。近义关系指意义相近，用法不完全相同的关系。
- 【暴力】①强制的力量；武力。②特指国家的强制力量。
- 【鞭炮】①大小爆竹的统称。②专指成串的小爆竹。

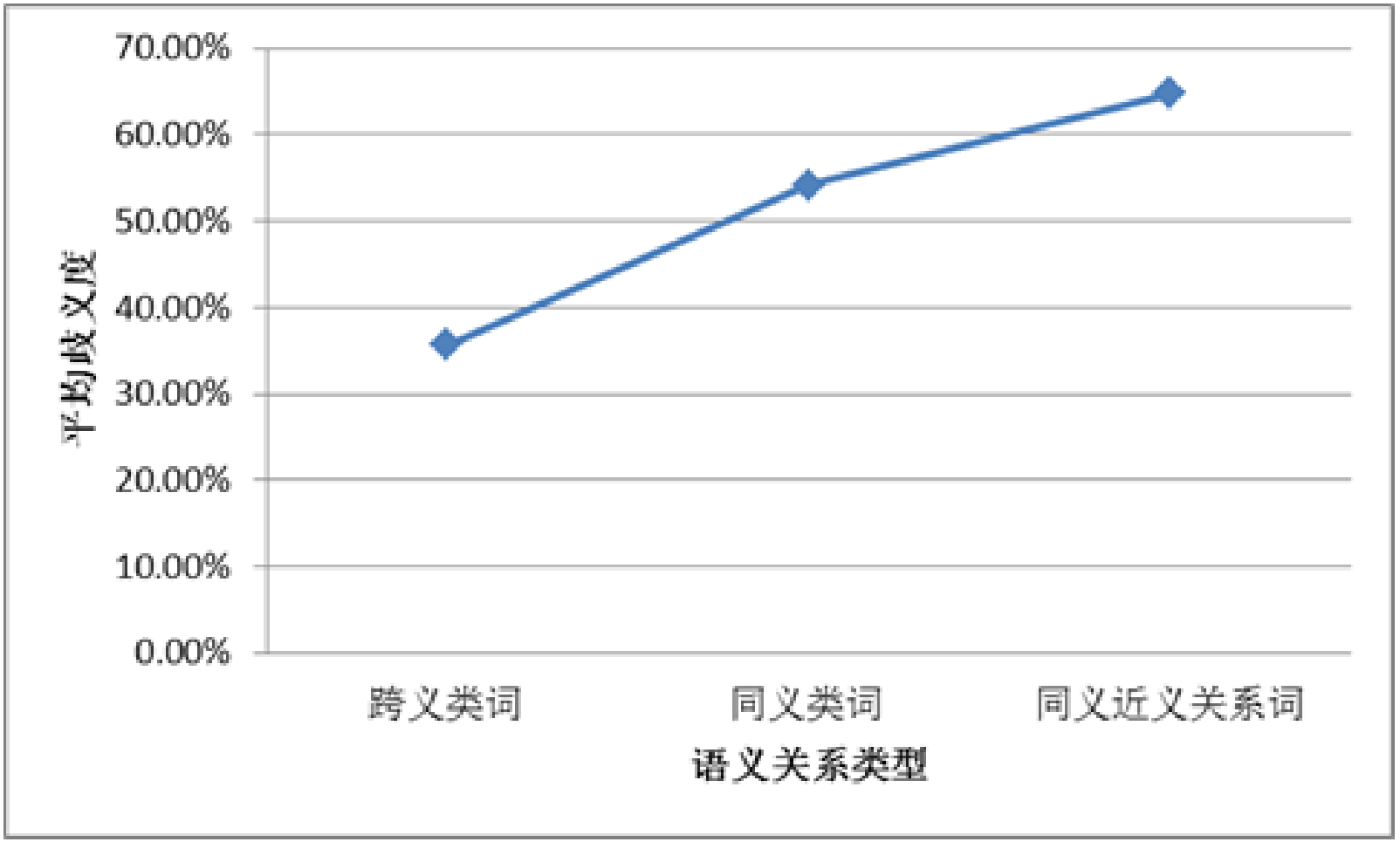
同类关系

- 各个义项属于同一基本层次范畴；语义距离为2或3。这类词“属”相同，但是“种差”较大。
- 【单间】①饭馆、旅馆内供单人或一起来的几个人用的小房间。②只有一间的屋子。

