DataFun Summit 2022
知识图谱在线峰会

知识图谱与金融科技论坛

# 老司机谈NLP半自动驾驶

## 从领域落地看
## 深层解析的符号模型
## 与深度学习的预训练模型

NLP Architect

李维（liweinlp.com）

03.12.2022

1. NLP历史和现状

2. 殊途同归的符号与神经

3. 低代码是趋势，也是王道

4. NLP “半自动驾驶”

# 1. NLP历史和现状

# AI 历史上的NLP两条路线

## 符号规则系统：

早已退出学界主流舞台
从来没有退出工业应用

## 机器学习模型：

传统机器学习
多层神经网络革命

# NLP 近代史

## 多层神经一声炮响：DEEP LEARNING

给NLP送来了监督学习的杀手级武器：典型案例 神经翻译

席卷AI监督学习全栈，从感知到认知：图像，语音，NLP

## 深度解析创新：DEEP PARSING

黑暗中摸索，虽九死其犹未绝

鬼子的进村，打枪的不要

# NLP 之痛：领域落地的知识瓶颈

事实：领域场景不缺原生数据，但常常没有带标数据

例如，金融知识图谱的自动构建

监督学习的知识瓶颈：

手工标注数据：大量低级劳动

符号模型的知识瓶颈：

手工规则代码：少量高级劳动

# NLP 现状：突破瓶颈的曙光

## 自学习的预训练模型：BERT / GPT3 / ...

从语言学习语言：LANGUAGE MODEL
预训练支持下游NLP落地
研究界热火朝天，尚未规模化领域落地


## 深度解析的符号模型：

以不变应万变的结构：结构主义是人类认知之花
下游NLP任务的低代码半自动化
冷启动道路已经打通，无论落地多语言还是多领域

# DEEP PARSING 是符号NLP应用的核武器

**Apple released iPhone 2.0. → logical form: release (Apple, iPhone 2.0)**

- **Passive voice:** iPhone 2.0 **was released** by Apple yesterday
- **Deverbal noun:** **The release** of iPhone 2.0 by Apple yesterday (was an instant success.
- **Conjoined patterns**
  - Apple **developed and released** iPhone 2.0 …
  - The release of **the new iPod series as well as the recent version of iPhone 2.0** (was well received).
- **Hidden Subjects/objects**
  - **Apple** was reported to have released iPhone 2.0 yesterday.
  - Released yesterday by Apple, **the iPhone 2.0** (was an instant success).
  - Having released iPhone 2.0, **Apple** (announced its plan for video game market).
- **Relative clause and emphatic pattern**
  - Apple designed iPhone 2.0 **which** was only to be developed and released an entire year later.
  - iPhone 2.0 **whose** release was announced yesterday
  - It is iPhone 2.0 that was released yesterday
- **Coreference and alias**
  - iPhone 2.0: **its** release was announced yesterday
  - iPhone 2.0 was the product that was released yesterday by Apple
- **Negation and factness**
  - * Apple **failed to release** iPhone 2.0 as planned.
  - *** If** Apple had released iPhone 2.0 a year before …

