# 关于自然语言处理的研究

**摘要**

数字、文字和自然语言一样，都是信息的载体，它们之间有着天然的联系。语言和数学的产生都是为了同一个目的——记录和传播信息。半个世纪前香农博士提出了信息论，人们开始把数学和信息系统联系了起来。语言的出现是为了人类之间的通信。字母、文字和数字实际上是信息编码的不同单位。任何一种语言都是一种编码方式，而语言的语法规则是编解码的方式，而语言的语法规则是编解码的算法。我们把一个要表达的意思，通过某种语言的一句话表达出来，用这种语言的编码方式对头脑中的信息做了一次编码，编码的结果就是一串文字。而如果对方懂得这门语言，他或者她就可以用这门语言的解码方式获得说话人要表达的信息。这就是语言学的本质，动物虽然也能传递信息，但是利用语言传递信息是人类的本质。1946年以后，现代电子计算机出现以后，人们开始尝试让计算机能够处理自然语言。关于这个问题涉及两方面的问题：第一，计算机能否处理自然语言。第二，如果能，那么它处理方法是否与人类处理的方法一样。

**引言**

1950年图灵在《思想》（Mind）杂志上发表了一篇名为“计算机的机器和智能”的论文。在论文中提出了验证人工智能的方法：让人和机器交流，如果人无法判断与自己交流的对象是否为机器，说明这个机器具有了智能，后人把这种方法称为图灵测试。自然语言处理是机器智能的一个重要分支。

I propose to consider the question, "Can machines think?" This should begin with definitions of the meaning of the terms "machine" and "think." The definitions might be

framed so as to reflect so far as possible the normal use of the words, but this attitude is dangerous, If the meaning of the words "machine" and "think" are to be found by examining how they are commonly used it is difficult to escape the conclusion that the

meaning and the answer to the question, "Can machines think?" is to be sought in a

statistical survey such as a Gallup poll. But this is absurd. Instead of attempting such a

definition I shall replace the question by another, which is closely related to it and is

expressed in relatively unambiguous words.

COMPUTING MACHINERYAND INTELLIGENCE

By A. M. Turing

**历史**

自然语言处理是人工智能的一个重要应用领域，它的主要目的是克服人机对话中的各种 限制，使用户能用自己的语言与计算机对话。目前，国外对这方面的研究已取得一定的成就，但就实用的系统来说，真正好的，能应用各种语法规则来分析用户输入的句子，当遇到语法未曾描述过的句子时，它亦能分析并确定其语义的系统还没有出现过。研究性的系统不仅能理解输入的语句，而且还能与用户进行对话，甚至能理解篇段内容。但对于汉语来说，由于它的文法不同于英语文法，具有相当的不确定性，字词搭配复杂它能首先对输入的句子加以标记，随后进行语法剖析，并理橢其语义。我们把这种方法用在一个小型的实用系统中，使语言的文法文字及词都有一定的范围。但我们的目的是要系统不仅能分析在文字、语法范围 内的句子，而且对系统规定范围之外的句子仍能进行有效的分析、理解。实验结果表明，对于语法正确的句子，系统能有效地理解其含义，对于文字或词（节点）不属于系统描述范围的 句子，系统也能较有效地理解它。但对语法结构（树结构）的限制仍然较大。我们试图寻找一种有效的方法（误差结构校正）来改进系统的运行。我们的实用系统能有效地应用于智能化计算机辅助数学中（ICAI）。

Lu Chuanxing,Gu Xinli(Chiua Research Center for Middle School Computet Education\nEast China Normal University)ABSTRACT.It has long been recognizeb that the ultimate effectiveness of system is lined to the amount of intelligence embodieb in programs, including perception systems (i.e. vision anb hearing),natural language processing and dialogue system etc. That is, a more intelligent program would be better able to formulate the questions and presentations, and it woulb be better equipped to understand a user's response, even to analyze and model the human knowledge。In this paper,we propose a simple Na tural Language Interface(NLI),Chinese。Now we are pursuing a prototype implem-entation of the NLI with PROLOG.\nKEY WORDS;choosing a topic, tokenization,syntax analysis,andesemantlc analysis.

计算机自然语言处理

占传兴 顾新理

最早的自然语言处理（NLP）是采用了One-Hot-Encoding的Bag of Word的形式来处理每个字。这样的形式自然是无法抓住文字之间的语义和更多有价值的信息。那么，之前Word2Vec的想法则是学习一个每个Word的Embedding，也就是一个实数的向量，用于表示这个Word的语义。当然，如何构造这么一个向量又如何学习这个向量成为了诸多研究的核心课题。

Word Embedding

Ben Athiwaratkun

自然语言处理经历60多年的发展，大体上可以分为两个发展阶段，即上世纪50年代到70年代，必须让计算机理解自然语言阶段。第二个阶段为上世纪七十年代至今，即采用给予数学模型和统计的方法来实现自然语言的处理。

