北京大学学报(自然科学版) 第 58 卷 第 1 期 2022 年 1 月
Acta Scientiarum Naturalium Universitatis Pekinensis, Vol. 58, No. 1 (Jan. 2022)
doi: 10.13209/j.0479-8023.2021.100

大规模中文具体度词典的构建及推理技术

谢志鹏† 毕冉

复旦大学计算机科学技术学院, 上海 200433; †E-mail: xiezp@fudan.edu.cn

摘要 针对中文词语具体度资源的匮乏, 提出一种自动的中文词语具体度词典构造方法。该方法充分利用已有的英文词语具体度资源, 基于在线翻译工具和预训练词向量, 训练出中文词语具体度的多层感知器回归模型, 构造大规模的中文词语具体度词典。为了评估该中文词语具体度词典的性能, 设计两项基本的具体度推理任务: 词语级具体度推理和句子级具体度推理, 并通过人工标注的方式构造相应的评测数据集。实验结果表明构造的词语具体度词典可以有效地完成这两项推理任务。

关键词 词语具体度; 具体度推理; 多层感知器; 自然语言处理

Construction and Inference Technique of Large-Scale Chinese Concreteness Lexicon

XIE Zhipeng †, BI Ran

School of Computer Science, Fudan University, Shanghai 200433; † E-mail: xiezp@fudan.edu.cn

Abstract To solve the resource-lack problem of Chinese word concreteness, this paper designs and implements an automatic method to construct Chinese concreteness lexicon. By making full use of the existing resource of English word concreteness, it builds up a large-scale Chinese concreteness lexicon based on pretrained word embeddings and an MLP concreteness regression model. In addition, it proposes the concreteness inference tasks on the word level and on the sentence level, and manually constructs the corresponding datasets for evaluation the performance of the Chinese concreteness lexicon on these tasks. Experimental results show that the constructed concreteness lexicon can perform the two inference tasks effectively.

Key words word concreteness; concreteness inference; multi-layer perceptron; natural language processing

在心理语言学中, 词语特性得到广泛的研究和使用。词语特性包括词语的具体度(concreteness)、抽象度(abstractness)、意象度(imagery)和主观度(subjectivity)等, 这些特性也称为词语量化(word norms)。在自然语言处理任务中, 词语特性有助于理解、抽取和表达句子或文本的含义, 具有重要的应用价值。词语具体度指词语所表示的概念可以被实体地感知到的程度。例如, “机器人”、“刺鼻”和“桌子”等词语的具体度较高, “女权主义”和“悲伤”等词语的具体度较低。抽象度是具体度的对立面, 指词语所表示的概念可以被非实体地感知到的程度。

在心理语言学领域, 国内外对词语具体度展开了很多研究。Barber 等[1]研究具体性对文字处理的影响, 发现相较于抽象词汇, 具体词语通常会引起更快的反应时间。Ferré 等[2]基于情感属性, 研究具体词语和抽象词语的区别。Altarriba 等[3]针对抽象词、具体词和情感词的各种词特征进行研究, 分析这 3 种词汇类型在词汇属性、联想强度和联想数量上的异同。Connell 等[4]打破传统认知(人们认为抽象概念与具体概念的区别在于抽象概念缺乏感知信息, 使得抽象概念被处理地更慢, 更不准确), 通过实验得出结论: 具体性来源于概念表征的知觉强度。Hanley 等[5]研究具体性对单词生成能力的影

国家重点研发计划(2018YFB1005100)和国家自然科学基金(62076072)资助
收稿日期: 2021−06−08; 修回日期: 2021−08−14

响。张钦等[6]通过词汇决定任务考察中文双字词的具体性效应, 发现对抽象词的判断时间显著长于具体词。王振宏等[7]使用词汇判定任务和愉悦度判断任务, 探讨情绪名词的具体性效应。

