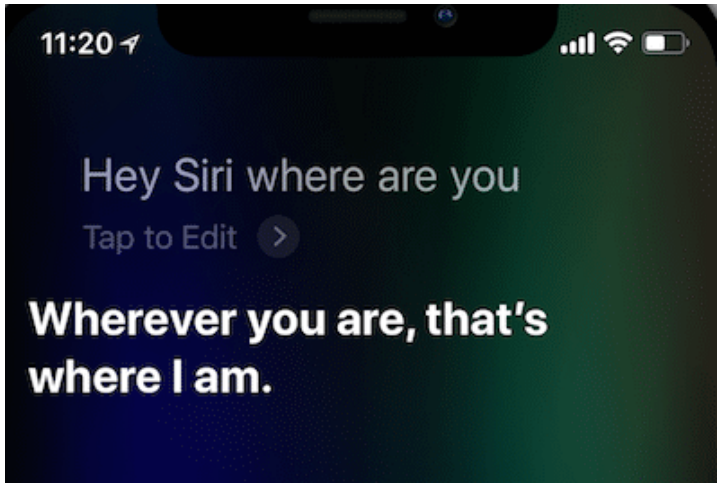
NLP 南京大学
自然语言处理研究组
Natural Language Processing Group

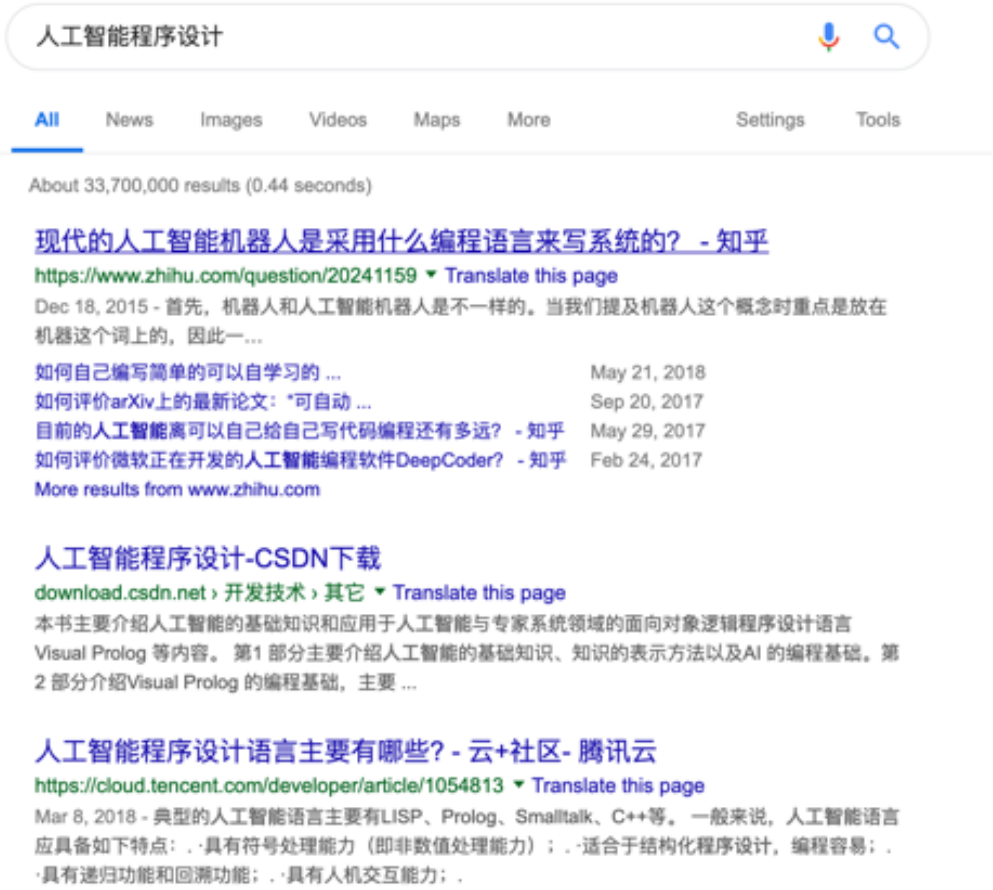
南京大学
NANJING UNIVERSITY

自然语言处理

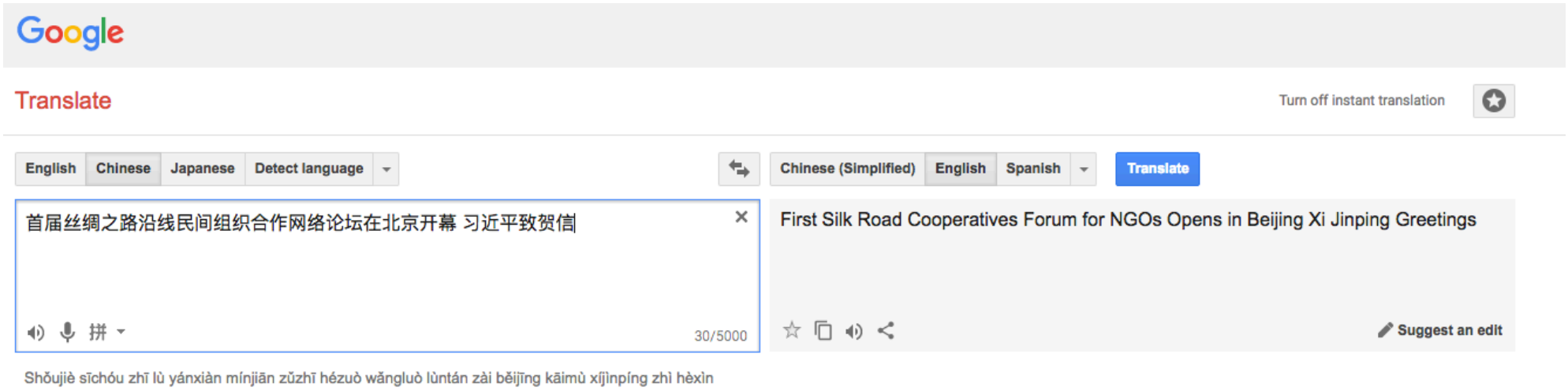
黄书剑

现实世界中的问题：


11:20
Hey Siri where are you
Tap to Edit
Wherever you are, that's where I am.


人工智能程序设计
All
News
Images
Videos
Maps
More
Settings
Tools
About 33,700,000 results (0.44 seconds)
现代的人工智能机器人是采用什么编程语言来写系统的？ - 知乎
https://www.zhihu.com/question/20241159 Translate this page
Dec 18, 2015 - 首先，机器人和人工智能机器人是不一样的。当我们提及机器人这个概念时重点是放在机器这个词上的，因此一...
如何自己编写简单的可以自学习的 ... May 21, 2018
如何评价arXiv上的最新论文：*可自动 ... Sep 20, 2017
目前的人工智能离可以自己给自己写代码编程还有多远？ - 知乎 May 29, 2017
如何评价微软正在开发的人工智能编程软件DeepCoder？ - 知乎 Feb 24, 2017
More results from www.zhihu.com
人工智能程序设计-CSDN下载
download.csdn.net › 开发技术 › 其它 Translate this page
本书主要介绍人工智能的基础知识和应用于人工智能与专家系统领域的面向对象逻辑程序设计语言 Visual Prolog 等内容。 第1 部分主要介绍人工智能的基础知识、知识的表示方法以及AI 的编程基础。第2 部分介绍Visual Prolog 的编程基础，主要 ...
人工智能程序设计语言主要有哪些？- 云+社区- 腾讯云
https://cloud.tencent.com/developer/article/1054813 Translate this page
Mar 8, 2018 - 典型的人工智能语言主要有LISP、Prolog、Smalltalk、C++等。 一般来说，人工智能语言应具备如下特点：.·具有符号处理能力（即非数值处理能力）；.·适合于结构化程序设计，编程容易；.·具有递归功能和回溯功能；.·具有人机交互能力；.

