DOI: https://doi.org/10.26549/jxffcxysj.v2i4.2111

Using Corpus and Deep Learning to Construct Personalized Online English Learning Course

Meihua Lu　Qiaoling Wang

Beijing Vocational College of Agriculture, Beijing, 102442, China

Abstract

This paper first introduces the concept of corpus and deep learning, and then discusses the functions of corpus and deep learning in dealing with English. According to the vocabulary grading function of corpus and the actual English level of students, it paper proposes to customize English learning courses for students and publish customized courses online through the existing software with network functions, and provides students with round-the-clock counseling through deep learning functions.

Keywords

artificial intelligence; corpus; deep learning; curriculum construction

Fund Project

Project Title: Development of Computer Aided Translation Teaching System for English Corpus in Higher Vocational Colleges (Project No.: 2017022).

利用语料库和深度学习建设个性化网络英语学习课程

卢美华　王巧玲

北京农业职业学院，中国·北京 102442

摘　要

本文首先介绍了语料库和深度学习的概念，然后分别论述了语料库和深度学习在处理英语上的功能。根据语料库词汇分级功能，结合学生实际英语水平，提出为学生量身定制英语学习课程，通过现有的具有网络功能的软件把定制化的课程发布在网上，通过深度学习功能为学生提供全天候的辅导。

关键词

人工智能；语料库；深度学习；课程建设

基金项目

课题名称：高职英语语料库计算机辅助翻译教学系统研制（项目编号：2017022）。

1 引言

随着互联网和人工智能的发展，中国先后发布了“互联网 +”规划和“人工智能”规划。为了克服英语教学投入产出比小的弊端，在全球信息化时代，结合大数据、语料库和人工智能为学生提供全天候、智能化、个性化的英语学习课程迫在眉睫、刻不容缓。

2 人工智能发展历史

人工智能 (Artificial Intelligence，英文缩写为 AI) 是计算机科学或智能科学的一个分支，涉及研究、设计及应用智能机器。其诞生于上个世纪 30 年代，经历了不同的发展阶段：（1）孕育奠基期（20 世纪 30 年代到 1956 年）；（2）形成发展期（1956 年到 20 世纪 60 年代末）；（3）低谷瓶颈期（20 世纪 70 年代到 80 年代初期）；（4）专家系统推广阶段（20 世纪 80 年代至 90 年代），这一时期内计算机识别、自然语言理解和处理、机器翻译等技术也取得了长足的发展；（5）深度学习引领发展阶段（21 世纪至今），中国于 2017 年印发《新一代人工智能发展规划》，分三步走，从人才培养、产业助推和企业扶持三方面促进人工智能理论、技术和应用发展，力争到 2030 年使中国成为全球主要人工智能创新中心之一。

当前，人工智能已经成为资本、技术、舆论追捧的热点，

教学方法创新与实践·第 02 卷·第 04 期·2019 年 08 月
DOI: https://doi.org/10.26549/jxffcxysj.v2i4.2111

甚至有人预测它将开启第四次工业革命的大门。在可以预见的未来，各国都将投入巨大的精力推动人工智能向更高层次迈进，人工智能的时代正在逐步向我们走来。

3 传统英语教材的弊端

英语教材是指英语教学中使用的教科书以及与之配套使用的教师教学用书、配套读物、各种练习册、手册、挂图、音像制品、计算机软件等教学资料。现行各级各类英语教材在发展性、思想性、科学性和系统性方面都有其优点，但是语言知识是动态发展的。国家语委语言应用研究所 2007 年对汉语新词语的调查数据显示：当代汉语每年出现 1000 个左右的新词语，平均每天出现 3 个，英语也不例外。而我们的英语课程要体现时代的变化，教材也必须是动态变化，随着时代的变化而变化。

众所周知，因材施教是自古到今的教学原则，但是在现有的教学体制和课堂教学方式下，要做到这一点是非常困难的，这也是目前大学、中学、小学学生出现各种学习障碍的原因之一，要真正做到因材施教就要结合现代化的信息技术，利用大数据、语料库和深度学习技术为学生提供定制化的学习课程。

4 语料库语言学及其在英语教学中的应用

语料库语言学（Corpus Linguistics）是 20 世纪中后期兴起的一门语言研究科学。关于语料库有三点基本认识：语料库中存放的是在语言的实际使用中真实出现过的语言材料；语料库是以电子计算机为载体承载语言知识的基础资源；真实语料需要经过加工（分析和处理），才能成为有用的资源 [1]。