具体度的研究在心理学领域逐渐流行的一个原因是大规模的词语具体度数据的出现。词语的具体度和抽象度通常通过数值进行量化, 现有的构建具体度数据集的方法大致分为两类: 第一类是通过人工对单词进行打分、评测和检验; 第二类是基于机器学习, 通过构建模型进行具体度评分。Paivio 等[8]1968 年构建一个由 925 个英文名词组成的数据集, 该数据集对单词的具体性、意象性和意义进行评级。Clark 等[9]收集并发布关于这 925 个名词的各种评级, 并对其进行拓展。Friendly 等[10]在 1982 年构建一个 1080 个英文单词的数据集, 将范围扩展到动词、形容词、名词和副词。Brysbaert 等[11]发布 4 万个英文词语的具体性打分词表, 该词典由超过 4000 人对这些英文词语进行 1~5 的具体性打分, 数值越大表示该词语越具体, 其中每个英文词语都最少由 20 个人进行打分。Soares 等[12]基于葡萄牙语单词, 由 2357 名大学生对 3800 个葡萄牙语单词的意象性、具体性和主观性进行打分。Brysbaert 等[13]构建 3 万个荷兰语词语的具体度词典。Charbonnier 等[14]基于已经发布的数据集, 使用支持向量机(support vector machine)构建回归模型, 基于词嵌入、词后缀和词性 3 种特征预测单词的具体性数值。Rabinovich 等[15]分别使用朴素贝叶斯(Naïve Bayes)、最邻近算法(nearest neighbor)和循环神经网络(recurrent neural network), 采用弱监督的方法预测单词的抽象值。

尽管词语具体度已经获得相当多的研究和关注, 但仍存在如下问题: 1) 这些研究针对的是西方语言, 面向中文词语具体度的相关数据较为匮乏; 2) 词语具体度的下游应用任务较为单一, 集中于隐喻检测任务。

针对中文词语具体度数据匮乏的问题, 本文提出一种根据已有英文具体度词典构造大规模中文具体度词典的自动化方法, 充分利用已有的英文具体度知识资源, 无须进行费时费力的人工标注, 最终得到的中文具体度词典涵盖 737531 个简体中文词语。为了评测所构造的中文具体度词典的性能, 分别设计词语级和句子级的具体度推理任务, 并手工创建相应的评测数据集, 用于评估该中文词语具体度词典的性能。本文构建的中文具体度词典和两个小型评测数据集已公开发布于 https://github.com/xiezp1976/chinese_concreteness/。

1 大规模中文词语具体度词典的自动构造

基于已有的英文词语具体度词典, 本文自动构建一个大规模的中文具体度词典, 步骤如下: 1) 基于英–中双语词典的中文词语具体度种子列表构建; 2) 基于预训练中文词向量库的中文词语具体度回归模型训练; 3) 基于具体度回归模型推理的大规模中文词语具体度词典构建。

词典的构建过程依赖于 3 个外部资源: 英文词语具体度词典、英–中双语词典以及预训练中文词向量库。步骤 1 和 2 如图 1 所示。

1.1 中文词语具体度种子列表的构建

英文词语具体度词典可以描述为二元组(V^e, γ),

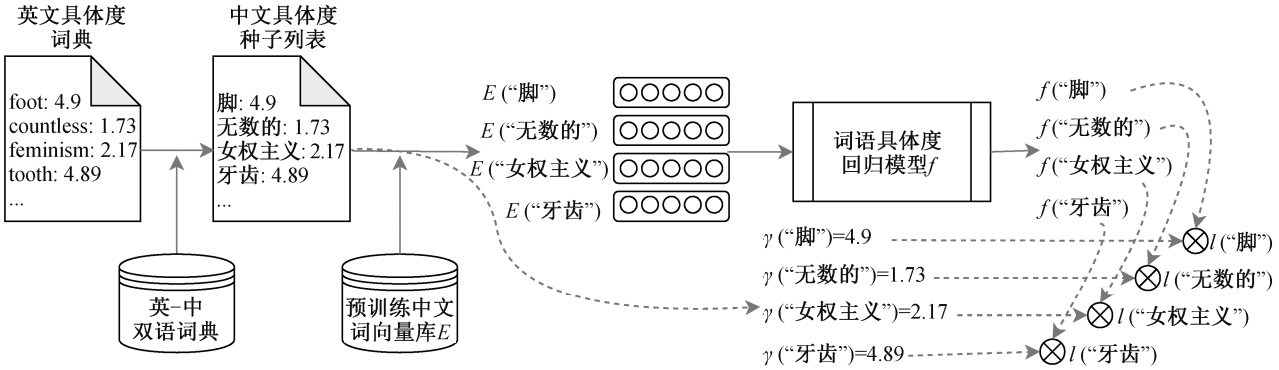

图 1 中文词语具体度种子列表构建以及具体度回归模型的训练

Fig. 1 Construction of seed list of Chinese word concreteness and the training of concreteness regression model