现实世界中的问题：


Google
Translate
Turn off instant translation
English
Chinese
Japanese
Detect language
Chinese (Simplified)
English
Spanish
Translate
首届丝绸之路沿线民间组织合作网络论坛在北京开幕 习近平致贺信
30/5000
First Silk Road Cooperatives Forum for NGOs Opens in Beijing Xi Jinping Greetings
Suggest an edit
Shǒujiè sīchóu zhī lù yánxiàn mínjiān zǔzhī hézuò wǎngluò lùntán zài běijīng kāimù xíjìnpíng zhì hèxìn


Bai du 翻译
人工翻译
下载翻译插件
下载翻译app
189*****197
检测到中文
英语
翻 译
人工翻译
首届丝绸之路沿线民间组织合作网络论坛在北京开幕 习近平致贺信
The first along the Silk Road folk organization cooperation Network Forum opened in Beijing Xi Jinping sent a congratulatory letter
双语对照

语言很难!

我 们 都 要 进 口 汽 车。

主 | 谓

状 | 中

动 | 宾

定 | 中

我 们 都 要 进 口 汽 车。

主 | 谓

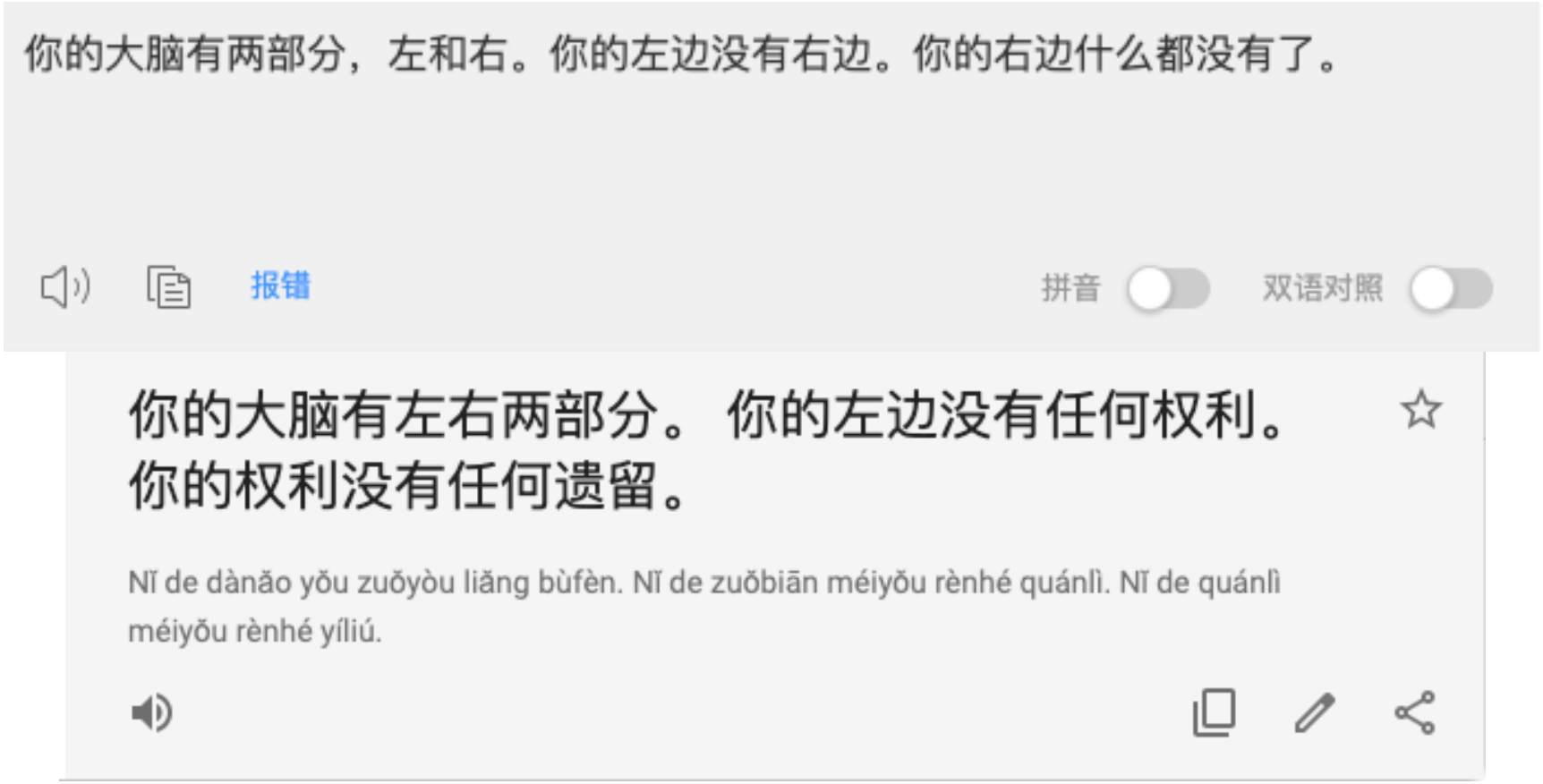
状 | 中

状 | 中

动 | 宾

语言很难!

- Your brain has two parts, left and right. Your left has nothing right. Your right has nothing left.



