相似词获取的集成方法*

石 静，邱立坤，王 菲，吴云芳
计算语言学教育部重点实验室（北京大学），北京 100871
北京大学 计算语言学研究所，北京 100871
E-mail: shijing09@pku.edu.cn

摘 要：语义相似度计算是自然语言处理领域的关键问题之一，在信息检索中的查询扩展、机器翻译中的模块识别，以及句法分析、词义消歧等任务中都发挥着重要的作用。本文将集成方法应用于基于大规模语料库的汉语语义相似度计算上，提出并实现了不同语域的集成方案。分别使用新闻语料和互联网语料，选取窗口大小为 2 或 3 的上下文词语特征、以上下文与目标词之间的互信息作为权值构建特征向量，计算向量之间的 cosine 夹角作为词语相似度，得到了三种语义相似度序列。对这三个相似度序列进行集成，使用了平均排名、调和平均排名和平均分数三种集成方法。对四种组合方式的集成结果进行了评测，实验结果表明，集成方法获取的语义相似度相对于单一方法准确率得到了提升，其中，与不同窗口的集成相比，本文提出的不同语域的集成，准确率提升更为显著。

关键词：语义相似度；相似词；集成方法；分布相似性

Ensemble Methods for Similar Word Extraction

Shi Jing, Qiu Likun, Wang Fei, Wu Yunfang
Key Laboratory of Computational Linguistics (Peking University), Ministry of Education, Beijing 100871
Institute of Computational Linguistics, Peking University, Beijing 100871
E-mail: shijing09@pku.edu.cn

Abstract: Automatic acquisition of similar words is one of the most crucial problems in natural language processing. It plays an important role in some application systems, such as query expansion in information retrieval, pattern identification in machine translation, parser analysis and WSD missions. In this paper, we apply ensemble methods to lexical semantic similarity computation, by combining three different distributional similarity lists. We choose news corpus and Web corpus, select 2 and 3 window context words as features, adopt PMI measure as feature weight, and use cosine algorithm to compute word similarity. This paper focuses on combining the three similarity lists using ensemble methods: mean rank, harmonic mean rank, and mean score. The experiment results show that ensemble methods outperform individual methods, and combining the two similarity lists from different domains gets the best results.

Keywords: semantic similarity; similar words; ensemble methods; distributional similarity

1 引言

词汇语义信息是自然语言处理中重要的资源，是进一步进行句法和语义分析的基础。在信息检索中的查询扩展、机器翻译中的模块识别等方面，相似词都是不可或缺的知识；在句法分析、词义消歧等信息处理任务中，词语相似度也发挥着重要的作用。而相似词词典的手工构建是一项费时费力的浩大工程，存在着不易更新、覆盖度不全等诸多缺陷。自动地获取相似词并得到相似度，使自动构建词典成为可能，不仅减少了工作量，还使词典资源能够自动更新和扩展。

语义相似度的研究可分为两大类方法：基于大规模语料库[1-4]和基于词典[4-5]。基于词典的研究多是利用词典中的语义类层次关系计算两个词的相似度，例如基于 WordNet 来计算英语词语的语义相似度，基于 HowNet 来计算汉语词语的语义相似度[5]；基于语料库的研究多是选取上下文特征，用向量来表征词语，再利用这些特征向量进行词语之间相似度的计算[1-3]。基于词典的方法需要有

* 基金资助：国家自然科学基金(60703063)；九十八年度蒋经国国际学术交流基金会项目“历代语言知识库建置计划”。

完备的知识库的支撑，依赖词典编撰者个人的经验知识，无法反映大规模语料库中词语真实的意义和用法。Agirre et. al (2009)的研究[6]表明，基于语料库的方法可以取得和词典方法相媲美的结果，而无需词典知识的支撑。

集成方法是一种常用的机器学习方法，将多种分类器的结果通过投票等方法进行集成，用以提升整个分类器的性能，不同分类器或学习算法不同，或特征表示不同，或训练数据不同。集成方法能够有效去除噪音和统计偏差，近年来已经成功地应用在大量的自然语言处理任务中，例如词性标注、词义消歧、句法分析等。本文将其应用在基于大规模语料库的相似词获取上，不仅可以将多种语料多种方法互补集成，提高准确率，还可以部分解决计算复杂度高的问题。