语料库可以分为：通用语料库（general corpus），专用语料库（specialized corpus），共时语料库（synchronic corpus），历时语料库（diachronic corpus），口语语料库（spoken corpus），笔语语料库（written corpus），本族语料库（native speaker’s corpus），学习者语料库（learner corpus），单语语料库 (monolingual corpus) 和平行 / 双语语料库（parallel corpus）和多语语料库（multilingual corpus）。

4.1 平行语料库标注示例

```
<s id="1">
  <w pos="r">我们</w>
  <w pos="ns">非洲</w>
  <w pos="c">及其</w>
  <w pos="p">在</w>
  <w pos="n">世界</w>
  <w pos="u">的</w>
  <w pos="n">地位</w>
  <w pos="d">正</w>
  <w pos="v">处在</w>
  <w pos="n">决定性</w>
  ……
  <w pos="w">。</w>
</s>
```

```
<s id="1">
  <w pos="PRP">We</w>
  <w pos="VBP"
 lemma="be">are</w>
  <w pos="IN">in</w>
  <w pos="DT">a</w>
  <w pos="NN">period</w>
  <w pos="IN">of</w>
  <w pos="JJ">decisive</w>
  <w pos="JJ">historical</w>
  <w pos="NN">significance</w>
  <w pos="IN">for</w>
  ……
  <w pos="w">.</w>
</s>
```

4.2 语料库的统计：概率和频率

概率（probability）是语料库语言学中最重要的基本概念之一。一个词的频率表示该词在语料库中出现的次数。如下是培根的《On Study》的概率统计：

序号	字词	出现次数	出现频率
1	,	11	7.2368
2	and	7	4.6053
3	for	7	4.6053
4	.	5	3.2895
5	;	5	3.2895
6	by	5	3.2895
7	in	5	3.2895
8	is	5	3.2895
9	the	4	2.6316
10	are	3	1.9737
11	much	3	1.9737
12	of	3	1.9737
13	too	3	1.9737
14	ability	2	1.3158
15	experience	2	1.3158
16	natural	2	1.3158
17	one	2	1.3158
18	ornament	2	1.3158
19	studies	2	1.3158
20	that	2	1.3158

4.3 语料库索引

索引（concordance）又称为“语境中的关键词”（Key Word in Context，KWIC) 指的是运用索引软件（concordancer）在语料库中查询某词或短语的使用实例，然后将所有符合条件的语言使用实例及其语境以清单的形式列出：

0: nationalities in Xinjiang **regard** the " Army as a great wall ， and soldiers as
1: society . It is extremely wrong to **regard** the " three-companionship"
2: representatives back then ， I **regard** the " Roundtable" as the biggest
3: still requires a Creator . Most proponents **regard** the " days" of creation as
4: find a number of Americans **regard** the (super collider) as a
5: and cannot **regard** the ' safe haven ' being constructed along the Turkish
6: does not **regard** the ' cessation of violence ' as affecting the methods it
7: protectionism . Yeltsin also expressed himself in this **regard** the other
8: both sides who **regard** the other point of view as ridiculous . People say '
9: source added As **regard** the other topics created by the said paper ，
10: TrueType hints ? most font vendors **regard** the more exotic capabilities of

4.4 搭配与类联接

早在上个世纪 50 年代，英国著名的语言学家 Firth 就提出了搭配（Collocation）的概念。Firth 认为，搭配是词语之间的“结伴关系”。为了更有效的分析词语搭配关系，人们提出了搭配强度（collocability）这一概念。概括的说，搭配并不是某个词语单方面的行为，我们至少应该从两个词语各自的出现频数（occurrences）和共现频数（co-occurences）两方面来考察搭配的强度 [2]。

如上图所示，regard 一词在右侧经常与介词 as 构成类联接。类连接可以看成是搭配的更高层次，与语言的句法有密切关系。

4.5 多词系列

多词系列（MWE/Multiword expressions）又称多词单位（MWU/Multiword Units），复现词组等，与此相关的说法（有的并不完全相同）还有词块（lexical chunks）、词簇（word clusters）、预制语块（prefabs or prefabricated chunks）、套语（formulatic sequences）、N 元组（N-grams）等，是近年来的一个研究热点。如下表：

