文章编号：1003-0077(2011)06-0026-07

# 基于互联网自然标注资源的自然语言处理

孙茂松

（清华大学 计算机科学与技术系，清华大学 智能技术与系统国家重点实验室，北京 100084）

**摘 要**：该文提出了“基于互联网自然标注资源的自然语言处理”的学术思想，并从自然标注资源的定义和基本类型、基于自然标注资源的计算、方法论层面上的初步思考等三个角度对这一学术思想进行了初步的阐发。最后指出了其中的一个基础问题：如果我们把全部自然标注资源所能提供的全部信息或知识都以一种系统的方式用到了极致，并且把它们最大限度地有机集成起来，机器能否如愿以偿地获得对自然语言一定深度的理解呢？

**关键词**：自然标注资源；用户产生数据；互联网；自然语言处理

**中图分类号**：TP391　　**文献标识码**：A

## Natural Language Processing Based on Naturally Annotated Web Resources

SUN Maosong

(State Key Lab. of Intelligent Technology and Systems　Department of Computer Science and Technology Tsinghua University, Beijing 100084, China)

**Abstract**: This article proposes an idea of “natural language processing based on naturally annotated Web resources”. The discussion is carried out from three perspectives: the definition and types of naturally annotated resources, naturally annotated resource-based computing, as well as several key points concerned at the methodological level. A fundamental problem is presented for further exploration at last: If we could explore and integrate all the information provided by all the available naturally annotated resourcesin different respectssystematically, can themachine, as expected, ultimatelyachieve some degree of deep understanding of naturallanguage?

**Key words**: naturally annotated resource; User generated data; web; natural language processing

自然语言处理自诞生之日起已先后发展出两种影响全局的主流研究方法(methodology)：理性主义方法和经验主义方法。总的来说，目前学术界应该形成了如下两点基本认识：第一，比较少地依赖统计的理性主义方法以及停留在 N-gram 之类比较表层语言单元的经验主义方法在全世界同行们的共同努力下，已经差不多做到了极致，需要谋求新的突破；第二，近中期的发展趋势是两种方法的融合，即多一点理性主义的经验主义，也就是说，研究超越了 N-gram 的、基于相对深层语言单元的经验主义方法。各种半监督的机器学习算法和基于结构的机器学习算法在其中发挥着越来越重要的作用[1-2]。

“行到水穷处，坐看云起时”。在这种大的态势下，自然语言处理研究下一步应该怎么走？这是每一位长期在这个领域辛勤耕耘的学者都不能不有所思考的问题。鉴于互联网上各类资源空前丰富的基本格局，本人不揣浅薄，于 2010 年 3 月 29 日在日本京都大学举行的“第二届清华—京都大学面向知识社会的智能技术与信息管理研讨会”(The Second Tsinghua University-Kyoto University Symposium on Intelligent Technologies and Information Management for Knowledge Society)上，首次给出了“基于极大规模自然标注语料库的自然语言处理”(Natural Language Processing Based on Huge-scale Naturally Annotated Corpora”的提法，并于 2011 年 5 月 6-8 日在香港教育学院举行的“语言语料库和

**收稿日期**：2011-10-20　**定稿日期**：2011-10-25

**基金项目**：国家自然科学基金资助项目(60873174)

**作者简介**：孙茂松(1962—)，男，教授，主要研究方向为计算语言学。

语料库语言学圆桌会议”(Round-table Conference on Linguistic Corpus and Corpus Linguistics)上，再一次做了“基于大规模自然标注 Web 资源的自然语言处理：一种可能的新的研究范式”(Natural Language Processing Based on Huge-scale Naturally Annotated Web Resource)的学术报告，进一步阐发了这一学术理念。