1. NLP历史和现状

2. 殊途同归的符号与神经

# 剪不断理还乱，符号、神经

**思考题**：上帝用符号还是向量？外星人呢？

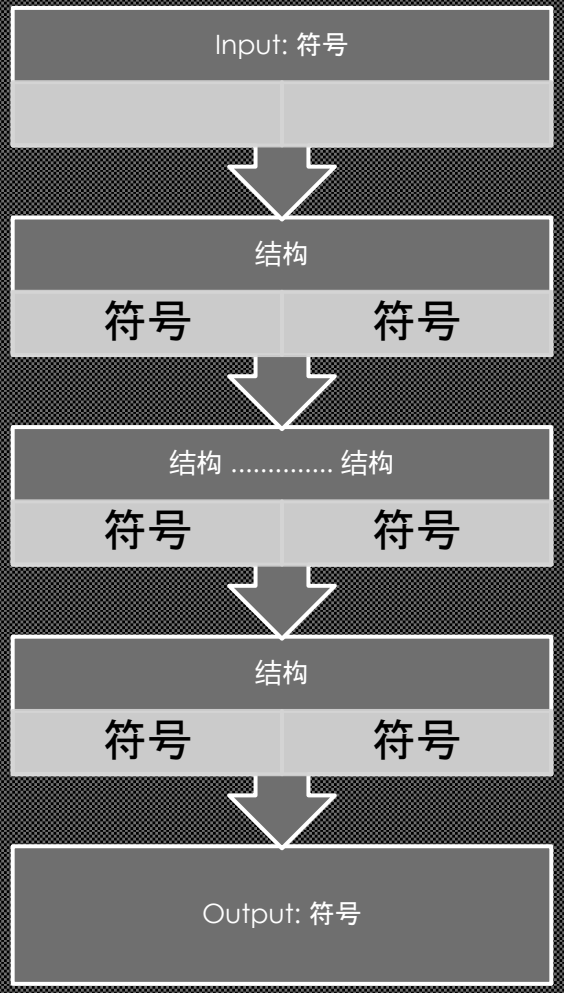
## 符号：人类智能的载体，文明的结晶

形式与意义的结合体（文本、语音、程序、词典...）
符号主义把NLP符号逻辑贯彻到底：所有模块符号化表示
透明化程序化，可解释性与定点纠错的保障

## 神经模型：符号不耐症

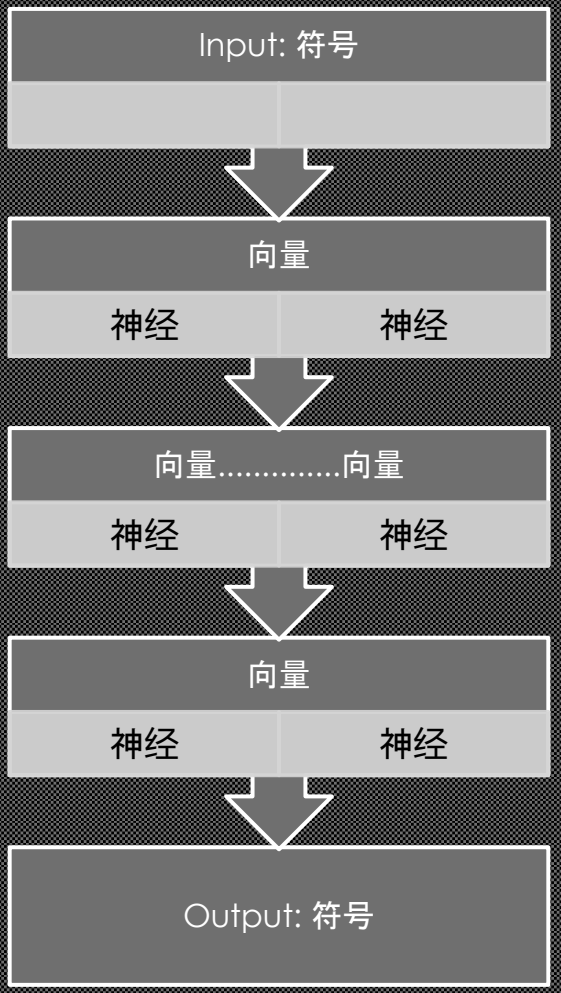
向量化，发挥可计算性：例如，词嵌入
预训练本质是深度的词嵌入，以多层向量捕捉隐藏结构

# 符号模型：结构表示




用符号结构表示非结构化符号（语言）

# 神经模型：向量空间




所谓端到端（end to end）：输入输出端符号，但满肚子向量

# 殊途同归：无论架构还是方法论

理性主义拥抱经验主义
**巧合还是天意？**

## 架构：都是多层/深层模型

我们的英文 PARSER 设计为 50层 左右
中文复杂一些，PARSER 有 100 层

自底而上，由浅入深，层层归约
前馈PIPELINE数据流，与多层神经模型本质一致
（同一个内部数据结构 TOKENLIST，涵盖线性结构与图结构）
**都是深层表示赋能下游NLP**

## 方法论：都是数据驱动（包括目标驱动、错误驱动）

## 目的：都是规模化落地应用，克服知识瓶颈

# 神经与符号各自的短板

## 神经系统

对标注数据的依赖

**出路**：自学习预训练！

对人不友好

**黑箱子、灰箱子**？可解释性差

雾里看花：隐藏层里面的秘密

隔靴搔痒：难以定点纠错（工程部署上有致命性）

## 符号系统

编码门槛高

**出路**：低代码、冷启动、半自动、流程化！

算法通用性差

符号NLP创新难以推广到语音、图像等，但不必了

# 多层符号模型对机器学习的补足

## 监督学习搞不定冷启动（但预训练初见成效）

多数场景缺乏标注，甚至 SPECS 也常是一个 MOVING TARGET
数据标注无论自产还是外包，质量控制是一个挑战
关系和事件标注，难度远远大于字段抽取标注

