自然语言理解

(1)

宗成庆

中科院自动化研究所

模式识别国家重点实验室

cqzong@nlpr.ia.ac.cn

http://www.nlpr.ia.ac.cn/English/cip/cqzong.htm

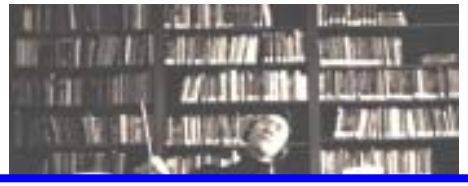
第一章 引言

1.1 基本概念

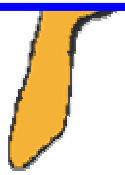
当我们从事任何一项研究的时候，总要关注两方面的问题：一是是什么，为什么？二是做什么，怎么做？这恰恰是科学与技术紧密相关的两个方面。

自然语言处理既是一项技术，又是一门学科。

1.1 基本概念



无处不在的文字，无处不在的声音，无处不在的网络，无处不在的通讯；如此繁多的文档，如此繁多的电话，如此繁多的邮件，让我哪里能够应付的了？



1.1 基本概念

信息的主要载体－语言

语言的两种形式－文字和声音

文字和声音作为语言的两个不同形式的载体，所承载的信息占整个信息组成的70%以上（文字：70%，图象：20%；其它：10%）

- 如何让计算机实现人们希望实现的语言处理功能？
- 如何让计算机真正实现海量的语言信息的自动处理和有效利用？

1.1 基本概念

- 基本定义－自然语言处理

自然语言处理（Natural Language Processing，简称 NLP）就是利用计算机为工具对人类特有的书面形式和口头形式的自然语言的信息进行各种类型处理和加工的技术。

-冯志伟《自然语言的计算机处理》

1.1 基本概念

- 其它名称
  - 自然语言理解（Natural Language Understanding）
  - 计算语言学（Computational Linguistics）

计算语言学是利用电子数字计算机进行的语言分析。虽然许多其它类型的语言分析也可以运用计算机，计算分析最常用于处理基本的语言数据－例如建立语音、词、词元素的搭配以及统计它们的频率。

—《大不列颠百科全书》

1.2 可以让计算机为我们做什么？

- 我们遇到的问题－机器翻译

（以下翻译结果来自 Systran：http://www.systransoft.com ）

Ex-1: What questions does the study of language concern itself with? (参考文献[12]，Page 8)

什么问题语言的研究有关本身与?

Ex-2: When our founders boldly declared America's independence to the world and our purposes to the Almighty, they knew that America, to endure, would have to change.（克林顿就职，1993）

当我们的创建者大胆地宣称美国的独立对世界和我们的目的对全能之神, 他们知道, 美国, 忍受, 会必须改变。

1.2 可以让计算机为我们做什么？

Ex-3: The spirit is willing, but the flesh is weak.
（心有余，而力不足。）

精神是愿意的, 但骨肉是微弱的。(Systran)

English-> Russian->English:

The wine is good, but the meat is spoiled.
(酒是好的，肉是馊的。)

1.2 可以让计算机为我们做什么？

Ex-4: Out of sight, out of mind.

眼不见，心不烦。）

出于视域, 在头脑外面。 (Systran)

From English to Russian:

又瞎又疯。

1.2 可以让计算机为我们做什么？

- 我们遇到的问题—信息检索

  http://www.google.com

  - 微软：2,060,000 条

    微软，亚洲研究院：19,100 条

    微软，亚洲研究院，研究方向：4,850 条

    微软，亚洲研究院，自然语言处理：695 条

⇒ 300多亿个网页，每天几百万增加

⇒ 获得的信息只有1%被有效利用

1.2 可以让计算机为我们做什么？

- 我们遇到的问题－语音识别

  输入：美欧贸易摩擦升级

  识别结果：美欧贸易摩擦生机

  输入：新技术的发展日新月异

  识别结果：新纪录的发展日新月异

1.2 可以让计算机为我们做什么？

- 不能想象的同音字识别

  — 施氏食狮史 （赵元任）

石室诗士施氏，嗜狮，誓食十狮。氏时时适市视狮，十时，适十狮适市，是时，适施氏适市，施氏视是十狮，拭矢试，使是十狮逝世，适石室，石室湿，氏使侍拭石室，石室拭，始食是十狮尸，始识是十狮尸，实十石狮尸，试释是事。

