● 人工智能-自然语言处理

基于统计与词汇语义特征的中文文本蕴涵识别*

刘茂福，李妍，顾进广

(武汉科技大学 计算机科学与技术学院，湖北 武汉 430065)

摘 要: 对中文这种意合型语言而言，为了进行文本内容理解和文本语义推理，必须首先识别文本间的蕴涵关系。针对中文文本，在文本预处理的基础上，提取中文文本的相关统计特征和词汇语义特征；然后基于获取的统计与词汇语义特征，使用支持向量机设计并实现分类器对中文文本对间蕴涵关系进行分类，最后的实验结果表明基于统计与词汇语义特征进行中文文本蕴涵关系识别是可行的。

关键词: 文本蕴涵; 统计特征; 词汇语义特征; 支持向量机; 矛盾

中图分类号: TP391.1 **文献标志码:** A **文章编号:**

Chinese textual entailment recognition using statistical and lexical semantic features

LIU Maof-fu, LI Yan, GU Jin-guang

(*College of Computer Science and Technology, Wuhan University of Science and Technology, Wuhan 430065, China*)

Abstract: In order to further analyze and understand the text, textual entailment recognition should be paid more attention to, especially the ones in Chinese text pair. The statistical and lexical semantic features, associated with Chinese text pair, are extracted after the Chinese text preprocessing, such as Chinese word segmentation and stop words removal. The textual entailment recognition is actually one classification task and the classification model based on support vector machine can be designed and implemented using the extracted statistical and lexical semantic features. The final experiment results demonstrate the effective and feasibility of the classification model using the textual statistical and semantic features.

Key words: textual entailment; statistical feature; lexical semantic feature; support vector machine; contradiction

0 引言

在篇章语言学的研究中，文本蕴涵关系有着非常重要的地位。文本蕴涵的出现主要是为了解决自然语言中的同义异形现象和多样性问题[1,2]，即同一个含义有多个相似语义表达与之对应，需要为这些同义或近义表达建立识别模型。同时，文本蕴涵识别实际上是一个语言基础研究，它在自然语言处理的很多应用中起着关键作用，如问答系统、多文档自动摘要、信息抽取、语义检索以及机器翻译评测等[3-5]。

与数理逻辑中的逻辑蕴涵关系相比，文本蕴涵的定义相对宽松，它实际上是个一种概率蕴涵[1,6]。文本蕴涵目前被普遍认同的定义为：给定文本对（*T*, *H*），在使*T*意义为真的所有可想象的条件下，如果*H* 的意义很大程度上也为真，也就是说*H*可以从*T* 的意义中推断出来，称为*T* 蕴涵*H*。该定义采用了逻辑蕴涵关系的模糊观念，即通过给一个蕴涵关系的实例标定一个概率得分，来评价蕴涵关系是否存在于这个特定的文本对中。

国外的许多学者积极从事文本蕴涵问题的研究，构造了不同的文本蕴涵推理模型和识别模型，并且还举行国际性的竞赛和评测，日本的国家情报研究所NII（National Institute of Informatics）组织的RITE（Recognizing Inference in TExt）就是其中之一，也是目前唯一评测中文文本蕴涵关系的组织[7]。RITE包含二分类BC（Binary-Class）和多分类MC（Multi-Class）两个任务。BC任务判断给定的文本对（*T*, *H*）间是否具有蕴涵关系，即文本*T*是否蕴涵*H*，实际上是一个典型的二分类问题；MC任务判断文本对（*T*, *H*）间是否具有正向蕴涵F（Forward entailment）、逆向蕴涵R（Reverse entailment）、双向蕴涵B（Bidirectional entailment）、矛盾C（Contradiction）以及独立I（Independence）五类关系中的一种，即典型的多分类问题。实际上，矛盾是正向蕴涵或者逆向蕴涵的否定，本质上也跟蕴涵关系有关。例 1 中的句子对都来自RITE的语料，其中*T1*和*H1*间是正向蕴涵关系，*T2*和*H2*间是双向蕴涵关系，而*T3*同*H3*之间则是矛盾关系。