序号	频数	3 元多词系列
1	2	the cash the
2	2	the trustees were
3	2	trustees were holding
4	1	and the cash
5	1	between the cash
6	1	cash the managers
7	1	cash the trustees
8	1	discovered there were
9	1	discrepancies between the
10	1	holding and the
11	1	in late October
12	1	it was discovered
13	1	it was only
14	1	late October this
15	1	managers thought the
16	1	October this year
17	1	only in late
18	1	that it was
19	1	the managers thought
20	1	there were discrepancies
合计	20	

从以上 3 元系列分析的统计结果可以看出，有些组合是有意义的。加大统计规模，有意义的组合出现的次数就会增加，像“in late October”、“trustees were holding”、“October this year”这些组合，也就是我们通常所说的词组，是要重点关注和学习的。

5 深度学习语料库中的语言知识

“深度学习”开源框架都可以用来构造自然语言处理（NLP）的语言模型，下面并举例说明“深度学习”在自然语言处理（NLP）中的应用场景。

5.1 词嵌入（word2vec）

自然语言是一套用来表达含义的复杂系统。在这套系统中，词是表义的基本单元。顾名思义，词向量是用来表示词的向量，也可被认为是词的特征向量。把词映射为实数域上向量的技术也叫词嵌入（word embedding）。近年来，词嵌入已逐渐成为自然语言处理的基础知识。

word2vec 里包含了两个模型：跳字模型（skip-gram）[1] 和连续词袋模型（continuous bag of words，简称 CBOW）[2]。

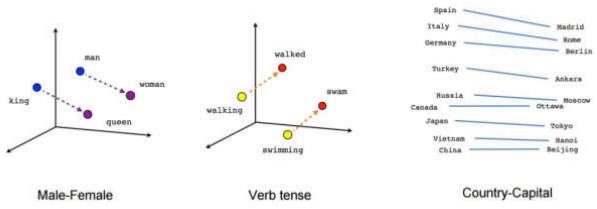

图 1 Word2vec 图形化

英语单词通常有其内部结构和形成方式。例如我们可以从“dog”、“dogs”和“dog-catcher”的字面上推测他们的关系，它们都有同一个词根 dog，但使用不同的后缀来改变词的意思。而且这个关联有着一定的推广性，例如“dog”和“dogs”的关系如同“cat”和“cats”，“dog”和“dog-catcher”的

教学方法创新与实践·第 02 卷·第 04 期·2019 年 08 月
DOI: https://doi.org/10.26549/jxffcxysj.v2i4.2111

关系如同“dish”和“dishwasher”。

使用预训练词向量求词与词之间的类比关系。例如，man（男人）:woman（女人）:: son（儿子）: daughter（女儿）是一个类比例子：“man”之于“woman”相当于“son”之于“daughter”。

“首都－国家”类比：“beijing”（北京）之于“china”（中国）相当于“tokyo”（东京）之于什么？答案应该是“japan”（日本）。

“形容词－形容词最高级”类比：“bad”（坏的）之于“worst”（最坏的）相当于“big”（大的）之于什么？答案应该是“biggest”（最大的）。

“动词一般时－动词过去时”类比：“do”（做）之于“did”（做过）相当于“go”（去）之于什么？答案应该是“went”（去过）。

5.2 中文分词、词性标注、实体名识别

中文分词：中文分词 (Chinese Word Segmentation) 指的是将一个汉字序列切分成一个个单独的词。分词就是将连续的字序列按照一定的规范重新组合成词序列的过程。我们知道，在英文的行文中，单词之间是以空格作为自然分界符的，而中文只是字、句和段能通过明显的分界符来简单划界，唯独词没有一个形式上的分界符，虽然英文也同样存在短语的划分问题，不过在词这一层上，中文比之英文要复杂得多、困难得多。

如中文句子：自然语言是人类交流和思维的主要工具，是人类智慧的结晶。

进过基于概率统计和深度学习的分词结果如下：自然 语言 是 人类 交流 和 思维 的 主要 工具 ， 是 人类 智慧 的 结晶 。

5.3 中文词性标注：给中文词语加上词类信息

如中文句子：周杰伦出生于台湾，生日为 79 年 1 月 18 日，他曾经的绯闻女友是蔡依林。

通过基于概率统计和深度学习的词性标注结果如下：

周杰伦 / 人名 出生 / 动词 于 / 介词 台湾 / 地名 ， / 标点 生日 / 名词 为 / 介词 79 年 / 时间短语 1 月 / 时间短语 18 日 / 时间短语 ， / 标点 他 / 人称代词 曾经 / 副词 的 / 结构助词 绯闻 / 名词 女友 / 名词 是 / 动词 蔡依林 / 人名 。/ 标点

