自然语言处理概论

常宝宝
北京大学计算语言学研究所

课程信息

- 名称: 自然语言处理
- 上课时间: 周五 8:30-11:20
- 上课地点: 3302

- 助教: 安凯凯 (ankaikai@stu.pku.edu.cn)
  胡鹤蓝 (huhelan@stu.pku.edu.cn)
  赵浩喆 (hanszhao@stu.pku.edu.cn)

- 讲义下载: 教学网(http://course.pku.edu.cn)

群聊: 自然语言处理 2024

该二维码7天内(2月28日前)有效，重新进入将更新

关于选课

- 课程特点
  - 基础性
  - 系统性
  - 专业性

- 请谨慎评估后选课。
- 选课后务请认真听课，认真完成大作业。

参考书目（一）

1. Speech and Language Processing, Jurafsky, D. and Martin, 2nd Edition, J.H., Prentice Hall, 2008
2. Neural Network Methods for Natural Language Processing. Yoav Goldberg. M&C publisher, 2017

参考书目（二）

- Foundations of Statistical Natural Language Processing, Manning,C.D. & Schütze,H., The MIT press, 1999 (有中译本)
- Statistical Language Learning. Charniak, E., The MIT Press. 1996.
- Natural Language Understanding, Allen, J., The Benjamins/Cummins Publishing Co., 1994 (有中译本)
- Natural Language Processing: An Introduction to Computational Linguistics, Gazdar, G. & Mellish, C., Addison-Wesley, 1989.
- Introduction to Natural Language Processing, Harris, M.D., Reston Publishing Co. , 1985

- Deep Learning, Goodfellow, I., Bengio, Y., Courville, A, MIT press, 2016

参考书目（三）

1. 统计自然语言处理，宗成庆，清华大学出版社，2008
2. 计算语言学概论，俞士汶主编，商务印书馆，2003
3. 自然语言理解，姚天顺，清华大学出版社，2002
4. 自然语言处理技术基础，王小捷、常宝宝，北京邮电大学出版社，2002
5. 计算语言学，刘颖，清华大学出版社，2002
6. 计算语言学基础，冯志伟，商务印书馆，2001
7. 计算语言学导论，翁富良、王野翊，中国社会科学出版社，1998
8. 自然语言的计算机处理，冯志伟，上海外语教育出版社，1997
9. 自然语言处理，刘开瑛、郭炳炎，科学出版社，1991

10. 神经网络与深度学习，邱锡鹏，机械工业出版社，2020
11. 机器翻译 基础与模型，肖桐、朱靖波，电子工业出版社，2021
12. 神经机器翻译，熊德意、李良友、张檬，电子工业出版社，2022
13. 自然语言处理导论，张奇、桂韬、黄萱菁，电子工业出版社，2023

相关学术期刊和会议

1. Transactions of the Association for Computational Linguistics (TACL)
2. Computational Linguistics (CL)
3. 中文信息学报 (中文信息学会)

1. Annual Meeting of the Association for Computational Linguistics (ACL年会)
2. Conference on Empirical methods in natural language processing(EMNLP)
3. Annual Conference of the North American Chapter of the Association for Computational Linguistics(NAACL)
4. International Conference on Computational Linguistics (COLING)
5. 全国计算语言学联合学术会议(CCL) (中文信息学会)
6. 自然语言处理与中文计算会议(NLPCC) (计算机学会)

课程概要

- 第一部分：基础
  - 自然语言处理概述
  - 机器学习基础
- 第二部分：经典任务和方法
  - n元模型
  - HMM模型
  - CRF模型
  - 依存分析

课程概要

- 第三部分：基于深度学习的自然语言处理
  - 自然语言处理中的文本例化
  - 神经网络基础
  - 词向量
  - 注意力机制和Transformer模型
  - 预训练语言模型及微调
  - 大语言模型

概要

- **自然语言处理概要**
- 自然语言处理的基本方法
- 自然语言处理的挑战性
- 自然语言处理的应用领域
- 自然语言处理发展简史

自然语言

- 自然语言(Natural Language)
  - 汉语、英语、德语……
  - 人类交流使用的语言
  - 人类交流实践中自然演化出来的语言

- 人造语言
  - 为了某种交流目标而设计的语言
  - 程序设计语言
    - C语言、Java语言

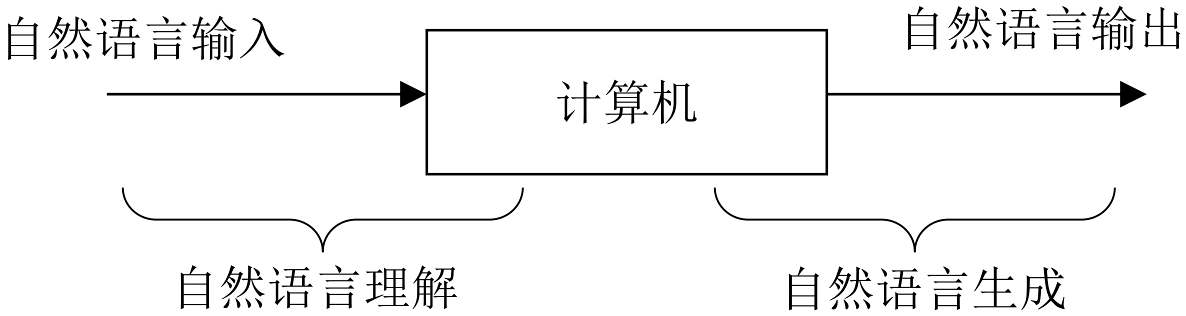
自然语言

- 人类语言能力
  - 理解语言：“听”与“读”
  - 生成语言：“说”与“写”
  - 利用语言实现交际意图

- 可否让计算机具有人类的语言能力？
  - 计算机理解自然语言
  - 计算机生成自然语言
  - 计算机和人类用自然语言交流并实现相应的交际意图

自然语言处理

- 什么是自然语言处理(Natural Language Processing, NLP)
  - 通过建立形式化计算模型来分析、理解和生成自然语言的学科
- 终极目标： 让计算机拥有自然语言交际能力