## 学习系统缺乏定点纠错能力

符号系统的多半错误可以在用户词典作定点纠错
规则定点纠错可以 PINPOINT 到问题所在

## 学习系统本性上缺乏可解释性

向量对人是天书
系统的不透明不利于取信于用户

# NLP 最激荡人心的今后 5-10年

钟摆还能摆得多高？

## 天问：神经可以终结符号吗？ 并非绝无可能，但值得谨慎怀疑。

先例：神经网络革命前，从未预计到神经翻译对规则翻译的终结。

如果预训练赋能NLP规模化领域落地也达到神经翻译的程度，那才真叫：

不废江河万古流

## 更大的两种可能性：

一、 花开两朵，各表一枝

二、 紧耦合范式转变：打开通用智能（AGI）之门？

1. NLP历史和现状

2. 殊途同归的符号与神经

3. 低代码是趋势，也是王道

# NLP 低代码潮流

## AI 开源平台的兴起

深度学习平台: TENSORFLOW / PYTORCH / KERAS / ……

NL P工具箱：NLTK / SCIKIT-LEARN / SPACY, / ……

无论传统的统计模型还是神经网络：都可以低代码搭建

越来越多的现存开源模型与数据资源可供调用

## 符号NLP多层平台 ……

半自动、冷启动、低代码、流程化

# 数据科学与工程的兴起

## 数据为王，数据工作者身处王室

哪怕是王室的一个园丁、管道工 ……

标志着 AI/NLP 从学术象牙塔走向各行各业

## 新生的所谓 DATA SCIENCE 学位从来没有如此火热

各大学、培训社区纷纷推出各种 PROGRAMS
批量化人才集训，火热直追 CS
课程设置与 CS 有很多交叉，但更“普、杂”
目标市场：知识工程师

1. NLP历史和现状

2. 殊途同归的符号与神经

3. 低代码是趋势，也是王道

4. NLP “半自动驾驶”

# 半自动符号NLP的设计哲学

## 符号NLP低代码的开发理念

全自动生成规则：SEED-BASED RULE GENERATION
半自动规则泛化：内嵌泛化路径，交互式开发环境
上下文泛化：结构泛化；窗口泛化（PROXIMITY）
词节点泛化（ONTOLOGY赋能，包括常识）

QA（质量控制）半自动：精度靠比对，召回靠统计
角色转变：码农转型为判官 (JUDGE NOT CODER!)
语言工程观：FOLLOW SOFTWARE ENGINEERING BEST PRACTICE
冷启动，**快速领域化**：**核心引擎赋能**，领域词典自学习，低代码落地

## 比较学习系统的训练

不依赖标注数据：开发集用非标注的原生数据，而不是标注训练集
泛化过程半自动：HUMAN IN THE LOOP 迭代式开发， 而不是梯度下降全自动

# 半自动符号NLP的实践

## 多语言、多领域落地

欧洲与亚洲主要语言

领域/场景涉及 金融、法律、电力、航空、水利、智能助理、客服、社会媒体

## 低代码的效能

以一当十：开发周期从月变为日

# 半自动符号NLP的实践

金融（案例采自将要出版的NLP新书）：句子长，关系错综复杂，但模式固定

截至2020年末，发行人资产总额1,750,604.68万元，负债总额669,613.94万元，所有者权益总额1,180,990.74万元，其中归属于母公司所有者权益总额1,080,033.73万元，资产负债率为34.51%。2020年度实现营业收入71,885.19万元，净利润7,330.27万元。2018年、2019年、2020年公司净利润分别为1.26亿元、0.69亿元和0.75亿元，2018年、2019年、2020年三年平均净利润为0.79亿元，公司盈利能力良好，具有较强的到期偿债能力。

截至2019年末、2020年末和2021年3月末，发行人资产总额分别为5,099,769.83万元、5,616,855.34万元和5,792,287.88万元，2019年、2020年、2021年内较为稳定。其中，流动资产金额分别为2,189,288.19万元、2,463,870.12万元和2,602,792.73万元，占总资产的比例分别为42.03%、43.77%和44.64%。2021年3月末，发行人流动资产较2020年末增加138,932.62万元，增幅5.84%，主要系存货、预付款项及其他应收款大幅增加所致。非流动资产金额分别为2,910,573.66万元、3,152,980.22万元和3,189,545.12万元，占总资产的比例分别为57.17%、57.13%和53.06%。2021年3月末，发行人非流动资产较2020年末增加36,515.90万元，增幅1.26%，主要系在建工程规模增加所致。

