基于视觉信息的汉语词汇语义习得

张春宇　张蔚　刘海鹏　于立平　王小捷　李睿凡
北京邮电大学计算机学院智能科学与技术中心 北京 100876
E-mail: cyzhang999@gmail.com; xjwang@bupt.edu.cn

摘　要：认知科学的研究表明，人类在语言习得过程中，其他认知通道(如视觉)的信息具有重要的辅助作用。本文描述了一个基于视觉信息的汉语词汇习得系统，系统基于一定规模的简单图像-句子描述对集，综合利用图像信息和词汇分布信息，获得了包括颜色在内的五类词汇范畴基于图像特征的意义表示。进一步，本文将所习得的词汇表示，用于为简单图像自动生成文本描述，实验表明，这种词汇表示具有一定的图像描述能力，这是基于符号的词汇表示方法所不具备的。

关键词：词汇语义，语义习得，视觉语义，图像描述

Visual Information Based Meaning Acquisition of Chinese Words

Chunyu Zhang, Wei Zhang, Haipeng Liu, Liping Yu, Xiaojie Wang, Ruifan Li
Center for Intelligence Science research, School of computer, Beijing University of Posts and Telecommunications, Beijing 100876
E-mail: cyzhang999@gmail.com; xjwang@bupt.edu.cn

Abstract: According to cognitive science research, the information from other perception channels, for example, visual information, are very useful in the process of human language acquisition. This paper introduces a visual information based meaning acquisition system of Chinese words. This system tries to learn from image-text pairs. By combining image information with word distribution information, the system tries to learn a kind of image based expression of word meaning for five different categories of words including color etc. Then, the learning result is used to automatically generate description for simple image. Experiments show the learned meaning expression of words can be used to describe image, although the performance should be improved.

Keywords: Word Meaning, Meaning Acquisition, Visual-based Meaning, Image Description.

1　概述

认知科学的研究表明，人类在语言习得过程中，其他认知通道(如视觉)的信息具有重要的辅助作用。母亲在开始教孩子说话时总会借助一些身边的玩具，比如一个皮球，在重复地和孩子说“球”等简单的指示性词汇时，也让孩子通过眼睛、耳朵、手等感觉器官进行看、听、摸等感知活动，帮助孩子获得对“球”这个词的理解，其中，视觉信息对语言习得的帮助尤其大。认知科学的研究表明，人们在提取物体语义知识时都会同时激活该物体的视觉表象[1]。随着自然语言处理研究的深化和其他各个模态信息处理技术的进步，出现了一些基于视觉信息的语言习得计算

模型的研究。

文献[2]是发表较早的一篇文章，文中希望通过“我们怎样学习描述我们看见的东西？”这个问题把认知科学中视觉感知、语言、推理和学习这几个分支统一起来，并认为这是认知科学中的一个新的试金石问题。Deb K. Roy 和他的 MIT 研究小组在关联视觉信息和英语词汇方面做了一系列的工作[3-7]。他们提出了一个多通道早期词汇学习模型 CELL(Cross-Channel Early Lexical Learning)以模拟母亲教儿童说话的场景，把机器想象成一个刚开始学习语言的儿童，通过机器视觉来观察周围环境的物体，通过研究候选词汇的声音和视觉信息间的互信息，使机器获得简单认知能力；构建了 DESCRIBER 系统，系统通过一定数量的静态场景及其对应的英语语句描述进行学习，获取描述人的词汇、句法以及空间关系概念。在学习后，系统能够对类似的新场景生成语言描述。[8]中描述了索尼公司的一款智能机器狗玩具 AIBO，它能够通过简单的人机交互学习到物体的英语名称。在德语方面，也有类似的研究，而基于视觉信息进行汉语语言习得的研究还很少。Zhao 等在[9]中提到了哈尔滨工业大学把自然语言描述转换成动画的研究进展。

本文构建了一个基于视觉信息的汉语词汇语义习得系统—ViMac(Visual Information based Meaning Acquisition of Chinese Words)。该系统具备较为初级的学习能力，能够学习与视觉相关的五个范畴的词，分别是颜色、形状、大小、方位和修饰颜色副词。本文还将系统习得的词汇用于自动生成对简单图像的描述。

本文的安排如下，第 2 节介绍基于图像信息的汉语词汇语义习得系统的结构及采用的主要关键技术。第 3 节介绍习得系统的一个应用，即把所习得的基于视觉信息的词汇语义用于图像描述的自动生成，第 4 部分介绍实验及评测，最后第 5 部分是本文的总结和展望。