本文在新闻和互联网两种不同语料上，使用了基于不同窗口的上下文特征，进行了词语相似度的计算，得到单一方法的相似度结果序列。从中选取了三种方法的结果序列，分析计算了它们之间的重合度。分别使用了三种不同的集成方法：平均排名、调和平均排名和平均分数，对这三个相似度序列进行了集成。Curran（2002）的研究[1]表明，不同窗口特征的集成可以提升相似度的计算结果。本文的实验结果表明，相比于不同窗口，不同语域的相似度结果差异性更大，对不同语域的集成可以取得更好的效果。

2 相似词获取方法

基于语料库的相似词获取是基于分布性假设[7]（Harris,1968）：语义相似的词语通常有着相似的上下文。上下文主要有两种选择方法：窗口上下文和依存关系上下文，由于中文依存分析器的准确率较低，因此本文选用窗口上下文。在特征权值的计算上，试验了目标词与上下文特征间的共现频次（tf）、是否共现（bool）、出现该上下文特征的目标词个数（idf）、tf*idf、互信息（PMI）五种计算方法，实验结果表明互信息效果最好，因此本文采用互信息作为特征权值。对每个目标词构建特征向量，利用向量间 cosine 夹角计算词语之间的相似度。

1）语料选择

分别选择了两种不同语域的语料。

国际语言资源联盟 LDC 提供的 Chinese Gigaword，选取其中的新华社语料。该语料库收录了 1991~2004 年共 14 年的新华社全部文本，共约 4.7 亿汉字。对语料进行前期处理： Unicode 到 GB 编码的转换；利用中科院计算所分词软件 ICTCLAS 对全部文本进行自动词语切分和词性标注。

搜狗实验室提供的互联网语料，从中提取 130000 个常用词所在的句子，每个词语抽取最多 1 万个句子，利用中科院计算所的分词软件 ICTCLAS 对全部文本进行自动词语切分和词性标注。

2）窗口上下文

选取一定上下文窗口内的词作为目标词的特征。原始语料经过切词、词性标注和断句后，取目标词前后窗口为 2 或 3 的上下文词、词性、相对目标词的位置作为目标词的特征。例如 “*黑龙江/ns 将/d 严格/ad 控制/v 森林/n 资源/n 消耗量/n。/w*”，目标词为“*控制*”，窗口为 3 的上下文特征 w_{-3},w_{-2},w_{-1},w_{+1},w_{+2},w_{+3} 分别为“*黑龙江/ns*”，“*将/d*”，“*严格/ad*”，“*森林/n*”，“*资源/n*”，“*消耗量/n*”。

3）特征的权值计算

特征权值的计算选取了互信息（PMI）方法，即以目标词 w_i 和上下文词 c_j 的互信息作为特征 c_j 相对目标词 w_i 的权值。其中，N 为语料中所有词的总个数，count(w)为语料中词 w 出现的频次。

$$PMI(W_i,C_j)=\frac{P(w_i,c_j)}{P(w_i)*P(c_j)}=\frac{tf(w_i,c_j)*N}{count(w_i)*count(c_j)} \quad \text{公式 1}$$

4）相似度计算方法

相似度计算方法使用了 cosine 夹角度量。其中 $\overrightarrow{V_1}$, $\overrightarrow{V_2}$ 是目标词 w1,w2 的特征向量，T(w)是出

现在 w 上下文的词集，c 为上下文词，weight 为计算得到的权值。

$$\cos(w_1,w_2)=\cos(\overrightarrow{V_1},\overrightarrow{V_2})=\frac{\overrightarrow{V_1}\cdot\overrightarrow{V_2}}{\left|\overrightarrow{V_1}\right|*\left|\overrightarrow{V_2}\right|}=\frac{\Sigma_{c\in T(w_1)\cap T(w_2)}(\text{weight}(w_1,c)*\text{weight}(w_2,c))}{\sqrt{\Sigma_{c\in T(w_1)}\text{weight}^2(w_1,c)+\Sigma_{c\in T(w_2)}\text{weight}^2(w_2,c)}}$$ 公式 2

