中文近义词辨析实验——机器学习程序与二语学习者的对比
(An Experimental Study on Discriminating Chinese Near Synonyms: Contrasts between Machine Learning Systems and Second Language Learners)

詹卫东
(Zhan, Weidong)
北京大学
(Peking University)
zwd@pku.edu.cn

曹晓玉
(Cao, Xiaoyu)
北京大学
(Peking University)
xiaoyu.cao@pku.edu.cn

崔巍
(Cui, Wei)
北京大学
(Peking University)
cuiw@pku.edu.cn

常宝宝
(Chang, Baobao)
北京大学
(Peking University)
chbb@pku.edu.cn

摘要： 本文将机器学习技术引入中文近义词辨析任务，与二语学习者在近义词辨析任务上展开了初步的实验对比研究。在近义词集的选取、测试题制作方面，本文遵循平衡与周全原则。测试结果显示：机器在中文近义词辨析任务上的表现与二语学习者有明显可比性，机器测试成绩与中级水平的汉语学习者测试成绩呈正相关。除近义词本身难度有别外，题型差异对测试成绩有显著影响。词语意义特征对机器近义词辨析的影响并不低于形式特征的影响，在辨析时机器和二语者对句法形式特征的把握比对搭配区别特征的把握更有效。

Abstract: In this article, machine learning technology is introduced into fill-in-the-blank (FITB) tasks involving the discrimination of Chinese near synonyms. A preliminary experimental study was carried out on said tasks between machines and L2 learners of Chinese. This study adheres to principles of balance and comprehensiveness in selecting synonyms and making test sets for the experimental research. The test results show that the performance of machines in discriminating Chinese near synonyms in FITB tasks is significantly comparable to that of human L2 learners. The score of the machine was also positively correlated with that of intermediate-level Chinese learners. In addition to the sets of near synonyms varying in difficulty, the difference of test question types also has a significant impact on test scores. The influence of lexical meaning features on the discrimination of near synonyms is no less than that of its formal features. Meanwhile, it is more effective for machines and L2 learners to exploit syntactic formal features rather than distinguishing collocation features in FITB tasks.

关键词：近义词、易混淆词、机器学习、Bi-LSTM 模型、近义词辨析实验

Keywords: Near synonyms, confusable words, machine learning, Bi-LSTM, FITB tasks

1. 引言

作为语言教学的重要组成部分，词汇教学，特别是其中的近义词辨析，一直是受到研究者、教师和学习者普遍关注的一个领域，无论在理论研究层面还是在实践运用方面，都有诸多挑战（Nation & Newton, 1997、洪炜，2012）。随着基于语料库技术的实证型研究在对外汉语教学研究领域的普及，以及二语教学理论和实践的不断丰富和发展，面向对外汉语教学的近义词研究和实践近年来取得了非常多的成果，比如赵新、刘若云（2005）对编写面向二语学习者的近义词词典的思考；张博（2007）、高再兰（2016）对面向母语者的近义词辨析和面向二语学习者的近义词辨析关系的深入讨论；洪炜（2017）有关课堂显性辨析对二语学习者近义词辨析的积极影响的实验研究。Hong（2014）、陆方喆（2016）基于语料库开展的近义词辨析研究，等等。可以说，当前学界对近义词辨析的探讨既有宏观视角的开拓，也有微观技术手段的不断更新，呈现出多维度、多层面、多种技术手段百花齐放的蓬勃发展态势。

与此同时，近年来自然语言处理（NLP）在以深度学习（deep learning）为代表的新一代基于人工神经网络（Artificial Neural Network）的机器学习技术推动下，也有不少研究开始涉足机器近义词辨析和同义词可替换性的探索。这方面主要包括两类代表性的任务：一是选词填空任务（Fill-in-the-blank，简称 FITB），由 Edmonds（1997）提出。二是在国际语义评测会议[1]（SemEval 2007）上提出的词汇替换任务（Lexical Substitution Task，简称 LST），可参见 McCarthy & Navigli (2007、2009)、Zhao 等（2007）、陈士婷等（2012）。其中前一种任务的研究工作更多一些。NLP 研究人员尝试了不同的机器学习模型来完成 FITB 任务。Wang & Hirst (2010) 基于隐性语义分析法（Latent Semantic Analysis，LSA）和支持向量机（Support Vector Model，SVM）分类器，在 7 组英语近义词上进行选词填空任务实验，达到了 74.5%的准确率。Yu 等（2011）将上述 7 组英语近义词翻译为 7 组中文近义词进行了中文近义词辨析实验，采用点式互信息（PMI）和 5 元模型（5GRAM）在选自台湾中研院 Sinica 语料库和 Chinese News 语料库的测试集上进行辨析实验，准确率为 68.07%。Yu & Chien（2013）在 LSA 模型基础上增加了独立成分分析法（Independent Component Analysis）来寻取影响近义词辨析的隐性因素，进而基于 SVM 分类器完成近义词辨

[1] 国际语义评测会议（SemEval）已有 20 年历史，致力于组织与计算机语义分析相关任务的评测和学术交流。其前身是 Senseval（词义消歧评测）会议。可参见 https://en.wikipedia.org/wiki/SemEval。

析任务，在中、英文语料上的准确率分别达到 82.57%和 76.78%。特别值得注意的是，Huang 等（2017）首次将机器近义词辨析任务跟二语学习结合起来。该文对比了高斯混合模型（GMM）和双向长短时记忆神经网络（Bi-LSTM）模型，在英语 24 组同义词辨析实验任务上，两种模型的准确率分别达到了 78.16%和 83.59%，并进一步针对二语学习者开展了计算机辅助近义词学习实验，发现尽管基于 Bi-LSTM 模型的机器评测效果好于 GMM 模型，但是 Bi-LSTM 是对整个句子的全局区别把握有优势，GMM 则是对近义词相关的特定区别特征把握有优势，因此，基于 GMM 模型来辅助二语者的近义词学习，效果反而优于 Bi-LSTM 模型。

已有的研究显示，基于机器学习技术的近义词计算机自动辨析，有可能对二语学习提供帮助。这方面的工作才刚刚开始，是一个值得探索的研究方向，特别是在中文近义词的辨析方面，还缺乏机器学习实验与二语学习者实验的对比考察。本文从这个角度出发，希望将机器学习技术用于中文近义词辨析任务中，并在大致相当的环境下，考察机器、母语者、二语学习者对同一套测试题的表现是否有可比性，通过分析初步的实验结果，为今后进一步的深入研究寻找方向。下面第 2 节简要介绍本文实验所用的机器学习模型；第 3 节说明实验的设计；第 4 节是实验结果及分析；第 5 节是结语。

2. 基于机器学习的近义词辨析方法

2.1 训练集的准备

在近义词辨析任务中，机器学习模型的目标是：在给定的若干个近义词中，选出一个合适的词语填入句中特定位置，相比选择其他候选词填入该位置，能够使句子更为合理通顺。为此，机器需要比对一个候选词集中所有的候选词的上下文语境特征，找出候选词与其上下文语境之间的选择规律。由于候选的近义词相对于词汇全集来说是非常相近的（或者说是比较容易混淆的），仅仅将自然语料中正确句子的相关信息作为正例不足以让模型充分学习到近义词之间的语境区别。为了区分，需要人为构造可对比的正例和负例，成对地来学习句子表示（sentence representation）。例如：对候选词集“不管、尽管”和语料中的句子“哈雷 _____ 不富裕，还是买了车。”，分别将候选词填入句子，会形成一组训练样例：<正例：哈雷尽管不富裕，还是买了车。负例：哈雷不管不富裕，还是买了车。>。如果候选近义词集有超过两个候选项，同样可以此方式形成多组正例和负例。例如：对候选词集“二、两、俩”和语料中的句子“她升入了 _____ 班。”，会形成两组训练样例：<正例：她升入了二班。负例：她升入了两班。>；<正例：她升入了二班。负例：她升入了俩班。>。直观而言，在大多数情况下，将真实语料句子中的一个词随机替换成另一个词，得到的新的句子将是一个错误的句子，在统计意义上，以这样的方式自动构造的负例样本是可用的。