自然语言输入

计算机

自然语言输出

自然语言理解

自然语言生成

- 两大问题
  - 自然语言理解(Natural Language Understanding, NLU)
  - 自然语言生成(Natural Language Generation, NLG)

自然语言处理

- 人工智能学科的分支学科
- 交叉学科
  - 计算机科学 (自然语言处理的研究工具)
  - 语言学 (自然语言是处理对象)
  - 数学 (自然语言处理的建模工具)
- 其他相关学科：认知科学、心理学、哲学、脑科学、神经科学
- 文理大交叉
  - 多学科知识需求

语言处理的机器模型

- 语言的机器理解和生成是一个复杂的问题
- 语言学家为理解建模– 还原主义
- 分而治之
  - 意义可以用符号化方法加以分层描述和表示
  - 分层进行、层层递进
  - 受语言学研究影响
  - 词法分析、句法分析、语义分析、语用分析
  - 逐步把“理解”推向深入

词法分析

- 汉语分词(Chinese word segmentation)
  - 汉语中词和词之间没有空格；识别出句子中的词。
  - 字符序列 ⇒ 词的序列

    我喜欢吃苹果。 ⇒ 我/喜欢/吃/苹果/。

    南京市长江大桥。 ⇒ 南京市/长江/大桥/。

    ⇒ 南京/市长/江大桥/。

    能穿多少穿多少。 ⇒能/穿/多少/穿/多少/。

    ⇒能/穿/多/少/穿/多/少/。
  - 藏文、蒙文、韩文、日文……
- 词的识别是一种简单的理解

词法分析

- 词类标注(Part-of-speech tagging, POS tagging)
  - 词类：名词、动词、形容词……
  - 兼类词

    选 他 当 代表 。⇒ 选/v 他/r 当/v 代表/n 。/w

    一 把 把 把 把 住 了 。 ⇒ 一/m 把/q 把/p 把/n 把/v 住/v 了/u 。/w

    book that flight . ⇒ book/VB that/DT flight/NN ./.

- 命名实体识别(Named Entity Recognition, NER)
  - 人名、地名、机构名、……

  1976年，**乔布斯** 和 朋友 在 **美国** 成立 **苹果公司** 。

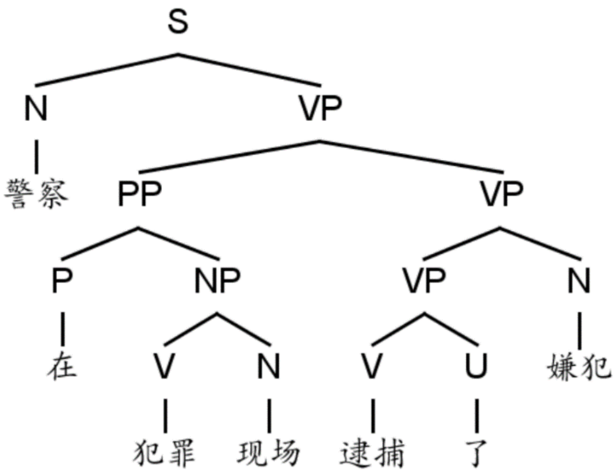
  我 也 想 过过 **过儿** 过 过 的 生活

- 判定词类、识别命名实体也是一种理解

句法分析

- 句法分析(syntactic parsing)
  - 分析句子，得到句子的句法结构
  - 通常是树结构
- 词组成短语、短语组成句子

警察 在 犯罪 现场 逮捕 了 嫌犯 ⇒



语义分析

- 词义标注(Word Sense Tagging)
  - 多义词

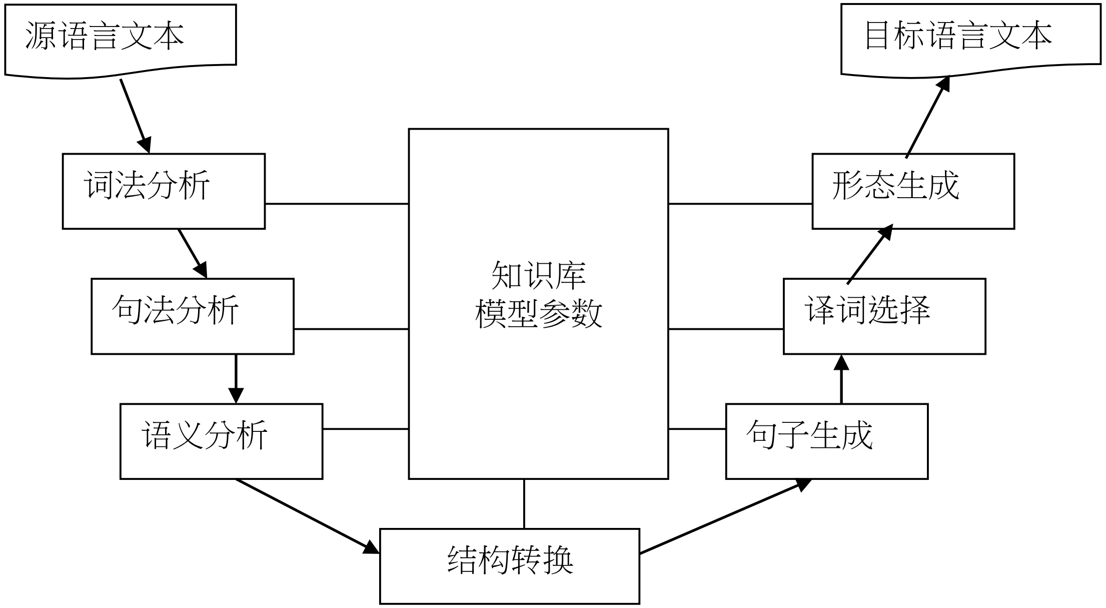
    写作前最好先**打**腹稿。

    I have very little money in the **bank**
- 语义角色标记(Semantic Role Labeling, SRL)
  - 识别谓词论元及其充当的角色
  - 部分句义分析、浅层句义分析