# 领域关系抽取

NLP核心引擎DEEP PARSER保持不变
领域化主要是领域词汇工作

从领域原始数据做新词发现：
聚类统计学到的领域近4万金融领域词库
用户词典：例如一级科目、二级科目词条

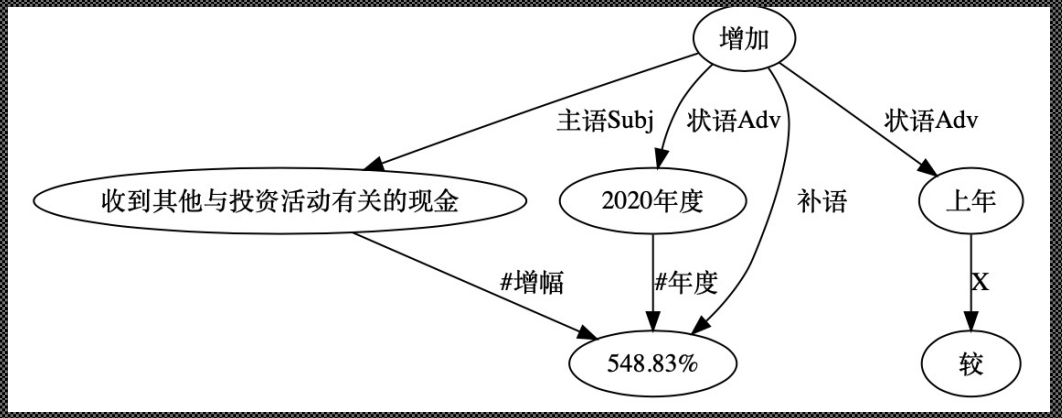
APPROACH：基于深度解析/PROXIMITY 的多层模式匹配

BALANCING ART IN RULE CODING：

结构解析赋能精准，PROXIMITY 增强鲁棒

# 关系抽取（例示）

Input：2020年度收到其他与投资活动有关的现金较上年增加548.83%




Output： 2020年度 #年度 = 548.83%

收到其他与投资活动有关的现金 #增幅 = 548.83%

```
E year %date% !comparee]~~[^Q.Link1]~~[^Q.Link1 DE percent|money:^0.#年度]
    ◦ 2020年度收到其他与投资活动有关的现金较上年增加548.83%
```

# 信息点多，关系错综复杂

Input:

2017-2018年末及2019年9月末，发行人流动资产分别为760,989.15万元、839,675.01万元和845,622.64万元，占资产总额的比重分别为99.19%、99.29%和99.32%；非流动资产分别为6,229.47万元、5,987.29万元和5,809.79万元，占资产总额的比重分别为0.81%、0.71%和0.68%。