例 1：

T1：与黑龙江与乌苏里江交汇处的抚远三角洲（中方称黑瞎子岛，俄称大乌苏里岛），两国对于该地区归属问题签定《中俄国界东段补充协定》。

H1：黑瞎子岛与大乌苏里岛指的是同一个地区。

T2：海底地震造成地层的大幅度陷落或抬升，是引发大海啸的主要原因。

H2：海底地震造成地层的大幅度下降或上升，是引发大海啸的主要原因。

T3：美国认为伊拉克具有生化武器，甚至可能发展核

收稿日期: **修回日期:**

基金项目：国家自然科学基金（面向自然语言文本生成的事件语义计算研究，Natural Language Generation Oriented Event Semantic Computation，61100133; 汉语文本推理的资源建设与统计分析研究，Resource Construction and Statistical Analysis of Chinese Textual Inferences，61173062）；**国家社会科学基金重大项目**（基于本体演化和事件结构的语义网模型研究，Semantic Web Model Based on Ontology Evolution and Event Structures，11&Z189）；**湖北省教育厅人文社科基金重点项目**（基于篇章语境认知的事件语义多面体构建，Event Semantic polyhedron Construction Based on Context Cognition，2011jyte126）

作者简介: 刘茂福 (1977－)，男，山东省单县人，博士，副教授，CCF 高级会员(E200013837S)，主要研究方向为自然语言处理、文本推理；李妍(1987-－)，女，湖北省武汉市人，硕士研究生，主要研究方向为自然语言处理；顾进广(1974－)，男，湖北省武汉市人，博士，教授，主要研究方向为语义网技术。E-mail: liumaofu@wust.edu.cn

武器。

H3：美国认为伊拉克具有核武器，甚至可能发展生化武器。

文本蕴涵关系识别本质上是分类问题，可以采用机器学习中常用分类方法来完成。分类前，首先要从文本对中获取文本特征，该文提到的中文文本蕴涵关系识别系统共混合了 11 种特征，其中包括 6 种统计特征和 5 种词汇语义特征。

1 系统描述

中文文本蕴涵关系识别系统由数据预处理、特征提取和 SVM 分类器三个主要模块组成，具体的系统结构如图 1 所示。

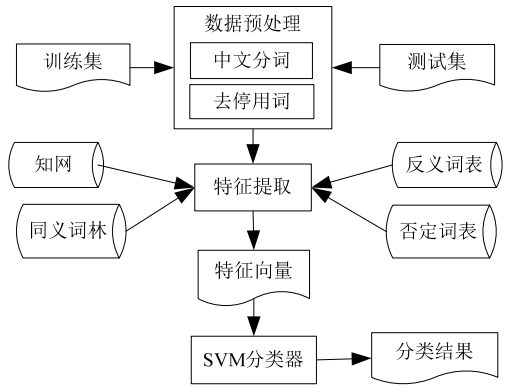

图 1 系统体系结构图

对于中文的书写习惯和标准而言，词间没有像英语中用于词间隔的空格，要想获取词这一信息单元，必须首先进行分词。因此，中文文本数据预处理主要包括对文本的分词处理和停用词过滤，实验中使用的文本分词工具来自中国科学院的ICTCLAS[1]，针对例 1 中的*T2* 和*H2*，做了分词处理和停用词过滤后，其结果如下所示，其中词间使用了分隔符“|”。

T2： 海底|地震|造成|地层|大幅度|下降|上升|引发|大海|啸|主要|原因

H2： 地震|造成|地层|大幅度|下降|上升|引发|大海|啸|主要|原因

本文使用支持向量机来解决文本蕴涵识别问题[8.9]。SVM最开始提出时就是用于解决二分类问题，因此可以直接用于RITE的BC任务。而对于RITE的MC任务而言，解决方法的主要思想是将一个多分类问题器分解成若干个二分类问题，从而可以使用多个二分类器模拟实现多分类器。该文选用“一对一”方法来完成多分类，如图 2 所示。