2.2 基于最大间距策略的训练过程

按照上述办法生成包含正例和负例的全部训练语料后，机器学习的过程就是训

练一个分类函数，让近义词辨析模型能最大程度区分填入正确候选词的句子与填入错误候选词的句子。具体方法是，采用 Bi-LSTM 对句子的信息进行捕捉，在观察大量的正例和负例句对过程中，学习一个分类函数，该函数为正例句子打出尽量接近 1 的分数，同时为负例句子打出尽量接近 0 的分数，使得正例句子的得分与负例句子的得分之差最大化[2]。在机器学习中，计算间距最大化（max-margin）的常用方法是采用折页损失函数（Hinge loss function）[3]。本文所用的函数形式为：**$L=max(0,1-(F_1-F_2))$**，其中 F_1 代表正例句子的综合评分，F_2 代表负例句子的综合评分。该损失函数的含义概括来说就是，如果填入候选词是正确的，则损失为 0，否则损失就是 1-（F1-F2）。为了使损失尽量小，就要求 F1 和 F2 的差距尽可能地大。

Bi-LSTM 的作用是训练一个词语用法模型（word usage model）。模型训练可以看做是记住了一个候选近义词的上下文特征，以及上下文中各个词语的权重，即上下文中各个词语对于选择某个特定候选词的影响大小。

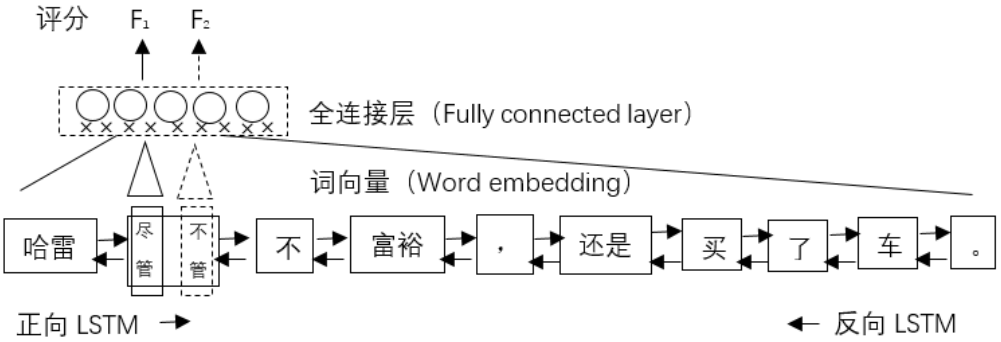

图 1 “不管、尽管”辨析的 Bi-LSTM 机器学习模型示意[4]

在图 1 所示的句子中填入“尽管”，为正例，计算综合评分 F1，填入“不管”，为负例，计算综合分值 F2，经过对包含“尽管”和“不管”的语料的充分训练后[5]，在图 1 所示的句例中，F1=0.9749742，F2=0.1040909。换言之，系统学习到一组参数，基于这些参数可计算得到“尽管”和“不管”在相同语境中的得分，且差距适当。

[2] 基于 Bi-LSTM 的机器学习模型在英语近义词辨析任务上取得了比较好的成绩，此外在构建基于文档的问答系统（Document-based Questions-Answer task，DBQA）等任务上也都有着不错的表现。

[3] 损失函数是机器学习训练参数的主要手段，即用损失函数来衡量模型学习到的参数的好坏。通常损失函数值为一个非负值。该值越小，表示模型的预测结果跟真实情况之间的差距越小，因而模型效果越好。

[4] Bi-LSTM 神经网络中，全连接层（Fully Connected Layer）也称为稠密层（Dense Layer）。该层的每一个节点都与前一层神经网络的每个节点连接，其作用简而言之就是把前一层学习到的分布式表示特征（比如一个句子的数百维向量表示）都综合起来，最终输出为一个值，这样便于完成分类任务。

[5] 在训练过程中，F1、F2 的值不断在变化。学习的过程就是通过 F1 和 F2 差值的变化，来调整句中词语的权重，从而找出最能影响正确选择候选词的上下文词语特征。比如机器通过观察大量语料，可能会发现规律：当上下文中出现了“还是”或者“不、没有”等否定词时，选择“尽管”的概率要高于选择“不管”。

2.3 对机器学习模型的测试

在测试阶段，将待测试的句子和候选近义词集作为机器学习系统的输入，机器学习模型就能够基于已有参数计算出候选近义词集中，哪一个更适合测试题的句子。仍以“尽管”和“不管”这组近义词为例，在本文实验中，测试题集第 70 题是“他没有来，_____ 我邀请了他。”，该题填入“尽管”后整句得分为 0.9999995，填入“不管”后整句得分为 0.0000134。机器学习系统将“尽管”作为答案输出。

3. 实验的准备

本文的主要目的是考察计算机辨析近义词的表现跟人类二语学习者的表现是否具有可比性，此外，通过比较机器学习系统在近义词辨析任务上的具体表现与人类二语学习者表现的相似或不同之处，探讨机器学习系统在二语教学中有可能发挥作用的途径。

因之前缺少在中文领域类似的实验研究可以借鉴，本文的工作属于初步尝试。在进行实验设计时，我们采取了在现有条件下尽可能平衡取样的做法，以便在控制实验成本的前提下尽量多地发现线索，为今后进一步深入研究打下基础。下面分别介绍近义词选词和测试题制作的情况，以及参与试验者样本情况和机器学习系统构建的情况。

3.1 近义词集的选样

前文提到的机器学习系统完成的近义词辨析实验往往只选取十组左右的近义词，而且通常仅限于实词性词语。这些研究聚焦在探讨机器学习系统的模型和算法，对近义词本身的语言学性质不太关注，此外，也较少从第二语言教学的角度去开展近义词辨析任务的相关研究。针对这些问题，本文在实验设计时主要考虑在以下三个方面有所加强：（1）要兼顾母语者视角和二语学习者视角的近义词，除一般的近义词（near-synonym words）外，还要选取部分易混淆词（confusable words）；（2）除一般实词外，还要选取一定比例的虚词；（3）尽量全面地考虑词语的频率、难度、义项数量等相关因素来平衡取样。

值得说明的是，关于易混淆词，对外汉语教学界有过不少探讨，特别是从二语学习者的视角来认识易混淆词，跟主要是从母语者视角认识的近义词（同义词）有很大的不同。张博（2007）指出，“‘易混淆词’与‘同义词’‘近义词’之间有交叉关系，而非包含关系或并列关系，因为它们是研究者站在不同的立场、以不同视角和不同标准归纳出来的词语类聚”。“易混淆词不仅体现为口头表达和写作中的词语混用，还体现为阅读和听辨中的词语误解。” 张博（2007）文中举出“往事-故事”“日前-目前”“有点儿-一点儿”“经验-经历”“乘（坐）-用”等多组易混淆词的例子，这些例子，都不是从母语者视角观察到的近义词，而是在对外汉语教学实践中发现的，由于各种原因（比如字形相近、母语影响等）造成的二语学习者容易混淆的词汇。

基于上述认识，本文选取了 67 个近义词（下文提及近义词时均含易混淆词，可认为是广义的近义词），共计 27 组（词集），作为辨析对象。下面表 1 和表 2 分别从词语类型分布和词集类型分布两个层面详细给出了本文所选词语和词集的统计信息。

