基于语义角色标注的新闻领域复述句识别方法

吴晓锋，宗成庆

（中国科学院 自动化研究所 模式识别国家重点实验室，北京 100190）

摘 要：复述（Paraphrase）句的识别可看作文本蕴含（Text Entailment）识别的一个子问题，传统的解决方法是通过词频或句法上的相似度来判断。可是哪怕用相同的文字书写的句子其含义也可能差别很大，而相同句法结构也不能保证意义一致。本文根据新闻语料的特点，提出了一种通过引入深层的语义角色标注来帮助识别新闻领域复述句的方法。该方法通过在语义角色这种结构化的含义表达形式中提取的特征来弥补传统方法的不足：先识别待判断的两个句子中所有谓词的语义角色，然后计算两个句子间对应语义角色的相似度，最后结合传统的句子相似度计算方法来进行相似性计算。实验证明，本文提出的方法能有效地提高复述语句的识别效果。

关键词：复述识别；语义角色标注；自然语言处理

中图分类号：TP391　　**文献标识码**：A

An Approach to News Paraphrase Recognition Based on SRL

Xiaofeng Wu,　Chengqing Zong

(National Lab of Pattern Recognition, Institute of Automation, CAS, Beijing 100190, China)

Abstract: Paraphrase Recognition can be regarded as a sub-problem of Text Entailment Recognition. This problem is hard in that simply using term frequency or syntax information is prone to error judgment. For even the same pack of words can cook up sentences with totally different meanings, while similar parsing trees can either have different meanings.　In this paper we present a new approach based on Semantic Role Labeling (SRL) to identify paraphrase. In our approach, we first label sentences with semantic role, then we get features that can partly represent the meaning of the sentence. By doing so, we also take the specialty of News sentences under consideration. Our experiment proved the effectiveness of our approach.

Key words: natural language processing, semantic role labeling, paraphrase recognition

1　引言

从某种意义上讲，复述（Paraphrase）可以看作文本蕴含（Text Entailment）的一个子问题。对于两个语言片段（短语、句子或篇章）A 和 B，如果能从 A 的语义中推理出 B，那我们说 A 蕴含 B，反之说 B 蕴含 A，而复述则可以看作 A 蕴含 B 且 B 也蕴含 A。

最早研究复述的文献见于[1]，复述在自然语言处理中有许多应用：如在机器翻译中可以借鉴复述识别中的技术处理实时处理遇到的未登录短语[2-5]；在自动问答系统中用于识别多种问句形式来提高系统性能[6]；在多文档自动摘要系统中用于句子生成、压缩、相似句子识别等等[7-8]。

作为一个独立的问题提出，句子级的复述识别（Paraphrase Recognition）一般指的是对于给定的两个句子，判断其是否在语义上一致，见下面的例子：

(1) *Amrozi accused his brother, whom he called "the witness", of deliberately distorting his evidence.*

(2) *Referring to him as only "the witness", Amrozi accused his brother of deliberately distorting his evidence.*

(1)和(2) 两个句子是互为复述句，直观的看，虽然单从用词的重合度上看这两个句子很类似，但从句法角度讲它们差别很大；

(3) *Armstrong, 31, beat testicular cancer that had spread to his lungs and brain.*

(4) *Armstrong, 31, battled testicular cancer that spread to his brain.*

(3)和(4)不是复述句，用词重合度也比较大，并且句法也类似，但是第一句包含有一些第二句没有的信息。

本文的工作主要致力于句子级别的复述识别方法研究。我们提出了如下方法：采用经过语义角色标注后的信息为特征，然后通过机器学习算法来识别复述句。虽然文献[9]也曾通过语义角色标注来识别复述句，但本文将从一个新的角度来获取特征,并考虑了新闻语句本身的特点。实验证明，本文的方法能够取得满意的结果。

论文其余部分如下组织：第二部分介绍前人在复述识别上的相关工作，第三部分介绍本文的设计思路及语义角色工具包 SRL，第四部分介绍相关语料和实验结果，最后一部分为本文的结论。