Who	at Where	did What	to whom?
警察	在 犯罪 现场	逮捕 了	嫌犯
Agent	Location	Predicate	Patient

案例

- 机器翻译系统由词法分析、句法分析等多个部件构成



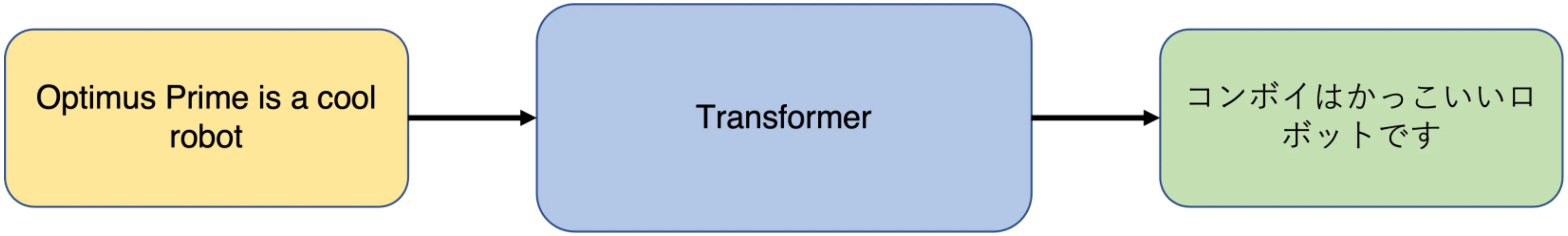
语言处理的机器模型

- 基于整体论的建模路线
  - 语言理解和生成的机制不清楚，是一个黑箱
  - 将自然语言处理建模为一个不可分割的整体
  - 以行为是否符合预期定义理解

- 端到端处理
  - 意义无需采用符号化方法显式描述和表示
  - 由语言输入直接生成语言输出
  - 理解体现为产生正确的输出

案例

- 利用transformer构建机器翻译系统



Optimus Prime is a cool robot

Transformer

コンボイはかっこいいロボットです

- 利用GPT等预训练模型构建各类应用系统
- 没有错误积累、没有中间结果

概要

- 自然语言处理概要
- **自然语言处理的基本方法**
- 自然语言处理的挑战性
- 自然语言处理的应用领域
- 自然语言处理发展简史

NLP的基本方法

- 基于规则的方法
  - 也叫符号主义、理性主义方法
  - 专家以规则的形式注入知识

- 基于统计的方法
  - 也叫机器学习方法、经验主义方法
  - 机器从语言样本中自动学习知识

基于规则的方法

- 研究人员(如语言学家)对语言的规律进行总结，形成规则形态的知识库
- 研制语言处理算法，利用这些规则对自然语言进行处理
- 研究人员根据处理结果，调整规则，改进处理效果。