1.2 可以让计算机为我们做什么？

- 信息过滤，信息安全
- 文摘生成
- 问答系统，人机交互
- 语言教学
- 文字输入，文字编辑与排版
- 语音翻译
- 网络内容管理与知识发现

… …

- 计算机不能为我们做什么？

1.3 关于“理解”的理解

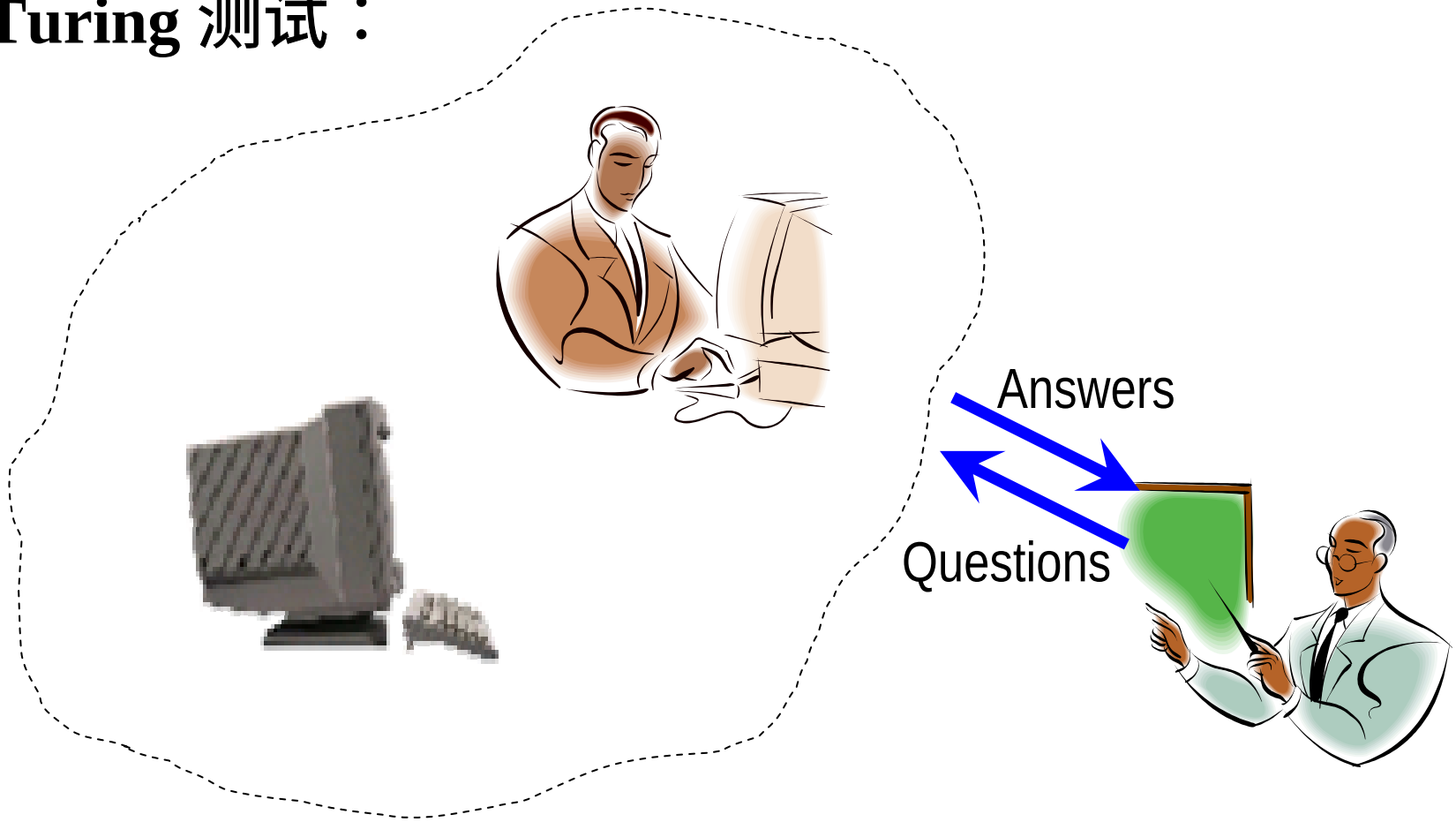
一个幽默片断：

他说：“她这个人真有意思(**funny**)”。她说：“他这个人怪有意思的(funny)”。于是人们以为他们有了意思(**wish**)，并让他向她意思意思(**express**)。他火了：“我根本没有那个意思(**thought**)”！她也生气了：“你们这么说是什么意思(**intention**)”？事后有人说：“真有意思(funny)”。也有人说：“真没意思(**nonsense**)”。