2　相关工作

我们将复述识别的方法大致分为三类，一种是基于词集信息的，第二种是基于句法信息的，第三种是基于深层语义的。下面从这三个方面简要介绍前人的相关工作。

2.1　基于词集信息的方法

虽然复述问题的提出是鼓励学者从语义角度寻找判断两个句子是否可以相互替代的方法，但因为语义角色标注正确率有待提高，所以从实用的角度，还是有许多学者尝试用表层信息来解决这个问题。基于词集的方法中向量模型是最主要的方法，这种方法广泛应用于文本分类，信息检索，以及句子相似度计算，在此方法上的改进一般包括：去停用词、词干化（stemming）、POS 标注以及词义扩展等。

文献[10]在词集的基本思想下，引入基于 WordNet[11]的世界知识，并且将传统的词到词的相似性计算方法扩展到文本到文本。其方法是将两段文本的相似性定义成了文本中名词和动词的加权函数，以及从别的语料中得到的倒排文档频率。该作者在有监督和无监督条件下分别测试了该扩展方法，取得了比较明显的效果。

在机器翻译中的自动评测启发下（以 BLEU[12]为例，其核心思想是计算 n 元文法匹配的对数几何平均），文献[13]采用机器翻译中常用的自动评测机制 BLEU、NIST、WER 以及 PER 来提取特征对分类器进行训练，并进行句子的语义相似性计算。

另外有基于隐含语义标题（LSI）[12]的计算相似度方法，但 LSI 缺乏合理的物理解释，并且计算复杂度较高。

基于词集信息的方法往往简单、实用，但是很明显，这样做违背了复述这个问题提出的初衷。单纯采用词集信息，不能也肯定不会达到很好的效果。

2.2 基于句法信息的方法

两个意思相近的句子虽然有可能在句法上有差异，但是其内部子结构往往能找到很多相似之处，而且，依存句法和语义表示有一定的相似性，所以有不少学者试图在句法层面上寻找复述问题的解决办法。

文献[14]提出了使用基于反向转换文法（ITG）[7]为复述问题建模的方法。简单讲，ITG 实际上是面向双语的上下文无关文法，其目的是让双语平行语料分析的鲁棒性最大化，而不是验证语料的合法性，从而应用于平行语料的标注，包括划界、对齐、切分等[12]。借助 ITG 模型，文献[14]没有采用任何外部知识，而是仅仅凭借句法层面上的相似就取得了不错的效果。

文献[15]提出了一种基于*词汇—句法*的方法，该方法要求判断蕴含时，不但要比较两句话词汇上的一致性，还要比较句法上的一致性。这种*词汇—句法*关系包括由词形变化引起的句法改变、动词的被动到主动的变化、共指等等。通过实验，[15]证明*词汇—句法*方法优于仅仅基于词汇的方法。

文献[16]系统归纳了 17 个特征用于训练分类器来识别复述句。这 17 个特征中前 9 个为词汇特征，10-15 为依存句法特征，16-17 为句长特征。

文献[17] 的理论假设是：如果两个句子是依存句，则它们的句法树虽然可以允许有不同，但总体应该对齐的比不是依存句的好。[17]结合词汇、句法信息，采用准同步依存文法（quasi-synchronous dependency grammar）建模。

文献[18]采用了依存句法分析将被动句式变为主动句式，然后采用如编辑距离、词汇相似度、以及类似于 BLEU 的 n 元文法作为特征进行训练。

从句法的角度尝试复述问题比单纯从词汇的角度有说服力，而且效果也普遍比后者好。可是句法层毕竟离语义层还有很大差距，使用句法特征也都只能集中在句法的某个片段上，无法从整体句法树上寻找句子间的相似性。