本文将围绕“基于互联网自然标注资源的自然语言处理”中的若干基本问题展开讨论，粗浅乃至不当之处，恳请同行们批评指正。

## 1　自然标注资源的定义和基本类型

什么是“自然标注资源”？其实，这个术语与“用户产生的数据”(user generated data)基本上是同义语，指互联网上各种用户出于各种交际目的而“制作”出来的各种资源，如网页、论坛、博客、微博、维基百科、社交网络、用户日志等等。如果我们从自然语言处理的角度而非从用户的角度以专业的眼光来看待这些数据，或者这些数据的某些部分，则它们实际上可以被视作已经加上了某些人工标注，而这些标注是可以为语言信息处理所利用的，虽然用户本人并没有意识到这一点(显然也没有必要意识到这一点)。换句话说，用户在无意中为自然语言处理研究的各种资源作了一定程度的义务“标注”——这正是“自然标注”的含义(说明性定义)。

我们举个例子说明一下。图 1 是 2011 年 10 月 7 日“百度新闻”中“财经新闻”的首页。在这个再普通不过的网页上，“空格”、“标点符号”和“句子开头或结尾”就是所谓的“自然标注”，能“透露”给我们不少关于词汇的信息。如：导航条中被空格两两隔开的“新闻 网页 贴吧 知道 MP3 图片 视频 地图”、“股票 大盘 新股 权证 板块 理财 基金 银行 黄金 保险 外汇 期货 案例 经济 民生 国内 国际 产经 公司 人物 消费 评论”等，都是词或短语很好的候选；双引号提示“开门红”也很可能是一个词或短语；标题“穆迪下调英国 12 家金融机构评级”显示“穆”可作为以“穆”开头的某个词或短语的左边界，“级”可作为以“级”结尾的某个词或短语的右边界，等等。我们不妨设想，如果把互联网上所有网页的相关信息都系统地汇集起来，将可能出现怎样的图景？!

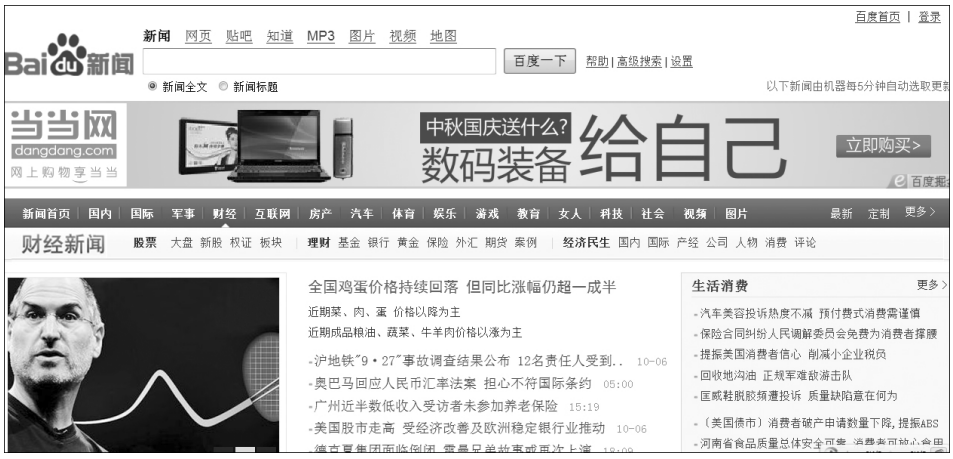


图 1　网页中的“自然标注”——“空格”、“标点符号”和“句子的开头或结尾”

“空格”、“标点符号”和“句子开头或结尾”属于“显式自然标注”(explicit naturalannotation)。与之对应的，是所谓的“隐式自然标注”(implicit natural annotation)。“隐式自然标注”需要借助某些知识予以揭示。如字符串“是一种”有助于在两个概念之间建立本体联系，可被视作一种“隐式自然标注”。在谷歌搜索引擎中键入“iPad 是一种”，会得到如下一些句子：

乔布斯本人宣称 iPad 是一种全新种类的产品
iPad 是一种娱乐加办公的时尚潮流产品
iPad 是一种简易的手持设备
iPad 是一种更浸入式的设备
iPad 是一种混合设备
iPad 是一种全新类型的电脑
iPad 是一种触摸屏平板电脑
……