基于规则的方法

- 句子的组成规则

S → NP + VP

NP → NP + PP

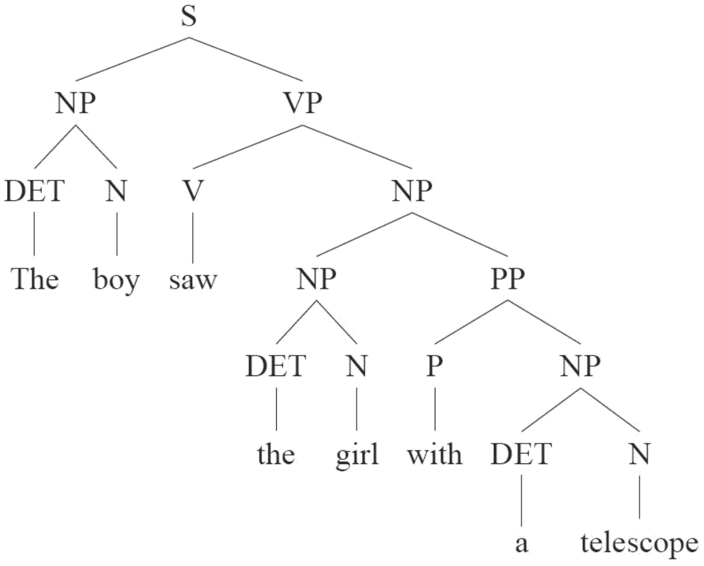
VP → V + NP

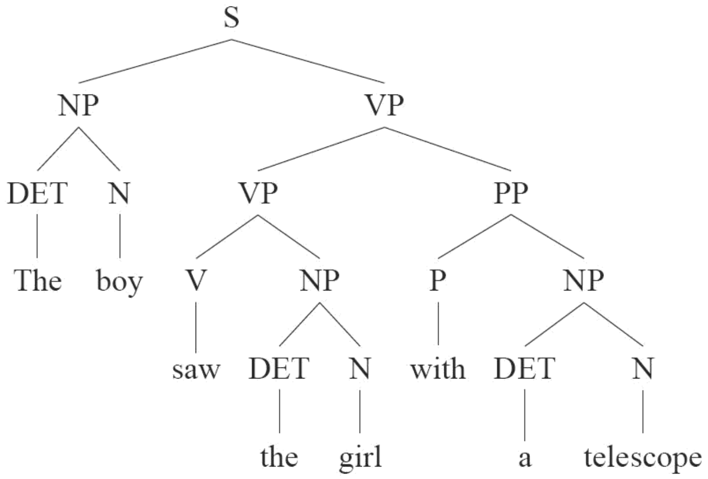
NP → DET + N

VP → VP + PP

PP → P + NP

The boy saw a girl with a telescope





基于规则的方法

- *All grammar leak.* (Edward Sapir, 1921)
  - 规则鲜无例外
  - 规则之中有着不规则
- 无法覆盖所有语言现象
  - 无法穷举所有规则
  - 不规范的语言使用(如社交媒体)
- 规则一致性难以保证
  - 按下葫芦浮起瓢

Toy system?
Toy syndrome?

基于统计的方法

- 建立可以反映语言使用情况的语料库
- 研究人员对自然语言进行统计建模
- 利用统计技术或机器学习技术，借助语料库训练语言模型
- 基于得到的模型设计算法对语言进行处理
- 根据处理效果改进模型，提高处理性能

基于统计的方法

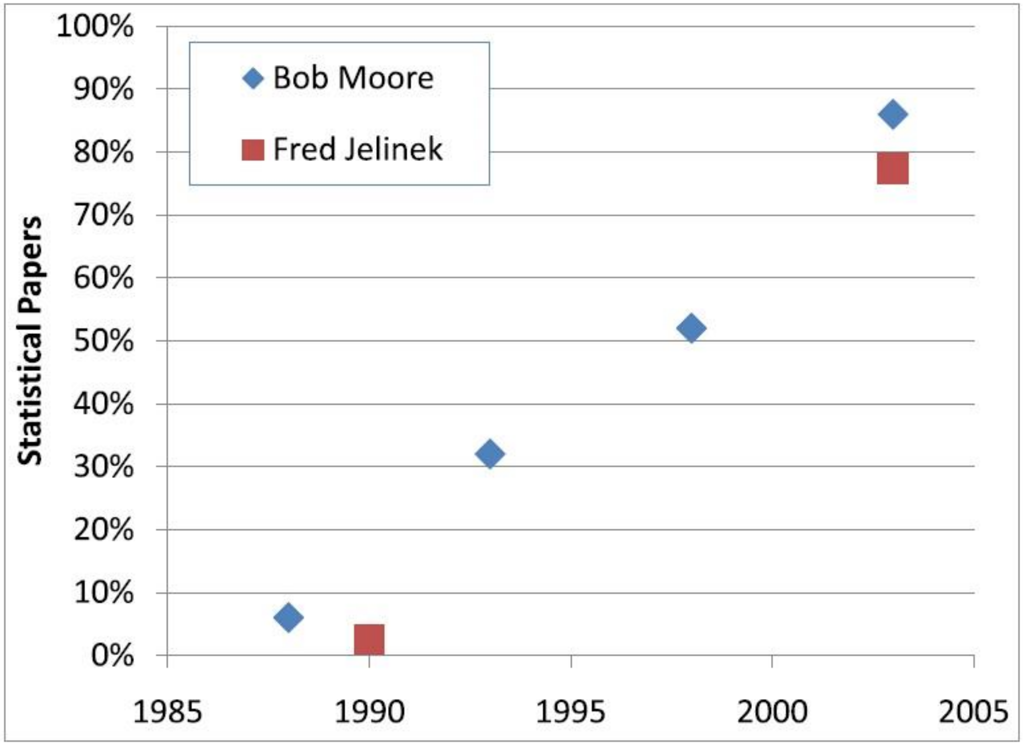
- 词袋模型
  - 语言就像掷骰子
  - 多项分布
- 马尔科夫过程
  - 句子是随机过程的产物
- 隐马尔科夫过程

- 分布参数从哪里来？

基于统计的方法

- *All models are wrong.* (George Box, 1978)
  - 模型只是现实世界的简化
- 需要大量的语言样本
  - 样本的代表性
  - 标注代价高昂
- 数据稀疏问题
  - 长尾效应
- 目前基于统计方法进展显著
  - 大数据、算力

从学术会议看NLP研究方法



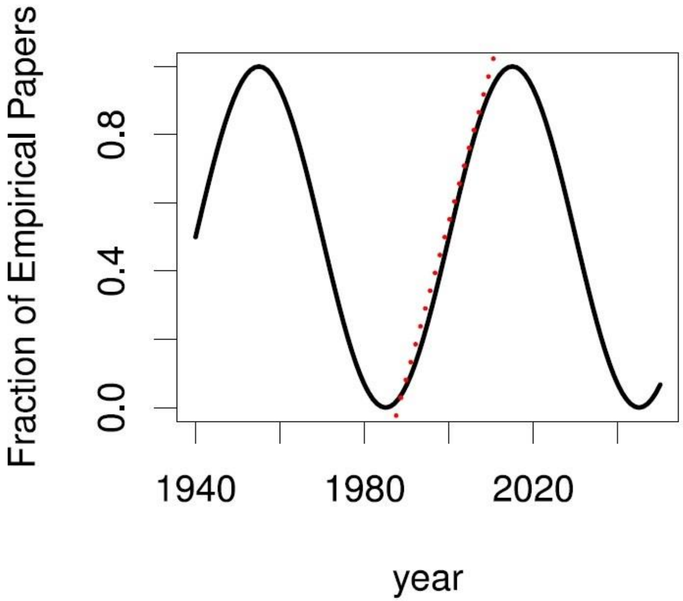
two independent surveys of ACL meetings by Bob Moore and Fred Jelinek