2.3 基于深层语义分析的方法

文献[9]是基于语义角色标注的复述识别方法，文中不再单纯用词片段或句法树片段作为信息单位，而是用基于 PropBank[19]的谓词论元结构（Predicate Argument Structure, 结构化地表示一个动词谓词及其参量——也就是论元，见图 1）。这种结构可以很好地表示动作、概念、及其相互关系。而单纯用动词、名词来表示这个含义有可能面临相同的词集但却不同意义的情况（如：“猫吃老鼠”和“老鼠吃猫”；意义正好相反）。

predicate: 吃	predicate: 吃
arg0: 猫	arg0: 老鼠
arg1: 老鼠	arg1: 猫

图 1 谓词论元结构

文献[9]的系统流程图见图 3。首先将句子对用 Charniak 句法分析器[20]进行句法分析，并用语义角色标注工具 ASSERT[21]进行语义标注，找到句子里的谓词论元结构；然后通过贪婪算法寻找匹配的谓词论元结构，在这个过程中用到了外部词库；没有匹配上的其它成分被送进一个有监督分类器，来判断这些成分是否为重要成分，并做出是否这两个句子互为复述的判断。

送入分类器的特征有两类：一类是句法树路径特征，指的是在句法树上的未匹配的单元与最近的有公共父节点的匹配单元之间的距离。[9]的作者认为这个特征可以在一定程度上表征这个未匹配部分的重要性。另一类是谓词论元结构中谓词，也就是动词之间的相似性。

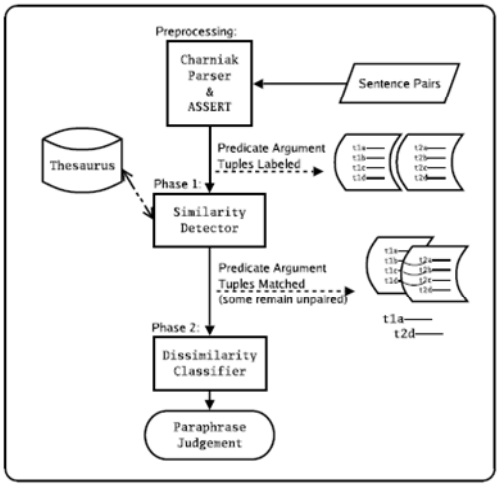

图 2 文献[9]方法流程

基于深层语义分析的方法相对较少，其原因之一是语义角色标注本身正确率有待提高。虽然如此，从语义层考虑复述问题，毕竟最合乎逻辑。我们认为，要想做到真正实用的句子复述识别，必须要从这个角度着手加以研究。

3 论文动因及设计思路

3.1 论文动因

我们分析了文献[9]中采用的方法，认为其主要针对情况是论元互换（类似于图 1），以及主动语态和被动语态等，这些现象从句法分析以及词汇分析上不易得出正确的语义。因为 ASSERT 语义角色工具包只能保证谓词的 arg0 和 arg1 的一致性，所以其方法仅仅能识别最基本的谓词以及这个谓词的发起者和接收者（arg0,arg1），并没有利用更多的语义角色标注信息识别句子里的其它成分。然而现实中很多复述问题要面临的情况非常复杂，比如一个单纯的补语或状语的不同，就有可能导致两个极为相似的句子不称其为复述：如“我在家吃饭”和“我在学校吃饭”。这些情况恰恰在新闻语料中非常常见，也是新闻语料的一个特点。

为了得到更丰富的语义角色，我们选用伊利诺伊斯大学认知计算组（Cognitive Computational Group，UIUC）的语义角色标注工具包 SRL。SRL 工具包有比较好的语义角色标注效果，对时间、地点、数

字、指代等都有很好的识别结果，其标注信息见表 1。

标注	解释	标注	解释
A0	主语	**A1**	宾语
A2	间接宾语		
AM-DIR	方向	**AM-DIS**	篇章标记
AM-EXT	范围	**AM-LOC**	位置
AM-MOD	一般修饰	**AM-NEG**	否定
AM-MNR	举止	**AM-PRD**	第二谓词
AM-PRP	提议	**AM-REC**	反义
AM-TMP	时间	**AM-ADV**	副词修饰

