自然语言处理——让输入法变得更聪明

王砚峰、贾剑峰、张扬

1. 引言

随着电脑的普及和互联网的发展，输入法已经成为了人们生活和工作中最不可或缺的工具软件，人们在网上聊天中的对话交流，以及工作中的电子邮件和文档报告，都是通过输入法来完成的。

用户使用输入法目的，是打出为了表达自己意图所需要的字词句，那么输入法的聪明程度，也就是输入法对用户意图的猜测能力，实际上决定了用户输入的效率；进一步的，从更大的视角来看，在以文字为交流载体的信息社会中，更聪明的输入法实际上代表了更先进的生产工具，它意味着更大的经济效益和社会价值。因此，虽然当前的输入法软件尚没有上好的商业模式，不能带来直接的商业利益，但还是各大互联网 IT 公司仍然不遗余力的投入到这个看似不大的战场上进行角逐，并且各自宣称，自己的输入法是“最聪明的”。

那么，究竟什么样的输入法是一个“聪明的”输入法呢？输入法又是如何变聪明的呢？文本就将带你走进自然语言处理技术的世界，为你揭开输入法智能性的神秘面纱。

2. 传统输入法与智能输入法

所谓**传统输入法**，是指智能输入法出现之前的输入法，典型的代表作就是“智能 ABC”。它的主要特点是：只能基于字和词的输入，每个拼音下面的汉字结果采用固定的排序，同时能直接打出的词的个数十分有限，词库中只有几千到几万个高频的常用词，而且通常都是经过某专业机构人工收集整理得到的。

显而易见，在使用这个输入法时是十分低效的。例如用户想输入一个宋词词牌“鹧鸪天”，这个词不是很常用，因此不会在词库中，同时“鹧”“鸪”都是相对很低频的字，因此用户想输入这个词的时候需要大量的翻页查找操作。另外，按照目前互联网上信息的膨胀速度，每天互联网上都会随着各种事件的发生产生相应的新词，比如“兽兽门”、“犀利哥”，这些网络新词就更是传统输入法所望尘莫及的了。

然而打开目前市场上流行的任意一款**智能输入法**，上面提到的词都会被轻轻松松的输出来；不仅如此，所有的智能输入法都支持用户短句级别以及句子级别的输入方法，并且能够保证相对较高的准确率；同一个拼音下，不同词语的排序也不再人为的固定，比如当你刚刚输入了“我的”的时候，再输入 daxue 的第一位会是“大学”，而如果你刚刚输入了“漫天”时，再输入 daxue 的第一位则是“大雪”，就像输入法真的可以读懂你的心思一样。

那么智能输入法是如何做到这一点的呢？答案就是基于统计的自然语言处理技术。

3. 给力的统计自然语言处理技术

统计自然语言处理，顾名思义，根本方法是概率统计。统计自然语言处理，就是利用统计的方法，得到自然语言处理中的各种语言现象的整体描述，并采用概率的方法，来解决语言现象中出现的歧义问题。比如一个经典的问题：文字“南京市长江大桥”，是指在南京市里面的长江大桥呢，还是说南京有个市长名字叫江大桥呢？

统计自然语言处理包括多种方向，比如配搭发现、语言模型、词性标注、语义消歧、词聚类、概率语法分析等，也应用于多个方向比如语音识别、机器翻译、分词、信息检索等。输入法本质上是语音识别，用户输入的拼音即可理解为语音，本文着重讲述统计自然语言处理技术在输入法方向中的应用。在这之前，我们作如下的假设：

1. 在统计自然语言处理中，一项最基本的前提条件是，必须有大量的语料供统计使用。实际上，搜狗输入法能够依托于搜狗搜索的网络爬虫，每天得到网上最新鲜的大量的语料。本文中假定我们已经获得了并且初步处理了这些语料。
2. 我们已经具有了一个人工整理过的常用词库，并且我们具有出色的分词工具，能够将我们的词库带入到语料中，得到基本正确的分词结果。
3. 在后面的叙述中，没有“字”的概念，只有“词”的概念，单个的字也成为“词”。

I. 互信息——发现成词的搭配

据搜狗输入法对样本用户输入的统计，每天用户输入中 93%的词来自于搜狗输入法系统词库，可见输入法词库之全；同时，现代的智能输入法每天都会给用户推送大量的网络新词，来满足用户实时性方面的输入需求，可见输入法词库之新。如此“全”而“新”之需求是如何得以满足的呢？