--- A PENDULUM SWUNG TOO FAR

- 上世纪90年代统计方法开始复苏
- 机器学习方法(经典机器学习→深度学习)现在是主流

研究方法演变

- 经验主义、理性主义的交替演变
  - 1950s: Empiricism (Shannon, Skinner, Firth, Harris)
  - 1970s: Rationalism (Chomsky, Minsky)
  - 1990s: Empiricism (IBM Speech Group, AT&T Bell Labs)
  - 2010s: A Return to Rationalism? *No*, *Deep Learning*☺



The oscillation
between Rationalism and Empiricism

--- A PENDULUM SWUNG TOO FAR

概要

- 自然语言处理概要
- 自然语言处理的基本方法
- **自然语言处理的挑战性**
- 自然语言处理的应用领域
- 自然语言处理发展简史

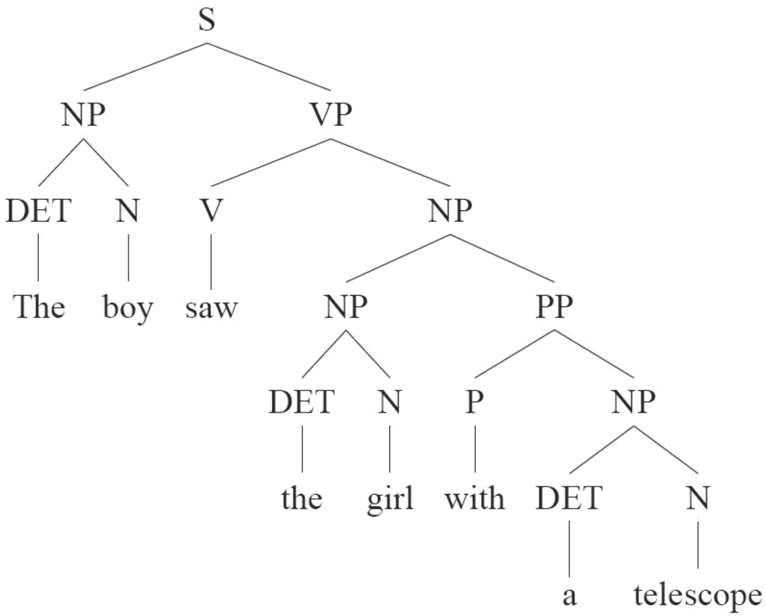
挑战性

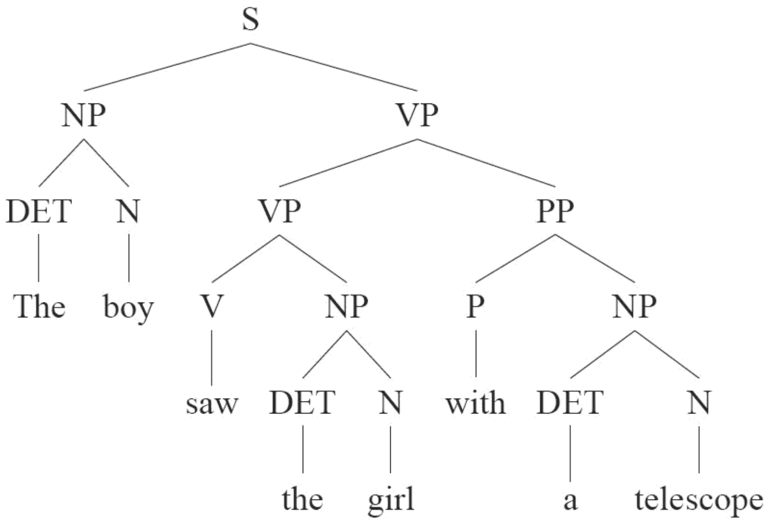
- 无处不在的歧义(Ambiguity)
- 歧义：对同一个语言形式有不止一种解读
- 分词
  - 能 穿 多 少 穿 多 少 ⇔ 能 穿 多少 穿 多少
- 词类
  - 一 **把** **把** **把** **把** 住 了
  - Time **flies** like an arrow ⇔ fruit **flies** like apple
- 词义
  - **打** 酱油 ⇔ **打** 毛衣 ⇔ **打** 篮球

挑战性

- 结构歧义

The boy saw a girl with a telescope.





挑战性

- **海量的知识需求**
- 语言知识(linguistic knowledge)
  - 词法、句法、语法、语义、语用……
- 领域知识(domain knowledge)
  - 物理、化学、生物、计算机….
- 世界知识(world knowledge)
  - 妈妈是女的、爸爸是男的
  - 父母的年龄比子女的年龄大
- 知识表示、知识获取、知识运用、知识更新存在巨大挑战

概要

- 自然语言处理概要
- 自然语言处理的基本方法
- 自然语言处理的挑战性
- **自然语言处理的应用领域**
- 自然语言处理发展简史

自然语言处理的应用

- 机器翻译
- 自动问答
- 自动摘要
- 信息提取
- 自动写作
- 情感分析
- 文本分类
- 垃圾邮件过滤
- 汉字输入技术
- 术语提取

自然语言处理有着广阔的应用领域

机器翻译

- 目标是研制能把一种自然语言翻译成另外一种自然语言的计算机软件系统。
  - 例如 汉英机器翻译系统
- 相关研究始于四十年代末（计算机诞生不久）。
- 机器翻译研究经历了曲折的历程，机器翻译研究导致了自然语言处理学科的诞生。