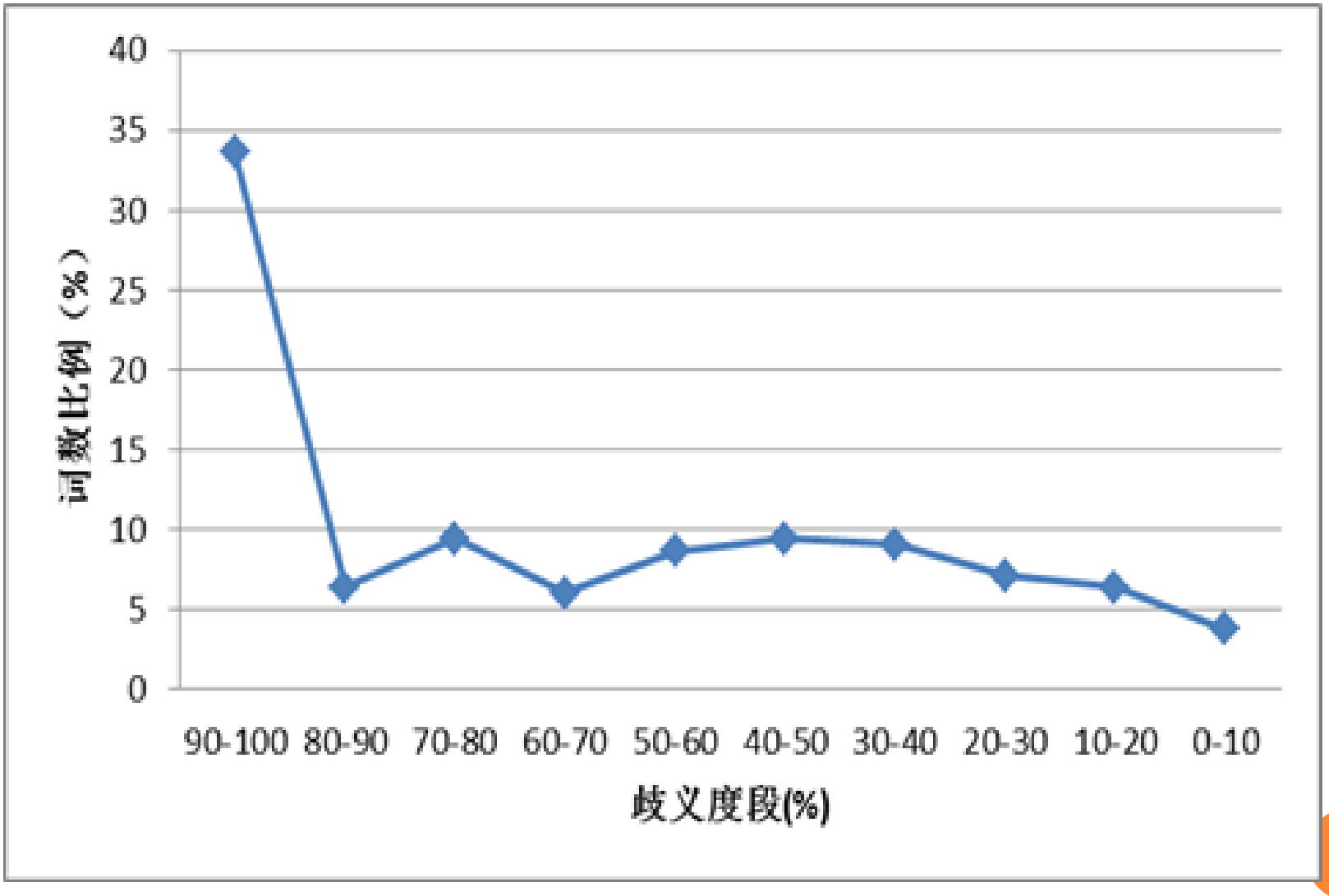
跨类关系

- 其义项不属于同一个基本层次范畴；义项间区别较大；从在词义系统的关系看，语义距离普遍为4、5或6，一个词的义项不同属于一个三级类或二级类甚至一级类。
- 【品质】①物品的质量江西瓷～优良。②行为、作风上所表现的思想、认识、品性等的本质道德～。