其中 V^e 是一个英文词表，$\gamma(W^e)$给出英文单词 $W^e \in V^e$ 的具体度取值。本文使用 Brysbaert 等[11]发布的英文词语具体度词典①，包含 37058 个英文单词和 2896 个双词表达。该词典由超过 4000 位参与者进行评级打分获得，使用 5 分制，1 表示最大的抽象程度，5 表达最大的具体度，每个单词至少由 20 位参与者进行打分。

基于英文词语具体度词典，本文使用英–中双语词典 T，将英文词语翻译成中文。给定英文词语 $W^e \in V^e$，对应的中文译词的集合为 $T(W^e)$。由于存在多义性、语言表达多样性以及中英文语言差异等问题，中英文词语之间普遍具有“多对一”和“一对多”的关系。“一对多”指一个英语词语被翻译成多个中文词语，例如“bank”被翻译成“河岸”或“银行”，“tooth”被翻译成“牙”、“齿”或“牙齿”。“多对一”指多个英文词语可以翻译成同一个中文词语，例如“desk”和“table”都可以翻译成“桌子”，“abbreviate”和“abbreviation”都会被翻译为缩写。针对此问题，我们将给定中文词语的具体度设定为其对应的所有英文词语具体度的算术平均，即对于给定的中文词语 W^c，其具体度 $\gamma(w^c)$为

$$\gamma(w^c)=\frac{1}{\left|T^{-1}(w^c)\right|}\sum_{w^e\in T^{-1}(w^c)}\gamma(w^e), \tag{1}$$

其中，$T^{-1}(w^c)$表示可以翻译成 w^c 的所有英文词语集合，即 $T^{-1}(w^c)=\{w^e|w^c\in T(w^e)\}$。本文使用有道翻译(https://fanyi.youdao.com)获得英文词语的中文翻译，经过式(1)处理，最终获得包含 25703 个中文词语的具体度种子列表。

1.2 中文具体度回归模型的训练

基于获得的中文词语具体度种子列表，训练一个具体度回归模型 f，学习中文词语到其具体度的映射关系。Harris[16]提出的语义分布假说认为，具有相似上下文的词语，其语义也相似。我们认为，具有相似上下文的词语，其具体度也相似(称为具体度分布假说)。本文采用 Mikolov 等[17]发布的 FastText 预训练中文词向量②作为词语的分布表征，并使用多层感知器网络，将词向量映射到具体度取值。

首先，根据预训练词向量库的词表，对中文词语具体度种子列表进行过滤，删除没有出现在预训练词向量库中的中文词语，确保数据集中所有词语都具有相应的预训练词嵌入。由此得到 14960 条数据，每条数据都是一个中文词语及其具体度取值。将这些数据按照 8:1:1 的比例切分，构建训练集、测试集和验证集，如表 1 所示。

然后，对于训练集中的任何一个词语 w^c，设 $E(w^c)$表示中文词语 w^c 对应的预训练词向量，经过 4 层感知器后的输出为

$$\begin{aligned}(w^c)=&\tanh\left(W_3\tanh(W_2\cdot\tanh(W_1\cdot E(w^c)\right.\\&\left.+b_1)+b_2)+b_3\right)。\end{aligned} \tag{2}$$

该 4 层感知器的第 1 层为词嵌入层(或输入层)，通过查表获得输入词语的词向量，将其作为输入；第 2 层和第 3 层为隐藏层，每一层的维度都比上一层降低一半，且都使用 tanh(·)作为激活函数；第四层为输出层，也使用 tanh(·)作为激活函数，输出为[–1, 1]中的一个实数值。由于词语具体度的值域在[1, 5]中，因此，4 层感知器的输出 $o(w^c)$经过式(3)的线性变换，作为模型输出 $f(w^c)$，以便与原始具体度取值范围保持一致。

$$f(w^c)=o(w^c)\times 2+3。 \tag{3}$$

这个多层感知器网络的训练则是通过 Adam 优化算法来最小化平方误差损失 Loss:

$$\begin{aligned}\text{Loss}&=\frac{1}{|\text{Train}|}\sum_{w^c\in\text{Train}}l(w^c)\\&=\frac{1}{|\text{Train}|}\sum_{w^c\in\text{Train}}(f(w^c)-\gamma(w^c))^2,\end{aligned} \tag{4}$$