2 基于图像信息的汉语词汇语义习得

基于图像信息的汉语词汇语义习得系统 ViMac 总体结构如图 2.1 所示。

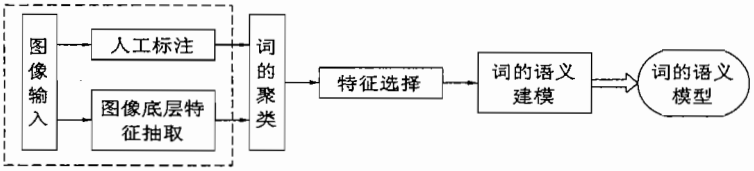

图 2.1 ViMac 系统学习总体结构图

系统的词汇习得是基于一定规模的图像-文本描述对。系统的目标是习得能直接与图像特征对应的颜色、简单形状、位置和大小等范畴的词汇。为回避图像分割的困难，本文采用的是计算机生成的图像，图像采用纯色背景，每幅图像中仅包含一个简单图形，包括正方形、圆形、正三角形、正五边形、正六边形。图形的颜色是随机生成的，也都是纯色，但与背景色不同。图形的大小和图形在图像中的位置都是随机生成的。本文采用了限定维度的语言对图像进行了人工标注。标注中包括所有要学习的五个范畴，不包含任何噪声词汇。针对要学习的五个范畴，抽取了 13 个维度的图像底层特征[10-12]，如表 2.1 所示。表中第三列是特征所对应的范畴，是人为选定的与各个范畴关联程度最高的特征。

汉语词汇语义的习得过程是，首先基于图像特征和词汇分布进行词聚类，通过特征选择获得词汇范畴与图像特征的对应，基于这种对应习得各个范畴中词汇的语义表达。

特征名称	描述	范畴
φ_1-φ_7	图形的七个不变矩	形状
X	图形中心点的 x 轴坐标	方位
Y	图形中心点的 y 轴坐标	
AREA	图形的面积	大小
R	RGB 颜色的红色分量	颜色、修饰颜色的副词
G	RGB 颜色的绿色分量	
B	RGB 颜色的蓝色分量	

表 2.1 ViMac 系统抽取的图像底层特征

下面首先介绍基于词汇和图像信息的词汇混合聚类方法。

通常，语言处理中的聚类利用词汇在语言中的分布信息来进行，本文还进一步利用词汇所在的描述语句与图像的对应关系。因此，本文综合利用这两种信息采用贪婪聚类算法进行词聚类。

算法中，类间距离 $d(C_i, C_j)$采用式（2.1）[6]。

$$d_{ds}(C_i, C_j) = \alpha d_d(C_i, C_j) + (1-\alpha) d_s(C_i, C_j) \qquad (2.1)$$

其中， $d_d(C_i, C_j)$为仅考虑词汇分布信息时的类间距离，$d_s(C_i, C_j)$为进仅考虑词汇与图像特征关联时得到的类间距离。下面分别介绍这两种类间距离的计算方式。

首先考虑词汇分布信息，一个基本假设是：处于同一类的词，不会出现在同一句子中。也就是说出现在同一句子中的两个词应属于不同的类。这个假设与儿童语言习得中的互斥偏置机制十分类似[13]。儿童在学习时不会为了同一个概念学习两个词，这个偏置使得有限的例子能得到充分的学习。

假设范畴 C_i 中有 N_i 个词，范畴 C_i 中的第 k 个词为 $C_i(k)$，则两个范畴 C_i 和 C_j 基于词分布的距离为：

$$d_d(C_i, C_j) = \frac{\sum_{k=1}^{N_i}\sum_{l=1}^{N_j} R(C_i(k), C_j(l))}{\sum_{k=1}^{N_i}\langle C_i(k)\rangle + \sum_{k=1}^{N_j}\langle C_j(k)\rangle} \qquad (2.2)$$

其中，〈 $C_i(k)$ 〉表示第 i 个范畴中的词 k 在语料库中出现的次数。$R(w_i, w_j) = \sum_{u \in U} \pi(u, w_i, w_j)$ 为词 w_i 和 w_j 在语料库中同时出现在对同一图像的描述语句 u 中的总次数。

