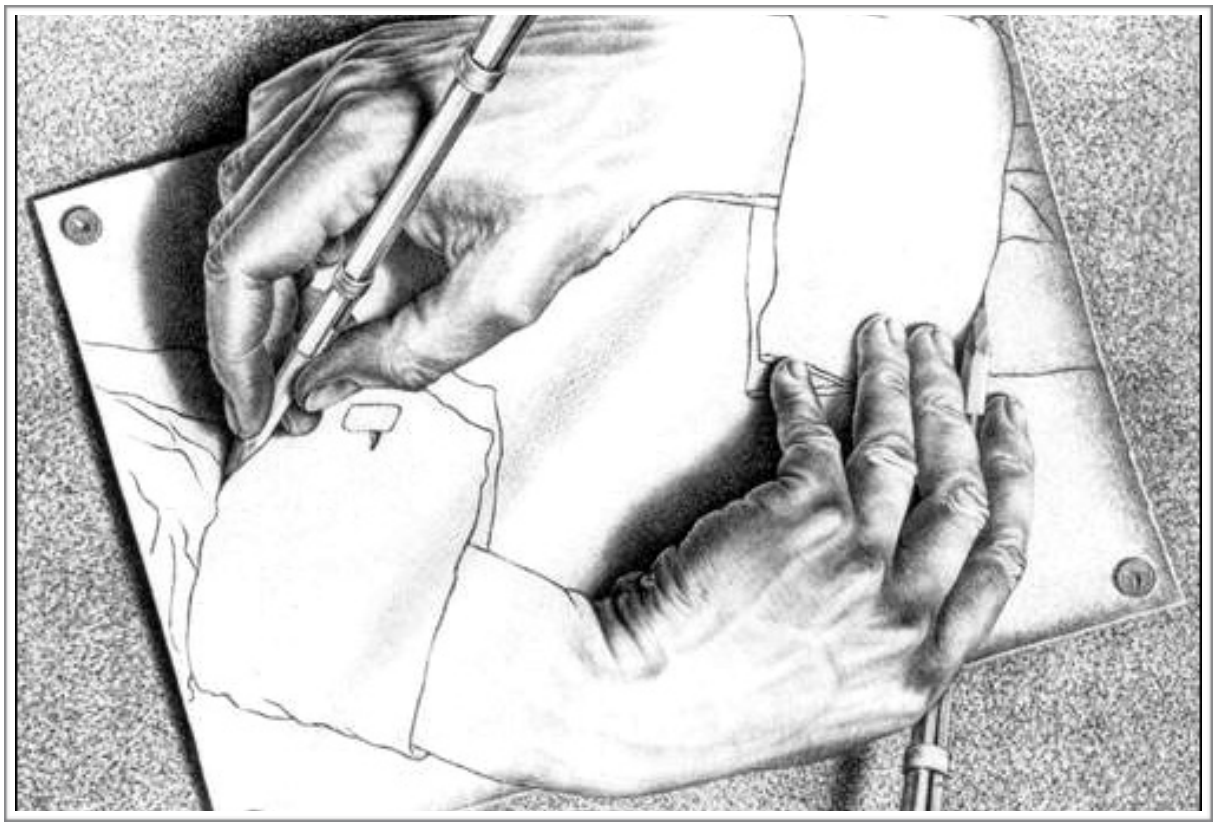
言语自然语言

Able was I ere I saw Elba.



朱胤恒

2015年冬季

言语自然语言

Able was I ere I saw Elba.

先讲个故事吧

很久很久以前，天下人都讲一样的语言。人类建巴别塔，要与 神比高低。神不满，变乱人类语言，巴别塔就建不成了。

可见，人若能通过语言完全无障碍地传递信息，人能够到达与神一样的高度。（这里的高度是双关：一指巴别塔的高度，二指能力的大小）

自然语言的缺陷

然而遗憾的是：自然语言是没有办法无损地传递信息的，在不同语言之间是这样，在同种语言内部也是这样。比如：中文中的“道”，在英文中是什么呢？再比如：我说“苹果”，你理解的是“🍎”，而我想表达的是“📱”。

若把形而上之“道”理解为信息，形而下之“名”理解为语言。那么“道可道，非常道。名可名，非常名”就可以非常好的描述这样的关系。我心中所思之“道”，不能被完全无损的表达出来。能被表达出来的，都不是原来那个“道”了。

若把人的对话理解为：“编码->信道->解码”的过程，那么在编码的过程中信息已经被扭曲，解码的时候信息也被扭曲。发送者所思与所说不同，接收者所听与所想不同。

“无上甚深微妙法，百千万劫难遭遇，我今见闻得受持，愿解如来真实义。”即便是人，想要理解“真实意”尚且不易，如今又想要让机器理解不更是难上加难。

撇开话题的机器翻译

为何我想讲一讲机器翻译呢？我觉得机器翻译是无法取代人工翻译的。因为人工翻译本质上是一种“再创造”的过程。比如：副标题中的“Able was I ere I saw Elba.”翻译成中文要怎么翻译？（注意这句话和标题一样是回文的），再比如“a friend indeed is a friend inneed”(注意这句话是押韵的)，再再比如外国小说的城市、地名要怎么翻译？（注意某些外国的城市是有特定的含义的），若是直译看似符合原文，却失去了原意。这些问题在侯世达（Douglas Hofstadter）教授的《哥德尔、艾舍尔、巴赫 : 集异璧之大成》讲的很清楚，事实上这本书的中文版就是采用意译的方式重新创作的。