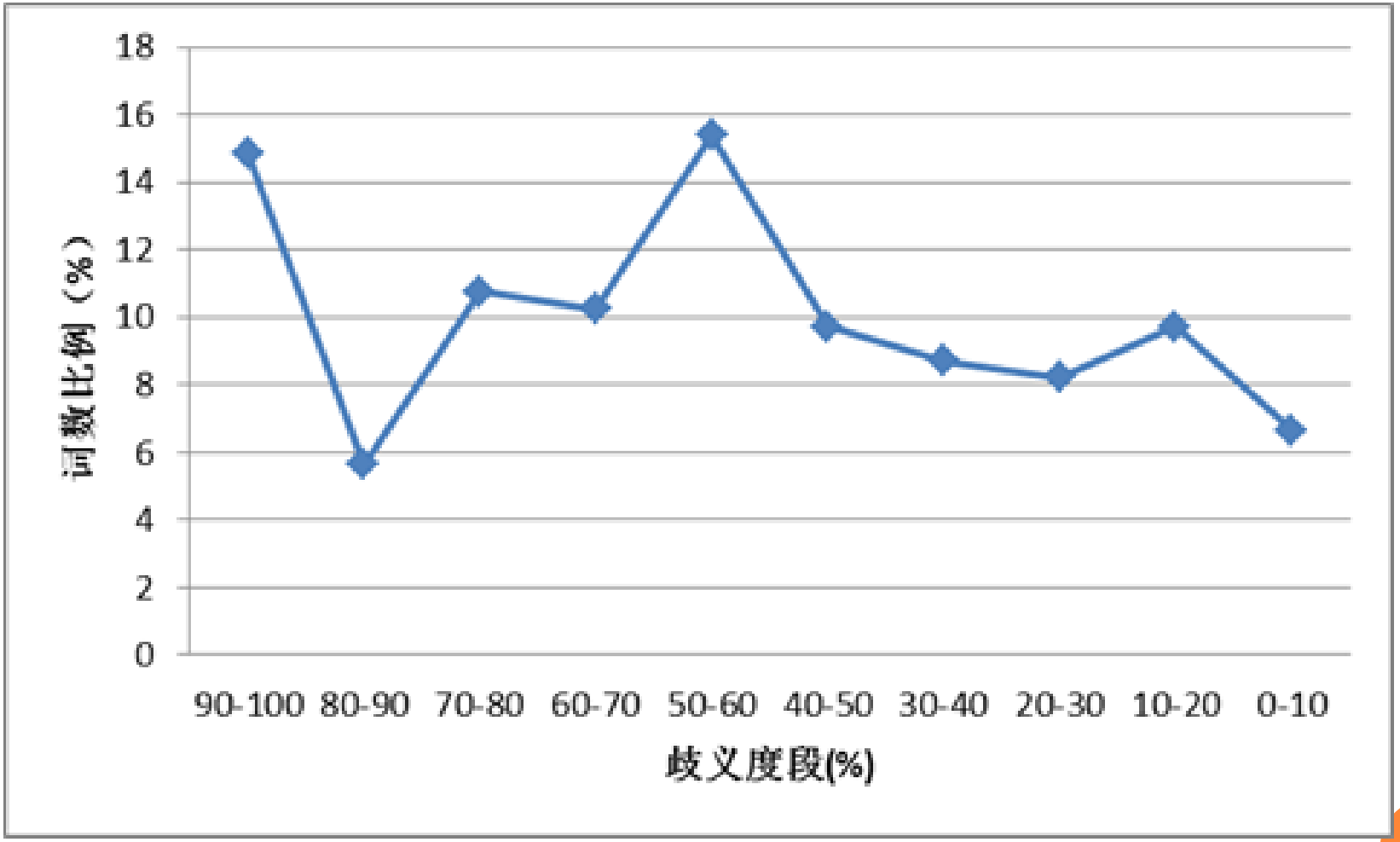
三种词歧义度差异明显


70.00%
60.00%
50.00%
40.00%
30.00%
20.00%
10.00%
0.00%
平均歧义度
跨义类词
同义类词
同义近义关系词
语义关系类型

