

从历史书中探索知识图谱

2026年03月19日

一、项目缘起：从语法高亮到知识重构

1.1 为什么要做《史记》知识库

首先感谢永启的热情介绍。

我已经有一年多没有来跟大家汇报了。

从两年前开始，我就退居二线，现在公司的全部领导层都是90后和00后，像我这样的70后已经回到自己应该做的事情上。

所以我现在是文因互联的一位普通工程师，也很感谢公司给了我很多自由，让我有时间去探索一些纯粹个人兴趣的事情。

我今天跟大家汇报的东西，一开始是一个完全不务正业的项目。

因为我非常热爱历史，但在阅读过程中，我感觉传统的历史书特别难读。

光《史记》我家里就有8个不同的版本，但没有1个版本让我觉得满意，所以我之前就有了1个想法：为什么不能做1个让我自己很满意的《史记》呢？

1.2 程序员的语法高亮启发

我的初衷是这样的：我们程序员写代码都有语法高亮。

在我十几岁刚开始编程的时候（**90年代中期**），那时候还没有**语法高亮**。

但后来上大学之后，基本上所有的编程环境都是**语法高亮**。

如果让一个**程序员**去读没有彩色的程序，我相信任何人都不能接受。

那么为什么普通人现在还能接受没有彩色的文本呢？

我相信如果去问**20年之后**的年轻人，他们会觉得：

"爸爸，你那个时候读的书全是黑白的吗？"

为什么普通人没有享受到**程序员**享受的这种高效认知辅助工具？

因为我们以前没有那么好的**人工智能**工具来帮我们做这件事情。

这是一个非常小的初衷，但在做这件事情的过程中，我**发现**越做越大，有了越来越有趣的结果。

今天跟大家分享一下，为什么这个完全不务正业的东西，最后反而真的**做成**了一个挺有意义的工作。

二、核心成果：从文本到**知识图谱**

2.1 数据规模统计

《**史记**》的正文是**52万字**，后来诸先生给他加了一些标注注释等等，最后变成**57万字**。

在**57万字**里面，我们**识别**出来了：

15,000个 实体

10万多个 标注

3,000多个 事件

7,600个 关系（**事件**之间）

但这只是挂一漏万，把这里面最核心的**语义结构 提取**出来。

在这之上我们做了很多工作。

从最基础的 **知识图谱 建模** 开始：**实体** 有 **属性**、有 **关系**，还有各种特殊的 **实体** 类型，比如人物、地名、职位、**事件** 等等。

事件 之间有 **关系**，人物之间也有 **关系**。

对于 **15, 000个 实体**，肯定需要进行 **概念分类**，所以我们又进行了 **本体 构造**，这都是非常经典的 **知识工程** 工作。

2.2 效率对比：AI 时代的 知识工程

考虑到我个人花在这个问题上的时间，前前后后大概也就 **70个小时**，就能 **构造** 出一个包含 **1万多个 实体** 的非常丰富的 **本体**。

相比之下，**20年前** 我还在读 **博士** 的时候，**2006年** 的暑假，我花了 **三个月** 的时间 **构造** 了一个可能是世界上第一个猪肉质量评估 **本体**，那个 **本体** 包含 **400个 类**，花了我整整 **3个月** 的时间。

那个时候我们的 **知识库** 编辑工具叫 **Protégé**，是 **斯坦福大学** 开发的一个 **本体编辑器**。

但我在做整个 **知识库** 开发的过程中，没有 **用** 到任何一个 **本体编辑器**，因为现在我们 **发现**，**IDE** 加上任何一个强大的 **AI**，就是最好的 **本体编辑器** 了，或者说可以是任何一种知识结构的编辑器。

所以你不再需要一个特殊的编辑器，这个结论可以推而广之。

2.3 AI 依赖的现实

在做这件事情的过程中，我 **发现** 如果现在离开 **Claude**，我基本上什么都干不了了。

我现在所有的工作都是在 **Claude** 的辅助下 **完成** 的，我所有的 **Word** 文档、所有的 **PPT**，包括今天这个 **PPT** 也都是 **Claude** 帮我 **写** 的，我大概花了 **30分钟** 的时间。

