计算机自然语言处理的机器翻译技术研究*

刘桂梅　吴佳蒂

（长沙幼儿师范高等专科学校　湖南长沙　410000）

摘　要：本文对自然语言处理的定义与内容进行介绍，然后阐述其在机器翻译中的应用要点，包括机器学习、机器翻译等方面，主要适用于情感等级分析、自动问答、全文自动检索等领域，最后分析自然语言处理KN模型在机器翻译中的实际应用，基于KN模型，以联合国会议记录为测试语料，通过对比验证KN模型的可行性。根据实验结果可知，该模型的应用可使英汉翻译更为准确，接近联合国翻译标准，值得推广使用。

关键词：计算机　自然语言处理　机器翻译　KN建模

中图分类号：G642；TP391.2　**文献标识码：**A

DOI：10.12218/j.issn.2095-4743.2023.10.146

引言

自然语言处理是在计算机科学、数学等多种学科基础上衍生的学科，是当前人工智能发展的主要研究方向之一。在进入大数据时代后，通信技术与互联网技术飞速发展，信息量爆炸式增长，国际间的关联日益密切，语言交流障碍问题越发凸显，对自然语言的处理需求更加迫切。通过机器翻译可进行语言处理和转换，使不同语言主体间的交流更加顺畅，且与人工翻译相比效率更高、成本较低、适用场景更广，拥有广阔的发展前景。

一、自然语言处理概述

该门学科是以计算为手段，针对自然语言进行探究与处理，学术上将其定义为探究人际交往和人机互动中的语言问题的学科，主要研究能描述语言能力、语言应用的模型，创建计算机框架，并不断采取相应措施，使语言模型不断完善。以语言模型为核心设计各类实用系统，对系统的评测技术优化升级。为了探究客观生活内计算机的自然语言处理情况，将其应用到多个场景下，主要处理流程如下。

第一步：先站在语言学立场，将自然语言处理抽象化，看成是语言问题。

第二步：将该问题形象化，以数学形式表现出来。

第三步：以严密规整的数学形式，根据算法创建计算模型，使其能够在计算机中得到处理。

当前自然语言处理在理论、技术等方面日益成熟，在语音自动识别、人机对话、信息检索等方面取得突破进展，以语言识别为例，借助计算机准确辨认语音，该技术可用于翻译的语音识别，还可应用到民航、铁路等问询系统内；在人机对话中，重点探究如何利用计算机理解和应用人类语言，并以对话形式回答用户所提问题，例如百度的“小度”、小米的“小爱同学”等，均是智能人机对话的代表；再如文字自动识别技术，可应用到扫描软件中，对印刷刊物或者手写文字进行识别，最后生成电子文档[1]。

二、自然语言处理在机器翻译中的应用要点

1．机器学习

自然语言处理中最典型特征在于利用机器学习采集语言知识，机器学习的作用在于探究如何利用计算方式和以往经验，使自身性能得到改善。其技术原理在于利用计算机从大量数据内获得模型算法，再将经验数据传递给计算机，由此创建新的模型。在数据更新后，计算机便可在现有模型基础上帮助用户作出相应判断。根据机器学习原理可知，先将大量训练数据传递给计算机，创建初始模型，即模型1，再利用测试数据检查，弥补模型的缺陷，便会获得训练完毕的新模型，即模型2，将新数据传递给新模型，便可借助计算机、模型做出相应预测，如图1所示。因需要检测的数据量不断增加，特征更加多样，可通过参数调整、算法性能表现提升等方式，使其得以优化。

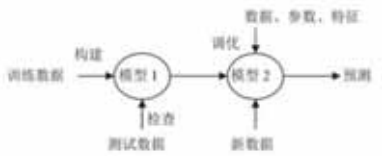

图1　机器学习工作原理图

2．机器翻译

机器翻译在语言处理中涉及许多典型问题，如词字切

*基金资助：本文系2021年度湖南省教育厅科学研究项目优秀青年项目《基于自然语言生成的网络推广管理系统研究与实现》（课题批准号：21B0945）阶段性成果之一。

分、句法分析、数据挖掘等等，根据翻译方式不同，可分成基于规则翻译和语料库翻译两种，又因建模不同，还可将语料库翻译分成统计翻译、神经翻译和实例翻译三种，均需要大量语料作为训练数据。因训练数据在语言语料规模上带有局限性，难以覆盖真实样例信息，可采用数据平滑算法，使语言模型更加成熟。许多自然语言处理都依托语言模型，生成文本，在前面出现的单词和语境，经过专业训练后，可预测后续出现的单词，在模型建成后便可为用户提供预测和判断依据，使翻译质量得到进一步提升[2]。以Good-Turing平滑算法为例，其技术原理在于利用频率类别信息来平滑频率，针对任何r次数的n元词，均假设其发生了r*次，定义如下：