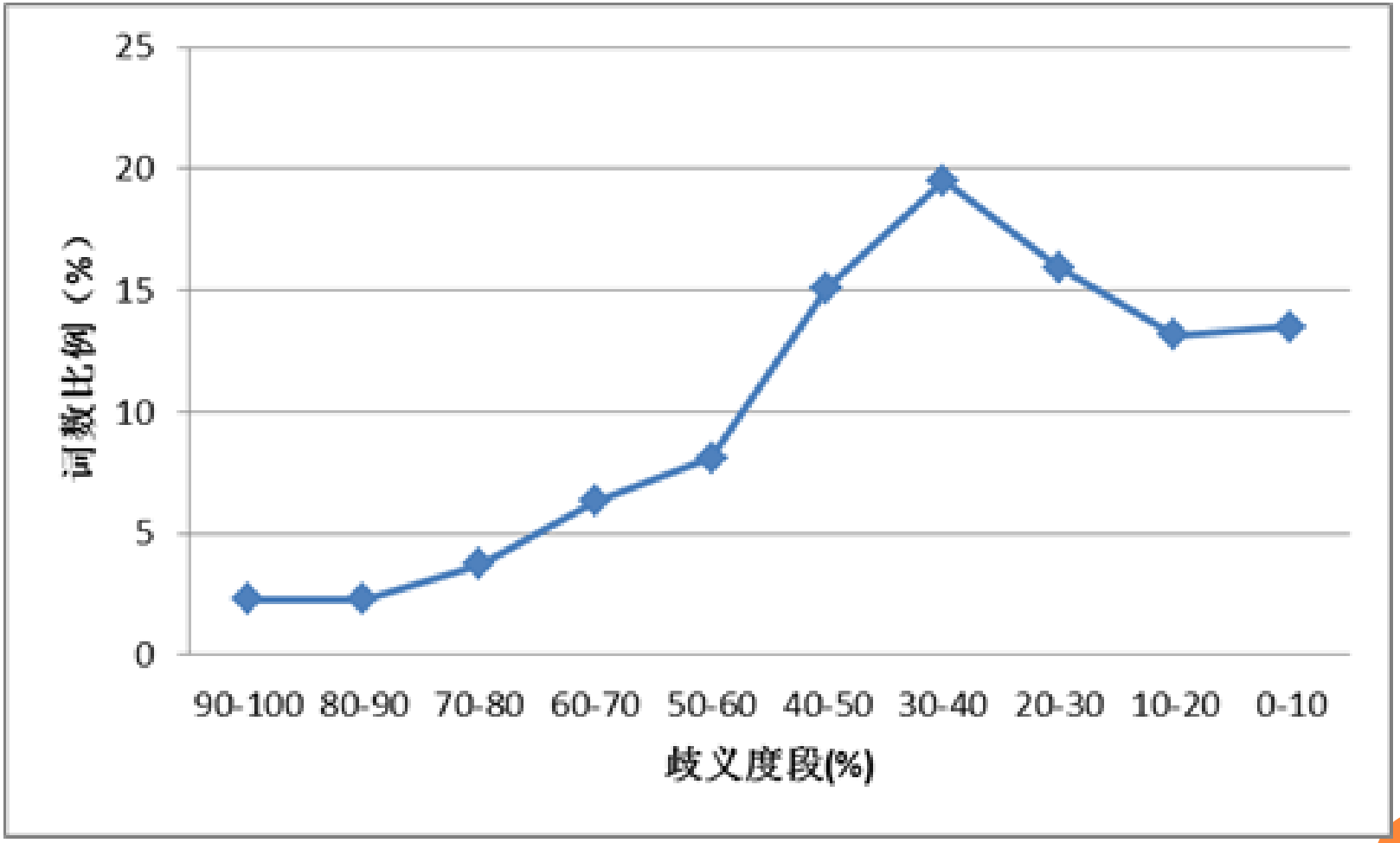
同义近义词的歧义度分布



同类词的歧义度分布


18
16
14
12
10
8
6
4
2
0
词数比例（%）
90-100
80-90
70-80
60-70
50-60
40-50
30-40
20-30
10-20
0-10
歧义度段(%)

跨义类词的歧义度分布


25
20
15
10
5
0
词数比例（%）
90-100
80-90
70-80
60-70
50-60
40-50
30-40
20-30
10-20
0-10
歧义度段(%)

同义类词与跨义类词歧义度成因不同

- 同义类词
  - 造成歧义度的原因为义项间共同义素的多少。
  - 消歧模式应以描写义项间的相同和不同义素为主。
- 跨义类词
  - 造成歧义度的原因是义项间意义引申模式或隐喻、暗喻的概念映射模式。