语言很难！

s***c

PLUS会员

★★★☆☆

相机不错，真贵，整天感觉还行

亮黑色　8GB+128GB　标准版　2019-05-01 07:45

洛***6

★★★☆☆

说实话，屏下指纹真的很不好用，其他的还可以。

亮黑色　8GB+128GB　标准版　2019-05-17 23:57

187*****453_p

★★★☆☆

手机没什么问题，很好很强大。就快递包装太简陋了

亮黑色　8GB+128GB　标准版　2019-04-25 14:28

哥哥的好大

★★★☆☆

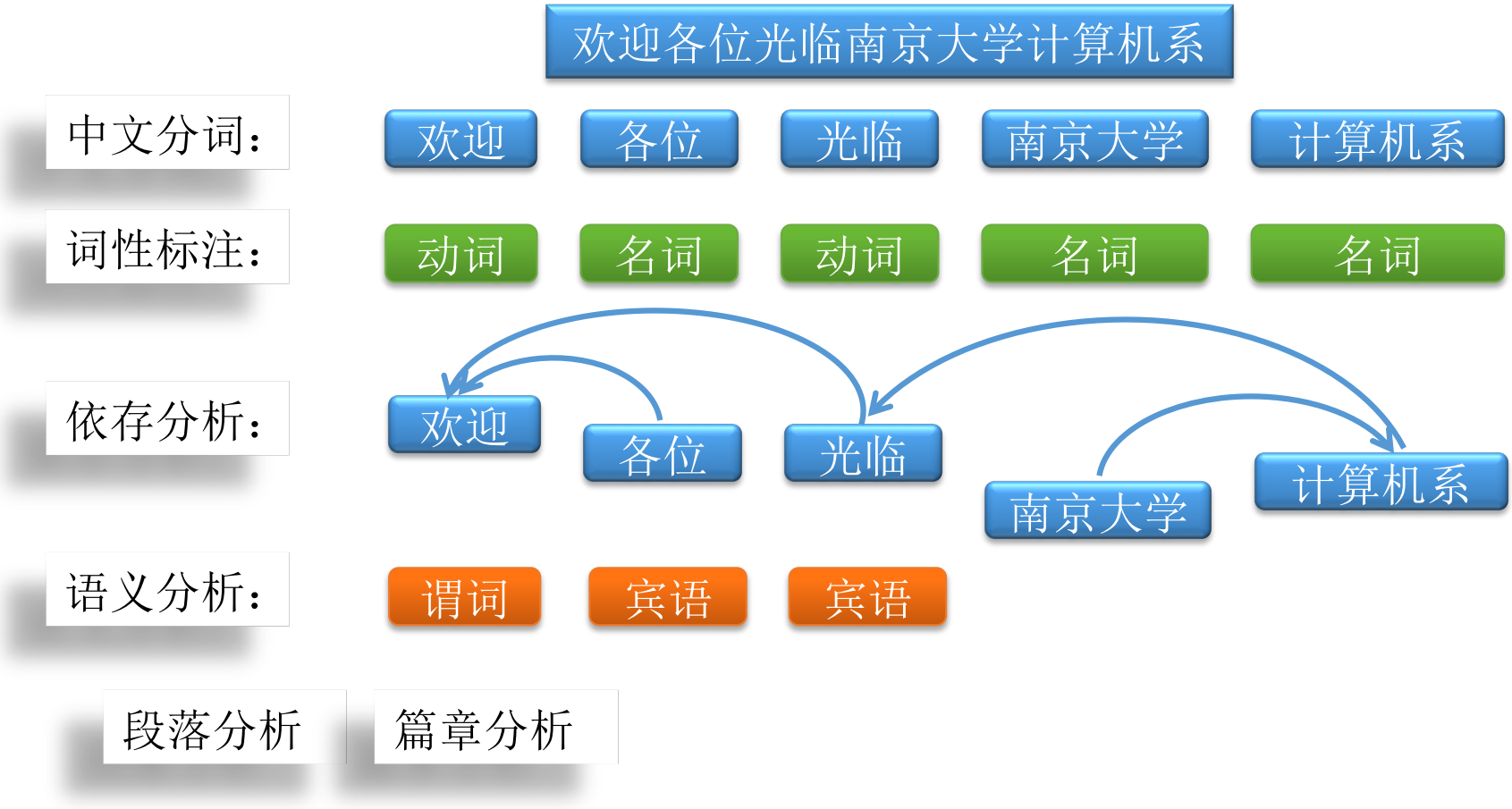
还行吧。除了速度流畅快，其他很普通，拍照特别自拍色差很大。

亮黑色　8GB+128GB　标准版　2019-04-26 15:22

自然语言处理

- 分析
  - 词、句子、篇章
- 生成
  - 翻译
  - 复述
  - 摘要
  - 自动写作

分析单个语言


欢迎各位光临南京大学计算机系
中文分词：
欢迎
各位
光临
南京大学
计算机系
词性标注：
动词
名词
动词
名词
名词
依存分析：
欢迎
各位
光临
南京大学
计算机系
语义分析：
谓词
宾语
宾语
段落分析
篇章分析

复述

类型	例子
细微变化	(a) Work at the office. Work at office.
同义词替换	(b) How can I build confidence. How can I develop confidence.
语序更换	(c) Yesterday, I got a present. I got a present yesterday.
句子拆分与合并	(d) I have a friend who is cute. I have a friend. She is cute.