根据实验结果，从中选取效果较好的三个相似度序列，作为集成方法的基础：Giga 新闻语料窗口为 3 的结果（**Giga**）；sogou 互联网语料窗口为 3 的结果（**sogou3**）；sogou 互联网语料窗口为 2 的结果（**sogou2**）。

3 重合度分析

为了评估三个相似度结果序列的互补程度，使用两种方法来计算不同相似度序列之间的重合度（overlap）。首先对相似词数据进行预处理：1）对每个目标词，只截取前 200 个相似词，2）从相似度序列中去掉目标词本身，3）去掉只出现在一种方法的目标词。设剩下的目标词集为 T。

1）数匹配

最简单的做法是直接计算两个序列的相交程度，即交集部分占各序列的比例。设目标词 w 在方法 m_1 的相似度序列中有 len_1 个相似词，在方法 m_2 的相似度序列中有 len_2 个相似词，两个相似度序列的交集有 n 个相似词，那么对目标词 w，两方法的重合度为公式 3。方法 m_1 和方法 m_2 总的重合度为 T 中所有目标词的平均。

$$\text{overlap}_{m_1,m_2}(w)=\frac{n}{\min(len_1,len_2)}$$ 公式 3

$$\text{overlap}_{m_1,m_2}=\frac{\Sigma_T\text{overlap}_{m_1,m_2}(w)}{|T|}$$ 公式 4

表 1 结果表明，不同语域（新闻语料 Giga 和互联网语料 sogou）的重合度低互补性较强，同一语料的不同窗口特征重合度高互补性弱。

表 1　不同相似度序列的数匹配结果

词集	重合度	总目标词数
Giga 和 sogou2	0.279	48343
Giga 和 sogou3	0.283	47663
sogou2 和 sogou3	0.671	47655

表 2　不同相似度序列的序相似结果

词集	重合度	总目标词数
Giga 和 sogou2	0.381	48343
Giga 和 sogou3	0.388	47663
sogou2 和 sogou3	0.784	47655

2）序相似

数匹配方法只是简单地计算交集部分所占的比例，对相似度序列中的所有相似词是同等看待的，而不同排名的相似词重要性不同，即语义相似度结果的排序很重要，所以进一步利用序相似性计算不同相似度结果的重合度（Weeds,2002）[8]。设目标词 w 在方法 m_1 中的相似度序列为（$w'_1,w'_2,w'_3,\ldots,w'_{len1}$），在方法 m_2 中的相似度序列为（$w'_{1'},w'_{2'},w'_{3'},\ldots,w'_{len2'}$），通过计算这两个序列的相似性得到重合度。以相似词 w'为特征，以序数的倒数作为特征权值（weight(w')），分别构成两种方法的权值向量 $S(w,m_1)$、$S(w,m_2)$，计算两个向量的 cosine 夹角，得到两个序列的重合度。

设目标词 w 的相似词集合为 C，相似词 w'的权值由公式 5 计算，两种方法在目标词 w 上的重合度由公式 6 计算。两种方法总的重合度为所有目标词的平均。

$$\text{weight}(w')=\begin{cases}\dfrac{1}{\text{rank}(w')}, & W'\in C\\ 0, & \text{else}\end{cases}$$ 公式 5

$$\mathrm{overlap}_{m_1,m_2}(W)=\cos(S(w,m_1),S(w,m_2))=\frac{\Sigma_{w'}\mathrm{weight}_1(w')*\mathrm{weight}_2(w')}{\Sigma_{\mathrm{len}_1}\frac{1}{i^2}*\Sigma_{\mathrm{len}_1}\frac{1}{j^2}} \quad \text{公式 6}$$

$$\mathrm{overlap}_{m_1,m_2}=\frac{\Sigma_T\mathrm{overlap}_{m_1,m_2}(w)}{|T|} \quad \text{公式 7}$$