表 1 本文选取的 67 个近义词的类型分布

<table>
<tr><td colspan="13">词语类型分布</td><td>合计</td></tr>
<tr><td rowspan="2">词类</td><td>动词</td><td>名词</td><td>形容词</td><td>副词</td><td>数词</td><td>数量词</td><td>量词</td><td>代词</td><td>介词</td><td>连词</td><td>助词</td><td>叹词</td><td rowspan="2">67</td></tr>
<tr><td>13</td><td>8</td><td>7</td><td>11</td><td>2</td><td>2</td><td>3</td><td>3</td><td>4</td><td>6</td><td>6</td><td>2</td></tr>
<tr><td rowspan="2">词汇等级1</td><td>HSK1</td><td colspan="2">HSK2</td><td colspan="2">HSK3</td><td>HSK4</td><td colspan="2">HSK5</td><td colspan="2">HSK6</td><td colspan="2">其他</td><td rowspan="2">67</td></tr>
<tr><td>13</td><td colspan="2">14</td><td colspan="2">7</td><td>17</td><td colspan="2">9</td><td colspan="2">1</td><td colspan="2">6</td></tr>
<tr><td rowspan="2">词汇等级2</td><td colspan="3">甲级</td><td colspan="3">乙级</td><td colspan="3">丙级</td><td colspan="4">丁级</td><td rowspan="2">67</td></tr>
<tr><td colspan="3">41</td><td colspan="3">14</td><td colspan="3">11</td><td colspan="4">1</td></tr>
<tr><td rowspan="2">义项数</td><td>单义</td><td>二义</td><td>三义</td><td>四义</td><td>五义</td><td>六义</td><td colspan="2">七义</td><td>八义</td><td rowspan="2">67</td></tr>
<tr><td>22</td><td>17</td><td>11</td><td>5</td><td>3</td><td>5</td><td colspan="2">3</td><td>1</td></tr>
</table>

表 2 本文选取的 27 组近义词集的类型分布

近义词辨析组的类型分布				合计
义项关系	单义组	单义-多义组	多义组	27
	3	13	11	
词类关系	词类相同	词类有同有异	词类不同	27
	12	13	2	
词语数	二词辨析	三词辨析	四词辨析	27
	16	9	2	
语素	语素完全不同	有相同语素	语素逆序	27
	13	13	1	

附录 1 进一步给出了 27 组近义词集的描述信息，其中包括易混淆词集 11 组（40.74%）。另外值得说明的是，无论汉语本体研究还是对外汉语教学研究，因都是面向人的词汇教学，在考虑近义词辨析的因素时，通常都会兼顾多个层面的特征，包括句法特征、搭配、词语意义（含概念义、感情色彩义等）、语用特征（如语体风格、方言用法等）。在教学实践中，往往也强调从多个维度展开对具体近义词的辨析。不过，本文实验的测试对象包括计算机，在面向机器考虑近义词辨析因素时，只能分为两类，即有形式区别特征（包括广义句法形式特征和狭义的搭配词等）与无形

式区别的近义词[6]。附录 1 对此进行了大致的分类标记[7]，在 27 组近义词集中，无明显形式区别（即对人而言是意义差异）的近义词集有 13 组（48.15%）。

3.2 近义词辨析测试题集的制作

按照上述指导思想甄选出近义词集后，我们通过多种渠道来制作测试句集，包括在语料库、词典、相关研究文献中搜集合适的例句，以及自拟测试句，主要是设计辨析词项可以替换的语境和不可替换的语境。为尽可能真实地反映人的词汇知识和词汇运用能力，全部测试题均为不定项选择题，即除一般的单选题外，还有意设计了一定比例的多选题，并在所有备选答案之外增加了“我不知道”和“都不可以”选项（避免二语学习者猜答案）[8]。为了避免因词语偏难造成二语者理解上的障碍，测试句中的词语难度等级都控制在 HSK5 级以下，题干字数尽量控制在 20 字以内。然后再核查每个测试句的自然度、语境的适切度。在测试集第一版试题完成后，我们以在线问卷形式进行了调研，在收集了近 200 位母语者的语感调查数据后，我们增删了部分近义词集，并对试题进行了调整，最终制作了 100 道测试题的近义词辨析问卷。附录 2 给出了这 100 道题的题干和参考答案[9]，并对每道题根据是否有形式区别特征做了标注。表 3 是 100 道题的区别线索分类统计信息。其中“假搭配”特征是指题干中有意设计了某个候选近义词常见的共现形式特征，但整句语义排斥该候选词填入其中。

表 3 近义词辨析测试集（100 题）根据是否具有形式区别线索分类的分布统计表

特征类型 / 题型	无特征	搭配	句法形式	假搭配	合计
单选题	44	19	11	4	**78**
多选题	9	9	0	0	**18**
选“都不可以”	1	2	0	1	**4**
合计	**54**	**30**	**11**	**5**	**100**

需要说明的是，机器学习系统要跟人类被试者一样完成这 100 道测试题。不过，

[6] 有意思的是，这里的所谓“无形式区别”又是从人的视角来说的，指没有“人通常理解的显式的语法形式区别”。按照人的语感，对这类近义词，一般是从意义的角度去把握其区别。但从计算机的视角来说，两个符号对象若存在区别，则一定是在形式上有区别。只不过，计算机察觉的“形式区别”，人不一定从“形式”的角度去看。对这类“形式区别”，通常笼统地概括为“意义的区别”。

[7] 这两类之间并无严格的界限。标注有一定的主观性。附录 1 中的分类标记是综合了多人标注的结果。

[8] 这些设置都是针对人类被试者的。目前的机器学习模型无法给出“我不知道”这个答案。而要针对多选和“都不可以”两种情况作答，需要对机器学习模型做重新设计，限于条件，本文实验所用的机器学习模型，仅实现了单选题的学习模型，无法像人那样回答答案是多选和“都不可以”这两种题型。

[9] 测试题的参考答案在进行评分时是作为标准答案看待的，即由程序自动比对被试答案与标准答案的异同来打分。不过，母语者语感调查显示，确实也有少数测试题的答案在母语者中也有语感差异，并不能做到完全一致。对于这些测试题，是按照少数服从多数原则，取多数母语者的答案作为参考答案。

机器学习系统是通过观察真实语料来学习近义词的语境分布差异，从而模拟人类的近义词辨析这一语言行为的。从统计学的角度讲，本文制作的 100 道题，并不是来自真实语料的样本，跟真实语料的分布有一定的差异，为了对比，同时也是为了更准确地反映机器学习系统在真实语料基础上所学习到的近义词辨析能力如何，我们在 100 题之外，还采用从真实语料中自动取样的大测试集对机器学习系统的近义词辨析能力进行了测试。参见下面 3.3 的具体说明。

3.3 被试者概况

实验的被试包括母语者（native speaker，记作 N）、二语者（second language speaker，记作 S），以及机器学习程序（Machine Learning System，记作 M）。其中二语者包括两组：一组是北京大学中文系的留学生（31 人），汉语水平达到 HSK6 级（记作 S_HSK6）；另一组是天津财经大学的留学生（28 人），汉语水平未达到 HSK6 级（记作 S_HSK6-）。他们来自韩国、日本、泰国、哈萨克斯坦、乌兹别克斯坦、赤道几内亚、加蓬、老挝、马来西亚、蒙古等 18 个国家。下面表 4（机器学习程序也当做一个被试看待）给出了本文被试的概况信息。

表 4 母语者、二语者、机器学习程序的基本信息表

组别	数量	年龄范围	平均年龄	学习时长	男 ： 女
N	20	21-56	29.85		11 : 9
S_HSK6	31	17-21	19.35	1.5 年-19 年	15 : 16
S_HSK6-	28	17-25	20	8 个月-2 年	12 : 16
M	3			126 小时 - 289 小时	

上表中 M 的数量标记为 3，指的是在程序开发和实验过程中，机器学习程序根据训练数据集大小不同有 3 个版本。机器学习程序的训练数据是从北京大学 CCL 语料库中抽取的包含 67 个近义词语的全部句子，经过自动分词和词性标注处理。训练集分为小、中、大三个版本，以观察数据规模变化对机器学习程序效果的影响。

表 5 近义词辨析机器学习程序的训练数据集统计信息

程序版本	训练数据集	字例数	字型数	词例数	词型数
M30	30 万句训练集	28,237,314	6,475	15,239,701	184,985
M100	100 万句训练集	89,674,369	7,286	48,760,493	293,659
M300	300 万句训练集	228,081,378	8,199	122,791,248	466,176