其中，$l(w^c)=(f(w^c)-\gamma(w^c))^2$ 是训练样例 w^c 的平方误差损失函数，$\gamma(w^c)$是式(1)计算得到的 w^c 的具体度。

表 1　具体度回归模型训练的数据集统计信息

Table 1　Statistics of datasets for training the concreteness regression model

数据集	词语数量
训练集	11968
验证集	1496
测试集	1496

① https://static-content.springer.com/esm/art%3A10.3758%2Fs13428-013-0403-5/MediaObjects/13428_2013_403_MOESM1_ESM.xlsx

② https://dl.fbaipublicfiles.com/fasttet/vectors-crawl/cc.zh.300.vec.gz

1.3 大规模中文词语具体度词典的构建

得到中文词语具体度回归模型后，根据词语的词向量，使用该模型进行推理，预测中文词语的具体度。我们选择 FastText 词向量库中的所有中文简体词语，构建最终的大规模中文词语具体度词典。

第一步，获取 FastTextt 中文词典中的所有中文简体词语列表。判断中文简体词语的方法是，要求该词语的每个汉字都是基本汉字(Unicode 编码在 4E00~9FA5 之间)，并且转换成简体汉字后应与转换前的汉字相同。

第二步，对于列表中的每个中文简体词语，获取其 FastText 预训练词向量，输入到 1.2 节习得的具体度回归模型中，获得模型的输出作为该词语的具体度。

通过这种方式，最终获得的中文简体词语具体度字典包含有 737531 个简体中文词语。下面针对该具体度词典在词语级具体度推理任务和句子级具体度推理任务上的性能分别进行量化评测。

2 实验评测与结果分析

2.1 具体度回归模型的实验评测

用皮尔逊相关系数(Pearson's correlation coefficient)和肯德尔秩相关系数(Kendall's rank correlation coefficient)测量正确数值和模型预测值之间的关联程度。除本文的多层感知器模型外，我们还实现一个基于 SVM 的回归模型作为基线方法。该 SVM 回归模型使用径向基(RBF)核函数，超参数设置为 C=1.0，γ=0.01。实验结果如表 2 所示，可以看出，基于多层感知器的回归模型在性能上明显优于基于支持向量机的回归模型。

2.2 词级别具体度推理任务

为了评测中文具体度词典在词级别推理任务中的性能，我们手工构建一个评测数据集。首先，从词典中随机抽取 200 个词对，要求这些词对具有适中的具体度差异(差异为 0.7~1.0)，再由 3 位标注人员分别独立地判断词对的具体度，并进行人工标注。为了衡量 3 位标注人员的一致性(inter annotator agreement, IAA)，我们使用 Fleiss Kappa 系数，得分为 68.35%。每两位标注人员的 IAA 用 Cohen Kappa 系数衡量，分别为 65.76%，74.44%和 64.93%。这 200 个词对中，3 位标注人员标注结果完全一致的有 154 个词对，占总体数据的比例为 77%。154 个词对及其标签就构成对应的评测数据集。

对于该评测数据集中的任一词对(w_1,w_2)，通过在中文具体度词典中查询 w_1 和 w_2 的具体度取值，并比较其高低进行判定。将结果与人工标注结果进行对比，可以获得在整个评测数据集上的准确率(Accuracy)，实验结果如表 3 所示。可以看出，基于多层感知器模型构造的中文具体度词典在词级别具体度推理任务中的性能明显好于支持向量机模型。

2.3 句子级具体度推理任务

句子级的具体度在对篇章的理解方面具有潜在的应用价值。以中文词语具体度词典为基础，我们设计一种简单的句子具体度计算方法，并手工创建一个数据集来评测句子级具体度推理的能力。