我花了非常多的时间导出我所有的数据，然后 **发现 AI** 基本上可以取代我们桌面案头工作的几乎全部环节。

我已经找不到任何事情是 **Claude** 不能帮我干的。

但这么做的一个代价是：有时候我的梯子挂掉了，或者 Claude 被封掉了，我突然发现我的大脑就已经被摘除了，什么都干不了。

三、实际应用：多维度的知识呈现

3.1 语法高亮阅读器

我们的核心工作是把《史记》的 知识库 开发出来，然后上面搭建了一些小规模的应用。

左边就是我最大的初衷：语法高亮阅读器。

我可以把《史记》当中所有句子的动词、名词和句子之间的结构，让 AI 帮我分析。

因为我们从小学中学的时候，学语文最重要的就是要把一个段落拆分成主要的句子。

如果一个句子有排比结构、层级结构、因果结构，我们是用大脑来做这些分析。

现在我是直接用 AI 帮我把所有底层的这些 语义结构 都分解出来了。

可以说，当我习惯了这种阅读方式之后，再去阅读那种完全线性的黑白文字，我自己现在已经无法忍受了。

这是一个单行道。

所以有很多书我以前一直都没能读完，太难懂了，比如《红楼梦》我一直都没能读完，因为人物太多、事件太复杂。

我下一步准备把《红楼梦》也按这种方式重新做一遍。

3.2 知识图谱可视化与查询

还做了各种事件的可视化，还有知识图谱的查询，以及在这之上的一些更深层次的东西。

3.3 从书到App：PDF to Skill

因为我们最近还在做另外一件非常有趣的事情：你把任何一本书给我，我把它转换成一个 **APP**。

一开始我们想的都是挺正经的应用，比如把 上交所 的《46号文》、《42号文》、《43号文》转换成一个科创板审核 **APP**，或者把《财务粉饰面对面》这本书转换成一个财务报表造假发现应用。

以前我们要用人工来做这些事情，需要很多 **业务经理** 去读这些文件，然后再把它写成规则。

我们在给 北交所 做 **利器系统** 的时候，他们给了我们 **100多兆** 的文件，我们的 **业务经理** 们有 **五六十** 人读了 **三个月** 才能读完，然后整理出来了 **几千条** 规则。

所以我们就想把这件事情 **自动化**。

当然我们探索了很多错误的技术方案，到了今年才算是找到了一个相对有效的方案。

现在基本上你丢任何东西给我，那个规则自己就出来了。

总结起来，现在的每一条规则都是一个 **Skill** 了，所以我们首先构造了一个叫 PDF to Skill 这样一个工具。

3.4 有趣的应用案例

然后我就跟同学们说，我们能不能做更多有趣的应用。

梦想男神生成器：其中一个女生做了一个“我梦想中的男神长什么样”的应用。她怎么做的呢？

她就丢了一本《相面书》（看相的书），然后出来了一个应用。我描述一下我的 **生辰八字**、我的 **星座**，它就可以画出来我的男神长什么样。

观鸟App：还有一个同学喜欢看鸟，他就把一本《观鸟图鉴》丢进去了，然后就出来了一个观鸟的 **App**，那是我们目前最有趣的一个观鸟 **游戏**。

侦探游戏生成器：还有一个同学喜欢 **桌游**，各种侦探 **游戏**。他就找了一本《侦探小说写作指南》，丢进去以后，它就可以生成无限多的各种侦探 **游戏**。

现在我们做了 **1200个** 这样的应用，《史记》是其中一个。

我们每天平均要要做好几百个（**200-500个** 之间）这样的 **游戏** 和 **APP**。

比如说最近还有一个我们现在最火的应用，我完全想不到，就叫**机修指南**，这是目前我们访问量最高的一个应用，就是把一本手册**转换**成一个指南。

还有一个是叫**妇科圣手**，是把一个**妇科医生**的《**妇科教科书**》放进去，然后病人在里边咨询非常多的东西。

3.5 《史记》的241项历史技能

我把自己的**知识库**放进去，然后把《**史记**》放进去，**构造**出来了**241项历史技能**，基本上就是跟打仗、经济、管理国家、管理一个家庭相关的所有这些东西。

