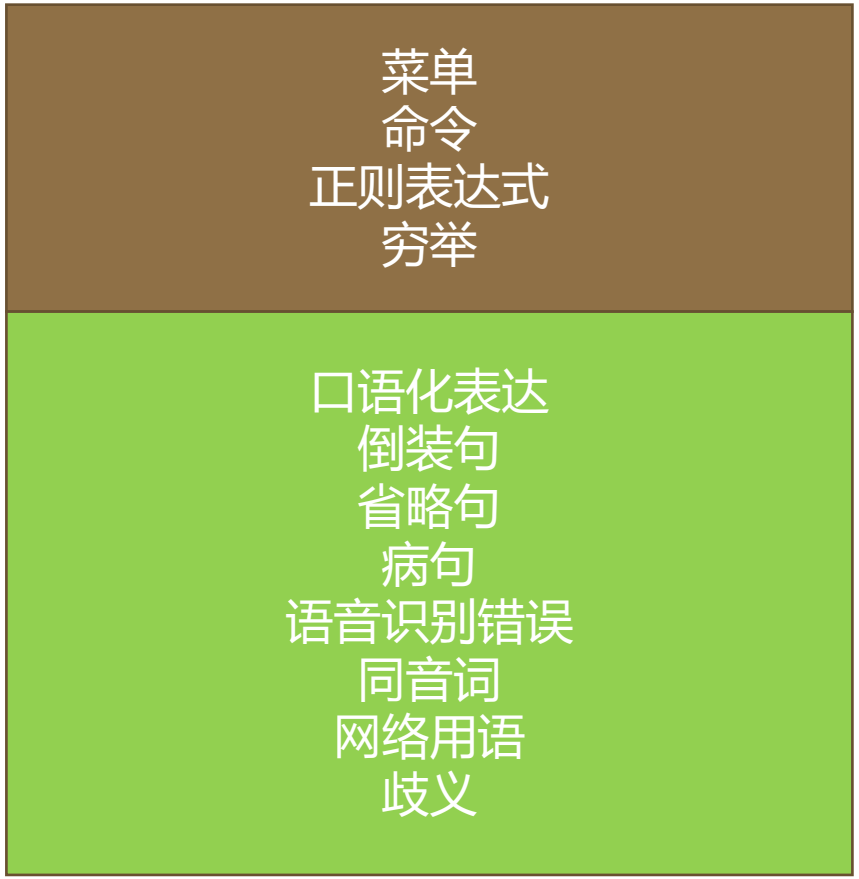
怎么让机器理解自然语言

氖星智能 彭军辉

什么是自然语言？



×这些都不是

√这些都是

一组自然语言：
苹果是什么垃圾
你说苹果是啥垃圾
橘子呢
平果是啥垃圾
平锅是啥垃圾
品国是啥垃圾
小苹果是啥垃圾
诺基亚手机是啥垃圾
苹果呢

自然语言的特点

不稳定

“中国足球谁都打不过。”是褒义还是贬义呢？
“吃了吗？”是吃药了吗？还是吃饭了吗？
同一句话一个字没变，不同情况下语义可能变了。

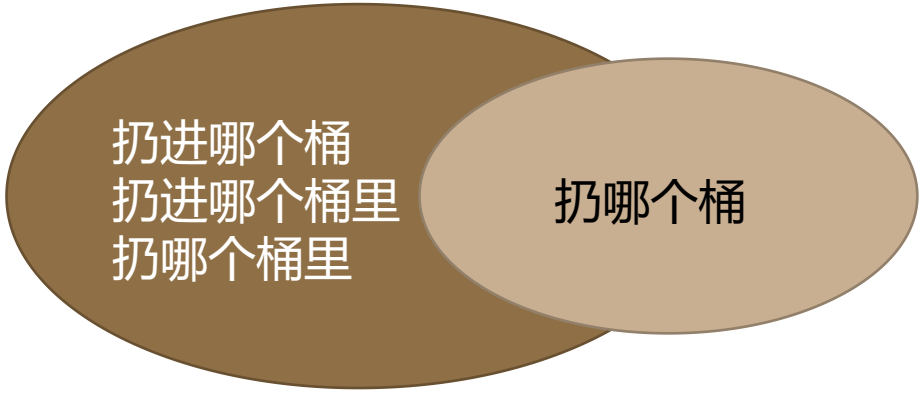
不规范

“他丢了狗”是狗丢了，不是他丢了。
“吃了吗你”是“你吃了吗”，不是“你被吃了吗”。
“我稀饭你”跟稀饭没关系。
“我对这个不感冒”，我没感冒。
语言是灵活的，语法不是生搬硬套的规则。