句子结构变换	(e) China grows fast in economy. China's economic growth is fast.
基于推理的复述	(f) Where is your hometown. What city is your hometown.

摘要

Source Text: Peter and Elizabeth took a taxi to attend the night party in the city.

While in the party, Elizabeth collapsed and was rushed to the hospital.

Summary: Peter and Elizabeth attend party city. Elizabeth rushed hospital.

- 对于中国这样的大体量经济体而言，依靠外需不足以支撑经济平稳平衡发展。扩内需并不仅仅就是扩消费，而是要培育新的消费增长点。信息消费是一个重要的增长点，是中国经济转型升级的重要支撑，对“稳增长、调结构、促改革”具有非常重要的意义。

- 扩消费重在培育新的消费增长点

世界杯决赛 法国打爆克罗地亚 摘得大力神杯

今日头条　2018-07-16 01:07:07

世界杯决赛 法国迎战克罗地亚，在北京时间2018年7月15日23时0分打响。 最终，法国4:2战胜克罗地亚，夺得大力神杯。姆巴佩，格列兹曼，博格巴为本队建功，曼朱基奇险送乌龙。佩里西奇为克罗地亚队挽回颜面。

此役法国使用了4-2-3-1的阵型。首发门将是洛里。瓦拉内和乌姆蒂蒂将坐镇后方，保卫后防线。卢卡斯和马图伊迪坐镇左翼。锋线上吉鲁伺机而动。 另一方面克罗地亚排出的阵型是4-2-3-1。后卫线方面洛夫伦和维达组合将肩负防守重任。莫德里奇构建中场屏障。斯特里尼奇和佩里西奇出现在左路。