第一阶段的科学家普遍认为要实现自然语言的识别要做好两件事，即分析语句和获取语义

对于语义的研究具有很多年代历史了，《圣经》的出现使得西方的语法研究有了规范，形成了十分完备的体系，学习西方语言都需要学习语法规则、构词法和词性。

在中国《说文解字》中便是对语义的研究。往往一个字在不同的环境中具有不同的意义，而这正是当时科学家急需解决的问题。

因为自然语言处理相对于语音和视觉来说是高度抽象化的表现，它不是信号，而是一些非常抽象化的理念。

大家都认为人类的语言有非常强的逻辑性，其实人类的语言逻辑性并不强。我给大家举一个例子，大家看这句话“我从来没说他偷过钱。”这句话有6种理解方法，我一一列出来了。

比如说，我可以这么来说：“我从来没说他偷过钱。”这个意思就是可能别人说过，但是我没有说。

第三个可以说“我从来没有说他偷过钱”，可能我确实没有说，但是我用其他的方式暗示过。

除了这6种以外，如果把这个句子加长的话，变成“我从来没说他偷过我的钱。”那么就有7种解释，不光有1到6，还有第7种解释，这个句子可以变得更长，这个歧义就会更多。

对于计算机来讲，如果单单给它这一句输入，要做到真正语境上的理解是不可能的事情。要做到真实语境上的理解可能需要更多的辅助信息和上下文的信息，不然是没有任何可能性的。

其次，我们要理解人的语言不能光靠逻辑，还要有非常强的知识库，要有很多知识才能正确理解人类语言。

我举个例子，下面两句话中，第一句话We gave monkeys the bananas because they were hungry。这个地方的they指猴子。第二句话We gave monkeys the bananas because they were over ripe。这个地方的they指香蕉。对于计算机来说这两句话看起来结构非常相似，句式也非常类似，所以计算机必须知道猴子饿了，香蕉不能饿，猴子不能烂的，香蕉才能烂，才能对这句话有一个正确的理解，不然是完全无法知道。

再次，人的语言还有一个非常大的特性即组合性。

我们通过字母组合成词，通过词组合成短语，短语组成句子、句子组成段落、段落组成文章。如果单单抽出里面一部分进行解析的话，比如说解析字母、解析词，我们就算理解了词的意思也不能表现出人本来的含义，因为单个抽出词是没有意义的，人的自然语言表达的含义往往就在这些组合当中，恰恰是学习这些复杂的组合对于计算机来说是一件非常难的事情。

最后，人类语言是非常灵活和开放的。

开放是什么意思？人的语言是随着时间而改变的，不停的有新词冒出来，以前词的意思也会随着时间有完全不同的意思。比如说“灌水”、“潜水”，这两个词在网络时代有了完全不同的含义。对于计算机来说怎么能够实时的学会这些新词、发现新的用法，也是非常有挑战性的。

TAB（Tech and Business）

孙琳

第二阶段，首先成功利用数学方法解决自然语言处理问题的是语音和语言处理大师贾里尼克 (FredJelinek)。当时贾里尼克在 IBM 公司做学术休假 (Sabbatical Leave)，领导了一批杰出的科学家利用大型计算机来处理人类语言问题。统计语言模型就是在那个时候提出的。

给大家举个例子：在很多涉及到自然语言处理的领域，如机器翻译、语音识别、印刷体或手

写体识别、拼写纠错、汉字输入和文献查询中，我们都需要知道一个文字序列是否能构成一

个大家能理解的句子，显示给使用者。对这个问题，我们可以用一个简单的统计模型来解决

这个问题。如 果 S 表示一连串特定顺序排列的词 w1， w2，…， wn ，换句话说，S 可以表示某一个由一连串特定顺序排练的词而组成的一个有意义的句子。现在，机器对语言的识别从某种角度来说，就是想知道 S 在文本中出现的可能性，也就是数 学上所说的 S 的概率用 P(S) 来表示。利用条件概率的公式，S 这个序列出现的概率等于每一个词出现的概率相乘，于 是P(S) 可展开为：P(S) = P(w1)P(w2|w1)P(w3| w1 w2)…P(wn|w1 w2…wn-1)