—《生活报》1994. 11. 13. 第六版

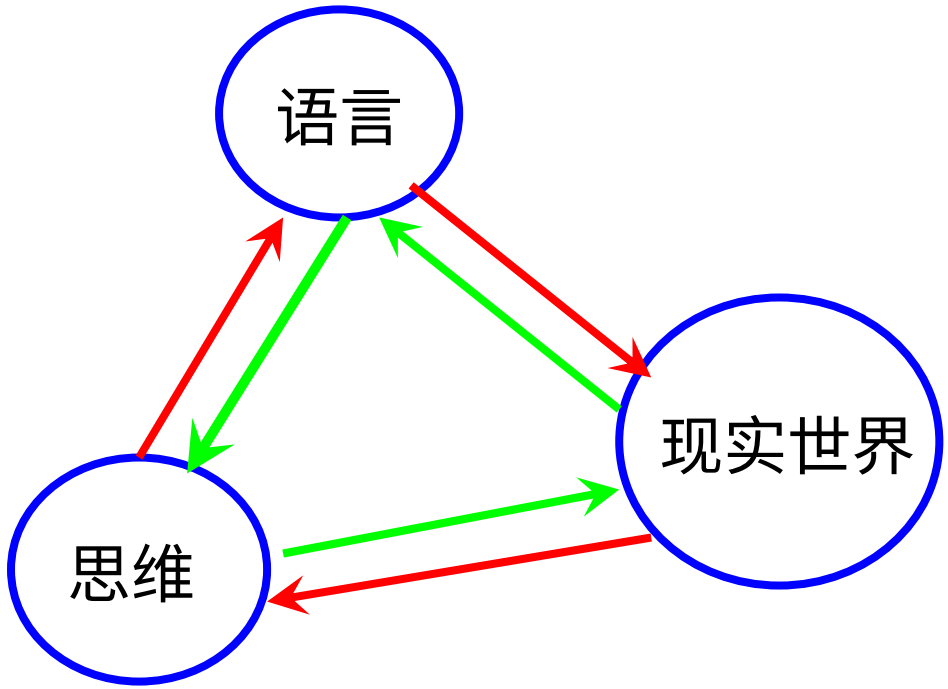
1.3 关于“理解”的理解

Turing 测试：



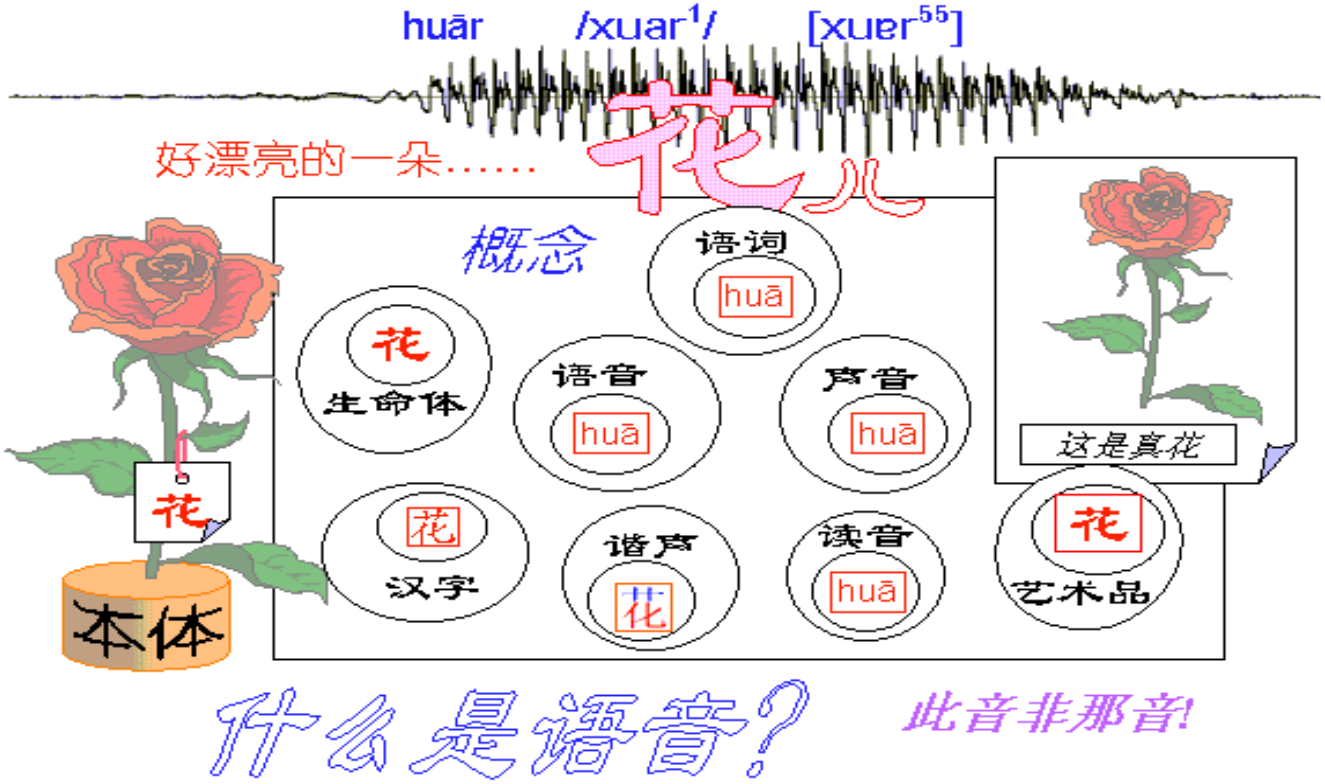
1.3 关于“理解”的理解

- 人脑对语言的理解是一个复杂的思维过程
  - 语言学
  - 语言心理学
  - 逻辑学
  - 计算机科学
  - 人工智能
  - 数学与统计学
  - …… ……