表 2 表明，序相似与数匹配的结果一致，不同语域的重合度低，同一语料的不同窗口重合度高。由此预测，对不同语域的相似度结果进行集成，性能的提升将优于不同窗口特征的集成。

4 集成方法

对于一个目标词 w，根据不同方法的相似度结果序列，得到相似词 w'_j 的得分及排名。依照集成算法计算新的分值或排名，对所有相似词依据新的分值或排名重新进行排序，得到目标词 w 集成后的相似度序列。本文主要使用了三种集成方法[1]。设目标词 w 的相似词 w'_j 在方法 1 中排名为 rank_1，得分为 score_1；在方法 2 中排名为 rank_2，得分为 score_2，新的分值（$\mathrm{score}_{new}(w'_j)$）或排名（$\mathrm{rank}_{new}(w'_j)$）计算方法如下，必要时进行归一化。

1）平均排名（mean rank）：

$$\mathrm{rank}_{new}(w'_j)=\frac{\mathrm{rank}_1+\mathrm{rank}_2}{2} \quad \text{公式 8}$$

2）调和平均排名（harmonic mean rank）：

$$\mathrm{rank}_{new}(w'_j)=\frac{2}{\frac{1}{\mathrm{rank}_1}+\frac{1}{\mathrm{rank}_2}} \quad \text{公式 9}$$

3）平均分数（mean score）：

$$\mathrm{score}_{new}(w'_j)=\frac{\mathrm{score}_1'+\mathrm{score}_2'}{2} \quad \text{公式 10}$$

5 实验结果

1）评测方法

选择了哈尔滨工业大学的《同义词词林扩展版》（《词林》）作为语义相似度计算结果的评测标准。《词林》共收录了 91114 个词，以 5 层树状结构组织，第 1 层有 12 个类，第 2 层有 95 个类，第 3 层有 1425 个类，第 4 层有 4229 个类。基于类别大小的考虑，使用第四层进行评价。

对目标词的相似词序列进行评估时，排名前 N 个词具有更大的实际意义，因此用 P@K 和 MAP 两个指标来评价。P@K 中，K 分别取 1，5，10，50，即排名前 K 个词对应到《词林》中目标词所在类别的准确率。MAP 则为所有目标词的平均 AP 值。

$$\delta(w_i\in ans)=\begin{cases}1, & \text{if } w_i\in ans\\ 0, & \text{else}\end{cases},\ P@K=\frac{1}{K}\Sigma_{i=1}^{K}\delta(w_i\in ans) \quad \text{公式 11}$$

$$AP(w)=\frac{1}{R}\Sigma_{i=1}^{N}P@i,\ MAP=\frac{1}{M}\Sigma_{j=1}^{M}AP(w_j) \quad \text{公式 12}$$

其中，N 为计算结果中目标词的相似词个数，R 为《词林》中目标词所在词类中词语总数，若一个词出现在多个词类中，ans 为这些词类的合集，M 为目标词个数。

本文对常用名词和常用动词进行实验，去掉了 Chinese Gigaword 语料中出现 10 次以下的低频词。因为语料不同导致目标词集不同，所以在评测时，选取了三种方法目标词的交集，以确保评测在同一词集上进行。取交集后，目标词数为：名词 26743，动词 13344。

2）单一方法的实验结果

对三种单一方法的相似度序列进行了评价：Giga 新闻语料窗口为 3（Giga）、sogou 互联网语料窗口为 3（sogou3）、sogou 互联网语料窗口为 2（sogou2）。实验结果如表 3 所示。

表 3　单一方法实验结果

语料	pos	P@1	P@5	P@10	P@50	MAP	pos	P@1	P@5	P@10	P@50	MAP
Giga	n	0.414	0.321	0.272	0.162	0.070	v	0.344	0.230	0.174	0.077	0.054
sogou2		0.473	0.370	0.314	0.187	0.084		0.364	0.242	0.185	0.082	0.058
sogou3		0.465	0.362	0.307	0.180	0.080		0.362	0.243	0.184	0.082	0.058