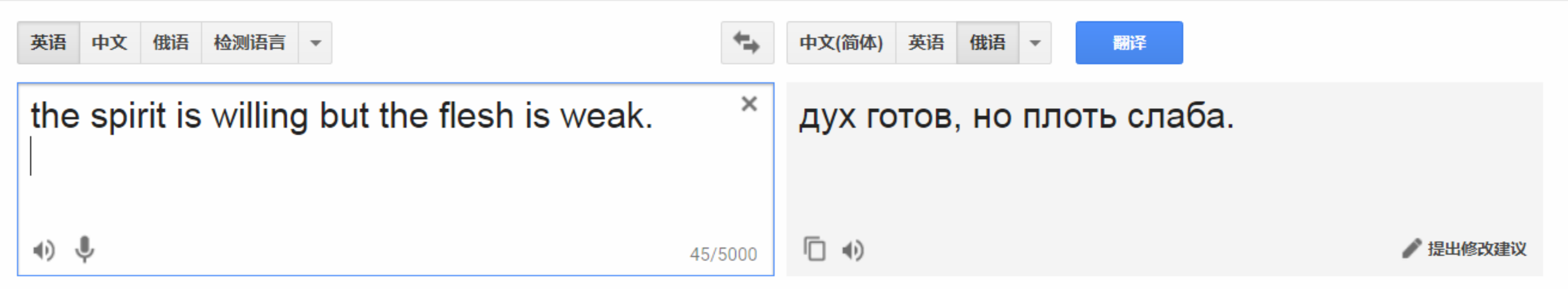
- 全自动高质量机器翻译(FAHQMT) – 尚须时日

机器翻译

- 著名的例子

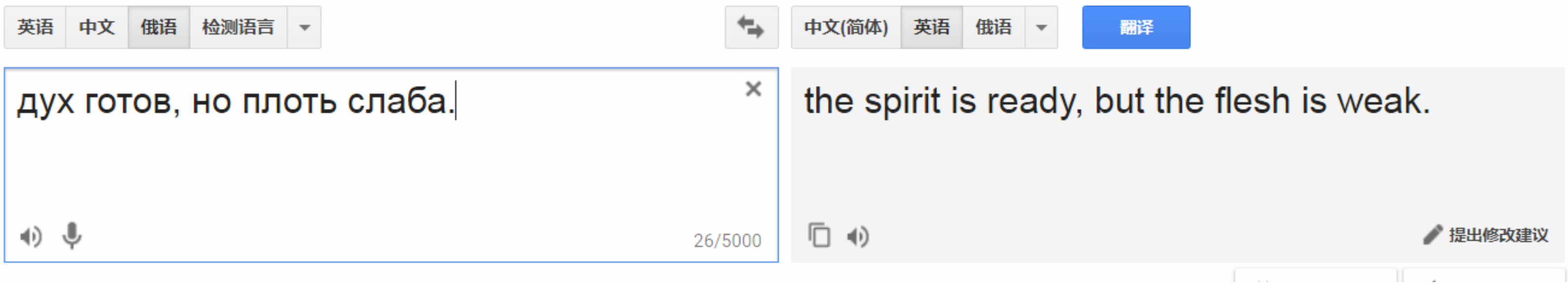
  the spirit is willing but the flesh is weak.

  the vodaka is good but the meat is rotten.



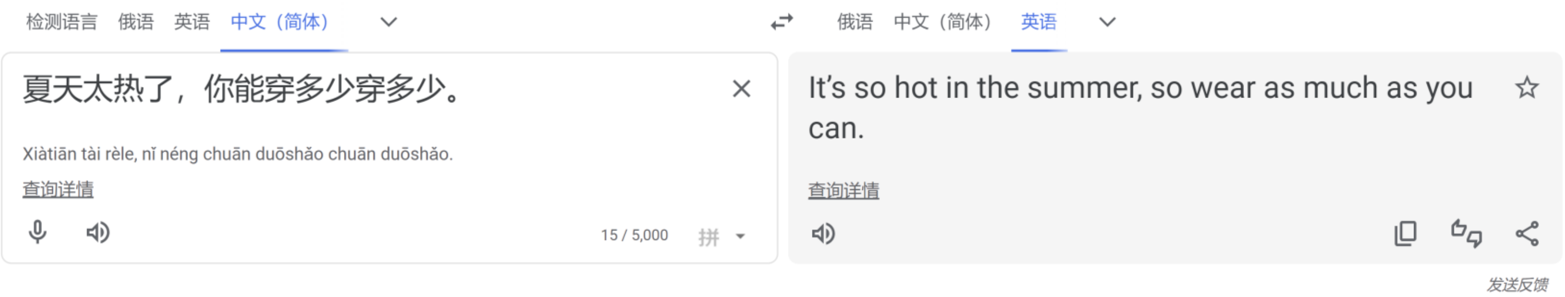
the spirit is willing but the flesh is weak.

дух готов, но плоть слаба.



дух готов, но плоть слаба.

the spirit is ready, but the flesh is weak.