$$r* = (r+1)\frac{n_{r+1}}{n_r}$$

式中，nr代表的是语料库内的r次n元词数量，利用近似计数，计算出概率。

在语言模型构建中，其本质在于自然语言处理，先以目标语言语料为训练数据，计算机先学习知识，在此基础上创建语言模型。以计算机编程语言为例，是针对某些特定目标而创造的语言。不同语料类型对应的训练模型不同，以目标语言为例，可用于训练语言模型、平行语料对应翻译模型等，依靠计算机将自然语言转变为自然语言，全过程自动完成，方便快捷。

3．应用领域

（1）情感分析

该领域中机器翻译的应用是为了借助机器分析获取人们对某种事物的看法，根据被分析文本粒度可将分析任务分成不同等级，即文档级、语句级、属性级。文档级是分析文档内容对目标体的情感倾向，属于最粗粒度分析。通常情况下，一段文档内带有多个目标体，在文档级分析中受到较大局限，所得结果的使用价值有限，与之相比，语句级的情感分析可用性较强，但在信息量方面较少，在一定程度上增加了分析难度。属性级分析以目标体、用户态度为核心，并非前两种是侧重于文本内容，因此应用价值相对较高。

（2）自动问答

该系统能够接受自然语言，并准确应答，主要完成信息采集、问题处理等工作。首先，该系统对原文本预处理，将文本内容按序列标准，停用词剔除，再明确原文本内词语的频率、位置信息，得出位置与词频因子，代入权重函数评分，最后对比分值获得最终结果。在领域的研究热点在于相关性推荐，也就是通过查询某个问题，获得一系列所需信息，一般利用TF-IDF算法评估文本内容相似性。此外，用户适应度评估也属于研究热点，根据用户人性化分析回答问题，但在准确度方面仍有很大上升空间。

（3）全文自动检索

在自然语言处理中，全文自动检索技术适用于主题词自动提取、文摘自动生成等需求。在主题词自动抽取中，根据文献所述与具体对象，为文献拟出恰当标题，使其有序存入检索库。主题自动抽取算法包括综合词频、位置等统计法。在信息提取过程中，通过一定算法筛选文档内较为关键的语句，将其称为主题句，在此基础上进行句法、语义分析，获得主题句中各部分关联的语义关系。在文摘自动生成中，是从语义与逻辑角度，将文档内容缩写成摘要，以简短语言描述文章内容，便于用户快速评价检索结果与检索需求的相关度。较为常用的生成技术是在统计基础上，先对全文自动分词，再统计文中各词出现频率、权重，根据某种规则确定关键词，将其从语句中抽离出来，根据语句权重计算综合权重，选出一组最能代表文献主题的语句，将其作为文摘句，实现自动生成文摘的目的。

三、自然语言处理KN模型在机器翻译中的实际应用

1.KN模型创建

在自然语言处理中，该模型可扩展绝对折扣，消除低阶模型对高阶插值时产生的影响。在机器翻译实验中，创建KN模型并进行平滑处理，可使实验结果更加高效。在估计高阶模型概率时，应综合分析低阶模型对其产生的影响，使数据稀疏问题得到解决，但可能会对高阶概率估计产生负面作用。例如，在创建二元模型时，语料库内某些词出现频率较高，以“Francisco”为例，该词一般在“San”后出现。因该词出现频率较多，使得统计频率增加，一元模型概率随之增加。在绝对平滑算法应用中，即便该词前面不是“San”，二元模型的出现概率仍会增加。对此，在KN平滑中，认为一元模型概率应是与其他邻接词组合，并非与本身出现次数为正比，由此减轻负面影响，模型如下：

$$P_{KN}(w_i) = \frac{N_1 + (\bullet w_i)}{N_1 + (\bullet\bullet)}$$

该模型中“●”为位置符号，在KN模型估计后，高阶平滑概率的分布应与训练集相对应。例如，二阶模型内的PKN应与一元模型语料库中的要求相符合，如下：

$$\sum P_{KN}(w_{i-1}w_i) = \frac{c(w_i)}{\sum_{iw} c(w_i)}$$

该等式左侧为KN模型平滑后的变量wi总概率分布，右

侧为变量wi在语料库内的频率，将全部与上述要求相符的PKN（wi）汇集起来，代入到高阶模型内，便可获得KN平滑模型。在该模型中，针对语料库内全部非零统计，利用单一折扣来完成，因每个折扣函数所依据的统计值不同，与之相对的定义与计算方式也不尽相同。为了便于执行，可选择常用的不同常数值为折扣，通过反复多次验证，使最终取值与实验要求相符[3]。