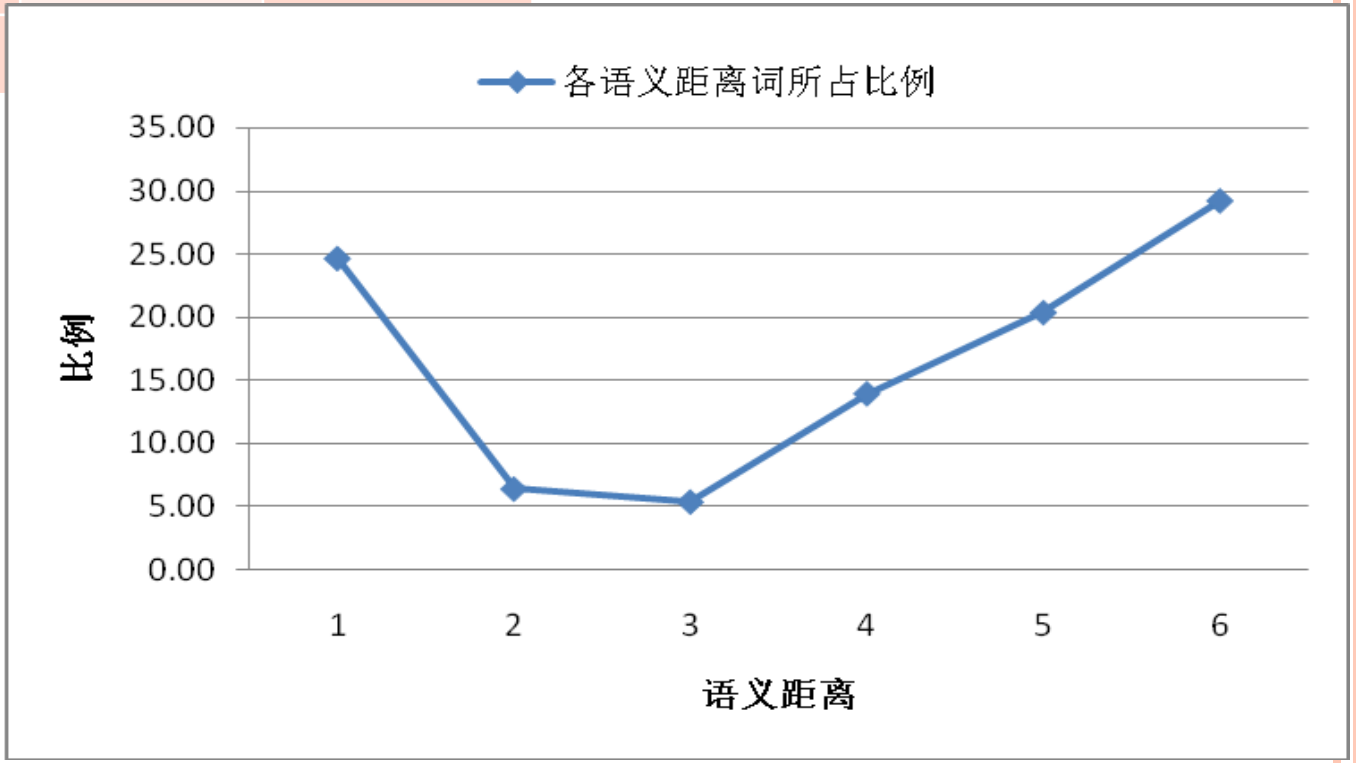
返回

词义消歧的词典问题

- 词义划分有很强的主观性。
- 义项区分粒度是词义消歧中的一个基础、核心问题。
- 目前既没有确定义项粒度的方法，也没有形成取舍的标准。
- 分析论和功能论无法单独解释义项粒度现象。

“现汉”有部分词义项粒度过小

语义距离	词数	比例(%)	累加(%)
1	265	24.65	24.65
2	69	6.42	31.07
3	58	5.40	36.47
4	150	13.95	50.42
5	219	20.37	70.79
6	314	29.21	100
合计	1075		



举例

- 【大雨】①指24小时内雨量达25－50毫米的雨。②指下得很大的雨。
- 【封面】①线装书指书皮里面印着书名和刻书者的名称等的一页。②新式装订的书刊指最外面的一层，用厚纸、布、皮等做成。③特指新式装订的书刊印着书刊名称等的第一面。也叫封一。
- 【砖头】①不完整的砖②砖。
- 【工钱】①做零活儿的报酬②工资。
- 【职工】①职员和工人。②旧时指工人。
- 【匪徒】①强盗。②危害人民的反动派或坏分子。

调整义项粒度对歧义度的影响

- **（一）合并语义距离为1的义项对歧义度的影响**
- **1、多义词减少24.1%**