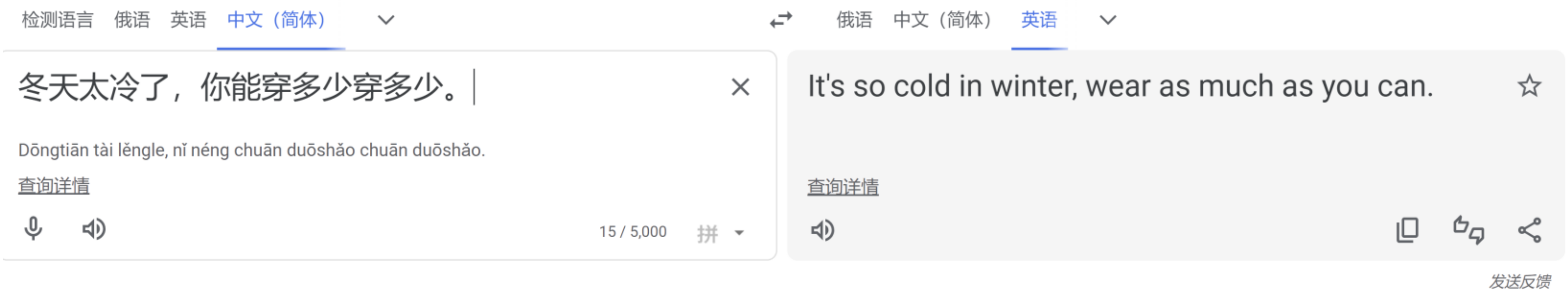
夏天太热了，你能穿多少穿多少。

Xiàtiān tài rèle, nǐ néng chuān duōshǎo chuān duōshǎo.

查询详情

It's so hot in the summer, so wear as much as you can.

查询详情



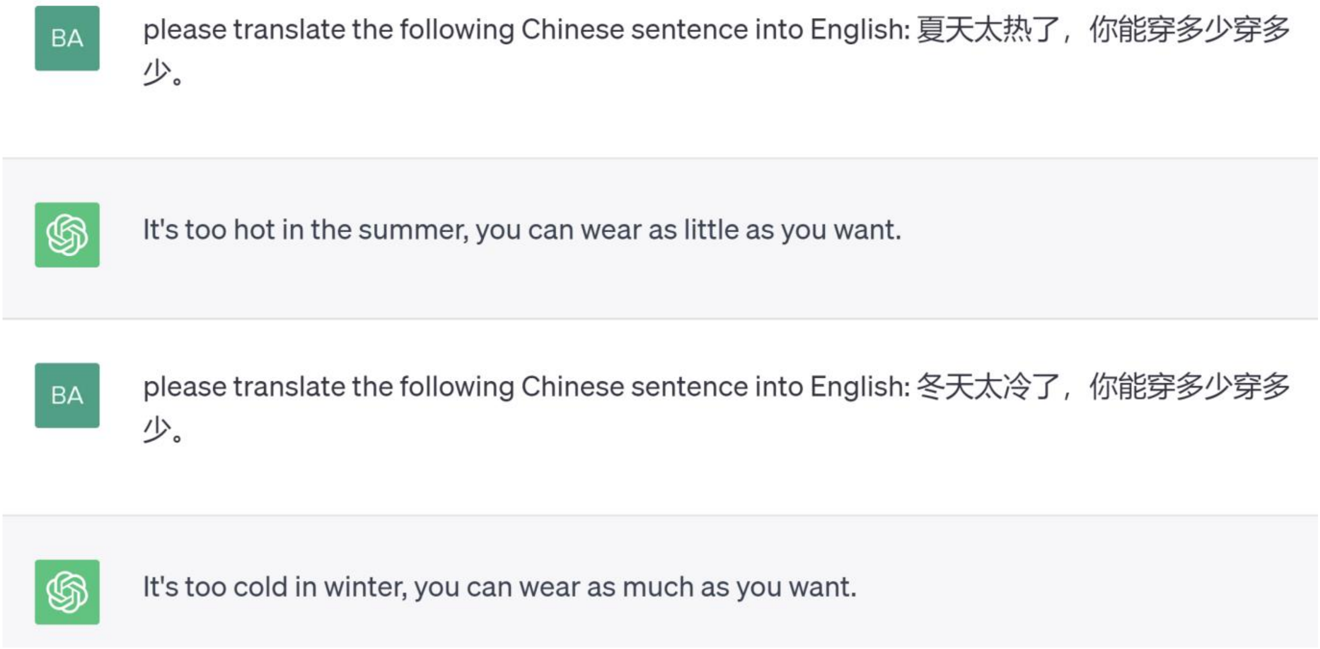
冬天太冷了，你能穿多少穿多少。

Dōngtiān tài lěngle, nǐ néng chuān duōshǎo chuān duōshǎo.

查询详情

It's so cold in winter, wear as much as you can.

查询详情



please translate the following Chinese sentence into English: 夏天太热了，你能穿多少穿多少。

It's too hot in the summer, you can wear as little as you want.

please translate the following Chinese sentence into English: 冬天太冷了，你能穿多少穿多少。

It's too cold in winter, you can wear as much as you want.