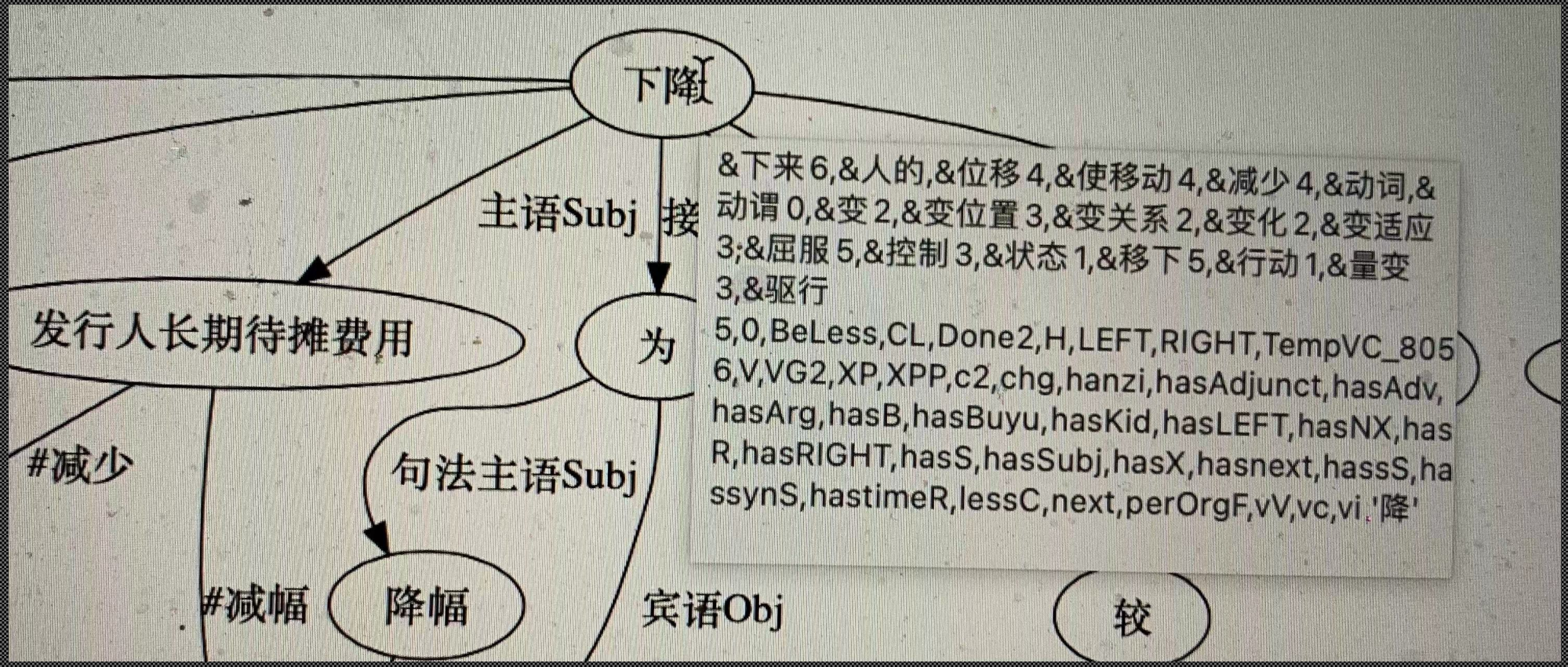
Output:（19个信息点，30个二元关系需要抽取）

2017 #年度 = 760,989.15万元 / 2017 #年度 = 6,229.47万元 / 2017 #年度 = 99.19% / 2017 #年度 = 0.81%
2018年末 #年度 = 839,675.01万元 / 2018年末 #年度 = 5,987.29万元
2018年末 #年度 = 99.29% / 2018年末 #年度 = 0.71%
2019年9月末 #年度 = 845,622.64万元 / 2019年9月末 #年度 = 5,809.79万元
2019年9月末 #年度 = 99.32% / 2019年9月末 #年度 = 0.68%
资产总额 #所占比例 = 0.81% / 资产总额 #所占比例 = 0.71% / 资产总额 #所占比例 = 0.68%
非流动资产 #为 = 6,229.47万元 / 非流动资产 #为 = 5,987.29万元 / 非流动资产 #为 = 5,809.79万元
非流动资产 #占比 = 0.81% / 非流动资产 #占比 = 0.71% / 非流动资产 #占比 = 0.68%
资产总额 #所占比例 = 99.19% / 资产总额 #所占比例 = 99.29% / 资产总额 #所占比例 = 99.32%
发行人流动资产 #为 = 760,989.15万元 / 发行人流动资产 #为 = 839,675.01万元 / 发行人流动资产 #为 = 845,622.64万元
发行人流动资产 #占比 = 99.19% / 发行人流动资产 #占比 = 99.29% / 发行人流动资产 #占比 = 99.32%