我们也分别为这三个版本的机器学习程序构造了大测试集，分别为 6700 句，20000 句和 67000 句。三个版本在大测试集上的表现呈线性增长。以 67000 句测试集（即每个近义词对应的测试题为 1000 句）为例，M30、M100、M300 在该测试集上的正确率分别是 65.2%、65.9%、66.6%。在人工设计的 100 题测试集上，M30、M100、M300 的表现同样也呈线性增长关系，得分分别是 54、56、58。显然，表现

最好的版本是训练数据量最大的 M300。下文在介绍实验结果和进行分析讨论时，所用数据均为 M300 的数据。

4. 实验结果及分析

为实验操作方便，汉语母语者参与测试采用的是网上调查问卷的形式。二语学习者的测试则是课堂环境中进行的。在测试前一周发放近义词辨析词汇表（27 组 67 个词），学生可以查资料准备。在测试当天，学生拿到 100 题纸质测试卷后，基本在一节课时间（45 分钟）内完成答题（答题为闭卷形式，时间上无严格要求）。回收试卷后，在 Excel 表中对原始数据进行录入、核对，并跟计算机答题的结果一道，在程序辅助下，开展进一步的数据统计和分析。下面主要从三个方面，对实验结果进行概要介绍和分析。4.1 节说明测试成绩总体情况，在宏观层面分析各组被试者的基本特点；4.2 节考察各组数据间的差异显著性和相关系数，进一步分析机器学习系统的成绩跟二语学习者之间的可比性。4.3 节比较机器评测数据和二语学习者评测数据在各组近义词集上的成绩分布差异。

4.1 测试成绩总体情况

上文已经提到，实验被试包括 4 组。主要的实验是由 4 组被试完成 100 道测试题，此外，机器学习系统还在大测试集上进行了 27 组近义词的测试。因此，本文实验得到的数据可以分为 5 组，除下面表 6 中的 N、S_HSK6、S_HSK6-这三组外，机器学习程序的成绩有两组：机器学习程序大测试集成绩（下文以 M'标识）和机器学习程序 100 题成绩（以 M 标识）。另外，为更全面对照机器和二语者，有时也把 S_HSK6 和 S_HSK6-合并为 S。下面表 6 呈现了 N、S（S_HSK6、S_HSK6-）、M 在 100 题测试集上的总成绩对比情况。

表 6 100 道测试题总成绩对比表（每题均计 1 分）

题型	题量	N		S		S_HSK6		S_HSK6-		M	
单选题	78	73.45	73.45	59.83	59.83	66.81	66.81	52.11	52.11	58	58
多选题	18	13.45	18	3.12	16.20	5.23	16.71	0.79	15.64	0	18
选“都不可以”	4	3.45	3.45	1	1	1.71	1.71	0.21	0.21	0	0
总计	100	90.25	94.90	**63.95**	**77.03**	73.75	85.23	53.11	67.96	**58**	**76**

表 6 中每组被试的成绩都分为左右两列，左列是从严评分的分值，右列是从宽评分的分值。从严指的是被试选项跟标准答案完全一致才算对。从宽是针对多选题而言，只要被试选项是标准答案的子集，就算对。按照从严标准，M 的成绩介于 S_HSK6 和 S_HSK6-之间。如果仅从单选题成绩来看，M 的成绩跟 S 的单选题成绩非常接近（58:59.83）。按照从宽标准，M 的成绩也跟 S 的成绩非常接近（76:77.03）。

可以认为，本文实验的机器学习程序在近义词辨析任务上的表现，跟二语学习者应该是具有相当高的可比性的。

表 6 所展示的各组被试的总成绩分布也说明了本文对近义词的选取以及测试题的制作是比较合理的。在 100 题测试集上的表现，HSK6 级二语者全部超过 60 分，平均分比 HSK6 级以下二语者高 20 分。如果机器学习程序参与二语学习者成绩排名，在全部 59 名二语者中排第 43 名，在 28 名 HSK6 级以下二语者中排第 12 名（达到中等水平）。

此外值得一提的是，按从严评分标准，母语者单选题得分百分制为 94 分（73.45/78），多选题得分为 75；二语者这两项的得分分别是 77 和 17；机器这两项的得分[10]分别是 74 和 0。从多选题与单选题平均分的巨大差距可以看到，题型对成绩的影响非常显著。

将 100 题原始成绩按组汇总后可以得到 27 组近义词的成绩对照（附录 3）。其中机器在 100 题测试集上的成绩方差很大，不如在大测试集上的表现稳定（M 的方差是 M'的 3.5 倍），这也显示出 100 题的样本性质跟真实语料测试题之间的明显差异。

表 7 六组测试成绩的均值、方差等基本统计量表

	M'	M	S_HSK6	S_HSK6-	S	N
平均值	0.662	0.608	0.751	0.558	0.659	0.905
方差	0.020	0.071	0.026	0.039	0.030	0.004
最高分	0.862	1.000	0.989	0.952	0.972	1.000
最低分	0.433	0.000	0.366	0.121	0.305	0.757
中值	0.665	0.667	0.763	0.518	0.648	0.920

母语者在 100 题上的成绩大致在 0.8 到 1 区间波动，方差仅为 0.004。显示母语者的近义词知识（语感）有很高共性，且表现稳定。各组数据按照降序排序后再加以对照，可以观察到不同被试组在近义词辨析表现上的稳定性有明显的差异（图 2）。从下降趋势线的斜率看，M'的数据最接近母语者，显示机器学习模型在大数据集上的训练和测试也具有较高的稳定性，但总体知识水平显著低于母语者，介于高级和中级二语学习者之间。二语者的下降斜率较大。而且高级水平二语者和中级水平二语者仅在中高分区域斜率相当，到低分区域，中级水平二语者的下降幅度显著高于高级水平二语者，显示在更为困难的近义词辨析任务上，两组二语者的语言能力差

[10] 本文的机器学习模型仅针对单选题进行训练，未考虑多选情况。因而对多选题也仅能当作单选题作答，包括零选题（答案为"都不可以"），也会选一个词作为答案输出。如按从宽评分标准，机器给出的多选题的答案，均在正确答案范围中，比二语者表现略好（机器与两组二语者的得分分别是 100,、90、87）。当然，18 道多选题中 2 选 2 的题有 8 道，3 选 3 的题 1 道（这意味着从宽评分标准下，可以白拿 9 分），3 选 2 的题 6 道，4 选 2 的题 3 道（后面这 9 道题使二语者和机器的得分稍微有了区别），从机器答题只选 1 个答案的角度看，机器按从宽标准计分也略占便宜（虽不容易选对，但也更不容易犯错）。

异较大。M 的下降趋势成阶梯状。这主要是因为机器在 100 题测试集上只是一个样本的成绩，跟其他组的数据为多个样本的平均分显著不同。从这点上说，可能 M’的成绩数据跟 S 更具可比性。

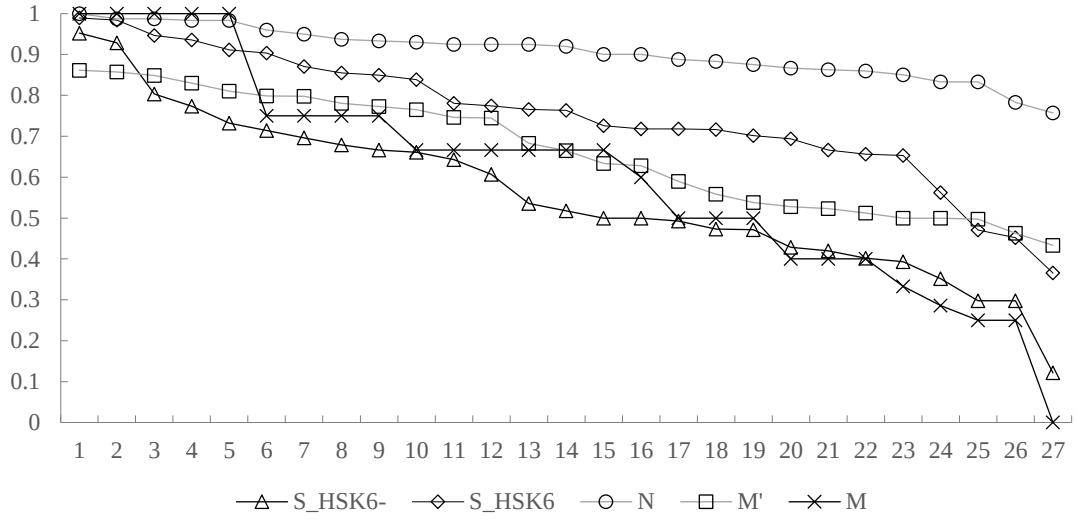

图 2　在 27 组近义词上测试得到的五组成绩数据按降序排序

4.2 各组测试成绩差异性与相关性分析

表 7 显示的机器测试成绩均值与二语者均值相对比较接近。但机器与二语者在本文测试集上的表现差异性到底如何，还需进一步进行统计检验。我们利用 Excel 内置的 T 检验功能（假设双样本异方差）[11]，对 6 组测试成绩数据进行了差异显著性分析，显著性水平 P 值设置为 0.05，结果如表 8 所示（表中数据为双尾 P 值，若 P>0.05，则表示差异不显著）：