表 1 RSL 中的标注

图 3 给出了采用 SRL 标注好的两个复述句。这两句话均来自微软的复述语料库（Microsoft Paraphrase Corpus，MSR[22]）。以图 3a 为例，SRL 针对每一个动词（包括 publish，offer，add）给出了谓词论元结构，也就是着色的三列，其中不同的颜色代表不同的论元。

如动词 publish，其主语 A0 为 They，宾语 A1 为 an advertisement on the internet。另外 SRL 还清楚地找到了时间状语 on July 4，以及状语 offering cargo for sale。

图 3b 给出了这句话的一个复述。可以看到，虽然这两句话句法结构有较大差异，但是在 publish 的论元中除了 A0 不一致外，其它语义角色都一致。

MSR 语料均来自于不同的新闻文稿，所以，虽然在 a 句中出现了无法消解的代词 They，以及后面的代词 he，但是因为新闻背景关系，在其它语义角色相同或大致相同的情况下，还是认为这两句话是复述——这在 MSR 语料中是很常见的现象。

从新闻语句自身特点看，很多新闻句可能主、谓、宾都一致，但是不同的时间、不同的地点以及不同的宾补成分，都可能使得这两句看似相似的句子实际上在诉说不同的事情。分析这种情况，我们认为，单纯从由谓词、A0、A1 论元组成的元组为单位寻找匹配并不能很好地解决新闻语句的复述识别问题，句子的各种修饰成分在复述句识别中同样占有很重要的作用。

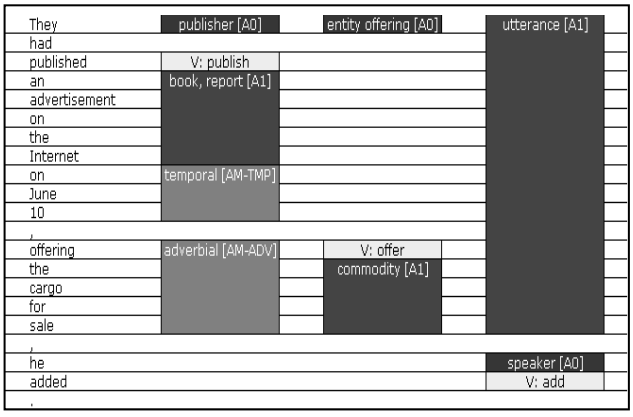

图 3 a SRL 例句

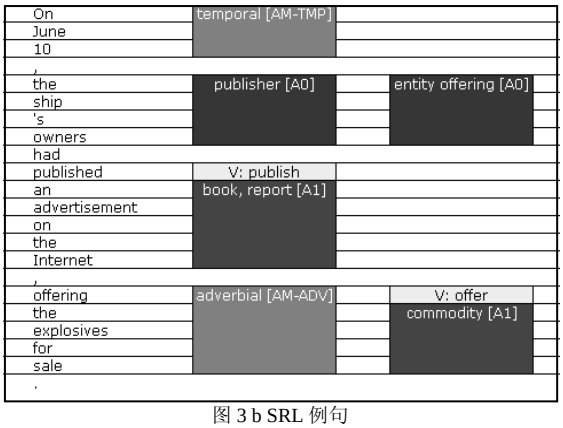

图 3 b SRL 例句

3.2 设计思路

复述句识别方法多数采用有监督的训练方法，事实上几乎所有人的工作都集中在如何找到最能反应这个问题本质的特征上。在文献[9]的基础上，我们针对新闻复述句的特点，提出了一种基于语义角色标注的有监督新闻复述句识别方法，其流程见图 4，下面介绍每个模块的功能，并给出选用的特征。