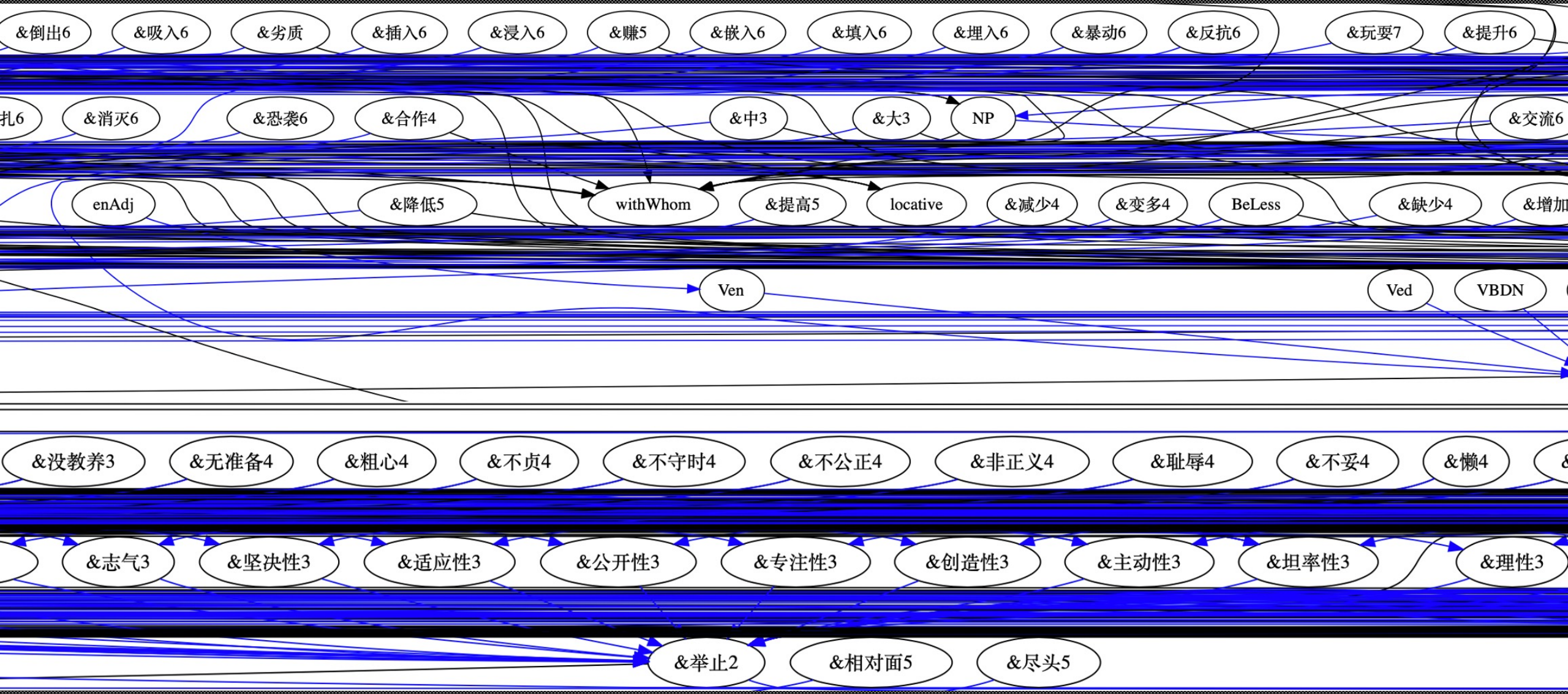
# 词节点泛化路径

## 泛化特征链条用数字后缀，方便规则调试：

下降：　&下来6 → &移下5 → &减少4 → &量变3 → &变2 → &行动1→&动谓0
&位移4 → &变位置3 → &变化2 → &状态1→ &动谓0



# HOWNET 本体知识库提供词泛化的层级支持



# 两类上下文约束条件的泛化

句法 Link-n 和窗口 Win-n

为规则上下文条件半自动泛化提供了提示的条件
最终是数据驱动，根据开发集中的表现来微调这些条件

[item]~~[Link1 为|合计]~~[Win5 占]~~[Win2 item]~~[Win9: #0.占比 #3.所占比例]

例如：X 占 Y percent

公司流动资产分别为3,962,405.59万元、3,871,949.11万元和3,507,986.88万元，分别占当期资产总额的48.96%、46.36%和43.77%。

[item]~~[~Subj 占]~~[Win2 item]~~[Link1 比重|比例]~~[Link2 percent: #0.占比 #2.所占比例]

例如： X占Y比重为percent

发行人流动资产（Subj）占总资产的比重分别为99.19%、99.29%和99.32%

# 泛化与约束

泛化能力加强召回（概括正例），约束条件加强精准（排除反例）：

正例：

发行人流动资产占总资产的比重分别为99.19%、99.29%和99.32%
发行人流动资产总额2017年末占总资产的比例为99.19%
发行人流动资产总额2017年末约占总资产比例99.19%
发行人流动资产不像~~固定资产~~，占总资产比重高达99.19%
与~~固定资产~~不同，发行人流动资产占总资产比重高达99.19%
……