表 8 六组测试成绩数据间差异显著性分析结果

	M’	M	N	S-HSK6	S-HSK6-
M	**0.36**				
N	9.02E-10	4.18E-06			
S-HSK6	0.035	0.022	0.00004		
S-HSK6-	0.03	**0.43**	6.91E-10	0.00024	
S	**0.95**	**0.41**	5.55E-08	-	-

从表 8 检验结果可知，统计意义上，M’和 M 无显著差异，M’、M 分别和 S 无显著差异，M 和 S-HSK6-无显著差异（表中 P 值大于 0.05，以黑体标识）。从这个结果可以推测，机器在大规模真实语料上的近义词辨析测试结果可以在相当程度上

[11] 我们在 SPSS22.0 中使用 Kolmogorov-Smirnov 检验方法，发现各组样本数据均呈正态分布。在 Levene 方差齐性检验中，发现这六组数据的方差不相等。

预测二语者某个特定近义词辨析词集上的表现，特别是中级（HSK6 级以下）二语学习者的水平。

在考察了各组被试成绩的差异性之后，本文又进一步考察了各组被试在 27 组近义词集上的测试成绩的相关性。利用 Excel 内置的 CORREL 函数计算 6 组数据的线性相关系数结果如下面表 9 所示：

表 9 六组测试成绩数据间相关系数

	M’	M	N	S-HSK6	S-HSK6-
M	0.2402				
N	0.2751	0.4085			
S-HSK6	0.1520	0.6256	0.7449		
S-HSK6-	0.1848	**0.6810**	0.5743	0.8886	
S	0.1742	**0.6737**	0.6741	-	-

结果显示，M 跟 S 两组数据的相关系数为 0.6737，经 T 检验公式计算（置信区间水平设为P=0.95）可知，M 跟 S 两组数据的 T 统计值为 5.8974，大于临界值 1.7081，因此，两组数据之间有较强的正相关。这在很大程度上说明，用机器学习系统进行近义词辨析测试所得的成绩数据，对于预估特定的近义词集和相应的一套测试题对于二语学习者的难度，有一定的参考价值。限于篇幅和数据处理成本方面的考虑，本文未进一步对近义词集分小类计算成绩相关系数，仅给出了总体相关性计算结果。如果分小类后再计算相关度，在某些具体类别的近义词集辨析任务上，机器学习系统的表现跟二语者的表现，可能会有更强的相关性。若据此为计算机辅助组织和实施近义词辨析的教学实践，应具有较高的参考价值。

4.3 机器与二语者答题分布差异考察与初步分析

近义词辨析，从名称来说，辨析的是词义的区别。就人的感觉而言，词义差别有的在用法上有形式区别，有的则仅有意义差别，而没有明显的形式区别。也正是基于这种语感，本文将 100 道测试题按照有无形式特征线索做了区分标记[12]（附录 2），以便考察被试在具体测试题上的表现跟近义词辨析形式线索之间的关系。我们统计了 100 题中得分大于 0.6 的题的形式特征标记分布情况，结果如下面表 10 所示。

表 10 从形式特征角度看各组被试的答题正确率分布

分值≥0.6 题数	N		S		M		总题数
总题数	98	98.00%	68	68.00%	58	58.00%	100
无形式特征题数	53	98.15%	39	72.22%	37	68.52%	54

[12] 本文所说的形式特征包括句法形式特征、搭配（一般是特定词语）特征。无形式特征指从人的角度来看，只能根据对整句意思的理解来从候选词集中选择合适词语填入。对每道测试题做特征标注时，只能从“句法特征、搭配特征、无特征、假特征”四项中选择其一进行标注。

搭配特征题数	29	96.67%	15	50.00%	11	36.67%	30
句法特征题数	11	100%	11	100%	8	72.73%	11
含假搭配特征题数	5	100%	3	60%	2	40%	5

需要说明的是，机器每道题的分值要么是 1（表示对），要么是 0（错），表 10 中 M 一列跟 N 和 S 两列的性质有所不同。N 和 S 两列的统计结果是基于所有参加测试的母语者和二语者在每道题上的平均分值。此外，上表是以全部 100 题作为统计对象的结果。其中二语者、机器的数据实际上仅限于单选题（二语者和机器多选题的分值均低于 0.6）。如前所述，二语者和机器在多选题上的正确率显著低于单选题，也远低于母语者正确率。对于母语者来说，如果以单选题作为统计范围，则各类题型的正确率均为 100%[13]。

从表 10 所示统计结果来看，人类二语者被试与机器学习系统对比有几个特点值得注意：（1）机器对“意义”的理解能力超出我们的预期。M 和 S 在“无形式特征”题中成绩超过 0.6 的题数数量接近，为 37:39。这意味着机器学习在没有明显形式特征提示线索的情况下，也能较好地模拟人的近义词辨析能力，做出比较恰当的选择。正如 Huang 等（2017）在英语近义词辨析任务实验时所发现的那样，Bi-LSTM 对句子之间的全局性差异有很好的把握能力，本文在汉语近义词辨析任务上的实验也展示了类似结果。基于神经网络的机器学习模型从句子中词语的分布表示（distributional representation）出发对句子建模，在一定程度上达到了从整体上把握句子“语义”的效果。（2）“搭配”特征在近义词辨析中通常作为主要手段进行分析和讲解，但在实际测试中，二语者和机器在有搭配线索题上的表现并不算太好。表 10 中的“搭配特征题”总数为 30，其中包含了多选题，如果去掉多选题，仅以 19 道单选题来说，在这些测试题上，得分超过 0.6 的比例，S 为 78.95%，M 为 57.89%。就机器来说，比例低于在无形式特征题上的表现。这提示我们，在统计意义上，搭配特征作为辨析线索，可能并不是以一种在统计上很凸显的模式出现，机器不见得能有效捕捉到[14]。（3）从“句法特征”与“搭配特征”的对比来看，机器在前一类测试题上的表现更好（72.73%>57.89%），这显示基于句法范畴的区别特征可能在分布上更具统计意义（或者说比“搭配特征”分布范围更广），机器更容易捕捉到这类信息。从二语者跟机器的成绩对比来看，二语学习者把握句法特征（或者说是“抽象形式”）的能力要强于机器。值得指出的是，含假搭配形式特征的题对于母语者没有起到“挖坑”效果，但对于二语者和机器，都造成了误判。不过由于样本量较小，不一定能说明问题，还有待今后设计更合理的实验来验证。

下面将附录 3 所示 27 组近义词测试成绩以折线图形式展示，以比较机器和二语者在具体的近义词集上的表现差异。

[13] 具体数据分别是：单选题总题数（78，100%），包括无形式特征题数（44，100%），搭配特征题数（19，100%），句法特征题数（11，100%），和假搭配特征题数（4，100%）