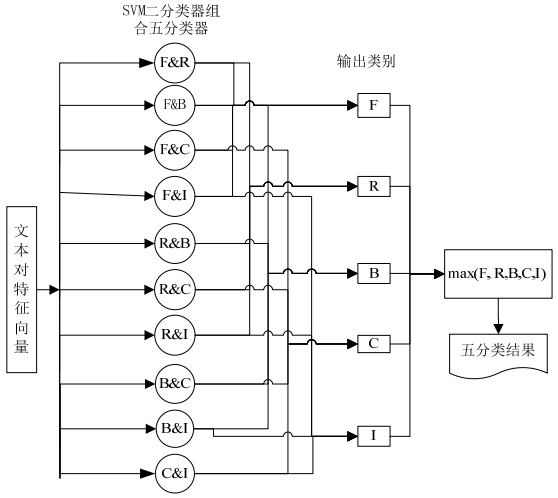

图 2 MC 任务 SVM 多分类器

在图 2 中，对于 RITE 的 MC 任务的五种类型的关系，需要将五类关系两两组合，从而需要训练 10 个二分类器。测试时，将测试数据对每个二分类器进行预测，分别得到一个类别，找出投票数最多的那个类作为最终的分类结果。

该文的中文文本蕴涵关系识别系统使用LIBSVM[2]，核函数使用径向基函数，针对BC和MC任务的训练集分别进行分类器训练，其中要训练的参数包含惩罚因子 C 和核函数因子 γ。

2 特征提取

对文本对进行预处理后，在进入基于支持向量机的分类模块前，还需要基于文本对提取相关的文本特征。中文文本蕴涵关系识别系统共使用了 11 种特征，其中统计特征 6 种，词汇语义特征 5 种。

2.1 统计特征

在 6 种统计特征中，有基于词汇的，如词重叠度特征；有基于文本长度的，如文本长度差和长度比特征等；有文本距离的，如 Jaro-Winkler 距离等；也有基于文本相似度的，如文本余弦相似度特征等。

（1）词重叠度

假定 *T* 和 *H* 中出现的相同词汇越多，*T* 和 *H* 的相似度越高，它们表示相同或相近意义的概率就越大。因此，使用词重叠度特征来表示 *T* 和 *H* 包含相同词汇的程度。

$$\mathrm{W}_{overlap}(T,H)=\frac{|\mathrm{Words}(T)\bigcap\mathrm{Words}(H)|}{|\mathrm{Words}(T)\bigcup\mathrm{Words}(H)|} \quad (1)$$

其中 $\mathrm{Words}(T)$ 是文本 T 包含的词汇集合。

（2）长度差

如果 *T* 蕴涵 *H*，那么 *T* 包含的信息量应该多于 *H*；在表达上，可以假定 *T* 要比 *H* 长。因此，使用 *T* 与 *H* 的长度差从表面上度量两者信息量的差异。

$$\mathrm{LS}(T,H)=|\mathrm{Len}(T)-\mathrm{Len}(H)| \quad (2)$$

其中函数 $\mathrm{Len}(T)$ 用来计算文本 T 的长度。

（3）长度比

相对于长度差，如果两者的长度差异太大，那么两者的相似度势必会降低，因此使用长度比来调和这一矛盾。

$$\mathrm{LR}(T,H)=\frac{\mathrm{Len}(T)}{\mathrm{Len}(H)} \quad (3)$$

（4）Jaro-Winkler 距离

Jaro-Winkler 距离是针对两字符串相似度的一个度量，Jaro-Winkler 值越大，表明两字符串的相似度越高。Jaro-Winkler 尤其适合短字符串相似度的度量，如专有名词中的人名、地名等。

$$\mathrm{JW}_{dis}(T,H)=\frac{m}{3*\mathrm{len}(T)}+\frac{m}{3*\mathrm{len}(H)}+\frac{m-t}{3*m}$$

$$\mathrm{L}_{JW}(T,H)=\frac{\max(\mathrm{len}(T),\mathrm{len}(H))}{2}-1 \quad (4)$$

其中 m 是文本 *T* 和 *H* 匹配的文本串的个数，这里“匹配”的含义是同一个文本串在指定的 L_{JW} 长度范围内同时出现在文本 *T* 和 *H* 中。