1.4 自然语言理解研究的基本问题

- 研究的层次
  - 语音学（Phonetics）：研究词及其语音的关联。



1.4 自然语言理解研究的基本问题

- 研究的层次

一形态学（Morphology）：研究词是如何由意义的基本单位一词素（morphemes）构成的。

词素（morphemes）→ 词 （word）？

⇓

词根、前缀、后缀、词尾

例： 人，蜈蚣

老虎 ← 老 + 虎；　　图书馆 ← 图 + 书 + 馆

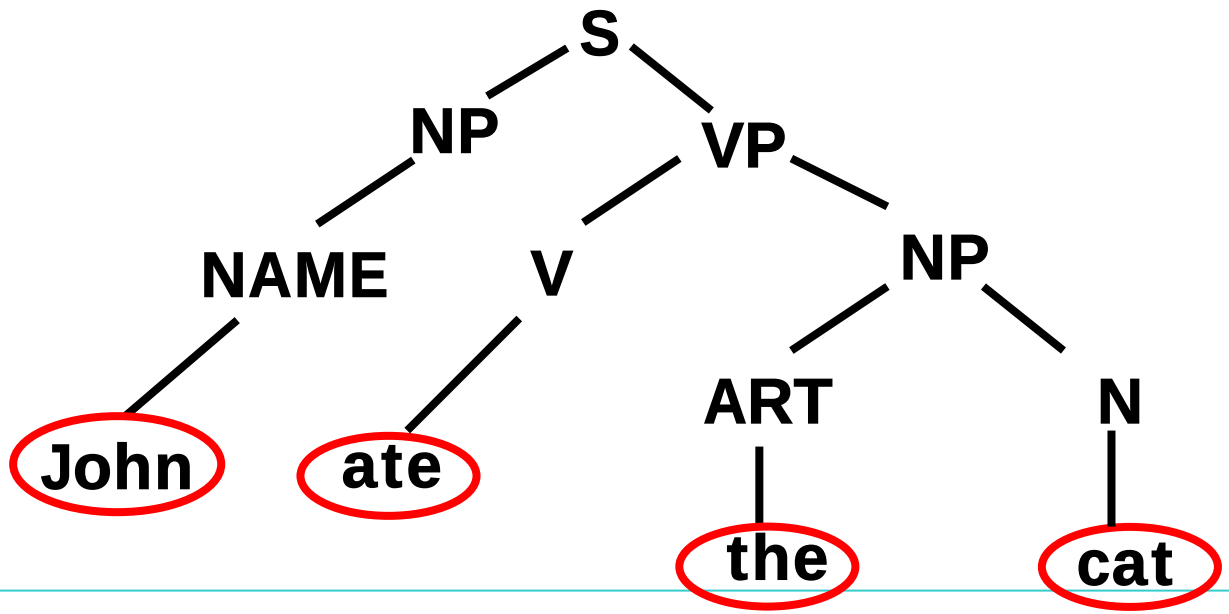
re + ex + port → reexport

1.4 自然语言理解研究的基本问题

- 研究的层次

一语法学（Syntax）：研究语句的组成结构，包括词和短语在语句中的作用等。

为什么一句话可以这么说也可以那么说？



1.4 自然语言理解研究的基本问题

- 研究的层次

—语义学（Semantics）：研究如何从一个语句中词的意义，以及这些词在该语句中句法结构中的作用来推导出该语句的意义。

这句话说了什么？

(1) 今天中午我吃食堂。

(2) 这个人真牛。

(3) 这个人眼下没些什么，那个人嘴不太好。

1.4 自然语言理解研究的基本问题

- 研究的层次

一语用学（Pragmatics）：研究在不同上下文中的语句的应用，以及上下文对语句理解所产生的影响。

为什么要说这句话？

（1）火，火！

（2）A：看看鱼怎么样了？

B：我刚才翻了一下。

1.5 不同语言的差异

❑ 分类

孤立语（分析语）：形态变化少，语法关系靠词序和虚词表示，如：汉语。

曲折语：用词的形态变化表示语法关系，如：英语。

黏着语：词内有专门表示语法意义的附加成分，词根或词干与附加成分的结合不紧密。如：日语。

❑ 基本单位

汉语：汉字（单音节，无空格）

英语：单词（多音节，有空格）

日语：字和词（多音节，无空格）

1.6 自然语言理解研究的基本方法

- 理性主义与经验主义方法的哲学分野

  之一：对语言知识来源的不同认识

理性主义认为：人的很大一部分语言知识是与生俱来的，由遗传决定的。

Chomsky 的内在语言官能(innate language faculty)理论被广泛接受。

人工编汇初始语言知识 ＋ 推理系统 ⇨ 自然语言处理系统

1960s – 1980s中期

1.6 自然语言理解研究的基本方法

经验主义认为：人的语言知识是通过感观输入，经过一些简单的联想(association)与通用化(generalization)的操作而得到的。

大量的语言数据中获得语言的知识结构。

1920s – 1950s，1980s中期-

1.6 自然语言理解研究的基本方法

- 理性主义与经验主义方法的哲学分野

  之二：研究对象的差异

理性主义：研究人的语言知识结构（语言能力，language competence），实际的语言数据（语言行为，language performance）只提供了这种内在知识的间接证据。