[14] 事实上，二语者的测试成绩也表明，学习者对搭配特征的把握效果并不理想。这说明从搭配的角度辨析近义词，对二语学习者也是比较大的挑战。

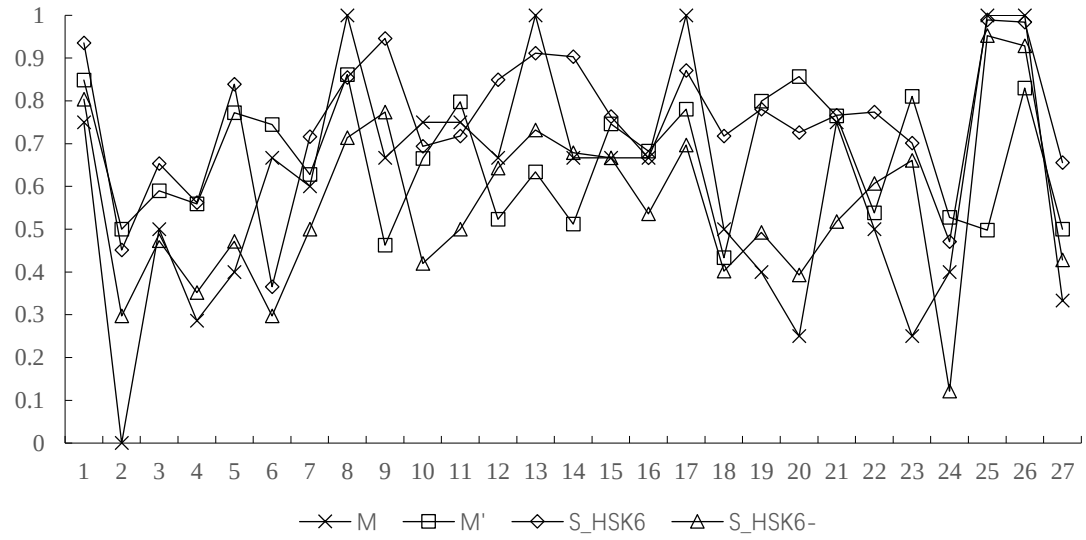

图 3　在 27 组近义词上测试得到的四组成绩数据对照图

整体而言，图 3 显示在大多数近义词集上，机器成绩与二语者成绩之间相差较大。像第 1 组这样成绩相对接近的近义词集占少数。就机器和二语者在每组具体的近义词集上的成绩表现而言，影响因素较多，较难概括出一般性的规律。比如像第 2、4、6、24 这四组近义词集，是二语者测试成绩排序最低的四组。机器成绩大多也较低，但跟二语者成绩并不相近。影响因素包括（1）近义词集中候选词较多（3 到 4 个）；（2）涉及题型包括多选题和答案为“都不可以”的题；（3）近义词本身辨析的形式线索较少或较难掌握，等等。机器成绩相对突出的有第 8、13、15、17、25、26 这几组，其中第 8 组“适合、合适”，第 15 组“有点儿、一点儿”，第 17 组“买、购买”，第 26 组“人、人们”的形式区别线索较为显著，机器和二语者得分均较高。这说明形式线索明确的近义词辨析任务相对容易完成。第 13（也、又、再、还）和 25（他、她、它）这两组，M 成绩很高，二语者成绩也相对较高，但 M’成绩却偏低，原因可能可以归结为人工测试题偏容易，但还有待进一步考察。类似的情形还出现在第 19（吗、吧、呢）、20（不管、尽管）、23（两、二、俩）这三组，M 跟 M’相差非常大，但关系倒过来，均为 M 远低于 M’。除 M 的题型难度更高（含多选、零选等）外，可能人工测试题跟真实训练语料的分布差异比较大也是影响因素，导致机器在真实语料测试集上表现正常，但在人工测试题上却容易出现误判。不过这也只是推测，有待进一步论证。

5. 结语

本文在前人有关近义词辨析的语言本体和教学研究基础上，将机器学习技术引入中文近义词辨析任务，对机器、母语者和二语学习者做了初步的实验对比研究。实验中所用近义词集的选择和试题的制作，均充分考虑平衡性和周全性原则。相比以往的类似实验，本文实验设计有两个特色：一是将机器测试与人类测试进行对照；二是将机器测试分为真实语料测试集和人工构造测试集进行对照。人工测试题的题

型设置比以往机器测试题的类型更丰富，更接近面向人类测试的题型设计。通过对测试成绩进行初步的统计和分析，我们得到了一些有参考意义的结论：（1）机器学习程序在中文近义词辨析任务上的总体表现，跟二语学习者具有明显的可比性。（2）机器学习程序跟二语学习者（特别是中级水平的汉语学习者）在近义词辨析任务上的总体表现有较强的相关性。（3）从二语者和机器学习系统在多选题与单选题上所得平均分的差距可以看到，题型对成绩的影响非常显著。（4）在影响近义词辨析的各项因素中，从“意义”（包括语体）角度辨析的重要性不亚于从形式区别特征的角度辨析，从抽象的句法形式特征角度进行辨析的重要性不亚于从搭配的角度进行辨析。

前人在英语近义词辨析任务上的实验表明，机器学习系统可以为二语者学习近义词提供辅助，为学习者自动推荐例句（Huang 等 2017）。本文在中文近义词辨析任务上的人机对照实验研究也有类似启示：机器学习技术有可能在汉语近义词辨析的教学方面起到辅助作用，可能的途径不仅包括推荐例句，还可以辅助教师构造近义词辨析测试题集，对测试集难度进行分级评估。此外，经过合理的系统设计，有可能形成人机互助的近义词辨析教学平台，由机器辅助人类教师构建近义词辨析知识库、题库，对近义词集的辨析难度，试题的难度进行分级，从而更为科学地组织和实施教学。从 NLP 的角度来说，目前近义词辨析任务无论是英文还是中文，都还缺乏广为接受的标准的可比测试集。本文在近义词集的选择和试题制作方面积累的经验，有可能为研制中文近义词辨析任务的标准测试集提供借鉴。

本文实验所用机器学习模型仅为基线系统（baseline system），即利用常用的深度学习技术构建的处理近义词辨析任务的程序。相比已有的在英文近义词辨析任务上的系统的表现来说，机器学习系统的设计还有优化空间，可以继续提升在中文近义词辨析任务上的性能表现。另外，在近义词集的选择、试题制作、试题类型标注等方面，此次试验均为全人工完成。未来可以更多地借助程序来进行辅助，包括内容设计和管理，针对某些特定类型的近义词集进行更为深入的专项测试等等，可以进一步提升实验的科学性和效率。

致谢：本文研究工作得到教育部人文社科重点研究基地重大项目（编号 13JJD740001、15JJD740002）和国家自然科学基金项目（编号 61876004）资助。北京大学中文系万艺玲老师和白一瑾老师为组织留学生参与此次实验提供了很多帮助。中文系现代汉语专业硕士研究生裴晓倩、施朝、柯丽珍等多人参与了母语者语感调查、近义词集筛选、测试题特征标注等工作，在此一并致谢！

参考文献

Chen, S., He, W., Prasetyo, P. K., & Yu, L. (2012, September). Skip N-gram Modeling for Near-Synonym Choice. In *Proceedings of the Twenty-Fourth Conference on Computational Linguistics and Speech Processing (ROCLING 2012)* (pp.163-175). [陈士婷,何维晟,关松坚,禹良治. (2012). 应用跳脱语言模型于同义词取代之研究.见第 24 届计算语言学和语音处理会议议程, 163-175.]

Edmonds, P. (1997, July). Choosing the word most typical in context using a lexical co-

occurrence network. In *Proceedings of the 35th annual meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics* (pp. 507–509). Madrid, Spain.

Gao, Z. (2016). A new probe into theory and method of comparing near-synonym: on teaching near-synonyms at these two levels. *Journal of Huaibei Normal University (Philosophy and Social Science)*, 2,141-145. [高再兰. (2016). 近义词比较理论与方法新探——兼谈两个层面的近义词教学.*淮北师范大学学报（哲学社会科学版）*, 2 , 141-145.]