不明确

“苹果是什么垃圾？”“木头？”
”苹果能吃吗?”“木头呢？”
两个“木头呢”语义不一样。省略掉上一句话，
两个“木头呢？”就无法理解了。

自然语言的本质是语义。文字是表达形式，不是语义理解的处理对象。

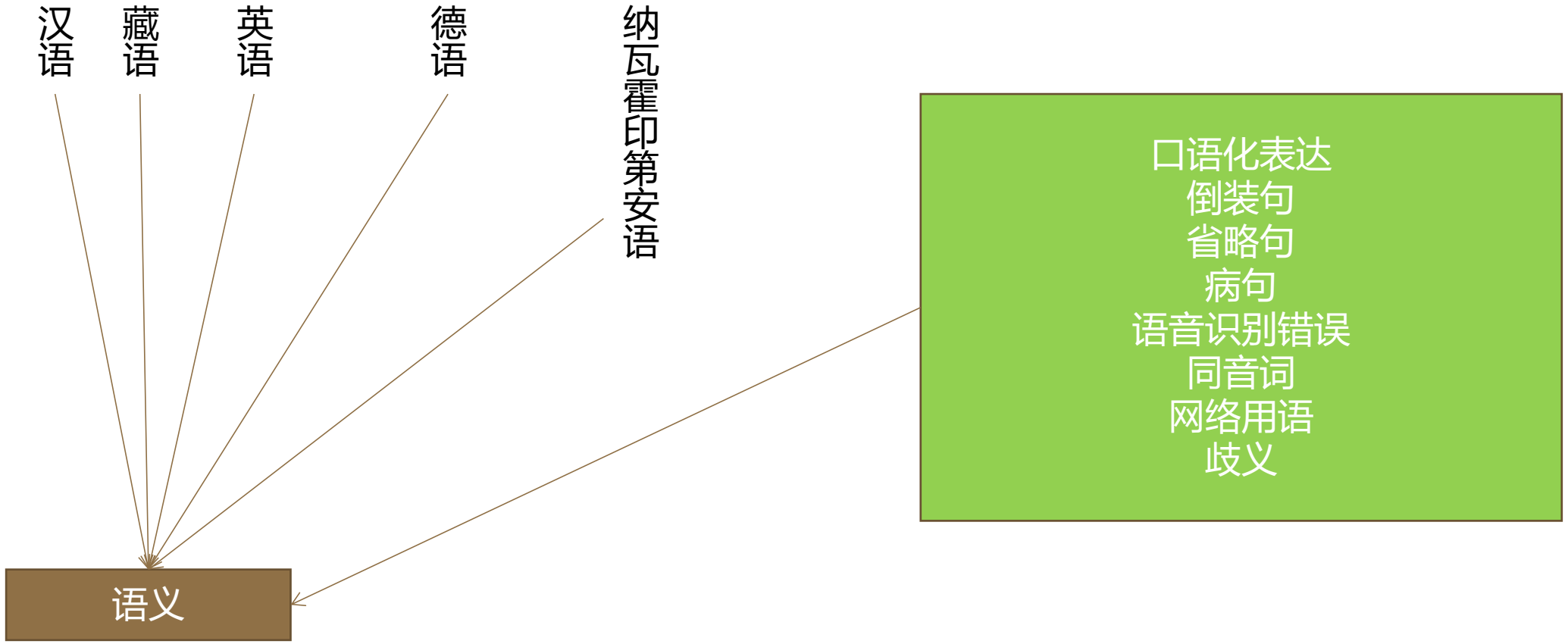


“扔哪个桶”和其他三个句子有时候语义是相似的，它是对其他三个句子的口语化表达；“桶”不是扔的宾语，要扔掉的是其他东西。

有时候它和其他三句没有语义相似性，“桶”是扔的宾语，是要被扔掉的东西。

如果仅仅处理文字，不会发现“扔哪个桶”这句话有歧义。

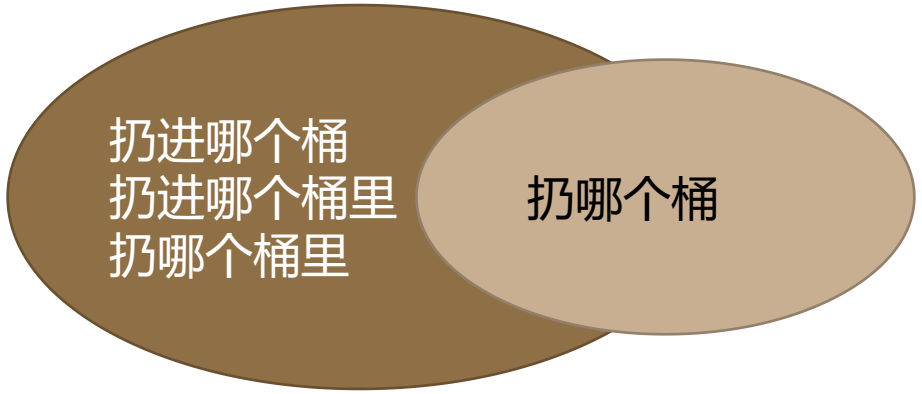