表 3 显示，在语义相似度计算上，互联网语料比新闻语料有着较大的优势。因为新闻语料针对特定领域，质量较高但多样性和覆盖率不足，词语的用法较受限，经常出现一些特定的句式；而互联网语料是不限领域的，多样性和覆盖率较高，词语用法丰富灵活。sogou2 的结果优于 sogou3，窗口大小取 2 在特征的区别度和稀疏度上取得了一个较好的平衡。对于名词和动词两个不同的词类而言，名词的语义相似度的准确率要远好于动词。后续集成方法的实验选择准确率最高的单一方法 sogou2 作为基准线（baseline）。

3）集成方法的实验结果

在表 3 所列的三个相似度序列：Giga、sogou2、sogou3 上进行所有可能的组合：Giga+sogou2，Giga+sogou3，sogou2+sogou3，Giga+sogou2+sogou3，并分别使用三种方法：平均排名、调和平均排名和平均分数，对相似度结果序列进行集成。表 4 到表 7 列出了不同组合方式的实验结果。

a）Giga 和 sogou2 集成

表 4　Giga 和 sogou2 集成实验结果

方法	pos	P@1	P@5	P@10	P@50	MAP	pos	P@1	P@5	P@10	P@50	MAP
baseline	n	0.473	0.370	0.314	0.187	0.084	v	0.364	0.242	0.185	0.082	0.058
mean rank		0.491	0.382	0.323	0.194	0.091		0.376	0.246	0.186	0.084	0.061
har_mean rank		0.475	0.370	0.317	0.194	0.092		0.383	0.263	0.203	0.092	0.067
mean score		0.495	0.392	0.333	0.199	0.095		0.387	0.261	0.198	0.086	0.065

b）Giga 和 sogou3 集成

表 5　Giga 和 sogou3 集成实验结果

方法	pos	P@1	P@5	P@10	P@50	MAP	pos	P@1	P@5	P@10	P@50	MAP
baseline	n	0.473	0.370	0.314	0.187	0.084	v	0.364	0.242	0.185	0.082	0.058
mean rank		0.484	0.376	0.318	0.190	0.090		0.380	0.249	0.188	0.085	0.062
har_mean rank		0.488	0.368	0.314	0.191	0.091		0.380	0.261	0.202	0.091	0.067
mean score		0.495	0.389	0.329	0.194	0.093		0.392	0.263	0.199	0.087	0.065

c）sogou2 和 sogou3 集成

表 6　sogou2 和 sogou3 集成实验结果

方法	pos	P@1	P@5	P@10	P@50	MAP	pos	P@1	P@5	P@10	P@50	MAP
baseline	n	0.473	0.370	0.314	0.187	0.084	v	0.364	0.242	0.185	0.082	0.058
mean rank		0.474	0.372	0.315	0.186	0.086		0.365	0.245	0.187	0.083	0.060
har_mean rank		0.474	0.373	0.316	0.188	0.087		0.364	0.247	0.188	0.084	0.061
mean score		0.477	0.374	0.317	0.187	0.087		0.369	0.247	0.188	0.083	0.061

d）Giga、sogou2 和 sogou3 集成

表 7　Giga、sogou2 和 sogou3 集成实验结果

方法	pos	P@1	P@5	P@10	P@50	MAP	pos	P@1	P@5	P@10	P@50	MAP
baseline	n	0.473	0.370	0.314	0.187	0.084	v	0.364	0.242	0.185	0.082	0.058
mean rank		0.494	0.384	0.326	0.196	0.093		0.381	0.249	0.188	0.085	0.063
har_ mean rank		0.489	0.378	0.324	0.197	0.095		0.379	0.265	0.205	0.093	0.069
mean score		0.512	0.401	0.340	0.200	0.098		0.397	0.265	0.201	0.087	0.067