Hong, J. (2014). Chinese Near-Synonym Study Based on the Chinese Gigaword Corpus and the Chinese Learner Corpus. In X. Su and T. He (Eds.). CLSW 2014, LNAI 8922 (pp. 329–340). Switzerland: Springer International Publishing.

Hong, W. (2012). A review of the synonyms studies in teaching Chinese as a second language. *TCSOL Studies*, 4, 44-51. [洪炜. (2012). 面向汉语二语教学的近义词研究综述.*华文教学与研究*, 4, 44-51.]

Hong, W. (2017). The effect of in-class differentiation on L2 acquisition of Chinese near-synonyms. *Chinese Language Learning*, 4, 84-91. [洪炜. (2017). 课堂显性辨析对汉语二语者习得近义词差异的影响.*汉语学习*, 4, 84-91.]

Huang, C., Chen, M., & Ku, L. (2017, April). Towards a Better Learning of Near-Synonyms: Automatically Suggesting Example Sentences via Filling in the Blank, *International World Wide Web Conference Committee* (IW3C2) (pp.293-302). Perth, Australia.

Lu, F. (2016). Corpus-based near-synonym teaching in TCSL. *Journal of Yunnan Normal University* (Teaching and Research on Chinese as a Foreign Language), 5, 49-56. [陆方喆. (2016). 基于语料库的对外汉语近义词教学.*云南师范大学学报（对外汉语教学与研究版）*, 5, 49-56.]

McCarthy, D. & Navigli, R. (2007, June). The English lexical substitution task. In *Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval-2007)* (pp. 48–53). Prague, Czech Republic.

McCarthy, D., & Navigli, R. (2009). The English lexical substitution task. *Language Resources and Evaluation*, 43(2), 139-159.

Nation, P. & Newton, J. (1997). Teaching vocabulary. In James Coady & Thomas Huckin. *Second Language Vocabulary Acquisition: A Rationale for Pedagogy* (pp.238-254). Cambridge: Cambridge University Press.

Wang, T. & Hirst, G. (2010, August). Near-synonym Lexical Choice in Latent Semantic Space. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)* (pp. 1182–1190). Beijing, China.

Yu, L., & Chien, W. (2013). Independent component analysis for near-synonym choice. *Decision Support Systems*, 55(1), 146-155.

Yu, L., Chien, W., & Chen, S. (2011, November). A Baseline System for Chinese Near-Synonym Choice. In *Proceedings of the 5th International Joint Conference on Natural Language Processing* (pp. 1366–1370). Chiang Mai, Thailand.

Zhang, B. (2007). Synonymy, near-synonymy and confusable word: A perspective transformation from Chinese to interlanguage. Chinese Teaching in the World,3,98-107. [张博. (2007). 同义词、近义词、易混淆词：从汉语到中介语的视角转移.*世界汉语教学*, 3, 98-107.]

Zhao, S., Zhao, L., Zhang, Y., Liu, T., &Li, S. (2007, June). HIT: Web based Scoring Method for English Lexical Substitution. In *Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval-2007)* (pp.173–176). Prague, Czech Republic.

Zhao, X., & Liu, R. (2005). Some basic problems in compiling *Practical dictionary of near-synonyms for foreigners. Lexicographical Studies*,4,57-67. [赵新,刘若云. (2005). 编写《外国人实用近义词词典》的几个基本问题.*辞书研究*, 4, 57-67.]

附录 1 本文实验所用的近义词（含易混淆词）组（27 组词）

序号	词语 1	词语 2	词语 3	词语 4	近义/易混淆	形式区别特征		
						词类	句法形式	搭配
1	为了	因为			易混淆			
2	难过	难受			近义			
3	一再	再三			近义		┬	┬
4	认识	了解	知道	理解	近义			
5	不	没			近义			
6	经验	经历			易混淆			
7	能	可以	会		近义			
8	合适	适合			近义	┼	┼	
9	结果	后果	成果		近义			
10	骄傲	自豪			近义		┬	
11	可惜	遗憾			近义		┬	
12	还是	要么	或者		近义		┬	┬
13	再	也	又	还	近义			
14	希望	愿望			近义	┬		
15	有点儿	一点儿			易混淆	┼	┼	
16	相信	信任			近义			
17	买	购买			近义		┬	┬
18	次	趟	遍		易混淆			
19	不管	尽管			易混淆		┬	┬
20	吗	吧	呢		易混淆		┬	┬
21	着	了	过		易混淆			┬
22	哎	唉			易混淆			
23	二	两	俩		近义	┬		
24	向	往	朝		易混淆			
25	他	她	它		易混淆			
26	人	人们			易混淆		┼	
27	千万	万万			近义		┬	

说明：表中“形式区别特征”栏的单元格中填“┼”，表示有较明显的差异（比如“适合”跟“合适”分别是动词和形容词，句法差异较大）；填“┬”表示有一定的差异（比如“一再”跟“再三”在状语位置区别不明显，但“再三”可以出现在动词后，如“考虑再三”。“一再”没有这种用法）。不填表示无明显的形式区别线索。主要是意义上的区别。

附录 2 近义词辨析测试题（27 组词语，共 100 题）

题号	组号	题干	特征	参考答案
1	1	我们（　　）解决这个问题采取了不少的方法。	无特征	为了
2	1	她来中国学习汉语，是（　　）她觉得汉语很有用。	无特征	因为
3	1	（　　）他一个人在国外学习，所以他的妈妈非常担心他。	搭配	因为
4	1	我们这次活动是（　　）保护环境，所以希望大家都能参与进来。	假搭配	为了
5	2	他们听说这件事以后，心里非常（　　）。	搭配	难过|难受
6	2	他生病了，总咳嗽，非常（　　）。	搭配	难受
7	2	她的脸色非常（　　），像是在生气。	搭配	都不可以
8	3	他（　　）表示今后一定努力学习汉语。	无特征	一再|再三
9	3	你的成绩为什么（　　）下降？	无特征	一再
10	3	他考虑（　　），最后决定开一家书店。	句法	再三
11	3	他（　　）再四地向老师表示感谢。	搭配	再三
12	4	要想真正地（　　）一个国家，最好去那个国家生活一段时间。	无特征	认识|了解
13	4	我完全不能（　　）你现在的心情。	无特征	理解
14	4	他的家人都不（　　）他的病情。	无特征	了解|知道
15	4	我（　　）小明什么时候回国。	无特征	知道
16	4	我们已经（　　）很多年了。	无特征	认识
17	4	我的家人都十分（　　）妈妈的脾气。	搭配	了解
18	4	员工们都十分（　　）公司目前的情况。	搭配	了解|理解
19	5	小明为什么（　　）去爬长城呢？	无特征	不|没
20	5	昨天下大雨了，所以我（　　）去爬长城。	无特征	没
21	5	在公共场所，请（　　）吸烟。	搭配	都不可以
22	5	明天我要去机场接朋友，所以（　　）去爬长城了。	句法	不
23	5	对不起，你说得太快了，我（　　）听懂你说的话。	句法	没
24	6	他的生活（　　）非常丰富。	搭配	经验|经历

25	6	她当了三十年的汉语老师，有着丰富的教学（　　）。	无特征	经验
26	6	他希望以后再也不要重复这段痛苦的（　　）了。	无特征	经历
27	7	现在我的牙不疼了，(　　)吃饭了。	无特征	能|可以
28	7	他特别（　　）睡，一睡就是一整天。	无特征	能
29	7	你有空儿的时候，（　　）看看电影放松放松。	无特征	可以
30	7	你们不要再吵了，这样下去我（　　）失眠的。	无特征	会
31	7	让他一个人出差（　　）行吗？	搭配	能
32	8	这件西装不大不小，正（　　）。	句法	合适
33	8	她非常喜欢吃辣的，四川菜正（　　）她的口味。	句法	适合
34	9	比赛刚结束，运动员都在等待比赛（　　）。	搭配	结果
35	9	根据最新研究，地下水一旦被污染，将会产生极其严重的（　　）。	搭配	后果
36	9	这项科技（　　）将运用到人们的生活中。	搭配	成果
37	10	我们为自己的祖国感到（　　）！	搭配	骄傲|自豪
38	10	虽然他得了好几次第一名，但他一点儿也不（　　）。	无特征	骄傲
39	10	这个机会增强了她的自信心和（　　）感。	无特征	自豪
40	10	蜿蜒万里的长城是中华民族的（　　）。	无特征	骄傲
41	11	这么好的机会，他都错过了，挺（　　）的！	无特征	可惜|遗憾
42	11	对于这一事件的发生，我们深表（　　）。	搭配	遗憾
43	11	这件衣服还能穿，扔了有点儿（　　）。	无特征	可惜
44	11	这是一次非常难得的机会，（　　）他没有抓住。	无特征	可惜
45	12	你想喝茶（　　）喝咖啡？	无特征	还是
46	12	你要么战胜困难，（　　）被困难战胜。	搭配	要么
47	12	我们坐地铁去（　　）打车去都可以。	无特征	或者
48	13	你刚才唱得真好听，（　　）唱一个吧。	无特征	再
49	13	这是老王的主意，我（　　）同意。	无特征	也