对这些句子进行适当的自动分析，不难得出如下 AKO(A Kind Of)关系：

iPad AKO 产品
iPad AKO 设备

iPad AKO 手持设备
iPad AKO 混合设备
iPad AKO 电脑
iPad AKO 触摸屏平板电脑
……

这些断言无疑有助于新词“iPad”在一个已有的本体体系内找到合适的位置。

自然标注资源可以远远超越网页这种基本形式，如博客作者可以自由地为自己写作的博文添加任意的“标签”，以表达作者的意图或者心情。标签是另一种典型的“显式自然标注”。图2给出“搜狐博客”上某位作者发表的一篇博文“喝什么温度的水最健康”：

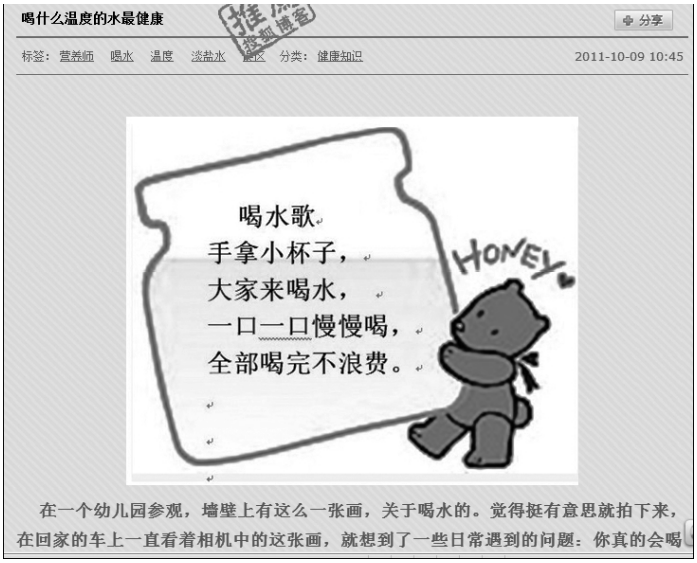


图2 博客文章中的“自然标注”——标签

注意左上角的“标签：营养师 喝水 温度 淡盐水 误区”。标签显然是高度个性化的，为个性化搜索创造的条件。网上庞大作者群标签的集合形成了所谓的folksonomy(大众分类体系)，又呈现出了很大程度的一般性。与通常主要依赖专家构造的taxonomy(分类体系)相比，folksonomy的“草根性”更能及时地抓住网民的网上生活“脉搏”(众所周知，大众分类体系在YouTube、Facebook等著名社交网站的图像和视频搜索中也扮演着十分关键的角色)。

Wikipedia是无数作者智慧的结晶，代表了“显式自然标注”资源的一个极致。观察中文维基百科中关于“乔布斯”条目的一个片段，见图3。

显然，其中的加重字体、蓝色字体(由于显示原因，此处为黑色字体)可被视作某种“显式自然标注”：一则为语言信息处理提供了大量有用的词语(把这些词语系统地收集起来，将会形成一个词语的宝库)，二则这些词语本身刻画了“乔布斯”的基本面貌，可作为特征用于词语语义相似度计算、命名实体排歧等任务中。从这个片段中，还可以很容易地得到“乔布斯”的中英文全名、简称等信息。

**史蒂芬·保罗·乔布斯**(英语：**Steven Paul Jobs**，**1955年2月24日—2011年10月5日**)，简称为**史蒂夫·乔布斯**(英语：**Steve Jobs**)，**苹果公司**的创办人之一，并曾任苹果公司的**董事会主席**、**首席运行官**，同时也是前**皮克斯动画工作室**的董事长及首席执行官(皮克斯动画工作室已于**2006年**被**迪士尼**收购[3])。乔布斯还曾是**迪士尼公司**的董事会成员和最大个人股东[9]。乔布斯被认为是电脑业界与娱乐业界的标志性人物，同是人们也把他视作**麦金塔**电脑、**iPod**、**iTunes Store**、**iPhone**、**iPad**等知名数字产品的缔造者[10]。**2007年**，史蒂夫·乔布斯被《**财富**》杂志评为了年度最强有力商人[11]。