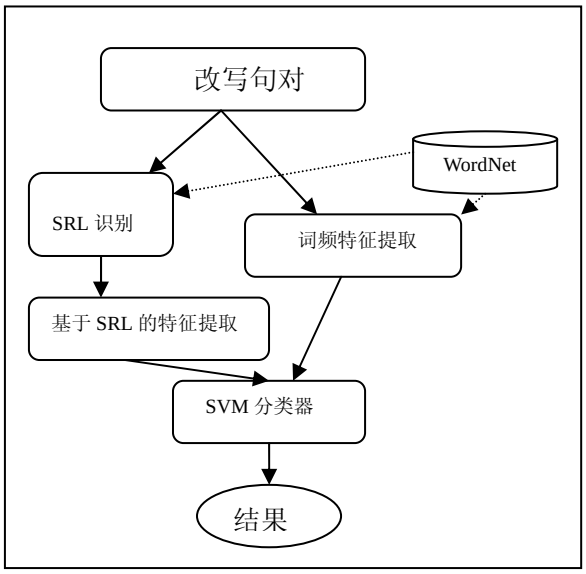

图 5 系统流程

首先将复述句输入词频特征提取模块。在词频特征提取模块之中，我们首先对句子进行了去停词、词干化的工作，然后我们考虑了如下几个特征，我们称之为基本特征：

基本特征（BF）

BF1 句长差特征：两个句子长度相减所得的差，可以为正值或负值。

BF2 绝对句长差特征：对句长差特征取绝对值。

这两个特征参考了文献[16]。

BF3 句子相似度特征：我们用如下公式(1)计算相似度

$$Sim_{S_1S_2} = \frac{2\bullet(S_1 \cap S_2)}{|S_1|+|S_2|} \qquad (1)$$

BF4 基于 WordNet 的相似度特征：对公式（1）中的分子，我们对词干化以前的词采用 WordNet3.0 计算相似度，为句子中的每个词寻找其对应复述句中相似度分值最高的词，然后将这些相似度叠加并乘以 2。

同时我们也将句子输入 UIUC 的 SRL 标注工具模块，进行语义角色标注。得到标注好语义角色的句子后，将其输入 SRL 特征提取模块，我们提取了如下一些特征，我们称之为语义特征。

语义特征（SF）

SF1 谓词数差特征：两句话的谓词数相减所得，可以为正值或负值。

SF2 匹配的谓词数量特征：两句话中有多少谓词可以匹配。我们经验地规定两个谓词在 WordNet 上的相似度超过 0.5 即为匹配成功，我们最多寻找三对最相似的谓词。

SF3 未匹配的谓词是否被其它谓词的论元覆盖：我们经验地判定一句话中越是重要的部分被各个谓词论元覆盖的次数越多。以图 3a 为例，‘They’这个主语被三个论元所覆盖：publish 的 A0，offer 的 A0，以及 add 的 A1，而对于谓词 add 以及其论元 A0 都只被覆盖了一次。

对于匹配的谓词，我们按照表 1 寻找是否有匹配的论元，并给出如下一组特征。

SF4 句子 1 中的论元特征：第一句话中是否包含某个论元成分。这是个 0、1 特征。

SF5 句子 2 中的论元特征：第二句话中是否包含某个论元成分，同上。这两个特征反应了论元的匹配情况。

SF6 匹配的论元成分的相似度：我们用公式 2 计算相似度。

$$P_1 = \frac{A_1 \cap A_2}{|A_1|}$$
$$P_2 = \frac{A_1 \cap A_2}{|A_2|} \qquad (2)$$
$$Sim = \frac{2P_1 \cdot P_2}{P_1 + P_2}$$

其中，$A_{i=1,2}$ 表示句子 1 和 2 在某个谓词下相对应的论元。和计算相似度特征的方法一样，我们这里也采用了 WordNet 外部知识。对于不存在的论元成分，这个特征取零。这个特征会产生 3×15 个子特征，它们的设立是考虑到不同的论元成分，哪怕是一些修饰成分，在新闻语料中都可能扮演很重要的角色这个特点。