50	13	你每年生日的时候，我都会送你一件生日礼物。明天（　　）是你的生日了，你想要什么生日礼物呢？	无特征	又
51	13	我记得你十几年前在中国学过汉语，现在（　　）会说吗？	无特征	还
52	14	我对未来充满了（　　）和信心。	无特征	希望
53	14	（　　）我们这次合作愉快。	无特征	希望
54	14	他通过努力终于实现了自己心中的（　　）。	无特征	愿望
55	15	他现在（　　） 累，想休息一会儿。	句法	有点儿
56	15	我觉得今年夏天比去年热（　　）。	句法	一点儿
57	15	这件衣服的质量（　　）好。	无特征	都不可以
58	16	我（　　）你一定能学好汉语。	无特征	相信
59	16	夫妻之间应该互相关心，互相（　　）。	搭配	信任
60	16	公司的老板非常（　　）他。	无特征	相信|信任
61	17	太饿了，我去超市（　　）个面包吃。	句法	买
62	17	本店新到一批服装，欢迎新老顾客前来（　　）。	无特征	购买
63	17	如果你同意（　　）房的话，我们可以去售楼处看看。	无特征	买
64	17	双方当场就签订了（　　）合同。	无特征	购买
65	18	这个月他去了三（　　）广州了。	无特征	次|趟
66	18	在这半年里，他先后动了三（　）手术了。	无特征	次
67	18	他花了十年的时间把图书馆里的书都读了一（　）。	无特征	遍
68	18	这（　　）开往北京的列车就要出发了。	搭配	趟
69	19	（　　） 遇到的困难大不大，我们都要完成这项任务。	句法	不管
70	19	他没有来，（　　）我邀请了他。	无特征	尽管
71	19	（　　）天冷天热，他一运动就会出很多汗。	句法	不管
72	19	（　　）怎么解决好大家的困难，是他最关心的。	假搭配	都不可以
73	19	（　　）遇到的困难很大，但我们也要坚持下去。	句法	尽管

74	20	你是新来的同学（　　）？	搭配	吗|吧
75	20	要是我不说，难道你就不认识我了（　　）？	假搭配	吗
76	20	我们一直在这儿等他，要是他不来了（　　）？	搭配	呢
77	20	我大概拿错书了（　　）？这好像不是我的书。	假搭配	吧
78	21	他看（　）半天菜单，一个菜都没有点。	无特征	了
79	21	爷爷喜欢在沙发上坐（　　）看报纸。	无特征	着
80	21	我从来没有在大草原上骑（　　）马。	假搭配	过
81	21	他关注和思考的是随着历史发展而不断变化（　　）的人和人的关系。	无特征	着
82	22	（　　），我倒是有一个好主意。	无特征	哎
83	22	（　　），他长长地叹了一口气。	无特征	唉
84	23	开学了，玛丽刚升到（　　）班学习汉语。	搭配	二
85	23	医生让他在家里好好儿休息（　　）天。	搭配	两
86	23	他有（　　）朋友，他们的关系非常好。	无特征	俩
87	23	老板，给我来（　　）斤苹果。	搭配	二|两
88	24	年轻人，你们正在走（　　）美好的未来，祝福你们！	搭配	向
89	24	他说的这些话都是气话，你可别（　　）心里去。	搭配	往
90	24	这位明星（　　）我们挥了挥手。	搭配	朝|向
91	24	这条小路通（　　）山顶。	搭配	向|往
92	24	小孩子们（　　）东边跑去了。	无特征	朝|向|往
93	25	这位姑娘的爸爸是一名医生，（　　）的工作很辛苦。	无特征	他
94	25	这位男护士的女朋友长得非常漂亮，可最近听说这位男护士和（　　）分手了。	无特征	她
95	25	他奶奶家里养着一条狗，（　）的毛是棕色的。	无特征	它
96	26	傍晚，很多（　　）在公园里散步。	无特征	人
97	26	（　　）都说他是一位好医生。	无特征	人们
98	27	他（　　）没有想到，苦心经营了五年的公司就这样倒闭了。	搭配	万万

99	27	你们在外面（　　）要小心，不要被骗了。	搭配	千万
100	27	这件事情特别重要，（　　）不可马虎。	搭配	万万|千万

问卷样题：
请在划线处填入合适的词语，可以填一个或多个。如果没有合适的词语可填，就填“×”（表示“都不可以”），如果不知道该选哪个词，就填“我不知道”。例如：

A 时候　　B 时间

（1）请不要忘记我们集合的__B__和地点。
（2）__A 、B__不早了，我该走了。
（3）刚到中国__×__，我一句汉语都不会说。
（4）初学汉语__我不知道__，我感觉很迷茫。

附录 3 近义词辨析实验成绩列表（成绩取值在 0-1 之间）

组号	近义词项	M’	M	S_HSK6	S_HSK6-	S	N
1	为了 因为	0.849	0.750	0.935	0.804	0.873	0.988
2	难受 难过	0.500	0.000	0.452	0.298	0.379	0.883
3	一再 再三	0.590	0.500	0.653	0.473	0.568	0.863
4	了解 理解 知道 认识	0.559	0.286	0.562	0.352	0.462	0.757
5	不 没	0.773	0.400	0.839	0.471	0.664	0.930
6	经历 经验	0.745	0.667	0.366	0.298	0.333	0.783
7	会 可以 能	0.628	0.600	0.716	0.500	0.614	0.920
8	合适 适合	0.862	1.000	0.855	0.714	0.788	0.925
9	后果 成果 结果	0.463	0.667	0.946	0.774	0.864	0.933
10	自豪 骄傲	0.665	0.750	0.694	0.420	0.564	0.888
11	可惜 遗憾	0.798	0.750	0.718	0.500	0.614	0.938
12	或者 要么 还是	0.523	0.667	0.849	0.643	0.751	0.900
13	也 再 又 还	0.634	1.000	0.911	0.732	0.826	0.925
14	希望 愿望	0.512	0.667	0.903	0.679	0.797	0.983
15	一点儿 有点儿	0.747	0.667	0.763	0.667	0.718	0.867
16	信任 相信	0.683	0.667	0.667	0.536	0.605	0.833
17	买 购买	0.781	1.000	0.871	0.696	0.788	1.000
18	次 趟 遍	0.433	0.500	0.718	0.402	0.568	0.900
19	不管 尽管	0.799	0.400	0.781	0.493	0.644	0.960
20	吗 吧 呢	0.857	0.250	0.726	0.393	0.568	0.925
21	了 着 过	0.765	0.750	0.766	0.518	0.648	0.988
22	哎 唉	0.538	0.500	0.774	0.607	0.695	0.850
23	两 二 俩	0.811	0.250	0.702	0.661	0.682	0.875
24	向 往 朝	0.528	0.400	0.471	0.121	0.305	0.860
25	他 她 它	0.498	1.000	0.989	0.952	0.972	0.983
26	人 人们	0.830	1.000	0.984	0.929	0.958	0.950
27	万万 千万	0.500	0.333	0.656	0.429	0.548	0.833