“要提高机译的译文质量，首先要解决的是语言本身问题而不是程序设计问题；单靠若干程序来做机译系统，肯定是无法提高机译的译文质量的。同时，他还指出：在人类尚未明了大脑是如何进行语言的模糊识别和逻辑判断的情况下，机译要想达到“信、达、雅”的程度是不可能的。”

——《机器翻译五十年》

幸运的是，我们并不需要机器翻译取代人工翻译。事实上，对于“功能性的材料”，只需要就像史晓东老师的作品“云译”——让机器辅助翻译。而对于“艺术性的材料”，就需要文学大家的再创作。就如草婴之如《复活》，傅雷之于《贝多芬传》……

镜子

很难讲是先有语言还是先有智能，是先会想还是先会叫。但是，我的直觉告诉我语言与智能密不可分。研究透自然语言，说不定能反过来对理解智能有帮助。

“认识你自己”，苏格拉底如是说。认识自己最好的方式不久是照镜子吗？幸运的是，人工智能就是那一片镜子。自然语言处理或许能让机器帮助理解我们的语言，我们的想法。比如，表述自己的情感（自然语言生成），读懂诗歌（自然语言理解）。不知道发现没有，偶然间，我们也变成诗人了。让机器让我们更加了解我们自己，可能是某个藏的很深很深的情感，深到我们自己都无法触碰，可能是某个很浅很浅的情愫，浅到自己都没办法发觉。让机器理解自然语言就是那一面镜子，在镜子的那一边是我们自己。

就像递归对吗？事实上，我不只在一本书中看到，有人猜测智能的突破口在于自我调用——递归。比如，本文最开始的埃舍尔的画、再比如图灵的停机问题、lamda演算……充满着矛盾却又和谐统一，蕴含美感。借用梅贻琦先生的话，调用自我“大概或者也许是,恐怕仿佛不见得”就是突破口，而悖论反而是最自洽的真理。

习器

讲了怎么多玄之又玄的东西，接下来就来一点干货吧。

如何在什么都不会的情况下了解一个领域呢？我觉得最好的方式是读一些科普类的书籍。想自然语言处理方面我就是听老师推荐读了《数学之美》。这样的科普书籍可以让我们对这一领域有一个直观、宏观的认识。

之后，便是在科普的骨架上，不断扩充。我一般是不求甚解、浅尝辄止。看不懂为什么就先跳过，去看看怎么用。

再之后就是玩。利用一些开源库：opencv、opennlp的sample，自己改装，发现问题，回去查找之前跳过的内容，解决问题。

在这期间，还少不了到论坛、知乎、博客上找教程、找心得、找经验、找推荐。还有在官方的Tutorials上的示范。

每天早上都是带着对问题的好奇起床。这样的生活真是充实且美妙。

炼法

当今天下，除了VCZH（陈梓瀚）谁还会自己造轮子？如此一来，会使用别人的轮子就显得十分重要。问题变的简单了，我们只需要学会使用各种库文件的接口，直接使用就可以了！然而，是不是这样就足够了呢？当然不是。我们还要理解轮子是怎么造出来的，因为不同的算法各有长处，各有其解决问题的使用范围。当现有的轮子都不能够解决现有的问题的时，这时候就需要我们自己造轮子！终于轮到我们自己造轮子了！

生物信息学大牛Shirley曾经给搞生物信息学的人分了个层。0级 (Level 0)：为建模、而建模（modeling for modeling's sake）1级（Level 1）：给数据、能分析。2级（Level 2）：想新招、玩数据。3级（Level 3）：玩数据、作发现。X级（Level X）：玩科学、讲政治。我觉得这可以推广到用统计学方法做分析的所有学科。再结合之前讲的轮子理论。路漫漫其修远兮，0级新手将上下而用轮子求索。

悟道

"A genius interests in everything"

这是我在新生诱导实验课的最大收获。保持对世界的好奇心，自然而然就是别人眼中的天才。当然比成为别人眼中的天才更重要的是不辜负了来这世上的百八十年。尼采曾说："每一个不曾起舞的日子都是对生命的辜负"，诚哉斯言。

再讲个故事吧

伊卡洛斯使用蜡和羽毛造的翼逃离克里特岛时，他因飞得太高，太接近了太阳，双翼上的蜡开始融化，而他并不自觉，知道背后的羽毛都散了，身体开始坠落了，他才开始挣扎。然而，为时已晚，最终溺死于万倾碧波。

"必须在半空中飞行。你如果飞得太低，羽翼会碰到海水，沾湿了会变得沉重，你就会被拽在大海里；要是飞得太高，翅膀上的羽毛会因靠近太阳而着火。"，还记得他的父亲曾这样对他说。

附件

由于自然语言的缺陷、和我个人表达能力有限，可能前半部分会读起来怪怪的，因此我列出激发我的文章，可能读了这些，就能理解我文中表达不清的地方。

- 《geb》——《蚂蚁赋格》、《序言》
- 《永恒金色对角线》—by刘未鹏http://www.360doc.com/content/07/0327/15/4910_416192.shtml
- 《心智、语言和机器》
- 《意向性与人工智能》
- 《老子他说》南怀瑾
- 《伊卡洛斯和代达罗斯》出自希腊神话
- 《巴别塔》出自《圣经旧约》
- 《如何成为顶级生物信息学家》http://blog.sciencenet.cn/blog-404304-834869.html

另外：为了迎合这篇文章的主题——回文，再附上一首卡农《螃蟹卡农》。

另外：推荐一部电影《鸟人》，呼应《伊卡洛斯和代达罗斯》的故事