本文把不在目前输入法词库中的词称为“未登录词”，我们的问题是：给定一个很大的

语料库（TB 级别），一个基本词库，以及一个分词工具，如何从语料库中找到更多的未登录词，以提高词库的覆盖率。一个直观的想法是，利用基本词库对语料库分词后，进行词对儿的统计，然后把高频词对儿提取出来，扔掉其中不符合词法规则（比如动宾，形名，动补等）的组合，剩下的就是未登录词。但这种方法一方面依赖于分词时词性标注的正确性，另一方面中华文字包罗万象，很多都不是词法规则可以涵盖的，因此效果并不理想。互信息是发现成词搭配的极佳方式，它的定义如下：

$$I(x,y)=\log_2\frac{p(x,y)}{p(x)p(y)} \quad (x,y)\text{为左右相邻的一对儿搭配}$$

只看 log 函数里面的部分，公式的含义是：一个搭配在语料中出现的概率，除以组成这个搭配的各个部分在语料中出现的概率。可以从这样的角度来理解互信息公式：如果假定搭配中的两个词是相互独立的，那么两个高频词同时出现的先验概率，是比两个低频词同时出现的先验概率要大的。但在实际的语言中，两个词很少是完全独立的，如果统计得到的两个低频词同时出现的概率，反倒大于了两个高频词同时出现的概率时，那么这个两个低频词就是非常的“不独立”，也就是有着紧密的搭配关系。因此一个搭配的互信息值越高，代表这两个词之间的搭配关系也越紧密，他们形成一个词的可能性也越大。比如“杯—具”之间的互信息要远远大于“的—了”，因此它作为词的可能性也就更大。

通过互信息度量，不必引入复杂的和不确定的词法规则，就可以自动的发现未登录词。并且系统可以采用迭代的办法，发现更长的未登录词。第一次迭代中，通过“鹧—鸪”间的互信息发现了“鹧鸪”这个词；接着在第二轮迭代中，又可以通过“鹧鸪—天”间的互信息进一步发现“鹧鸪天”这个词；最后再看“鹧鸪”是否只出现在“鹧鸪天”单一语境中：如果是，那么“鹧鸪”只是“鹧鸪天”的一个片段（或称“碎词”），可以不作为未登录词发现的结果。实际上，搜狗输入法的新词发现，以及大词库制作，都是采用互信息为度量，并且得到了令人满意的结果。

II. 统计语言模型——神奇的马尔科夫链

起初，在解决语音识别问题的时候，人们倾向于使用语言学中的语法规则来解决句子识别的问题，但在几十年间从来就没有取得过任何突破，直到贾里尼克（Fred Jelinek）把“统计语言模型”带入这个领域。统计语言模型假定：一个句子中某个位置出现的词，只受到这个位置之前出现的 k 个词的影响。那么一个句子：

$$S = w_0\ w_1\ w_2\ w_3\ \dots\ w_{N-2}\ w_{N-1}\ w_N$$

在 k=3 的时候，概率计算公式就是：

$$P(w_0,\dots,w_N) = P(w_0)\ P(w_1|w_0)\ P(w_2|\ w_0w_1),\dots,\ P(w_N|\ w_{N-2}w_{N-1})$$

这种句子的概率计算方法称之为 n 元文法，当 k=3 时便为三元文法。三元文法是研究界采用的普遍模型，因为三元文法可以比较好的克服局部最优的缺点，同时数据稀疏问题不那么严重。可以直观的想象，n 越大，越接近全局最优解，但统计组合空间也越大，相应的数据稀疏问题也越严重。出于存储与性能的考虑，目前桌面智能输入法多采用二元文法，即当前位置出现的词只受前一个位置的词影响，而云输入法采用三元或者更长的语言模型。至此，智能输入法在处理用户短语及句子输入时的方法已经浮出水面。

例如，当用户输入“wohenfanganta”的时候（很大的概率他是要“我很反感他”），会有多种词汇的组合，比如“我恨-方案-他”、“我恨-反感-他”、“我很-方案-他”以及“我很-反感-他”等。通过二元文法，“我很—反感”与“反感—他”这些常用的语言搭配现象就会以概率的方式贡献到整体句子的概率中，从而得到正确的结果。回到本文第二章提出的问题，用户上一次输入了“漫天”之后，为什么输入“daxue”的时候，返回的首选是“大雪”而不是“大学”？显而易见，“漫天大雪”比“漫天大学”更常见的出现于我们日常的语言中，那么通过“漫天”来预测“大雪”是非常合理的。