后面我自己也搞了个小**App**，我的**App**比**00后**那帮人做的差远了，他们做得真棒。

但我做所有这些事情，其实**我一行代码都没有写**，我所有的这些工作都是用**Skill**来**完成**的。

四、技术架构：**四层语义模型**

4.1 知识处理的复杂流程

整个**Skill**的过程是不断增长的，因为一开始我**写**了一个**规范文件**，就是一个**Spec**。

也不是我写的，是用机器帮我写的，我就跟它对话，然后它写。

可能一开始就**200行**，然后越写越多。

因为你做这件事情——把一本书**转换**成一个**知识库**——其实是个很复杂的流程。

比如说**版本**问题，我到底是用**中华书局**版还是维基文库版，其实有很多差距，所以我要进行**校勘**。

然后第二步我要进行**结构分析**，就是我要把一本书**拆分**。

《**史记**》**130章**那是没问题的，但章里面要分小节，然后我处理的时候要给每一个小节加上一个标号。

然后它里面还有各种各样的韵文，什么太史公曰、赞，还有大段大段的引用文字，其实都是要进行**结构分析**的。

4.2 实体识别与多轮反思

把这些一点一点拆开来之后，下面就要进行**实体构建**了，有**实体**的**发现**、**实体**的**消歧**。

特别是在这个过程中，很多事情不是一下子就能做成的，要**反思**很多轮。

比如说**秦国**有一个【；大将】叫**白起**，我们在第一轮标注的时候把“白”**标注**成了**白起**，机器自动标的。

但是后来我们有了更多知识之后，因为我把所有的人名字拿出来先做了一个人名字表，然后又对每一个人做了一个生卒年表。

有了这些知识之后，我让它再**反思**一轮，结果它在下一轮就**发现**：

“噢，这个**公子白**是生活在**公元前4世纪**的人，而**白起**是**公元前3世纪**的人，这肯定不是同一个人。

”

所以它就把第二次标注的时候，把**公子白**的那个“白”字**标注**成了**公子白**而不是**白起**。

所以这个过程要经过**很多轮的反思**。

4.3 事件抽取与时间推理

后面差不多整个过程都是类似的。

实体当中有一种特殊的**实体**叫**事件**，因为《**史记**》是关于历史的，最后我们**抽取**出了**3000多个事件**。

事件里最难的地方就是要**确定**每一个**事件**发生在哪一年。

因为《**史记**》中直接标注了哪件事情发生在哪一年的只有**700多例**，其他的**2000多例**都要靠我们用**推理**手段把它**推理**出来，所以我们前后做了**5轮推理**，基本上把这**3000多个**时间全都**推理**出来。

特别是《**史记**》里还有表格，《**史记**》里有**10张**年表，那里面也有**事件**，从那里面**抽选**出了**800多个事件**，跟正文里出现的**事件**也要对齐。

很多**事件**在不同的来源要进行融合，所以有很多这些审查的工作。

类似的还有 **关系构造**、**本体构造** 等等，就不详细说了。

4.4 **四层语义结构** 理论

整个过程，我们整体上来说，底层的理论基础就是叫 **四层语义结构**：

第一层：**结构语义**

就是我们要把文字 **切分** 成段落、**切分** 成句子，然后在每一层面上都要去 **理解** 文本单元之间的 **语义结构** 是什么。

比如说我们把所有的太史公曰（**130章** 里面其实有 **125个** 太史公曰）全部 **提取** 进去，所以这就是对每一个区块要给它做标签，还有就是每个句子之间 **发现关系**，因为这样帮助我们 **构造** 阅读器。