2．实验过程

本实验的测试语料是在训练集基础上创建，以联合国会议记录的中文版本为原始数据，以与之对应的英文版本为标准翻译。为了获得备选翻译，在互联网上较为成熟的网站进行英汉翻译，如有道翻译。

原始会议记录："事实证明，很难得到实际交付的武器数量、交付日期与进口港等文件信息。"

联合国标准翻译："It proved more difficult to obtain specific doumentation an information on the actual number of weapons and ammunition delivere，the date of delivery and the port of entry."

有道翻译："It has proved difficult to obtain the number of weapons actually delivered，the date of delivery and the port of import."

人工翻译：The fact proves，very difficult to obain actual delivered quantity of weapons and ammuntion，delivery date and entry port's specific file data and information.

为了检验模型准确性，从会议中选取句子对训练完毕的KN模型进行检测，对训练语料、测试语料分行处理，每个语句单独成行，便于程序读入，获得输入变量后，便可利用KN模型估算概率公式编写程序，估算句子概率，根据困惑度定义进行计算，公式如下：

$$Perplexity = 2^{Hp(T)}$$

因句子困惑度与概率具有反比关系，当句子概率越大时，困惑度便越低。但部分概率受语句长度影响，用困惑度检验效果最佳。在翻译试验中，第一行是程序初始文本，第二行输入汉语句子，第三行及以后是KN模型对测试语料中句子估计的结果，将本句中后三个单词作为输出，可使每个句子的标记均清楚简洁。然后对该句中各项数据进行评估，包括对数结果、单词数、句子所在行数、最终困惑度等，在所有句子的输出参数中找出困惑度最低的一个，如若为标准英译版本，则验证成功。选出100个句子，实验过程便是KN模型的完善过程[4]。

3．实验结果

针对联合国会议记录中的语句，分别用有道、人工翻译与联合国标准翻译对比，实验结果如下：

（1）在输出数据中，最后一项为困惑度，将不同句子困惑度对比，发现第一句的困惑度最低，为54.42012，这意味着该组中第一句最符合训练语料库的结构，也就是联合国提供的标准翻译，该句验证成功。

（2）观察输出结果，发现有道与人工翻译的概率基本相同，二者仅相差0.8，但困惑度差别将近70，这说明有道翻译的句子长度为31词，而人工翻译句子长度为28词，根据困惑度计算特点，长度在定义式的分母上，语句越短，困惑度便越高，二者概率差别较小，困惑度差异较大[5]。

（3）在机器翻译KN模型优化期间，尽管筛选结果与实际值不同，但整体效果较好。缺陷在于KN模型仍会受训练语料库的限制，在模型应用中，从备选句中选出"最佳"的一个，但有时并非是真正的最佳，而是最符合语料库中常见语法规则的一个，忽视了语义在句子内的应用价值。对此，可采用平滑模型，准确辨别句子差异，在分词时将语义考虑进来，便会取得更加理想的翻译效果。

结语

综上所述，自然语言处理是计算机人工智能领域的主要科研方向之一，涵盖诸多学科领域，在深度学习计算逐渐成熟下，机器翻译任务量不断增加，对准确度的要求也不断提升，需要不断地更新理论、完善技术，基于KN模型进行句法剖析，充分发挥机器翻译的预测和判断作用，为人们生活与工作带来更多便利。

参考文献

[1]赵铁军,朱聪慧.世界最大的自然语言处理和语音技术实验室——哈尔滨工业大学语言语音教育部-微软重点实验室[J].计算机教育,2019(11):13-16.

[2]孙茂松,周建设.从机器翻译历程看自然语言处理研究的发展策略[J].语言战略研究,2019(6):0025-0027.

[3]陆正扬.基于计算机自然语言处理的机器翻译技术应用与简介[J].科技传播,2019,11(22):12-14.

[4]赵铁军,曹海龙.以机器翻译技术为核心的多语信息处理研究[J].中文信息学报,2021,25(6):10-12.

[5]肖桐,朱靖波.基于树到串模型强化的层次短语机器翻译解码方法[J].计算机学报,2019,39(4):14-16.

作者简介

刘桂梅（1983.11—），女，汉族，湖南娄底人，硕士，研究方向：软件工程。

吴佳蒂（1983.05—），女，汉族，湖南汨罗人，硕士，研究方向：网络工程。