反例：

~~固定资产~~部分出售后，占总资产比重达99.19%的是发行人流动资产 （“固定资产”不是主语）
发行人流动资产占总资产的比重持续下降，占比再也没有恢复到90%以上 （“比重”与90%没有直接或间接关系：Link2）
……

# 小结：主要优点

不依赖标注数据，冷启动低代码开发

- 弥补监督学习不足：监督学习无法做“无米之炊”
- 适用于缺乏标注的领域场景：“无米可以，有稻就成”

普适性、跨领域、跨语言平台

基于结构和理解（包括本体常识）

- 可以领域化到任何领域场景的文本数据

无论抽取字段、关系或事件，具有自带的泛化能力

- 词汇特征泛化、多层结构泛化/窗口泛化

# 小结：主要缺点

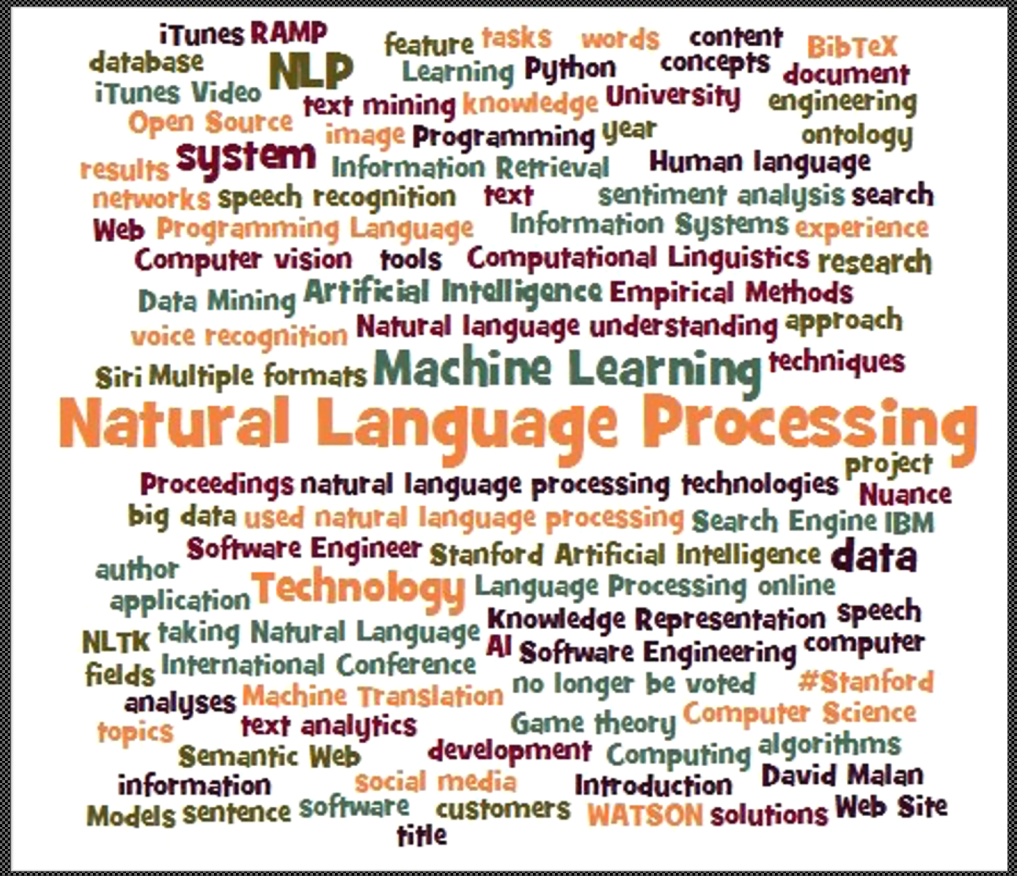
需要某些代码技能

低代码，但不是无代码

- 解析引擎代码可以复用（80+%），
- 直接针对项目需求的代码需要半自动调试（20-%）

开发效率决定于培训和经验

- 一旦上路，开发效率会不断提高


NLP 小书（广告？）

1. 李维 郭进《自然语言处理答问》
   （商务印书馆 2020）

2. 预告：李维《巴别塔影：符号自然语言处理之旅》
   （人民邮电出版社 2022）

3. 预告：李维等 《知识图谱：演进、技术和实践》
   （机械工业出版社 2022）

liweinlp.com (立委NLP频道)
liweinlp@gmail.com