SF7 代词特征：我们从 SRL 中得出的结果并没有考虑指代消解，我们用这个特征给出某个论元成分中是否包含代词。给出这个特征，是因为考虑到新闻语料中的复述句往往来自不同的新闻机构或者不同的日期，单纯从孤立的句子上看，会有一些无法消解的指代，比如图 3a 中的 They 和 he。图 3a 和 3b 这两句话在新闻背景下是复述句，但是严格意义上说，它们不应该是互为复述句的。

SF8 扩展的句子相似度特征：在语义角色标注后，我们引入了一个新的计算句子相似度方法，其公式如（3）

$$extSim = \frac{\sum_{w \in \{S_1 \cap S_2\}} (N_1^w + N_2^w)}{\sum_{w \in \{S_1\}} N_1^w + \sum_{w \in \{S_2\}} N_2^w} \qquad (3)$$

公式(3)中的 N_i^w 表示有多少个谓词论元成分覆盖了这个词，如图 4a 中的 offering 为 3，added 为 1。选用这个特征也是基于这样直观的假设：越是重要的成分被覆盖的越多。

这些特征的设计都是充分考虑了新闻领域中语句识别复述与普通领域的语句复述的不同。

我们将得到的特征直接送入 SVM 分类器进行训练和测试，并由 SVM 输出最终结果。

4 实验

这一部分介绍我们采用语料和实验结果。

4.1 MSR 语料介绍

我们的实验里采用的是 MSR[22]语料。MSR 是一个用于一般用途的复述语料库，它来源于在线新闻网站。其创建分两步，第一步是由仅仅依靠编辑距离过滤词汇上不相似的句对，这样共得到了 5801 个句子对；第二步为人工标注，其中 3900 个句子对标为正例，其余为反例。这样这个语料中的句子单从词汇上看句对间相似度很大，给复述识别任务提出了更高的要求。第一部分中的例子里给出的两组句子均来自 MSR，其中第二个例子稍微做了一点改动。

4.2 评测标准

为便于比较，我们的评测采用 F1 值作为评测标准，和文献[9]的评测方法一致，见公式（4）。

$$F1 = \frac{2 \times \text{Precesion} \times \text{Recall}}{\text{Precesion+Recall}} \qquad (4)$$

4.3 实验及结果

在 MSR 语料中，训练集有 2753 个正例，1323 个反例；测试结中有 1147 个正例，578 个反例。从训练集抽取特征后我们用 SVM 分类器进行训练，然后在测试集上做测试。

我们选用基于语义角色标注的方法 Qiu[9]和基于句法的方法 Zhang[18]以及 Wu[14]作为 baseline。在实验中，我们先测试了传统的基于基本特征（BF）的结果，然后给出基于语义特征（SF）的结果，最后将这二者结合给出 SF+BF 的结果。实验结果及 baseline 在表 2 给出。

方法	准确率	召回率	F1
Zhang	***0.743***	*0.882*	*0.807*
Wu	*0.723*	*0.925*	*0.812*
Qiu	*0.725*	***0.934***	*0.816*
BF	0.725	0.864	0.789
SF+BF	0.739	0.915	**0.818**

表 2 试验结果

4.4 结果分析

从表 2 我们看到，当取依赖于词频等表层信息的基本特征 BF 时结果为最低 0.789，而加入语义特征 SF 后，系统性能得到很大改善，取得了 3.7%的相对提高，达到 0.818。这证明我们采用的语义特征效果比较明显。

我们的结果比依赖句法和词集特征的 Wu 和 Zhang 的方法都有较明显提高。前面提到 Zhang[18]的方法依赖依存句法分析对句式做了调整，如将被动句式变为主动句式等。这种方法效果并不理想，仅取得了 0.807 的 F1 值。这个结果正反映出新闻语料自身的特点，即是否是复述句很大程度上依赖于句子中某些状语、补语等修饰成分。我们的方法和 Qiu[9]给出的方法相比提高虽然并不明显，但是我们的方法是在完全不依赖句法特征的条件下做出的，如何能将句法特征有效的融入我们的方法，也是我们下一步将要进一步研究的问题。