2.3.1 句子级具体度推理任务的评测数据集

本文将 2008—2019 年上海市语文考试试卷阅读材料作为原始语料，随机抽取 185 个句对，并进行人工标注，判断哪一条句子的具体度较高。在标注过程中，如果一个句对中的两条句子都很具体或者都很抽象，标注人员难以进行准确的判断，则会跳过。无法判定的句对如表 4 所示，可以看出，句对 1 的句 1 中"企冀"、"奢望"、"抑制"、"罢休"等词语的抽象程度高，句 2 中"有幸"、"可贵"、"信赖"、"情愫"等词语构成的句子也同样抽象，难以进行准确的判断。类似地，句对 2 中的两条句子的具体程度较高，也难以判定。

经过人工标注，185 个句对中，标注人员无法判断的有 88 对，可以判断的有 97 个，占全部数据的 52.4%。97 个句对构成句子级具体度推理任务的评测数据集。

2.3.2 句子具体度计算方法

以词语具体度为基础，本文采用一种简单的方法来计算出给定句子的具体度。首先，针对给定句

表 2 中文具体度回归模型的实验结果比较

Table 2 Experimental results of Chinese concreteness regression models

回归模型	皮尔逊相关系数 r	肯德尔秩相关系数 τ
支持向量机(SVM)	0.78	0.58
多层感知器(本文方法)	0.82	0.63

表 3 词级别具体度推理任务的准确率结果

Table 3 Accuracies of word-level concreteness inference task

回归模型	准确率 Acc/%
支持向量机(SVM)	76.0
多层感知器(本文方法)	90.3

表 4　标注过程中人工难以判定的句对示例

Table 4　Hard-to-judge sentence pairs in the human-annotating process

句对	句子	实例
句对 1	句 1	二三十年不见鹭鸶, 早已不存再见的企冀和奢望, 一见便不能抑止和罢休
	句 2	谁家有幸得此可贵的信赖情愫呢
句对 2	句 1	金星、牛郎星都是星星, 而太阳、月亮不是, 其中的道理是明显的
	句 2	我们看到水往低处流, 火焰向上窜, 那就是水往低处流, 火焰向上窜

表 5　句子级具体度推理的错误推理样例分析

Table 5　Examples of wrong predictions in the sentence-level concreteness inference task

句对	句子	样例	标签	预测
句对 1	句 1	简单的书写或皴擦、普通的黑白两色, 竟然以简驭繁、以静寓动, 胜过了许多复杂的艺术	2	1
	句 2	原本的“墨色”, 居然可使人感受到“五光十色”		
句对 2	句 1	这类作品应归入儿童文学的范畴	1	2
	句 2	像左思《娇女诗》、李商隐《骄儿诗》等以儿童入诗, 情感真挚、刻画逼真、手法多样, 但这不是儿童文学, 因为这些诗只是以成人的视角来审视孩子, 用成人的笔触来描写孩子, 表达的是成人内心的情感活动, 唤起的是与诗人身份、环境相近的那些人的共鸣		

子进行中文分词操作, 使用 Jieba 分词器[①]进行分词处理; 接着, 对分词处理后的句子进行去除停用词操作, 使用哈工大停用词表[②]去除无意义的停用词; 最后, 通过查询中文具体度词表, 获得该句子中所有词语的具体度取值, 进行取平均操作, 将平均值作为该句子的具体度。

2.3.3 实验结果与分析

实验结果表明, 句子具体度计算方法在句子级具体度推理任务中的准确率为 75%。通过对发生错误预测的测试句对进行分析, 我们整理出一些原因, 并在表 5 中展示一些典型例子。第一类错误是由于句子中包含成语或文言文用法, 如句对 1 中, 句 1 包含“以简驭繁”和“以静寓动”两个成语, 具有较低的具体度, 分词工具无法整体地切分出该成语或整体切分出的成语未出现在中文具体度词典中, 导致该句的预测具体度偏高。第二类错误在于某些句子较长, 结构复杂, 同时混杂抽象子句和具体子句, 使用取平均方法很可能受句中其他子句的影响, 导致在预测句子具体度数值时出现偏差。例如句对 2 的句 2 中, “情感真挚、刻画逼真、手法多样”、“表达的是成人内心的情感活动, 唤起的是与诗人身份、环境相近的那些人的共鸣”比较抽象, 中间的子句“但这不是儿童文学, 因为这些诗只是以成人的视角来审视孩子, 用成人的笔触来描写孩子”则比较具体, 从而提高了整个句子的具体度。