图3 中文维基百科中的“自然标注”——字体加重与变色

维基百科关于条目的分类体系是“显式自然标注”的一种“高级”形式(图4)。纳入这个分类体系的条目正文可被直接用于文本自动分类的训练和测试。

自然标注并不限于上述种种。当我们从搜索的前台——网页走到搜索的后台“用户查询日志”，便会体会到这种日志达到了“自然标注”的一个新的“形态”。自然标注也可以超越文本，如网页之间的链接，微博的“粉丝”关系和“关注”关系等。

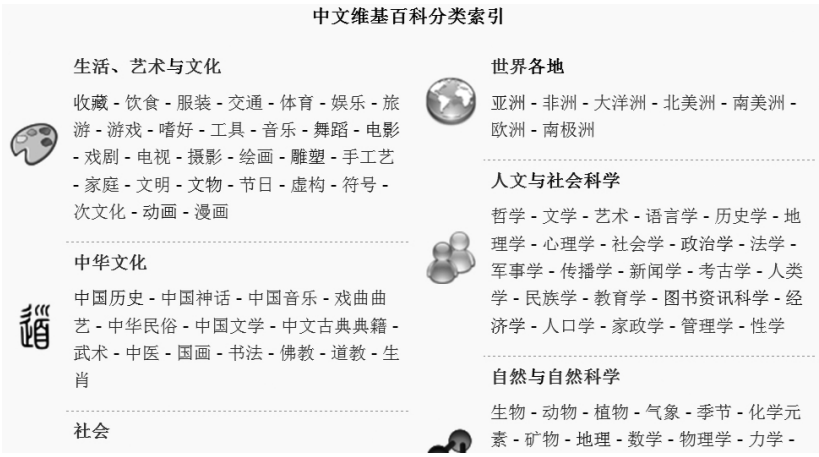


**图 4　中文维基百科中的另一种“自然标注”——分类体系**

## 2　从若干案例看基于自然标注资源的计算

近年来，国内外已经陆续有不少相关的研究工作，只不过研究者并没有意识到自己实际上是在“基于自然标注资源的自然语言处理”的框架下开展研究工作的。篇幅关系，这里不打算对这些工作做系统性的评介，而只是信手拈来几个例子，能够说明问题即可，点到为止。

文献[3]注意到了标点符号和句首、句尾等“显式自然标注”对中文自动分词可能的影响，基于超大规模中文 Web 语料库（目前是千亿字规模），利用最大熵模型，对任意汉字位置建立了关于此类“显式自然标注”的概率分布模型。实验结果初步验证了这个想法的可行性。

文献[4]利用引号从互联网新闻文本中自动抽取作者称之为 meme 的流行语句，通过对这些 meme 的追踪来定量、及时地把握美国政治、经济、文化等生活。“Lipstick on a pig”（猪的口红）即是一个典型 meme，反映了美国总统选战中得一个有趣的侧面。这个思路简单而巧妙：一方面妙在以“流行语句”作为各种社会热点问题的指标，另一方面使得计算变得简单。可以设想，如果不借助引号这个“显式自然标注”，要在海量文本中自动发现长度可变的流行语句，将是一件多么困难的事情！

文献[5]通过“We feel”“I feel”之类的“隐式自然标注”从互联网文本中抽取 Feeling 句子集合（如“We feel happy”），然后根据 happy 之类的情感词，运用情感分析技术来把脉用户的情绪。由于所设计的系统涵盖了实时社交网络，所以相关的时间信息、位置信息基本上都可以获得。以不断实时抽取的大规模 Feeling 句子集合为基础，系统做到了可在全球范围内从不同的角度动态地了解形形色色网民的情绪变化，建立起作者所谓的“情绪互联网”。配以生动的可视化技术，这个工作在 WSDM2011 国际会议上引起与会者的普遍兴趣。