- 研究范围内语义距离为1的词有259个，占双义项名词数的24.1%，也就是合并这部分义项后，多义词会减少24.1%。
- **2、歧义度为100%的词减少79.22%**
- 歧义度为100%的词为词义消歧中最难处理的部分，双义项名词中共有77个，其中语义距离为1的部分有61个，占到总数的79.22%，如果合并这部分义项，将同时排除这部难以辨析的词。
- **3、总体平均歧义度降低6个百分点**
- 双义项名词平均歧义度为46.20%，语义距离为1的词平均歧义度为64.76%，如果排除这一部分，双义项名词歧义度将降低6个百分点，变为40.54%。
- **（二）合并同义类词义项对歧义度的影响**
- **1、多义词减少42.79%**
- 本文研究范围内同义类词和跨义类词总数为1075个，其中同义类词数量为460个，如果合并同义类词多义词数量将减少42.79%。
- **2、歧义度为100%的词减少94.81%**
- 歧义度为100%的同义类词有73个，占全部歧义度为100%的同义类词和跨义类词的94.81%。
- **3、总体平均歧义度降低**10.52%
- 同义类词的平均歧义度为60.26%总体上是比较高的，如果去掉同义类词部分，平均歧义度将由46.2%降为35.68%，降低10.52个百分点。

谢谢！