透过文字（还有语音）处理语义就是语义计算

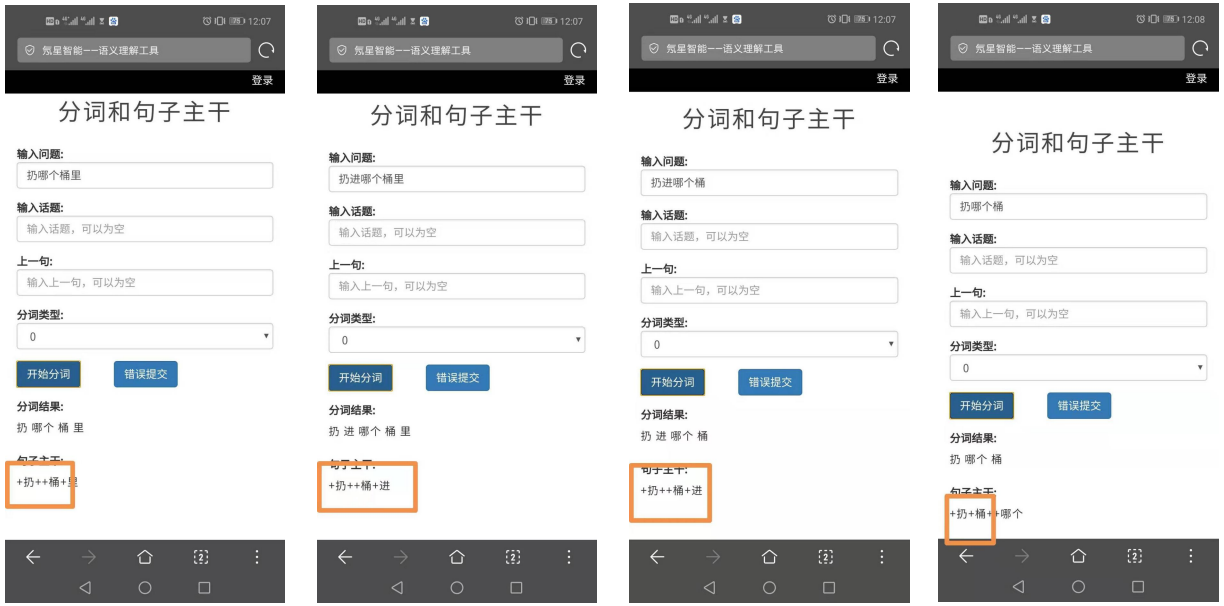


因为语义相同，我们能处理各种语言现象。一个语义的各种不同表达，在知识库里只存一条知识。
因为语义相同，我们能把一种语言翻译成另一种语言。

怎么表示语义？怎样计算语义相似性？

- 语义=主语+谓语+宾语+其他+其他。主谓宾就是句子主干。
- 口语表达、倒装句等语言现象我们计算的句子主干相同。
- 我们使用句子补全技术，对省略句的主干进行补全。
- 主谓宾相同，语义有相似性。我们通过计算两个句子的主干相似性来计算句子语义相似性。