学术界还有若干利用“A is a B”“A，such as B”“A and B”等“隐式自然标注”的有趣研究工作，这里不一一赘述。

文献[6]利用机器学习的算法，从大规模带有社会标签的文本中自动学习文本中词语与社会标签集合之间的统计关联，然后对任意输入的没有标签的文本，自动打上若干标签。例如，输入一句短文本：

我真的很喜欢你

机器会自动赋予如下标签（注意：文字长度甚至超过了输入）：

希望、情感、日记、对不起、心情、眼泪、我、勇气、答案、分手

看到这个结果时，我曾质疑其是否有道理。学生们则会心一笑，答曰非常符合现在的年轻人传达相关情感时的表达方式，捕捉到了言外之意。这个例子生动地显示了规模化了的“显式自然标注”folksonomy 所内敛的力量。

文献[7]则是应用来自搜索后台的“自然标注”——用户查询日志的一个经典案例。作者根据 2003 年至 2008 年谷歌公司在美国本土的 5 000 万个高频用户查询记录，自动挖掘出某些词语与流行性感冒的对应关系，并据此发出警告。与美国疾病控制和预防中心（CDC）以及欧洲流行性感冒检测计划（EISS）所采取的传统方法相比，这个工作在精度没有差别的条件下，将警告延迟从 1～2 周大大缩短到了 1～2 天。

以上给出了若干个从不同侧面(列举的侧面很不完全)利用“自然标注”进行计算的案例。可以看出,贵在“思路”,即如何巧施妙手,使自然标注能够为我所用。需要强调一点:“自然标注”貌似简单,但真正要把它挖掘出来、用起来,相关专业知识的指导,或者说从相关专家的研究成果中汲取养分,是极其重要的。举个例子:众所周知,汉语中双音节动补结构的紧密程度存在很大的差异,有些倾向于复合词,有些倾向于短语。一个极端是结合非常紧密者,如“扩大”“延长”,不能插入中缀“得”“不”;结合中间状态是结合比较紧密者,如“打碎”“杀死”,能加入中缀“得”“不”而有限制地扩展;另一个极端是结合非常松散者,“挖浅”“买长”,能比较自由地扩展[8]。“中缀”可被视作一种“隐式自然标注”,靠能否插入中缀能够轻易地将“结合非常紧密”这一极端与其他两种情况区别开来,但要区别后两种情况,仅靠中缀有时可能并不足够。文献[9]提供了另一个鉴别依据:结合紧密的双音节动补结构一般可以在后面带宾语,也可将宾语移至动词前,如“打碎”,可以说“打碎了玻璃”,也可以说“玻璃打碎了”;与此形成对照的是,结合松散的双音节动补结构在后面带宾语就十分勉强,一般只能摆在动词前,如“挖浅”,说“坑挖浅了”而几乎不说“挖浅了坑”。来自搜索引擎的检索结果验证了这个鉴别依据的有效性,如图5所示(注意两对检索结果次数的对比)。这个变换式的模板是一种“隐式自然标注”,直接得益于语言学家研究成果的启发(当然,这需要你具备从成果宝库中挖掘出可用“线索”的敏锐力和眼光)。