裁判吹响开始的哨声。 第18分钟，曼朱基奇乌龙！ 1-0！ 第27分钟，坎特动作太大，被黄牌警告。 第28分钟，佩里西奇把握良机，为克罗地亚攻入一球，扳平比分。 1-1！ 点球！第38分钟，裁判对克罗地亚处以极刑。格列兹曼点射破门！ 2-1！ 几分钟后，主裁对频频犯规的卢卡斯出示黄牌！ 法国球员在不停的传导，试图攻破这个铁甲阵。

词法分析

- **形态还原（针对英语、德语、法语等）**
  - 把句子中的词还原成基本词形。
- **词性标注**
  - 为句子中的词标上预定义类别集合（标注集）中的类。
- **命名实体识别**
  - 人名
  - 地名
  - 机构名
- **分词（针对汉语、日语等）**
  - 识别出句子中的词。

形态还原（英语）

把句子中的词还原成原形，作为词的其它信息（词典、个性规则）的索引。

- **构词特点**
  - 屈折变化：词尾和词形变化，词性不变。如：
    - study, studied,studied,studying
    - speak,spoke,spoken,speaking
  - 派生变化：加前缀和后缀，词性发生变化。如：
    - friend,friendly,friendship,...
  - 复合变化：多个单词以某种方式组合成一个词。
- **还原规则**
  - 通用规则：变化有规律
  - 个性规则：变化无规律

形态还原规则举例

- 英语“规则动词”还原
  - *s -> * (SINGULAR3)
  - *es -> * (SINGULAR3)
  - *ies -> *y (SINGULAR3)
  - *ing -> * (VING)
  - *ing -> *e (VING)
  - *ying -> *ie (VING)
  - *??ing -> *? (VING)
  - *ed -> * (PAST)(VEN)
  - *ed -> *e (PAST)(VEN)
  - *ied -> *y (PAST)(VEN)
  - *??ed -> *? (PAST)(VEN)

- 英语不规则动词还原
  - went -> go (PAST)
  - gone -> go (VEN)
  - sat -> sit (PAST) (VEN)

形态还原算法

1. 输入一个单词
2. 如果词典里有该词，输出该词及其属性，转4，否则，转3
3. 如果有该词的还原规则，并且，词典里有还原后的词，则输出还原后的词及其属性，转4，否则，调用<未登录词模块>
4. 如果输入中还有单词，转(1)，否则，结束。

尝试： 实现一个英语单词还原工具。

汉语分词（切分）

- **词是语言中最小的能独立运用的单位，也是语言信息处理的基本单位。**
- **分词是指根据某个分词规范，把一个“字”串划分成“词”串。**
  - 难以确定何谓汉语的“词”
    - 单字词与语素的界定：猪肉、牛肉
    - 词与短语（词组）的界定：黑板、黑布
  - 信息处理用现代汉语分词规范：GB-13715（1992）
  - 具体应用系统可根据各自的需求制定规范
- **分词带来的问题**
  - 丢失信息、错误的分词、不同的分词规范

切分歧义及歧义字段的种类

- **交集型歧义字段**
  - ABC切分成AB/C或A/BC
  - 如：“和平等”
    - “独立/自主/和/平等/独立/的/原则”
    - “讨论/战争/与/和平/等/问题”
- **组合型歧义字段**
  - AB切分成AB或A/B
  - 如：“马上”
    - “他/骑/在/马/上”
    - “马上/过来”
- **混合型歧义**
  - 由交集型歧义和组合型歧义嵌套与交叉而成
  - 如：“得到达”（交集型、组合型）
    - “我/今晚/得/到达/南京”
    - “我/得到/达克宁/了”
    - “我/得/到/达克宁/公司/去”