“扔哪个桶”和其他三个句子的主谓宾就是不一样的。

我们突破了关键技术——句子主干提取

我们使用语言学方法，能从各种不同句型各种语法现象中准确提取句子主干。

如果句子主干是个模型，我们的模型是事先设置好了的。

原句	我们提取的句子主干	别家提取的关键字
猎人的狗咬死了它	狗+咬+它+猎人+死	猎人+咬+狗+死+它
猎人的狗咬了它的	狗+咬++猎人+它	猎人+咬+狗+它+了
猎人的狗咬了它	狗+咬+它+猎人+了	猎人+咬+狗+它+了
是他的狗咬死了猎人	狗+咬+猎人+他+死	猎人+咬+狗+死+他
他的狗咬死了猎人	狗+咬+猎人+他+死	猎人+咬+狗+死+他
是它咬死了猎人的狗	它+咬+狗+猎人+死	猎人+咬+狗+死+它
它咬死了猎人的狗	它+咬+狗+猎人+死	猎人+咬+狗+死+它
它是咬死了猎人的狗	它++狗+猎人+死	猎人+咬+狗+死+它
咬死了猎人的狗是它的	狗+++猎人+咬	猎人+咬+狗+死+它
咬死了猎人的狗是它	狗++它+猎人+咬	猎人+咬+狗+死+它

怎样计算句子主干？我们的核心价值是什么？

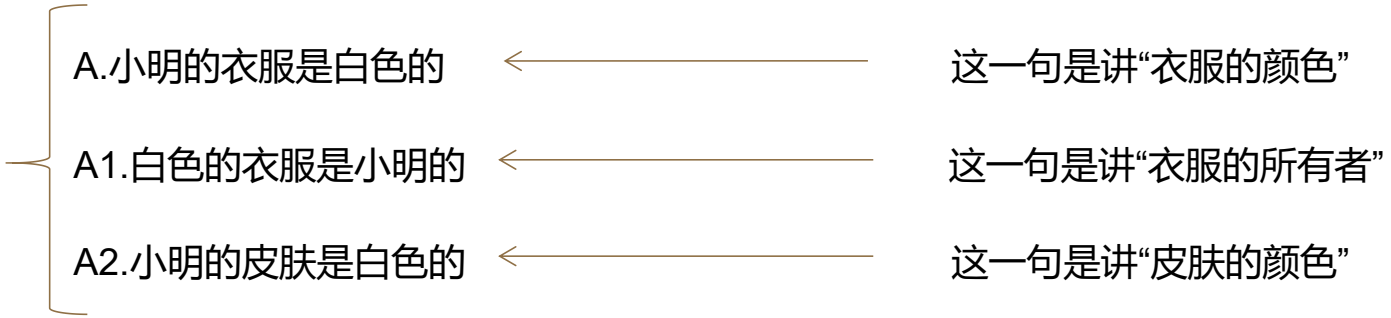
我们根据句型（句子结构）计算句子主干。
句型经常是由谓语动词决定的。（当然，还有其他的规律，都是我们在实践中总结出来的。这里不便赘述。）
比如“吃”“牛”“青草”，组成的句子只能是“牛吃青草”，因为，“吃”做谓语动词时，“青草”是不能做它的主语的，“牛”可以。
所以，计算出谓语动词就能计算句子的结构，从而计算出句子主干。

通过提取句子主干，把自然语言这种非结构化数据变成了结构化数据，是我们的核心价值。

搜索技术做不了语义理解，做不了机器人大脑。

- 搜索和问答是NLP（自然语言处理）的两个不同领域，它们有着巨大差别。
- 很多公司用搜索技术做问答，甚至用考核搜索的指标考核问题。
- 搜索处理的不是语义，只是关键字，是符号。问答处理语义，句子主干代表语义。主干相近，语义相近。