（5）余弦相似度

两文本的向量余弦相似度越高，它们之间存在蕴涵关系的概率就越大。在向量空间中，两文本的向量余弦相似度可以使用公式（5）来计算。

$$\mathrm{Sim}_{cos}(\mathbf{t},\mathbf{h})=\frac{\sum_{i=1}^{n}t_i*h_i}{\sqrt{\sum_{i=1}^{n}t_i}*\sqrt{\sum_{i=1}^{n}h_i}} \quad (5)$$

[1] http://ictclas.org/

[2] http://www.csie.ntu.edu.tw/~cjlin/libsvm/

其中 $\mathbf{t}$ 和 $\mathbf{h}$ 是相对于文本 *T* 和 *H* 的向量，*n* 是向量的维度，每个分量使用传统的 TF*IDF 方法计算。

（6）最长公共子串相似度

两文本存在的最长公共子串占比越大，它们之间存在蕴涵关系的可能性就越大。文本对最长公共子串相似度可以使用公式（6）来计算。

$$\mathrm{Sim}_{LCS}(T,H)=\frac{\mathrm{len}(\mathrm{LCS}(T,H))}{\min(length(T),length(H))} \quad (6)$$

其中 $\mathrm{LCS}(T,H)$ 用来计算文本 *T* 和 *H* 的最长公共子串。

2.2 词汇语义特征

在 5 种词汇语义特征中，有基于词义计算文本相似度的，也有为了准确判断矛盾关系的反义词特征和否定词特征。统计特征中的词重叠度只考虑了文本对中相同的词汇，而忽略了同义词和近义词，词汇语义特征中的知网语义相似度和同义词林相似度考虑了文本对中词汇语义上的同义和近义。词汇语义特征中的反义词特征和否定词特征主要用来判定文本对间的矛盾关系。

（1）基于知网语义相似度

文本对间词汇的语义相似度越高，它们之间存在蕴涵关系的可能性越大。文本 *T* 和 *H* 基于知网的语义相似度特征可以使用公式（7）来计算得到。

$$\mathrm{LS}_{Hownet}=\frac{1}{2}\left(\frac{1}{m}\sum_{i=1}^{m}\max\{\mathrm{sim}_w(w_{1i},w_{2j})\,|\,1\le j\le n\}+\frac{1}{n}\sum_{i=1}^{n}\max\{\mathrm{sim}_w(w_{1i},w_{2j})\,|\,1\le i\le m\}\right) \quad (7)$$

其中 $\{w_{1i}\,|\,1\le i\le m\}$ 和 $\{w_{2j}\,|\,1\le j\le n\}$ 分别用于表示文本 *T* 和 *H* 的词汇集，$\mathrm{sim}_w(w_1,w_2)$ 则用来计算两个词之间的语义相似度。

（2）基于同义词林语义相似度

假设文本 *T* 和 *H* 中的同义词可以提高识别双向蕴涵关系的识别率，它们基于同义词林的语义相似度特征可以使用公式（8）来计算得到。

$$\mathrm{LS}_{Cilin}=\frac{1}{2}\left(\frac{1}{m}\sum_{i=1}^{m}\max\{\mathrm{sim}_w(w_{1i},w_{2j})\,|\,1\le j\le n\}+\frac{1}{n}\sum_{i=1}^{n}\max\{\mathrm{sim}_w(w_{1i},w_{2j})\,|\,1\le i\le m\}\right) \quad (8)$$

其中词义相似度 $\mathrm{sim}_w(w_1,w_2)$ 的计算采用文章[10]中提到的方法。

（3）反义词特征

文本 *T* 和 *H* 中反义词对在一定程度上可以反映出两者的矛盾关系，利用反义词特征可以提高矛盾的识别率，系统利用公式（9）来计算两文本间的反义词特征。

$$f_{antonym}=\begin{cases}1 & (n\neq 0)\\ 0 & (n=0)\end{cases} \quad (9)$$

其中 *n* 为文本 *T* 和 *H* 间出现的反义词对的数目。

（4）否定词特征

除反义词对外，文本 *T* 和 *H* 中出现的否定词也可以在一定程度上反映出两者间的矛盾关系，系统使用公式（10）来计算两文本间的否定词特征。

$$f_{neg}=\begin{cases}0 & (n1\%2=n2\%2)\\ 1 & otherwise\end{cases} \quad (10)$$

其中 $n1$ 和 $n2$ 分别指文本 *T* 和 *H* 中否定词的数目。

（5）词义重叠比

在下面的例 2 中，文本 *T4* 和 *H4* 相同的词汇并不多，但这种类似的文本对往往被判断为双向蕴涵，而它们之间真正的关系却是“独立”，即没有任何蕴涵关系存在。