其次考虑图像特征信息。本文采用对称 KL 距离[6]来描述词与各个图像特征的关联程度。若设 $x(j)$ 是图像特征向量 x 的第 j 个分量，$p(x(j)|w)$ 表示关于单词 w 的词条件概率分布，$p(x(j)|\overline{w})$ 表示单词 w 的背景概率分布，则词 w 与各个图像特征的关联向量为：

$$s(w) = \left(KL_2(p(x_1|\overline{w}) \| p(x_1|w)),\ \ldots\ ,KL_2(p(x_n|\overline{w}) \| p(x_n|w)\right)$$

其中 $KL_2(p(x|\overline{w}) \| p(x|w))$ 为 $p(x|\overline{w})$ 与 $p(x|w)$ 的对称 K-L 距离。

用负点积计算两个范畴之间的距离，则两个范畴间的距离 d_s 可表示为：

$$d_s(C_i, C_j) = \frac{\sum_{k=1}^{N_i}\sum_{l=1}^{N_j} -[s(C_i(k))]^T s(C_j(l))}{N_i N_j} \qquad (2.3)$$

完成对语料中的词范畴的划分后，针对各个范畴进行特征选择。

对各个范畴进行特征选择，目的是选出与这个范畴最相关的特征。特征选择是从单个词开始，对范畴中每个词选出与之最相关的视觉特征，然后对各个词的视觉特征做析取得到整个范畴的特征。最终将范畴对应的特征作为范畴中词的特征。

对于单个词的特征选择，是选择使得 2.1 小节中提到的对称 KL 距离最大的特征。在 2.1 节中用到的是一维的对称 KL 距离，在本节中应用的是多维的对称 KL 距离[6]。

单个词的特征选择使用贪心算法，从仅有一个特征开始，然后迭代加入个特征，选择使多维对称 KL 距离最大的特征子集。每次迭代后，用选择到的特征数归一化对称的 KL 距离。加入新的特征后，如果归一化后的对称 KL 距离仍然增大时，迭代过程继续，直到归一化后的对称 KL 距离不再增大为止。

通过特征选择，可以得到词范畴与人工规定的 13 维特征中的某几个维度相对应，比如，通过算法，颜色范畴与其中的三个维度相对应。

特征选择后，利用范畴所对应的特征对范畴中的每个词进行语义建模，本文中利用多维的高斯分布进行语义建模。

这样，特定范畴中的具体词就可以用一个特征子向量的分布进行表示，比如，颜色范畴中的“红色”，可以用颜色范畴对应的三个维度的概率分布来表示。

至此，汉语词汇的语义已经习得，为评估所习得的基于图像的词汇语义表示，本文将习得的词汇表示应用到图像描述的生成中。

3 图像描述的自动生成

在结合图像信息和文本标注信息进行词汇表示习得的同时，基于标注句子，可以进行句子中词汇使用顺序的习得。词汇顺序模型的习得是基于词汇范畴的二元模型。利用 2.1 节中的聚类结果，从人工标注句子中计算出词汇范畴间的转移概率。本文所生成的图像描述要包含尽量多的词的范畴，但同一范畴中只出现一个词汇。所以，在句长固定的情况下，利用范畴间的转移概率做全局搜索，搜索使得式（3.1）最大的路径。也即最终习得的词汇顺序模型。

$$\gamma(Q) = \log P(C_{q1} \mid \text{开始}) + \sum_{t=2}^{T} \log P(C_{qt} \mid C_{qt-1}) + \log P(\text{结束} \mid C_{qT}) \qquad (3.1)$$

基于已习得的词的语义模型和词汇顺序模型生成对图像的句子描述，其过程如图 3.2 所示。

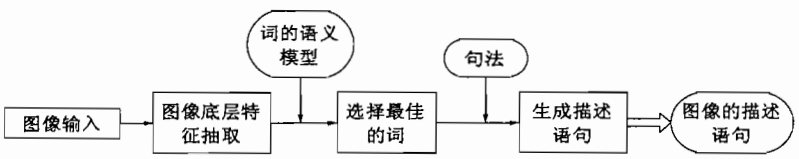

图 3.1 图像描述系统结构图

在接受一个图像输入后，首先对该图像抽取表 2.1 所示的 13 维特征。然后利用词的语义模型，从各个范畴中选出最适合该图像的词。最后利用词顺序模型生成最终的语句描述。

4 系统的实验及评测

本文中的实验包括两部分，一是学习部分的结果，即词的范畴及与其关联的特征。二是把学习部分应用到图像的文本描述自动生成系统中，本文借鉴机器翻译的评测方法，提出了一种对生成的描述进行评测的方法。