第二层：**图谱语义**

在句子之上，我们 **找到** 人物、地点、**事件** 等等，我们一共 **定义** 了 **18种** 不同的 **实体**，到这个级别，我们传统意义上的 **知识图谱** 就 **构造** 好了。

第三层：**知识语义**

大家可能说：**鲍** 同学，**知识图谱** 不一定有知识。

那不是 **知识图谱** 没有知识，我们传统意义上的 **知识图谱** 是只有图谱没有知识的。

你要有知识的话，首先你得有 **分类体系**，你得有 **推理规则**，你还得有我们后来自己 **定义** 的一个词叫 **标准知识单元**，这是从电商里学来的。

这里叫 **标准知识单元**（Standard Knowledge Unit，SKU），因为我们会 **发现** 知识传播的时候有一个颗粒度。

一个 **三元组** 在传统图谱上面的颗粒度是一个 **三元组**，这个颗粒度太小了。

但是如果你把一整章或者一整段的东西都放在一起，这又太大了。

所以比如说一个**事件**，其实它就是一个很合适的颗粒度。

我们有**3000多个事件**，所以这就是一个**知识单元**。

又比如说一个人的完整的家庭关系，这也是一个**知识单元**。

所以我们会把知识再**包装**一层，就相当于我们给知识装上了集装箱，然后按集装箱的元数据来进行知识分发。

第四层：**应用语义**

最上面一层就是**应用语义**，就是各种**历史推理**，比如说年代的**推理**、人物的**推理**、历史矛盾的**推理**等等。

但我最终想做的、真的正在做的一个叫**历史叙事的重构**，比如说刚才那个历史**事件**罗生门就是我正在做的一个早期的尝试。

五、方法论：**元技能（Meta-Skills）**体系

5.1 **27个管线 Skills**

我们具体怎么做？

其实就是我们**设计**了大概到目前为止（今天早上）是**26个27个 Skill**。

注意，所有这些**Skill**是怎么**造**出来的：**没有一个 Skill 是我手写的**，因为我在处理的每一步过程中，都留下了大量的**日志**，所以每一个中间环节都有**日志**、都有错误，这些**Skill**都是从这些**日志**当中**总结**出来的。

所以基本上我们每一个**Skill**都会包括这样的核心规范：

你的输入输出是什么

你怎么做标注

你怎么做 **质检**（**质检** 极其重要，真正的业余玩家和专业选手之间的区别，其实就在于 **质检** 工具怎么 **设计**）

然后你要做很多轮，要做 **多轮反思**

比如说对于任何一种有动词标注或者 **实体** 标注，你都要先 **做** 按章标注，然后再 **做** 按类型标注，再跨章一次性标注。

如果做 **一轮** 不够，你可能做 **两轮**、**3轮**、**5轮** 慢慢做。

还有就是在这个过程中不能完全依靠 **大模型** 来做，后面还有很多规则要去做。

5.2 **OTF+JIT+Bootstrap** 三原则

大家可能很好奇这 **20多个 Skill** 怎么 **设计** 出来的？

它也是这些 **Skill** 本身，最后是由其他一组 **Skill** **做** 出来的，所以我们 **总结** 出来了叫 **元Skill**（**Meta Skill**）来做这件事情。