3 结语

针对中文词语具体度资源匮乏的问题, 本文提出一种自动的中文词语具体度词典构造方法, 并充分利用已有的英文词语具体度资源, 构造一个大规模的中文词语具体度词典。本文还设计两项基本的具体度推理任务(词语级具体度推理和句子级具体度推理)来评估该具体度词典的质量, 并通过人工标注的方式构造相应的评测数据集。实验结果表明, 本文构造的中文具体度词典可以有效地进行这两项具体度推理任务。希望本文研究结果可以推动中文词语具体度的研究及其在自然语言处理领域的应用。

对于词语具体度下游应用任务较为单一的问题, 我们认为词语具体度理应在自然语言处理任务中扮演更重要的角色。例如, 现有的词向量预训练方法没有区别地对待具体词语和抽象词语, 将词语具体度集成到预训练模型中则将有助于建立抽象概念与具体概念之间的区分和关联。在图说(image captioning)任务中, 现有研究没有考虑训练数据中文本信息的具体度, 具体度较高的词语显然更容易从图像中直接生成, 抽象度较高的词语则需要经过

① https://github.com/fxsjy/jieba

② https://github.com/goto456/stopwords/blob/master/hit_stopwords.txt

推理, 由具体词语生成。如何拓展词语具体度在自然语言处理下游任务中的应用, 将是我们未来的研究课题。

参考文献

[1] Barber H A, Otten L J, Kousta S T, et al. Concreteness in word processing: ERP and behavioral effects in a lexical decision task. Brain and Language, 2013, 125 (1): 47–53

[2] Ferré P, Guasch M, Moldovan C, et al. Affective norms for 380 Spanish words belonging to three different semantic categories. Behavior Research Methods, 2012, 44: 395–403

[3] Altarriba J, Bauer L M, Benvenuto C. Concreteness, context availability, and image ability ratings and word associations for abstract, concrete, and emotion words. Behavior Research Methods, 1999, 31(4): 578–602

[4] Connell L, Lynott D. Strength of perceptual experience predicts word processing performance better than concreteness or imageability. Cognition, 2012, 125: 452–465

[5] Hanley J R, Hunt R P, Steed D A, et al. Concreteness and word production. Memory & Cognition, 2013, 41: 365–377

[6] 张钦, 张必隐. 中文双字词的具体性效应研究. 心理学报, 1997, 29(2): 216–224

[7] 王振宏, 姚昭. 情绪名词的具体性效应: 来自 ERP 的证据. 心理学报, 2012, 44(2): 154–165

[8] Paivio A, Yuille J C, Madigan S A. Concreteness, imagery, and meaningfulness values for 925 nouns. Journal of Experimental Psychology, 1968, 76: 1–25

[9] Clark J M, Paivio A. Extensions of the Paivio, Yuille, and Madigan (1968) norms. Behavior Research Methods, Instruments, and Computers, 2004, 36(3): 371–383

[10] Friendly M, Franklin P E, Hoffman D, et al. The Toronto Word Pool: norms for imagery, concreteness, orthographic variables, and grammatical usage for 1,080 words. Behavior Research Methods & Instrumentation, 1982, 14(4): 375–399

[11] Brysbaert M, Warriner A B, Kuperman V. Concreteness ratings for 40 thousand generally known English word lemmas. Behavior Research Methods, 2014, 46: 904–911

[12] Soares A P, Costa A S, Machado J, et al. The Minho word pool: norms for imageability, concreteness, and subjective frequency for 3,800 Portuguese words. Behavior Research Methods, 2017, 49: 1065–1081

[13] Brysbaert M, Stevens M, Deyne S D, et al. Norms of age of acquisition and concreteness for 30,000 Dutch words. Acta Psychologica, 2014, 150: 80–84

[14] Charbonnier J, Wartena C. Predicting word concreteness and imagery // Proceedings of the 13th International Conference on Computational Semantics—Long Pa-pers. Gothenburg, 2019: 176–187

[15] Rabinovich E, Sznajder B, Spector A, et al. Learning concept abstractness using weak supervision. EMNLP 2018: 4854–4859

[16] Harris Z. Distributional structure. Word, 1954, 10: 146–162

[17] Mikolov T, Grave E, Bojanowski P, et al. Advances in pre-training distributed word representations [EB/OL]. (2017–12–26)[2021–04–08]. https://arxiv.org/abs/1712.09405