表 4 结果表明：在 Giga 和 sogou2 的组合中，三种集成方法的结果较基准线都有提升，平均分数方法对名词、调和平均排名方法对动词的提升幅度较大。

表 5 结果表明：在 Giga 和 sogou3 的组合中，三种集成方法的结果较基准线都有提升，平均分数方法对名词、调和平均排名方法对动词的提升幅度较大。

表 6 结果表明：在 sogou2 和 sogou3 的组合中，三种集成方法的结果较基准线都有提升，但幅度较小，其中，平均分数的方法效果较好。

表 7 结果表明：在三个相似度序列的组合中，三种集成方法的结果较基准线都有明显的提升，其中以平均分数方法效果最好。

综上四种组合的实验结果，可以发现，集成结果相对于基准线都有提高，其中，调和平均排名方法优于平均排名方法，而平均分数方法效果最好。原因分析如下。

平均排名方法的数值相对粗糙，区分度不大，而调和平均排名方法对顺序更敏感，即相对于在两个序列中都排名中间的词，更重视那些在一个序列中排名靠前而在另一个序列中排名靠后的词。例如一个相似词 w'_i 在两个序列中的排名分别为 50、50，另一个相似词 w'_j 在两个序列中排名分别为 2、99，使用平均排名方法计算，w'_i、w'_j 的新排名分别为 50 和 50.5，集成后 w'_i 排在 w'_j 之前；而使用调和平均排名方法计算，w'_i、w'_j 的新排名分别为 50 和 3.92，集成后 w'_i 排在 w'_j 之后，与平均排名方法刚好相反。实际上，排名越靠前的可靠性更高，所以调和平均排名方法优于平均排名方法。

平均得分方法对得分数值更敏感，在实验数据中，相似度的得分差是呈现由疏到密的过程，例如第 1 名与第 2 名的差距为 0.01，而第 30 名和第 40 名的差距为 0.001，因此使用得分数值，除了与调和平均排名方法一样，能区分相似词的重要性之外，更能符合数据分布，所以效果最好。

e）四种组合方式比较：

将 1)~4)四种组合方式中最好的相似度结果列于下表中，以更直观地比较各组合方式的优劣。

表 8　四种组合方式比较

集成	pos	P@1	P@5	P@10	P@50	MAP	pos	P@1	P@5	P@10	P@50	MAP
baseline	n	0.473	0.370	0.314	0.187	0.084	v	0.364	0.242	0.185	0.082	0.058
Giga+sogou2		0.495	0.392	0.333	0.199	0.095		0.383	0.263	0.203	0.092	0.067
Giga+sogou3		0.495	0.389	0.329	0.194	0.093		0.380	0.261	0.202	0.091	0.067
sogou2+sogou3		0.477	0.374	0.317	0.187	0.087		0.369	0.247	0.188	0.083	0.061
all		0.512	0.401	0.340	0.200	0.098		0.397	0.265	0.201	0.087	0.067

结合表 8 与表 3 数据，发现两个有趣的现象：1) 在表 3 的单一结果中，sogou 的两种窗口方法都优于 Giga，而在表 8 中，Giga 与 sogou 的集成结果却优于 sogou2 和 sogou3 的集成结果；2) 各组合集成的结果相对基准线都有提升，其中三种序的组合提高最多，相比基准线的 P@1，在名词和动词词类上分别提高了 3.9 和 3.3 个百分点，MAP 上提高了 1.4 和 0.9 个百分点。

这两个现象与重合度分析的结果相符，Giga 与 sogou 因为语料语体不同，重合度低但互补性强，所以集成后结果提升较明显，高于单一结果序列中最好的 sogou2；sogou2 与 sogou3 窗口特征不同，但由于语料语体相同，所以重合度较高互补性偏弱，集成后虽较单一结果序列中最好的 sogou2 有提高，但幅度较小提升有限；Giga、sogou2、sogou3 之间都有一定的互补性，所以三者集成在一起的效果最好。