南京市长江大桥…

南京市长江二桥…

- **伪歧义与真歧义**
  - 伪歧义字段指在任何情况下只有一种切分
    - “挨批评”只有一种切分
    - 根据歧义字段本身就能消歧
  - 真歧义字段指在不同的情况下有多种切分
    - “从小学”可以有多种切分：
      - “从小/学”，如：“从小/学/电脑”（“从小”是切分成“从小”还是“从/小”要根据分词规范！）
      - “从/小学”，如：“他/从/小学/毕业/后”
    - 根据歧义字段的上下文来消歧

分词方法

一般通过分词词典和分词规则库进行分词。主要方法有：

- 正向最大匹配(FMM)或逆向最大匹配(RMM)
  - 从左至右(FMM)或从右至左(RMM)，取最长的词
  - “幼儿园 地 节目”或“幼儿 园地 节目”
- 双向最大匹配
  - 分别采用FMM和RMM进行分词
  - 如果结果一致，则认为成功；否则，
  - 采用消歧规则进行消歧（交集型歧义）：
- 正向最大、逆向最小匹配
  - 发现组合型歧义
- 逐词遍历匹配
  - 在全句中取最长的词，去掉之，对剩下字符串重复该过程
- 设立切分标记
  - 收集词首字和词尾字，把句子分成较小单位，再用某些方法切分
- 全切分
  - 获得所有可能的切分，选择最大可能的切分

基于规则的歧义字段消歧方法

- **利用歧义字串、前驱字串和后继字串的句法、语义和语用信息：**
  - 句法信息
    - “阵风”：根据前面是否有数词来消歧。“一/阵/风/吹/过/来”、“今天/有/阵风”
  - 语义信息
    - “了解”：“他/学会/了/解/数学/难题”（“难题”一般是“解”而不是“了解”，另外，还有“学会”）
  - 语用信息
    - “拍卖”：“乒乓球拍卖完了”，要根据场景（上下文）来确定
- **规则的粒度**
  - 基于具体的词（个性规则）
  - 基于词类、词义（共性规则）

尝试：实现一个基于词典与规则的汉语自动分词系统。

词性标注任务描述

- **什么叫词性?**
  - 词性又称词类，是指词的语法分类，或者说是按照其各自的语法功能的不同而分出来的类别
- **划分词类的依据**
  - 词的形态、词的语法功能、词的语法意义

英语词的分类

- 开放类（open class）
  - Nouns
    - 句法上：可作物主、可有限定词、有复数形式
    - 语义上：人名、地名和物名
  - Verbs
    - 句法上：作谓语、有几种词形变化
    - 语义上：动作、过程（一系列动作）
  - Adjectives
    - 句法上：修饰Nouns等
    - 语义上：性质
  - Adverbs
    - 句法上：修饰Verbs等
    - 语义上：方向、程度、方式、时间

- 封闭类（closed class，function words）
  - Determiners
  - Pronouns
  - Prepositions
  - Conjunctions
  - Auxiliary verbs
  - Particles（if、not、...）
  - Numerals

词性标注任务描述

- **词性标注：给某种语言的词标注上其所属的词类**
  - The lead paint is unsafe.
  - The/Det lead/N paint/N is/V unsafe/Adj.
  - 他有较强的领导才能。
  - 他/代词 有/动词 较/副词 强/形容词 的/助词 领导/名词 才能/名词。

词性标注歧义（兼类词）

- 一个词具有两个或者两个以上的词性
- 英文的**Brown**语料库中，**10.4%**的词是兼类词
  - The back door
  - On my back
  - Promise to back the bill
- 汉语兼类词
  - 把门锁上，　买了一把锁
  - 他研究与自然语言处理相关的研究工作
  - 汉语词类确定的特殊难点
- 对兼类词消歧 – 词性标注的任务

词性标注常见方法

- **规则方法：**
  - 词典提供候选词性
  - 人工整理标注规则
- **统计方法**
  - 寻找概率最大的标注序列
  - 如何建立统计模型
  - HMM方法
  - 最大墒方法
  - 条件随机场方法
  - 结构化支持向量机方法
- **基于错误驱动的方法**
  - 错误驱动学习规则
  - 利用规则重新标注词性