例 2：

T4：网易首席技术总裁丁磊。

H4：丁磊是网易公司首席构架师。

为了解决这一问题，提出了词义重叠度这一词汇语义特征，如公式（11）所示。

$$\mathrm{CWR}(T,H)=\frac{|\mathrm{SW}(\mathrm{Words}(T),\mathrm{Words}(H))_{similarity=1}|}{\min(\mathrm{len}(T),\mathrm{len}(H))} \quad (11)$$

其中 $\mathrm{SW}(\mathrm{Words}(T),\mathrm{Words}(H))_{similarity=1}$ 为出现在文本 *T* 和 *H* 中相同词和同义词集合，同义词基于同义词林来认定。

3 实验结果与分析

RITE 的 BC 和 MC 任务的训练集和测试集都是 407 对文本，使用训练得到的支持向量机分类器对测试集进行分类预测，系统的整体性能使用正确率（Accuracy）指标来进行度量，其具体计算如公式（12）所示。

$$Accuracy=\frac{1}{\#\,pairs}\sum[output\ \ label\ \ is\ \ correct] \quad (12)$$

其中 $\#\,pairs$ 指 RITE 任务测试集中文本对的数目。

针对具体的文本蕴涵关系的类别，该文给出每一类别的准确率（Precision）和召回率（Recall）来进行度量，其具体计算如公式（13）和（14）所示。

$$Precision=\frac{TP}{TP+FP} \quad (13)$$

$$Recall=\frac{TP}{TP+FN} \quad (14)$$

其中 TP（True Positives）和 FP（False Positives）分别指测试集中该类所有正例被正确和错误分类的数目；而 FN（False Negatives）则指测试集中该类所有负例被错误分类的数目。

该文的基于统计与词汇语义特征的中文文本蕴涵关系识别系统标记为 WUST，该系统利用 BC 和 MC 任务的训练集对分类器进行训练。在两个任务的训练中，惩罚因子 C 的 log 值都为 10，而核函数因子 γ 的 log 值皆为-9；BC 任务的训练正确率为 76.9%，MC 任务训练正确率为 57.74%。WUST 系统给出了对 BC 和 MC 任务测试集的分类结果，RITE 组织方最终的评测结果如表 1 所示。

表 1　RITE 组织方 BC 和 MC 任务最终评测结果

BC Runs	**Accuracy**
UIOWA-CS-BC-01	0.9750
UIOWA-CS-BC-03	0.9631
UIOWA-CS-BC-02	0.9361
ICRI_HITSZ-CS-BC-03	0.7764
FudanNLP-CS-BC-02	0.7617
ICRI_HITSZ-CS-BC-02	0.7568
FudanNLP-CS-BC-01	0.7469
WHUTE-CS-BC-03	0.7371
NTU-CS-BC-01	0.7346
WHUTE-CS-BC-02	0.7322
WUST-CS-BC-01	0.7248
Baseline	0.7617
MC Runs	**Accuracy**
UIOWA-CS-MC-01	0.8919
UIOWA-CS-MC-02	0.8919
UIOWA-CS-MC-01	0.8870
ICRI_HITSZ-CS-MC-03	0.6413
ICRI_HITSZ-CS-MC-02	0.6241
ZSWSL-CS-MC-02	0.6192
WHUTE-CS-MC-02	0.6093
III_CYUT_NTHU-CS-MC-02	0.5897
FudanNLP-CS-MC-02	0.5848
WHUTE-CS-MC-01	0.5823
WUST-CS-MC-01	0.5823
Baseline	0.5315