学习部分的实验结果分为两个部分，一是聚类的结果，也就是对汉语词汇范畴的划分。二是特征选择的结果，为各个范畴选择与之最相关的特征。学习部分的实验结果如表 4.1 所示。

范畴号	词	特征
1	蓝色、绿色、黄色、粉红色、紫色、紫红色、肉色、红色、棕色、天蓝色、灰色	φ_2~φ_7 Area R G B
2	的	φ_1~φ_7 X Y Area R G B
3	正方形、圆形、三角形、五边形、六边形	φ_1~φ_7
4	浅、深、暗	φ_4 φ_5 φ_6 φ_7 G B
5	大、小	φ_5 φ_6 φ_7 X Area
6	下边、左边、右上方、右下方、右边、左下方、 上边、左上方	φ_2~φ_7 X Y
7	亮	φ_5 φ_7 G

表 4.1 学习部分的结果

表 4.1 中是学习部分的结果，与表 2.1 中的预期结果比较可以看到，第 7 个范畴中“亮”应该是第 4 个范畴的，这里单独聚成了一类。通过聚类已经基本把标注语料中的词划分成需要的范畴，这也是本文系统中最终应用的词范畴。各个范畴对应的特征在表中的第三列。显然，这个结果与表 2.1 中人工规定的对应关系存在一定的差异，4.2 节中将对这种差异对图像描述自动生成的影响进行评测。

下面介绍对图像描述系统的评测。基于图像信息习得的词汇语义的评测具有极强的主观性，对于同一幅图像不同的人会有不同的描述，即使是同一个人在不同的时间也可能有不同的描述。目前，为图像生成句子描述的任务还没有公认的评测指标和方案。本文借鉴机器翻译系统的 BLEU(BiLingual Evaluation Understudy)[14]评测方法提出了图像描述任务的评测方案。本部分先介绍评测方案，之后给出系统评测结果。

本文的评测方案包括如下的几个部分：

测试语料及性能上界：选定 N 幅图像作为测试集，由三位以汉语为母语、具有正常认知能力的人分别独立对这 N 幅图像进行描述。在描述时，要求他们从颜色、形状、大小、方位和修饰颜色的副词这五个范畴中各选择一个词对图像对象进行描述，每个词均在系统习得的各范畴词集中选取。在所有标注生成后，本文基于不同描述之间的一致性来计算图像描述任务的上界性能。

本文分别考虑词和句子两个层次的性能上界。对于每个描述，本文只关注颜色、形状、大小、方位和修饰颜色的副词这五个范畴中的词。在词汇层上，对于每一个图像的三个描述句子，考察它们在这五个范畴中用词的一致性。对于每一个范畴的词都进行一致性比较，如果有两个或两个以上的描述在这个范畴上的用词是一致的，则认为该用词是一致的，否则是不一致。在句子层上，如果对同一个图像有两个或两个以上的描述中在五个范畴上的用词都是一致的，则认为对该图像的描述是一致，否则是不一致。

评测性能指标：本文用两个指标来度量描述性能。其一是词错误的比例 E_W。对描述句子中

的颜色等五个范畴中的词分别评估，并统计各个范畴的错误率。其二是描述句子错误比例 E_S，对一个图像的描述中只要有一个特征词出错，则整句标注即为错误。

基于上述的评测方法，本文共给出了六个实验结果，分别是图像描述人工标注在词汇级和句子级的不一致的比例，在系统自动选择特征的情况下和人工选择特征的情况下自动标注在词汇级和句子级的错误率。对人工标注评测的目的是得出图像描述的上限值，比较两种自动标注的目的是考察特征选择对整个系统的影响。

在词汇级，五个范畴分别统计，统计结果见表 4.2。表中第二列为人工标注的错误率，第三列为自动标注的错误率。

评测对象	人工标注的错误率	自动标注的错误率
颜色	9%	26%
形状	0%	1%
方位	11%	57%
大小	3%	24%
修饰颜色副词	24%	51%
整句	40%	89%

表 4.2 评测结果

评测对象	自动标注的错误率	
	基于聚类的特征	人工规定的特征
颜色	26%	23%
形状	1%	1%
方位	57%	50%
大小	24%	32%
修饰颜色副词	51%	52%
整句	89%	87%