第一个：**On the Fly**（及时总结原则）

我们不是立即开始 **设计**，在做的过程中我们不断地改，我们 **记录 日志**，然后不断地改。

第二：**Just in Time**（实时交付）

就是实时 **交付**，从来不是说先去想清楚怎么做，而是先做一个最小可用的版本，然后不断地去 **迭代** 它。

第三：**Bootstrap**（自举）

最后你要自举。

就是我们不依靠任何人、不依靠任何专家，所有的知识都是 **AI** 自己 **发现**、自己 **积累** 的，只是人类在某些时候给它一些对错的 **判断**。

所以让知识自动增值，这就是我们最核心的一个原则，叫好东西都是总结出来的。

5.3 14个元技能总结

后来我们总结了14个元技能，就是系统地讲到底怎么做这种反思。

在我的网站上有一本书，一共有400多页就是讲这个，欢迎大家回头到我GitHub上去下载。

这里面比如说反思就是一种常见算法，在任何一个上下文上面我们都需要反思。

还有多样的方法是指我们每一个工序其实都需要多种方法的综合混用，所以并不是说我们有了大模型之后就完全依靠大模型，因为归根结底我们要省钱。

我到目前为止在这件事情上大概花了1万块钱，但未来我们一定是可以把这个成本降下来。

如果我们不用这一整套规范方法的话，我觉得几十万都打不住做同样的工作。

六、深层推理：历史谜团的探索

6.1 复杂的逻辑推理任务

最后就是要进行各种各样的逻辑推理，比如说刚才提到了人物生卒年份推理，因为在《史记》当中，你要把每一个人出生和死亡的哪一年精确地发现出来，其实是很难的事情。

因为绝大多数人物的生卒年份《史记》上是没有说的，但是你要根据他跟其他人的互动来判断他的出生和死亡到底在哪一个区段之内，这就是一个推理过程。

又比如说每一个人的姓氏，因为在先秦的文献当中，人基本上都只是说一个单名，他不会说姓、不会说氏，因为先秦时代姓和氏是不一样的。

所以对《史记》当中的4000多个人物，你想让他知道每一处上下文情况下他到底姓什么、属于哪个氏族，其实是一件很麻烦的事情。

我到目前为止做了1800个人，其他还有2000多个人还没做，这也是一个很复杂的逻辑推理过程。

6.2 历史事件罗生门

但后面还有更多的逻辑推理过程，比如我正在做的一件事情叫事件罗生门。

因为同一件事情，比如说“鸿门宴”，在几个不同的章节当中都会出现，但是它的描述不完全一样。

同一场战役，往往会在人物本传当中偏向这个人、给这个人说好话，但是到了另外一个人传记里面，就知道了：“原来其实是那个样子的。”

所以它是有矛盾的。

你把所有这些矛盾汇总在一起，你会发现这里面有一些事情是我们可以发现不合理的地方，甚至背后会有一些秘密。

6.3 秦始皇的配偶之谜

比如说每一个君主应该是有配偶的，这个是没毛病的常识，但秦始皇没有配偶，我们就发现这里有一个非常反常的事情。

那背后到底有什么原因呢？

他为什么没有配偶？

我们要继续再推，就是他的后族集团、后宫里面到底有来自哪个国家的女人，以及根据王族的这些事迹来推断他们跟哪个国家是有关系的。

因为我们至少有一个常识，就是王族同姓不婚。

所以说秦始皇的配偶一定来自于异姓诸侯国，那我们就可以推断反正只有6个选择，到底是哪一个，这就是非常有趣的一些推理工作。

七、跨领域应用：方法论的迁移

7.1 金融领域的应用

而且这一整套 Skill 其实是可以用在其他的场景上面的，虽然我只是在《史记》上做，但是在我们的群里面有非常多的网友们拿它来处理《唐诗》、处理《资治通鉴》，还有人拿它处理手册，因为它是一套跨领域的技能，尤其是这中间的语言技能，所以你自己拿出来就可以用。

我们也实验了，我这套方法论用在我们金融语料上面，调4大类参数很快就能跑出来，所以我们也把这套东西放在我们自己的产品里面了。

这是我们金融公告抽取的 Skill，这是其中的一个小应用。

其中一个产品正在做异常波动逻辑分析，这也是它自动做的。

7.2 其他领域的探索

其他应用场景包括：

《唐诗》处理

《资治通鉴》知识化

技术手册结构化

财务报表审核

科创板审核应用

所有的资源都在这一页上面，下面有网址大家可以自己去看。

八、总结与展望

虽然这是一个非常微小的工作，但我相信现在这种可能性，我们有一个坚定的信心，就是非常大规模的知识图谱的构造，是不用人写代码就可以完成的。

好东西都是总结出来的。

[回到主页](#)