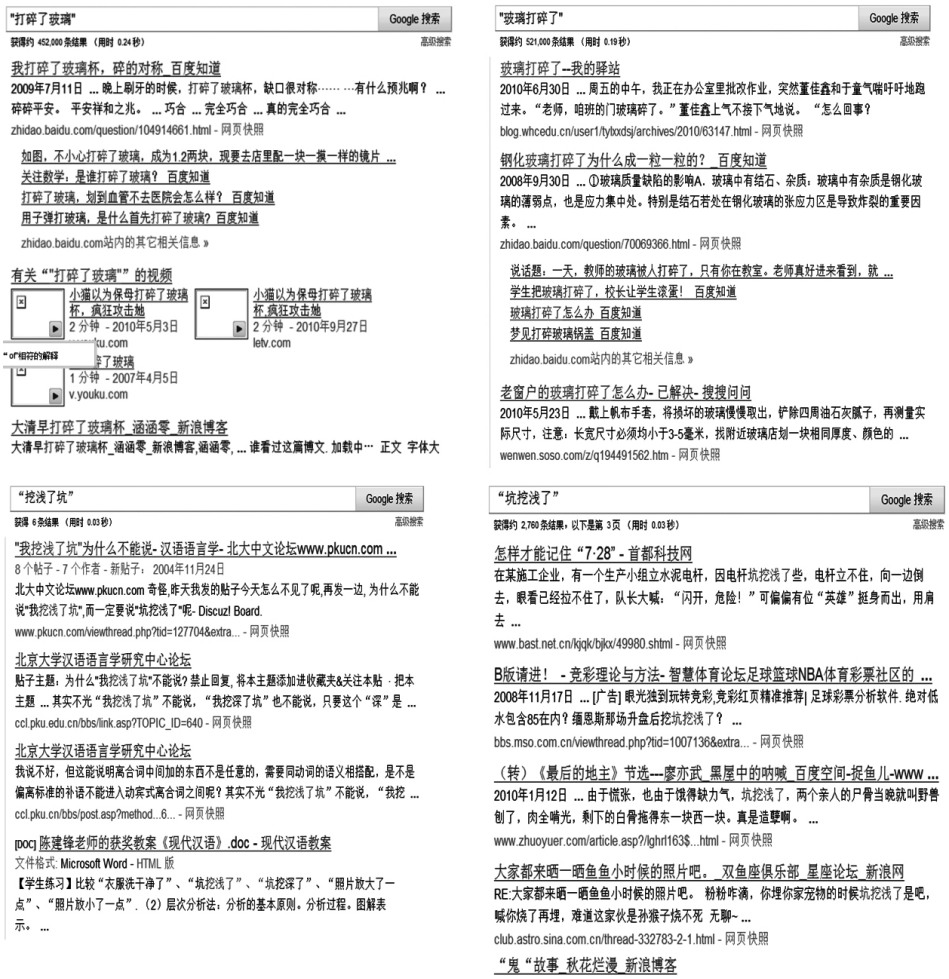


图5 挖掘“隐式自然标注”:“打碎”与“挖浅”后带宾语与前置宾语的量化对比

有了思路,大路数就对了,但在具体计算时,还是要有很多因具体任务而异的技术细节需要认真处理。如文献[4]指出:“Lipstick on a pig”在所考察的语料库中具有较为复杂的61种变形,需要进行normalization的处理,才能更为有效地追踪其轨迹,见表1[4]。

表 1　“Lipstick on a pig”的各种变形

| 长度 | Meme 的不同表达形式 | 频度 |
|---|---|---|
| 4 | lipstick on a pig | 11 235 |
| 7 | you can put lipstick on a pig | 1 051 |
| 12 | you can put lipstick on a pig but it’s still a pig | 967 |
| 11 | you can put lipstick on a pig it’s still a pig | 793 |
| 14 | i think they put some lipstick on a pig but it’s still a pig | 604 |
| 5 | putting lipstick on a pig | 542 |
| 25 | it’s still a pig you can wrap an old fish in a piece of paper called change it’s still going to stink after eight years | 400 |
| 5 | but it’s still a pig | 367 |
| 4 | it’s still a pig | 303 |
| 32 | you can put lipstick on a pig it’s still a pig you can wrap an old fish in a piece of paper called change it’s still going to stink after eight years | 241 |
| | …… | |

## 3　方法论层面上的一些初步思考

关于自然标注资源用于自然语言处理，在大的层面上，我想至少可以设想出以下几点：

（1）极大规模（massive scale）是一个资源能够成为可利用的“自然标注资源”的重要条件

由于“自然标注”具有“攻其一点，不及其余”的特点，所以如果规模上不去，往往会因存在严重的数据不足而遗漏一些本应出现的现象。在极大规模资源上（极大规模资源，舍互联网其谁也），“东方不亮西方亮”，对某种现象的观察总有希望能够做得尽量全面。