6 结语

本文实验分析了基于大规模语料库的汉语语义相似度计算的方法，着重比较分析了对不同相似度序列集成的方法。基于新闻语料和互联网语料，以一定上下文窗口内的词语、词性及位置作为特征，以互信息（PMI）作为特征权值，构成特征向量，对词语的特征向量求 cosine 夹角值，得到词语间的相似度。本文使用不同语料和不同窗口取值，得到三种词语相似度结果序列，对三种结果序列进行了重合度的分析计算，以各种组合方式集成，主要使用了平均排名、调和平均排名、平均分数三种集成方法。

本文基于《同义词词林扩展版》对实验结果进行了评测，使用了 P@K 和 MAP 两个指标。实验结果表明：集成后的结果比任何单一相似度序列准确率都有提升，集成方法中平均分数方法效果最好，调和平均排名方法优于平均排名方法；集成后对结果的提升幅度与重合度计算结果一致，即重合度越低提升越大，重合度越高提升越小，不同语域之间区别度大重合度低，集成后提升效果更为显著；其中三个序列组合的集成结果最好，因为它们很好地发挥了彼此间的互补作用。

参 考 文 献

[1] James R. Curran. Ensemble Methods for Automatic Thesaurus Extraction. Proc. Of the Conference on EMNLP, 2002.

[2] Lin, D. 1998. Automatic Retrieval and Clustering of Similar Words. In Proceeding of COLING/ACL 1998, 768-774.

[3] James R. Curran and Marc Moens. Improvements in automatic thesaurus extraction. Proc. of the 7th Conference on Natural Language Learning, 2003, 164-167.

[4] Hua Wu and Ming Zhou. 2003. Optimizing synonym extraction using monolingual and bilingual resources. Proc. of the Second International Workshop on Paraphrase: Paraphrase Acquisition and Applications(IWP2003).

[5] 刘群，李素建. 基于《知网》的词汇语义相似度的计算[A]. 第三届汉语词汇语义学研讨会，台北，2002.

[6] Agirre, E., Alfonseca, E., Hall, K., Kravalova, J., Pasca, M. and Soroa, A. 2009. A study on similarity and relatedness using distributional and WordNet-based approaches. In Proceedings of HLT-NAACL, Pp. 19-27.

[7] Harris, Z. 1968. Mathematical structures of language. Wiley, New Jersey.

[8] Julie Weeds, David Weir and Diana McCarthy. Characterising Measures of Lexical Distributional Similarity. COLING '04 Proceedings of the 20th International Conference on Computational Linguistics.

附 录

词语相似度示例（前 10 个相似词）

案犯：疑犯-0.9 嫌疑犯-0.86 嫌犯-0.859 凶犯-0.85 嫌疑人-0.847 犯罪分子-0.842 主犯-0.796 罪犯-0.766 涉案人员-0.749 凶手-0.718

桌子：凳子-0.979 椅子-0.968 茶几-0.936 书桌-0.921 办公桌-0.868 柜子-0.82 沙发-0.819 课桌-0.805 木桌-0.796 方桌-0.78

专刊：特刊-0.962 专版-0.916 专栏-0.883 副刊-0.826 刊物-0.807 简报-0.709 杂志-0.689 会刊-0.68 增刊-0.679 季刊-0.677

事实：现实-0.937 证据-0.856 史实-0.815 事情-0.772 真相-0.763 情况-0.645 情形-0.594 结论-0.586 案情-0.573 事例-0.566

谈天：叙谈-0.893 叙旧-0.834 闲聊-0.769 聊天-0.766 闲谈-0.71 喝茶-0.681 谈心-0.676 拉家常-0.667 聊聊天-0.636 喝酒-0.635

责备：责怪-1.0 责骂-0.817 埋怨-0.804 训斥-0.761 斥责-0.701 怪罪-0.692 责问-0.657 责难-0.65 指责-0.633 数落-0.63

节育：避孕-0.87 晚育-0.773 绝育-0.753 生育-0.716 婚育-0.671 晚婚-0.634 优生-0.548 结扎-0.521 生殖-0.49 防疫-0.465

荣登：蝉联-0.952 问鼎-0.91 雄居-0.819 荣膺-0.811 高居-0.787 位列-0.744 跃居-0.734 夺得-0.703 跻身-0.692 位居-0.66