经验主义：研究对象直接是这些实际的语言数据。

1.6 自然语言理解研究的基本方法

- 理性主义与经验主义方法的哲学分野

  之三：运用不同的理论

理性主义：通常基于 Chomsky 的语言原则（principles），通过语言所必须遵守的一系列原则来描述语言。

经验主义：通常是基于 Shannon 的信息论。

1.6 自然语言理解研究的基本方法

- 理性主义与经验主义方法的哲学分野

  之四：采用不同的处理方法

理性主义：通常通过一些特殊的语句或语言现象的研究来得到对人的语言能力的认识，而这些语句和语言现象在实际的应用中并不常见。

经验主义：偏重于对大规模语言数据中人们所实际使用的普通语句的统计。

1.6 自然语言理解研究的基本方法

- 理性主义方法与经验主义方法的融合

  符号智能 ＋ 计算智能

  理性主义研究方法 — 符号处理系统

  经验主义研究方法 — 基于语言数据的计算方法

  理性主义与经验主义的合谋 — 融合方法

1.7 自然语言理解研究现状

- 实用和半实用技术已经得到广泛运用
  - 文字处理器
  - 文字输入
  - 网络搜索引擎
  - 辅助翻译、电子词典
  - 语音合成

  …… ……

在一定程度上满足了人们的某些需要，但离真正实用的目标还有相当的距离。

1.7 自然语言理解研究现状

- 计算机对语言理解的能力到底有多大？
  - 计算机存储容量 100G ⇒ 10万本书
  - 计算机速度比人脑快 10^{12} 倍
  - 计算机智能 ⇒ 几岁小孩？

1.8 课程安排

第一章　引论
第二章　数学基础
第三章　形式语言与自动机
第四章　语料库语言学
第五章　概率语法
第六章　词法分析技术
第七章　句法理论与句法分析
第九章　计算语义学
第十章　应用系统介绍－机器翻译、语音翻译、文本分类、信息检索、对话系统等。

1.8 参考文献

❑ 专著

[1] 瓮富良，王野翊，计算语言学导论，中国社会科学出版社，1998。

[2] 刘颖，计算语言学，清华大学出版社，2002。

[3] 冯志伟，自然语言的计算机处理，上海外语教育出版社，1996。

[4] 姚天顺，自然语言理解——种让机器懂得人类语言的研究，清华大学、广西科技出版社，2002（第二版）。

[5] 赵铁军，机器翻译原理，哈尔滨工业大学出版社，2000

[6] 王小捷，常宝宝，自然语言处理基础，北京邮电大学出版社，2002。

1.8 参考文献

❑ 专著

[7] 张维明主编，语义信息模型及应用，电子工业出版社，2002。

[8] 陈小荷，现代汉语自动分析，北京语言文化大学出版社，2000。

[9] 詹卫东，面向中文信息处理的现代汉语短语结构规则，清华大学、关系科技出版社，2000。

[10] 侯敏，计算语言学与汉语自动分析，北京广播学院出版社，1999。

1.8 参考文献

❑ 专著

[11] James Allen, Natural Language Understanding. The Benjamin/Cummings Publishing Company, Inc. 1995.

[12] Christopher D. Manning, Hinrich Schute, Foundations of Statistical Natural Language Processing. The MIT Press. 1999.

[13] Rens Bod, Jennifer Hay et al. Probabilistic Linguistics. The MIT Press. 2003.

1.8 参考文献

- 期刊

1) Computational Linguistics

2) Machine Translation

3) 中文信息学报

4) 计算机学报、软件学报、计算机研究与发展

1.8 参考文献

- 会议论文集

[1] Proceedings of ACL (Annual Meeting of the Association for Computational Linguistics)

[2] Proceedings of COLING (Inter. Conf. on Computational Linguistics)

[3] Proceedings of IJC-NLP (Inter. Conf. on Natural Language Processing)

[4] 全国计算语言学联合学术会议论文集

小结

- 对自然语言理解的基本认识
  - 基本概念
  - 基本任务和目标
  - 研究方法
  - 课程安排和参考文献

习题

1-1. 说明如下句子有多少种不同的含义？

(1) Time flies like an arrow.　(2) He drew one card.

(3) 咬死猎人的狗。

1-2. 试比较汉语和英文句子中地点状语位置的差异。

1-3. 思考一下，如果用计算机编译技术中程序设计语言的某一句法分析方法直接解析普通的英文句子，会存在什么问题？

1-4. 思考一下，你的大脑理解一个英文句子的基本过程。

Thanks

谢谢!