表 1 只列出了比该文系统评测结果好的系统编号，RITE 中文文本蕴涵关系识别的 BC 和 MC 任务共有 33 个系统。从表 1 可以看出，WUST 系统初始评测结果的 BC 和 MC 任务皆处于第 11 位。UIOWA 系统结果明显高于其它系统，最主要的原因是 UIOWA 系统使用了众包（Crowdsourcing）方法，有人工参与其中。该文系统评测结果同基准测试结果的对比如图 3 所示。

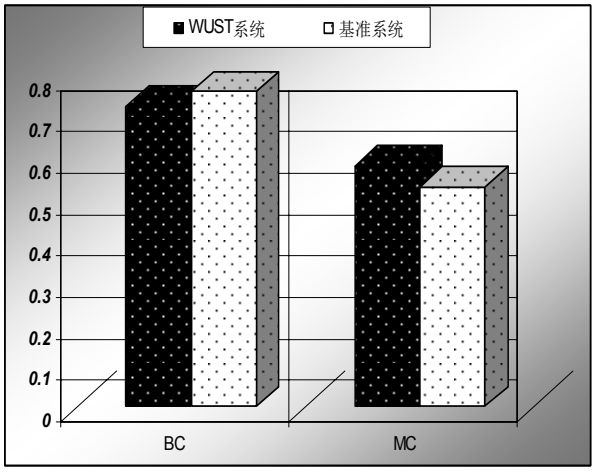

图 3 WUST 系统的总体评测结果

WUST和很多其它小组的系统在BC任务的表现不如基准测试，其中最可能的原因是基准测试使用的是字符重叠度特征[7]，而考虑到是对中文进行处理，WUST系统自然而然的使用了词重叠度特征。另外， WUST系统中BC和MC任务都使用了相同的特征集，而BC和MC任务还是有本质区别的，针对BC任务，将来应多考虑二值特征。

RITE 组织方为该文的中文文本蕴涵关系识别系统返回的混淆矩阵（Confusion Matrix）如图 4 所示。

---- Confusion Matrix ----

Cols: true labels
Rows: system labels

	Y	N
Y	223	72 | 295
N	40	72 | 112
	263	144 |

---- Confusion Matrix ----

Cols: true labels
Rows: system labels

	F	R	B	C	I	
F	69	0	3	6	15	93
R	0	75	4	5	21	105
B	17	9	63	54	8	151
C	3	1	1	7	3	15
I	12	6	0	2	23	43
	101	91	71	74	70	

图 4 BC 和 MC 任务的混淆矩阵

使用混淆矩阵可以计算每个类别的准确率和召回率，BC 任务的结果在表 2 中，而 MC 任务的结果见表 3。

表 2 BC 任务各类别评估结果

类标签	准确率	召回率
Y	0.75593	0.8480
N	0.6429	0.5

表 2 中类标签“Y”表示文本对间有蕴涵关系，而类标签“N”则表示没有蕴涵关系。从表 2 可以发现针对类标签“N”，召回率明显低于准确率，表明 WUST 系统在分类时，将更多的测试集里不具有蕴涵关系的文本对误判成了蕴涵关系。

表 3 MC 任务各类别评估结果

类标签	准确率	召回率
F	0.7419	0.6832
R	0.7143	0.8242
B	0.4172	0.8873
C	0.4667	0.0946
I	0.5349	0.3286

从表 3 可以发现，WUST 系统在正向蕴涵 F、逆向蕴涵 R 的判断上表现都不错；但双向蕴涵 B 却准确率远远低于召回率，表明测试集中很多不具有双向蕴涵的文本对被 WUST 系统误判成了双向蕴涵；跟其它类别相比，矛盾关系 C 的准确率和召回率都很低，可能因为 WUST 系统中针对矛盾关系识别的特征还远远不够，这跟 BC 任务的结论类似。在表 3 中，还可以发现 21.4%和 30%的独立 I 分别被错判为正向蕴涵和逆向蕴涵，更有 11.4%的独立被误判成双向蕴涵；这样，共有近 62.8%的独立被误判成蕴涵关系。

对 WUST 系统进一步的检测发现，MC 任务中，在基于知网和同义词林计算文本相似度时没有将变量及时归零，导致这两个特征值的结果错误，修改了这个错误后，MC 的评测结果如表 4 所示。