（2）计算的基本基调：极大规模资源上的浅层处理（shallow processing）

显而易见，大多数“自然标注”（尤其是显式自然标注）本身就是“浅层”的。利用其的技术手段往往也是比较“浅层”的（如引号及“是一种”之类的模板，依赖的手段多是字符串匹配），这样易于保证相关信息获取的可行性。我们知道，任何深层语言自动分析方法，到了互联网上，其可行性基本都不成立。这也是我们不得不更多地依赖浅层处理的一个原因。

（3）“能够帮助人的电脑，需要人的更多帮助”

这是钱钟书先生对“电脑”一句鞭辟入里的断言。在利用“自然标注资源”时，这句断言仍然适用，必须融入人的智慧。虽然计算多在“浅层”进行，但为了充分揭示所关注的问题，一般需要精心设计多个可相互印证的观察角度（如设计多个不同的模板），并且进一步整合诸多角度，这无疑需要仰仗人的专业知识，正如图 5 所示的情形一样。

（4）对根据“自然标注”所获取数据应进行“去粗取精”“去伪存真”的处理

这是由基于“自然标注”的计算的基本基调所决定的。如基于字符串匹配查到的相关句子会由于层次不分而产生“噪音”，干扰判断。需要设计合理的过程最大限度地抑制此类干扰。

（5）对一个语言资源尽可能进行多个不同角度的“自然标注”分析及整合

这种分析和整合有可能使从单一角度看或属于“弱可用”的资源升华为整体上看的“可用资源”。每一个“自然标注”的角度都似乎在织一张网的某个局部，有无可能把这些局部合理地拼装起来，以产生某种全局的效果？

当然，这些思考还很不成熟，有待于更多实验的检验或验证。

## 4　结语

本文冒昧提出了“基于互联网自然标注资源的自然语言处理”的学术思想。听起来有点像一种研究范式（paradigm），其实无非是每天忙碌于“低头拉车”之余，偶尔“抬头看路”的某种思索而已，或许不是没有一点道理，希望能起到抛砖引玉之效。这里面存在一个现在我们并不清楚的基础问题（funda-

mental problem)：**如果我们把全部自然标注资源所能提供的全部信息或知识都以一种系统的方式用到了极致，并且最大限度地有机集成起来，能否最终如愿以偿地使机器获得对自然语言一定深度的理解呢？或者说能否真的对自然语言处理产生某种实质性的帮助和影响呢？如果达不到这个境界，退而求其次，穷能力之所及，沿着这条路我们又能走多远呢？要把这个“终极”问题弄个“水落石出”，必须付出不懈且有新意的探究。**

## 参考文献


[1] Steven Abney. Semisupervised Learning for Computational Linguistics [M]. 2007. Chapman and Hall/CRC.

[2] Noah Smith. Structured Prediction for Natural Language Processing [C]//A Tutorial Presented at ICML, Montr al, Qu bec. 2009.

[3] Zhongguo Li andMaosong Sun. Punctuation as Implicit Annotations for Chinese Word Segmentation [J]. Computational Linguistics, 2009, 35(4): 505-512.

[4] Jure Leskovec, Lars Backstrom and Jon Kleinberg. Meme-tracking and the Dynamics of the News Cycle [C]//Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2009.

[5] Sepandar D. Kamvar and Jonathan Harris. We Feel Fine and Searching the Emotional Web [C]//Proceedings of the Fourth ACM International Conference on Web Search and Data Mining. 2011.

[6] Xiance Si, Zhiyuan Liu andMaosong Sun. Modeling Social Annotations via Latent Reason Identification [J]. IEEE Intelligent Systems, 2010, 25(6): 42-49.

[7] Jeremy Ginsberg, Matthew H. Mohebbi, Rajan S. Patel, Lynnette Brammer, Mark S. Smolinski and Larry Brilliant. Detecting Influenza Epidemics Using Search Engine UueryData [J]. Nature, 2009, 457(19).

[8] **梁银峰. 汉语动补结构的产生与演变[M]. 2006. 上海学林出版社.**

[9] **陆俭明. “VA了”叙补结构语义分析[M]//陆俭明自选集. 1993. 河南教育出版社.**