表 4.3 两组特征用于图像描述系统的评测结果

由表 4.2 可见，人工标注在词汇层面上也不是完全一致，尤其是修饰颜色的副词，但在形状范畴达到了完全一致，而这恰恰说明了对于图像描述中词汇选择的主观性，形状完全一致是因为它不受主观性的影响。

自动标注无论是在词汇层面上还是在句子层面上都与人工标注有较大的差距，说明 ViMac 系统的学习能力还比较弱。但这其中受到评测方式中主观因素的影响比较明显。

进一步，本文分别采用基于聚类的特征和人工规定的特征进行了图像描述自动生成的实验。结果如表 4.3。可见，两组方法产生的特征无论在词汇级还是在句子级其系统性能都没有显著的差异。这个现象的出现可能由两个原因造成，一是人工规定的特征本身不能够恰当的表征学习到的词汇，二是对词的语义建模方法不适合本系统。究竟是哪个原因造成的还需要更多研究。

5 总结

在本文中介绍了基于视觉信息的汉语词汇习得系统 ViMac，并把习得的结果应用于一个图像自动描述系统，并对系统结果提出了一种评测方式。

显然，本文中介绍的汉语词汇习得能力还比较初级，用于生成图像描述的性能还比较低，仅能描述简单的图像。基于图像信息的汉语词汇习得有很多工作需要进一步深入，也有很多方向可以扩展。比如，学习的方法，学习的对象，学习的场景等方面都有很大可扩展的空间。

本文采用的混合聚类和特征选择的技术有不少可能的替代技术，需要进行进一步的验证和选择。本文中主要学习的是名词，而形容词、动词等更多、更复杂范畴的词汇的习得将带来更大的挑战。本文为避免图像分割的困难，图像采用的是简单背景，图像对象也是直接指定的，对于复杂图像、真实场景，必然带来一系列的相关问题，是基于图像的词汇习得所需要解决的。系统的

评测也是一个很重要的问题，如何减小评测中的主观性因素，真实反应出系统的性能，还需要更加深入的研究。这些都是本系统需要进一步开展的下一步工作。

6 致谢

本文得到了高等学校学科创新引智计划（项目编号：B08004）、国家支撑计划项目(项目编号：2007BAHo5B02-04)的支持。

参 考 文 献

[1] 金花,刘鹤龄,杨娅玲,莫雷. 语义知识神经表征的fMRI研究：通道特异性或类别特异性？，心理学报，2005，37(2):159~166.

[2] Jerome A. Feldman. Miniature language acquisition: a touchstone for cognitive science. FELDMAN, LAKOFF, STOLCKE & WEBER, 1990.

[3] Deb Roy and Alex Pentland. Multimodal adaptive interfaces. Technical Report 438, MIT Media Lab Vision and Modeling Group, 1997.

[4] Deb Kumar Roy. Learning words from sights and sounds: a computational model. PhD thesis, MIT Media Laboratory, 1999.

[5] Deb Roy. Learning audio-visual associations using mutual information. International Conference on Computer Vision, Workshop on Integrating Speech and Image Understanding, 1999.

[6] Deb Roy. Learning visually grounded words and syntax of natural spoken language. Evolution of Communication, 2000/2001, 4(1).

[7] Deb K. Roy. Learning visually-grounded words and syntax for a scene description task. Computer Speech and Language, 2002, 16(3):353~385.

[8] Luc Steels and Frederic Kaplan. AIBO's first words. The social learning of language and meaning. in: H. Gouzoules(Ed), Evolution of Communications, 2001, 4(1).

[9] ZHAO Tiejun. Recent advances on NLP research in Harbin Institute of Technology. Frontiers of Computer Science in China, 2007, 1(4):413~428.

[10] H. Lilian Tang, Rudolf Hanka. Histological image retrieval based on semantic content analysis. IEEE Transactions on Information Technology in Biomedicine, MARCH 2003, 7(1):26~36.

[11] Yong Rui, Thomas S. Huang. Image retrieval: current techniques, promising directions and open Issues. Journal of Visual Communication and Image Representation, 1999, 10(1):39~62.

[12] Teh, C. H. and Chin, R. T. On image analysis by the methods of moments. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1988, 10:496~513.

[13] Ellen M. Markman. Categorization and naming in children. MIT Press, Cambridge, MA, 1991.

[14] Kishore Papieni. BLEU: a method for automatic evaluation of machine translation. ACL 2002, 311~318.