5.4 实体名识别：识别人名、地名、机构名等

如中文句子：詹姆斯·默多克和丽贝卡·布鲁克斯、鲁珀特·默多克旗下的美国小报《纽约邮报》的职员被公司律师告知，保存任何也许与电话窃听及贿赂有关的文件。

通过基于概率统计和深度学习的词性标注结果如下： 詹姆斯·默多克 = 人名，鲁珀特·默多克旗 = 人名，丽贝卡·布鲁克斯 = 人名， 纽约 = 地名， 美国 = 地名

5.5 自动文摘

摘要是“一段从一份或多份文本中提取出来的文字，它包含了原文本中的重要信息，其长度不超过或远少于原文本的一半”。自动文本摘要旨在通过机器自动输出简洁、流畅、保留关键信息的摘要。

自动文本摘要通常可分为两类，分别是抽取式（extractive）和生成式（abstractive）。抽取式摘要判断原文本中重要的句子，抽取这些句子成为一篇摘要。而生成式方法则应用先进的自然语言处理的算法，通过转述、同义替换、句子缩写等技术，生成更凝练简洁的摘要。

5.6 神经机器翻译（NMT）

机器翻译的发展如下图：

图 2 机器翻译发展历史

上世纪 80 年代之前，机器翻译主要依赖于语言学的发展，分析句法、语义、语用等；之后，研究者开始将统计模型应用于机器翻译，这种方法是基于对已有的文本语料库的分析来生成翻译结果；2012 年至今，随着深度学习的兴起，神经网络开始被运用在机器翻译上构造神经网络机器翻译（Neural Machine Translation，NMT），并在短短几年取得了非常大的成果。

6 定制个性化和信息化的英语学习课程

网络化学习平台目前有蓝墨云班课和社交软件微信，我们需要做的是根据学生不同的基础制定相应的学习课程，并能提供全天候的指导。

6.1 分层分级

新生入学时，对全校学生进行摸底考试，目前是把学生

分成基础班和提高班。其实，经过多年实践，分班后也存在基础参差不齐的现象，有的学生词汇量欠缺，有的学生阅读能力欠缺，有的学生口语能力欠缺，有的学生听了欠缺，等等。

6.2 学习材料分级

我校大部分学生来自普通高中，普通高中生完成全部高中阶段英语学习，应该可以掌握 2000 个左右的词汇量。然而实际上很多学生并没有达到这个要求。到了高职阶段，就要根据学生掌握的词汇量进行分级。根据学生实际掌握的词汇量水平，从语料库抽取对应高一层次的词汇量的学习材料，避免难度过高的学习内容，这样学生才有可能跟上学习进度，保持对英语学习的兴趣。

6.3 通过蓝墨云班课或者微信推进学习内容

随着智能手机和互联网的普及，学生通过网络获取信息的渠道非常畅通，从网上购物、网上聊天可见一斑。通过蓝墨云班课或者微信把以上分级的英语学习课程分发给学生，并在网上提供相应的慕课、微课和翻转课堂，这样学生就可以各取所需，真正做到“按部就班”的学习。

6.4 深度学习语言知识是学生全天候的、永不疲倦的良师

在学习过程中难免有各种各样的问题，帮助学生解决学习过程中出现的问题刻不容缓。而传统班级课堂教学要面对全班几十个学生，有些学生的学习问题得不到及时的解决，打消了学生学习的积极性。深度学习能够学习到各种语言知识，能够及时、不厌其烦的解决学生提出的问题，为学生提供永不下课的课堂。

参考文献

[1] 马润聪 . 基于汉语常识知识库的词语语义相似度衡量研究 [广西师范大学硕士论文] . 广西 : 广西师范大学 ,2015.

[2] 刘雪扬 . 基于 BNC 语料库分析英语同义词搭配特征——以 able 和 capable 为例 [J]. 科教文汇 (下旬刊),2019(03):182-184.