表 4 修改程序错误后 MC 任务各类别评估结果

类标签	准确率	召回率	系统正确率
F	0.7865	0.6931	
R	0.7080	0.8791	
B	0.4248	0.9155	0.6069
C	0.5	0.0541	
I	0.6364	0.4	

修改了错误后，除逆向蕴涵指标稍有降低外，其它四类的测试结果都有明显提高，系统的整体正确率也有改善。

4 结束语

数理逻辑中的推理理论基础是逻辑蕴涵关系，要分析与理解人类用于交流思想的书面文本，研究文本内容的句间、段间甚至文本间的蕴涵关系至关重要。文本蕴涵关系具体表现为文本正向蕴涵、逆向蕴涵、双向蕴涵（也称复述）、矛盾等。

该文使用如词重叠度、长度比、长度差、距离等文本统计特征以及否定词、反义词、文本相似度等词汇语义特征，针对 RITE 任务提供的中文文本训练集和测试集，基于支持向量机学习方法设计并实现分类模型。从 RITE 组织方对 WUST 系统的评测结果看，基于统计特征与词汇语义特征进行中文文本蕴涵关系识别是可行的。

对实验结果的进一步分析发现，RITE 的二分类问题和多分类问题应该使用一些不同特征，尤其是二分类问题，可以尝试增加二值特征；同时，针对矛盾关系，应该使用更多与之相关的特征来提高其准确率和召回率。

参考文献:

[1]DAGAN I., DOLAN B., MAGNINI B. And ROTH D. Recognizing textual entailment: rational, evaluation and approaches [J]. Natural Language Engineering, 2009. 15(4):1-17.

[2]IFTENE A., and BALAHUR-DOBRESCU A. Hypothesis transformation and semantic variability rules used in recognizing textual entailment[C]. In Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing, 2007, Prague, Czech Republic, 125–130.

[3]ANDROUTSOPOULOS I. and MALAKASIOTIS P. A survey of paraphrasing and textual entailment methods[J]. Journal of Artificial Intelligence Research, 2010, 38:135–187.

[4]DUBOUE P. A. and CHU-CARROLL J. Answering the question you wish they had asked: The impact of paraphrasing for question answering[C]. In Proceedings of the HLT Conference of NAACL, 2006, 33-36.

[5]NIELSEN R., WARD, W. and MARTIN J. Recognizing entailment in intelligent tutoring systems[J]. Natural Language Engineering, 2009, 15(4):479–501.

[6]YUAN Y. L. and WANG M. H. The Inference and Identification Models for Textual Entailment[J]. Journal of Chinese Information Processing, 2010, 24(2):3-13. [袁毓林, 王明华. 文本蕴涵的推理模型与识别模型[J]. 中文信息学报, 2010, 24(2):3-13.]

[7]SHIMA H., KANAYAMA H., LEE C.-W., LIN C.-J., MITAMURA T., MIYAO S. S. Y., and TAKEDA K. Overview of ntcir-9 rite: Recognizing inference in text[C]. In Proceedings of NTCIR-9 workshop meeting, Tokyo, 2011, 291-301.

[8]ZANZOTTO, F. M., PENNACCHIOTTI, M., and MOSCHITTI, A. A machine-learning approach to textual entailment recognition [J]. Natural Language Engineering, 2009, 15(4):551–582.

[9]MALAKASIOTIS, P. Paraphrase recognition using machine learning to combine similarity measures[C]. In Proceedings of the 47th Annual Meeting of ACL and the 4th International Joint Conference on Natural Language Processing of AFNLP, Singapore, 2009, 27-35.

[10]TIAN J. L. and ZHAO W. Words Similarity Algorithm Based on Tongyici Cilin in Semantic Web Adaptive Learning System[J]. Journal of Jilin University (Information Science Edition), 2010, 28(6):602-60. [田久乐, 赵蔚. 基于同义词词林的词语相似度计算方法[J]. 吉林大学学报(信息科学版), 2010, 28(6):602-608.]

联系方式： liumaofu@wust.edu.cn

邮编： 430065

地址：湖北省武汉市武汉科技大学黄家湖校区计算机科学与技术学院

联系人： 刘茂福

联系电话： 18971374322