差异点	搜索	问答
关键技术	关键字相关性查询	语义相似性计算
输出结果	结果列表	唯一答案
关键指标	召回率和准确率	差异性、同一性、模糊性、一致性
应用方向	搜索引擎、大数据	机器人



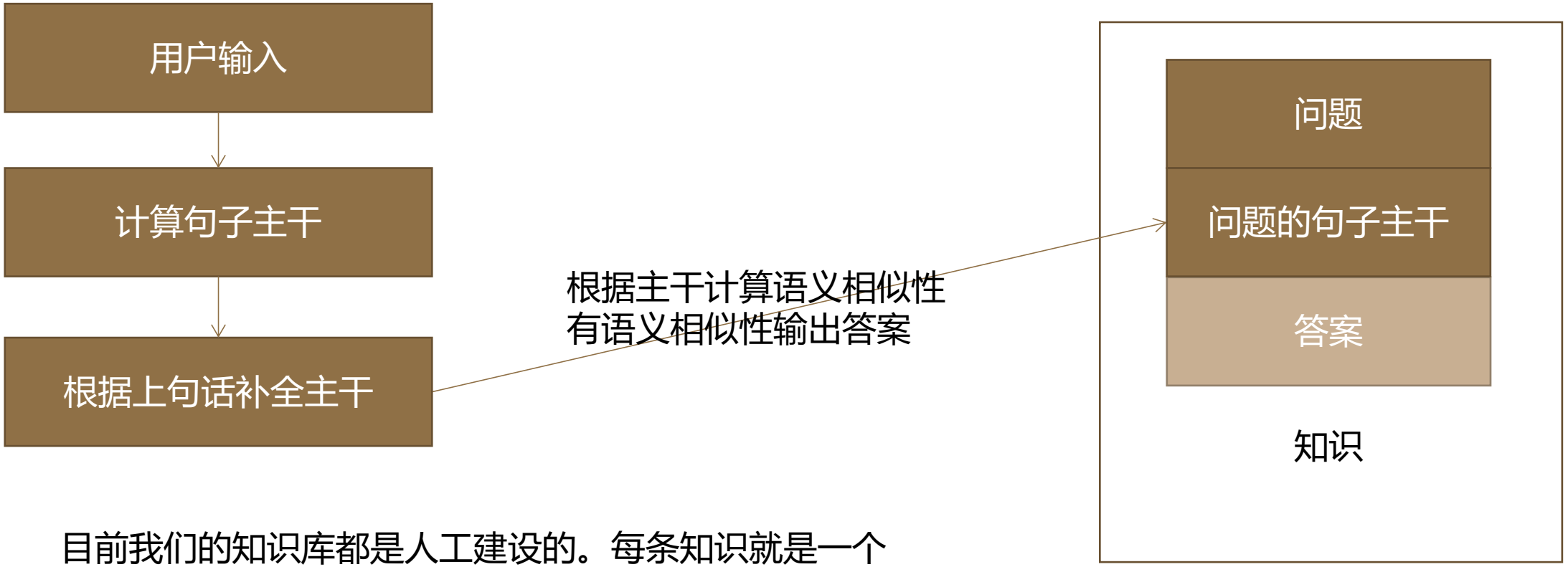
在搜索技术看来，A1和A2与A都是高度相关的。
而在问答技术看来，A1和A语义更接近，都是讲衣服的。而A2和A则完全没有语义相似性。

我们的技术标准


它们不一样
咬人的是他
咬人的是他的
同时能理解语义的差异性和同一性才是语义理解。
差异原则
分析语义细微差异
同一原则
处理一个语义的多种表达
它们都一样
你们公司在哪里
贵公司在啥地方
你公司在哪里
模糊原则
容忍用户输入的小错误
一致原则
说话不前后矛盾
前进
千金
前景
它们都是前进
我是男的
我不是女的
我性别是男
它知道自己在说啥

怎么构建知识库

- 知识=问题+答案
- 计算用户输入句子主干和知识中的问题的句子主干语义相似性。句子主干是格式化数据。



目前我们的知识库都是人工建设的。每条知识就是一个QA。以后计划引入机器学习构建知识库。

谢谢观看！