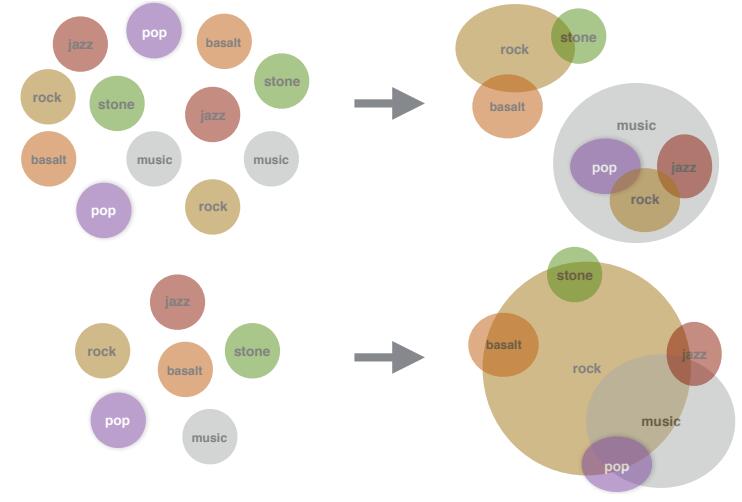
其 中 P (w1) 表示第一个词 w1 出现的概率；P (w2|w1) 是在已知第一个词的前提下，第二

个词出现的概率；以次类推。不难看出，到了词 wn，它的出现概率取决于它前面所有词。

从计算上来看，各种可能性太多，无法实现。因此我们假定任意一个词 wi 的出现概率只同它前面的词 wi-1 有关(即马尔可夫假设），于是问题就变得很简单了。现在，S 出现的概率

就变为：P(S) = P(w1)P(w2|w1)P(w3|w2)…P(wi|wi-1)… (当然，也可以假设一个词又前面 N-1 个词决定，模型稍微复杂些。） 接 下来的问题就是如何估计 P (wi|wi-1)。现在有了大量机读文本后，这个问题变得很简单，只要数一数这对词（wi-1,wi) 在统计的文本中出现了多少次，以及 wi-1 本身在同样的文本中前后相邻出现了多少次，然后用两个数一除就可以了,P(wi|wi-1) = P(wi-1,wi)/ P (wi-1)。

在ICLR 2015会议上，来自UMass的Luke Vilnis和Andrew McCallum在“Word Representations via Gaussian Embedding”文章中提出了用分布的思想来看待这个实数向量的思想。具体说来，就是认为这个向量是某个高斯分布的期望，然后通过学习高斯分布的参数（也就是期望和方差）来最终学习到Word的Embedding Distribution。这一步可以说是扩展了Word Embedding这一思想。然而，用一个分布来表达每一个字的最直接的缺陷则是无法表达很多字的多重意思，这也就带来了这篇文章的想法。文章希望通过Gaussian Mixture Model的形式来学习每个Word的Embedding。也就是说，每个字的Embedding不是一个高斯分布的期望了，而是多个高斯分布的综合。这样，就给了很多Word多重意义的自由度。在有了这么一个模型的基础上，文章采用了类似Skip-Gram的来学习模型的参数。具体说来，文章沿用了Luke和Andrew的那篇文章所定义的一个叫Max-margin Ranking Objective的目标函数，并且采用了Expected Likelihood Kernel来作为衡量两个分布之间相似度的工具。



通过UKWAC和Wackypedia数据集学习了所有的Word Embedding。所有试验中，文章采用了K=2的Gaussian Mixture Model（文章也有K=3的结果）。比较当然有之前Luke的工作以及其他各种Embedding的方法，比较的内容有Word Similarity以及对于Polysemous的字的比较。

**发展**

目前我们已经有非常好的语音识别系统了，现在基本上达到了人类的水平，在理想环境里可以达到95%以上的正确率。同样我们也有比较正确的机器翻译系统，正确率换算过来也可以有70%到80%，虽然离人的水平还有一定的差距，但是已经是可用的状态。除了这两个以外，自然语言处理（NLP）的应用目前进展不大。举一个最简单的例子，比如词性标注，在一个句子当中，动词、名词、形容词，这个任务是非常简单、非常基础的任务。但是句子级别（一句话一个词不错才算对）目前的正确率只有57%，而且从2009年到2017年间正确率提高了不到1%，无论使用深度学习、各种模型、各种方法，花了八年时间也是只是提高了不到1%。

另外一个例子是句法分析，就组合式句法分析来说，我们今天没有比十一年前做得更好，无论是用深度学习还是其他任何方法，十一年没有改进过。谷歌在去年推出了谷歌SyntaxNet，号称是世界上面最优秀的句法分析器，其实对比四年前最好的系统也只提了2%，当然谷歌用了目前最好的深度学习技术，也仅仅做到了这样。还有多轮对话系统，目前正确率最多只能做到60%，这其实是完全不可用的状态。深度学习的模型，其实在NLP的各个领域都取得了成功，不是说不成功，只是没有取得在视觉、语音领域那么大的成功。

**发展方向**

基础研究：

词法与句法分析

语义分析

篇章分析

语言认知模型

语言表示与深度学习

知识图谱与计算

应用研究：

文本分类与聚类

信息抽取

情感分析

自动文摘

信息检索

信息推荐与过滤

自动问答

机器翻译

社会媒体处理

语音技术语音合成

语音识别

文字识别

多模态信息处理

医疗健康信息处理

少数民族语言信息处理

《中文信息处理发展报告2016》

中文信息协会