5 结论

本文介绍了一种新的基于语义角色标注的新闻复述语句识别方法。针对新闻语句不容易从词频、句法等信息做出复述判断的特点，对经过语义角色标注的新闻语句，我们提取了能反映句子不同成分匹配情况的更为丰富的特征。从实验结果看，我们的结果比依赖词集信息，以及句法信息的方法都有明显的提高。

参考文献

1. McKeown, K. Paraphrasing using given and new information in a question-answer system. 1979: Association for Computational Linguistics.
2. Callison-Burch, C., P. Koehn, and M. Osborne. Improved statistical machine translation using paraphrases. 2006: Association for Computational Linguistics Morristown, NJ, USA.
3. 宗成庆, et al., 面向口语翻译的汉语语句改写方法. 汉语语言与计算学报(Journal of Chinese Language and Computing), 2002. **12**(1): p. 63-77.
4. Zong, C., et al., Approach to Spoken Chinese Paraphrasing Based on Feature Extraction, in In Proceedings of the 6th Natural Language Processing Pacific Rim Symposium (NLPRS). 2001: Tokyo, Japan. p. 551-556.
5. 车万翔, et al., 基于改进编辑距离的中文相似句子检索. 高技术通讯, 2004. **14**(7): p. 15-19.
6. Harabagiu, S. and A. Hickl. Methods for using textual entailment in open-domain question answering. 2006: Association for Computational Linguistics Morristown, NJ, USA.
7. Barzilay, R., K. McKeown, and M. Elhadad. Information fusion in the context of multi-document summarization. 1999: Association for Computational Linguistics Morristown, NJ, USA.
8. 秦兵等, 多文档自动文摘综述. 中文信息学报, 2005. **19**(6): p. 14-20.
9. Qiu, L., M. Kan, and T. Chua. Paraphrase recognition via dissimilarity significance classification. 2006: Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing.
10. Corley, C. and R. Mihalcea, Measuring the semantic similarity of texts. Ann Arbor, 2005.
11. Fellbaum, C., WordNet: An electronic lexical database. 1998: MIT press Cambridge, MA.
12. 宗成庆, 统计自然语言处理, 清华大学出版社, 2008.
13. Finch, A., Y. Hwang, and E. Sumita. Using machine translation evaluation techniques to determine sentence-level semantic equivalence. 2005.
14. Wu, D., Recognizing paraphrases and textual entailment using inversion transduction grammars. Ann Arbor, 2005.
15. Bar-Haim, R., I. Szpektor, and O. Glickman, Definition and analysis of intermediate entailment levels. Empirical Modeling of Semantic Equivalence and Entailment, 2005. **100**: p. 55.
16. Wan, S., et al., Using dependency-based features to take the "para-farce" out of paraphrase. Proc. of ALTW, 2006.
17. Das, D. and N. Smith, Paraphrase identification as probabilistic quasi-synchronous recognition. Proc. of ACL-IJCNLP, 2009.
18. Zhang, Y. and J. Patrick. Paraphrase identification by text canonicalization. 2005.
19. Paul Kingsbury, M.P., and Mitch Marcus, Adding semantic annotation to the penn treebank. In Proceedings of the Human Language Technology Conference, 2002.
20. Charniak, E. A maximum-entropy-inspired parser. in In Proceedings of the First Annual Meeting of the North American Chapter of the Association for Computational Linguistics (NAACL'2000). 2000.
21. Pradhan, S., et al. Shallow semantic parsing using support vector machines. in In Proceedings of HLT/NAACL. 2004. Boston, USA.
22. Dolan, B., C. Quirk, and C. Brockett. Unsupervised construction of large paraphrase corpora: Exploiting massively parallel news sources. 2004: Association for Computational Linguistics Morristown, NJ, USA.

作者简介：

吴晓锋，性别：男，出生年月：1976.04，博士研究生，现就读于中科院自动化所模式识别实验室，中文信息处理组。

宗成庆，性别：男，出生年月：1963. 07，模式识别国家重点实验室副主任，研究员，博士生导师。