到这里，其实敏感的读者已经发现，N 元文法其实就是一条马尔科夫链，足见信息论的祖师爷香农（Claude Shannon）对人类在人工智能方面的贡献是何等的伟大。

III. 词聚类——预测未知世界

数据挖掘中的聚类，提出的目的本来是用于给数据降维，或者对数据进行一个高层次的全局描述。但在自然语言处理领域，经常这样使用聚类：首先把某种度量函数下特征很相似的个体们看成一个整体，然后把整体中大部分个体都具有的特征作为整体自身的特征，最后利用整体的特征对个体特征形成反馈，弥补个体由于数据稀疏问题造成的特征缺失。

层次化聚类

前文讲述的二元文法中的搭配关系，有一定的数据稀疏问题。比如用户想要输入“铁板鳕鱼”的时候，如果“铁板—鳕鱼”之间的二元关系不在语料中出现，那么最终输入法给出的结果会是“铁板雪域”。但如果事先能把能够很多食物类词语聚成一个类，我们可能会得到“铁板—类$_{食物}$”这样的二元关系，同时因“鳕鱼”属于“类$_{食物}$”，我们就可以预测“铁板”

与“鳕鱼”之间有一定的二元关系。

不难发现，我们此时使用的聚类特征是“上下文”，处理基于上下文的聚类的时候，一个常用的方法是层次化聚类。层次化聚类的算法思想是：先找出特征最相似的个体进行合并，合并后的整体看成是新的个体，继续参加合并，直到我们认为整体数目已经足够小，或者找不到我们认为可以合并的个体对为止。

迭代聚类

与层次化聚类相对的，迭代聚类是首先随机指定一种聚类方式，并且定义一种代价函数，然后通过移动聚类中的元素到其他类，不断地优化代价函数，最后得到一个近似全局最优代价值，同时得到聚类的结果。典型的代表就是 LDA（Latent Dirichlet Allocation）聚类，它主要利用词与词的同现，进行同一主题词的聚类。

通过这种把同一主题的词聚成一类的方法，我们可以形成对应主题类的分类词库。比如大的游戏类下，可能有魔兽词库、天龙八部词库、三国词库；语言类下，可以聚出各个地方不同的方言词库；即使都是关于汽车，也会被聚成汽车品牌词库和汽车术语不同的分类词库。得到这些分类词库以后，针对一个用户，看该用户是否已经使用了词库中的一些词，以此推测这个用户于这个主题之间潜在的关联度。如果关联度比较大，那么可以把这个词库中的其他该用户没有使用过的词加载给用户，方便用户未来的使用。

例如，我们发现一个用户输入了“二校门”、“照澜院”这样的词，那么我们有理由相信，用户在未来很可能输入“紫荆”，“西操”这样的词，因为看起来他像是一个清华院内的用户。

IV. More and more…

统计自然语言处理技术在输入法中的能施展的功夫远不止上面提到的，比如通过实体命名识别提高输入法对实体词的输入准确率；利用信道模型结合语言模型进行用户拼音纠错；通过上下文无关文法分析和依存语法进一步提高输入法整句输入的正确率等等，这里就不一一描述了。

可以说，以中文输入法为代表的语音识别领域，涵盖了大部分自然语言处理理论要解决的问题，一个公司本身输入法产品的好坏，也代表着这个公司中文处理技术方面的能力。

4. 总结

文本讲述了智能输入法相比于传统输入法的优点，并简要介绍了智能输入法中使用的自

然语言处理技术。

总之，有了统计自然语言处理技术，才有了智能输入法的出现与不断地发展；输入法的聪明程度也会随着自然语言处理理论本身的进步而不断提高。

参考文献：

[1] Hinrich.: Foundations of Statistical Natural Language Processing, Electronics Industry, 2005
[2] Hastie, Trevor; Tibshirani, Robert; Friedman, Jerome: The Elements of Statistical Learning (2nd ed.). New York: Springer. 2009
[3] Peter F. Brown,Peter V. Desouza,Robert L. Mercer,Jenifer C. Lai: Class-Based n-Gram Models of Natural Language, Computational Linguistics, Vol. 18, No. 4. pp. 467-479, 1992
[4] Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent Dirichlet allocation. Journal of Machine Learning Research, 2003
[5] Church, Kenneth Ward, and Patrick Hanks: Word association norms, mutual information and lexicography. In ACL 27, pp.76-83, 1992
[6] 吴根清，统计语言模型研究及其应用，清华大学博士论文，2004.

作者列表

王砚峰，搜狗公司研究员，硕士，目前研究方向为自然语言处理，wangyanfeng@sogou-inc.com
贾剑峰，搜狗公司副研究员，硕士，目前研究方向为自然语言处理，jiajianfeng@sogou-inc.com
张扬，搜狗公司副研究员，硕士，目前研究方向为自然语言处理，zhangyang@sogou-inc.com