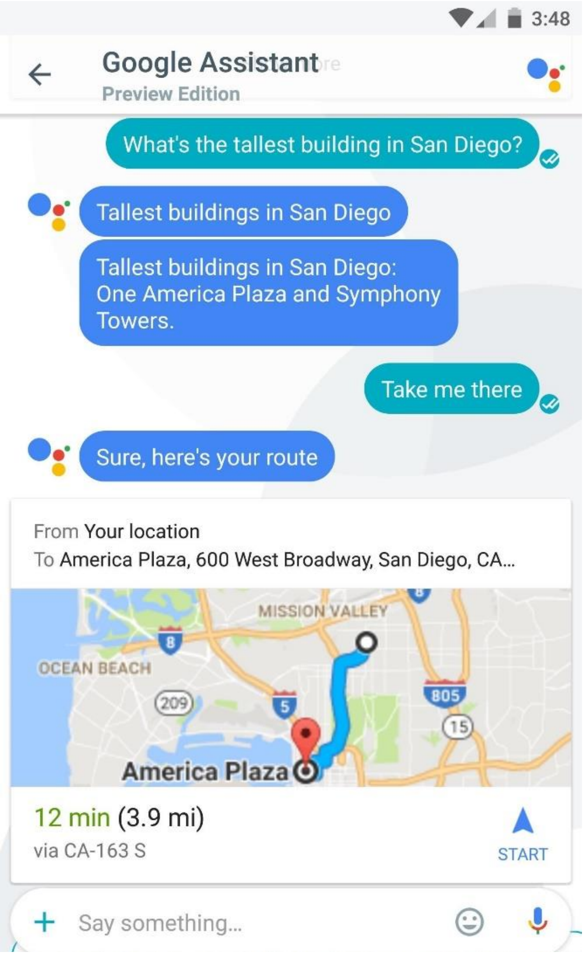
人机对话

- 科幻电影中的人机对话
- 2001:A space
  1968年奥斯卡奖
- HAL9000
  Dave: Open the pod bay doors, HAL.
  HAL: I’m sorry Dave, I’am afraid I can’t do that
  Dave: What’s the problem?
  HAL: I think you know what the problem is just as well as I do.

聊天机器人(Chatbot)

- 1966年，ELIZA系统(MIT)
- Google Assistant
- 苹果Siri
- 百度小度
- 微软xiaoice、微软Cortana

Google Assitant



What's the tallest building in San Diego?

Tallest buildings in San Diego

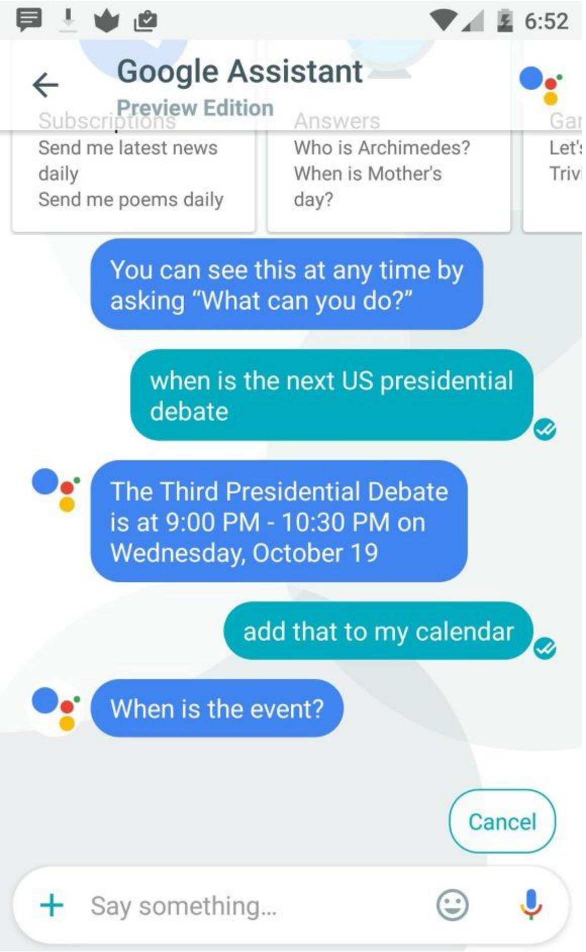
Tallest buildings in San Diego: One America Plaza and Symphony Towers.

Take me there

Sure, here's your route

From Your location
To America Plaza, 600 West Broadway, San Diego, CA...

12 min (3.9 mi)
via CA-163 S



Subscriptions
Send me latest news daily
Send me poems daily

Answers
Who is Archimedes?
When is Mother's day?

Gar
Let's
Triv

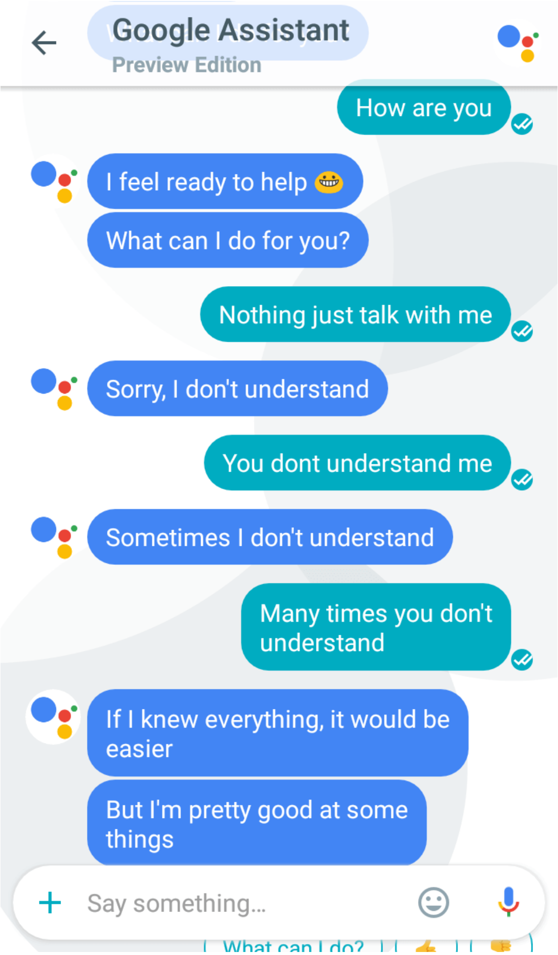
You can see this at any time by asking "What can you do?"

when is the next US presidential debate

The Third Presidential Debate is at 9:00 PM - 10:30 PM on Wednesday, October 19

add that to my calendar

When is the event?



How are you

I feel ready to help 😁

What can I do for you?

Nothing just talk with me

Sorry, I don't understand

You dont understand me

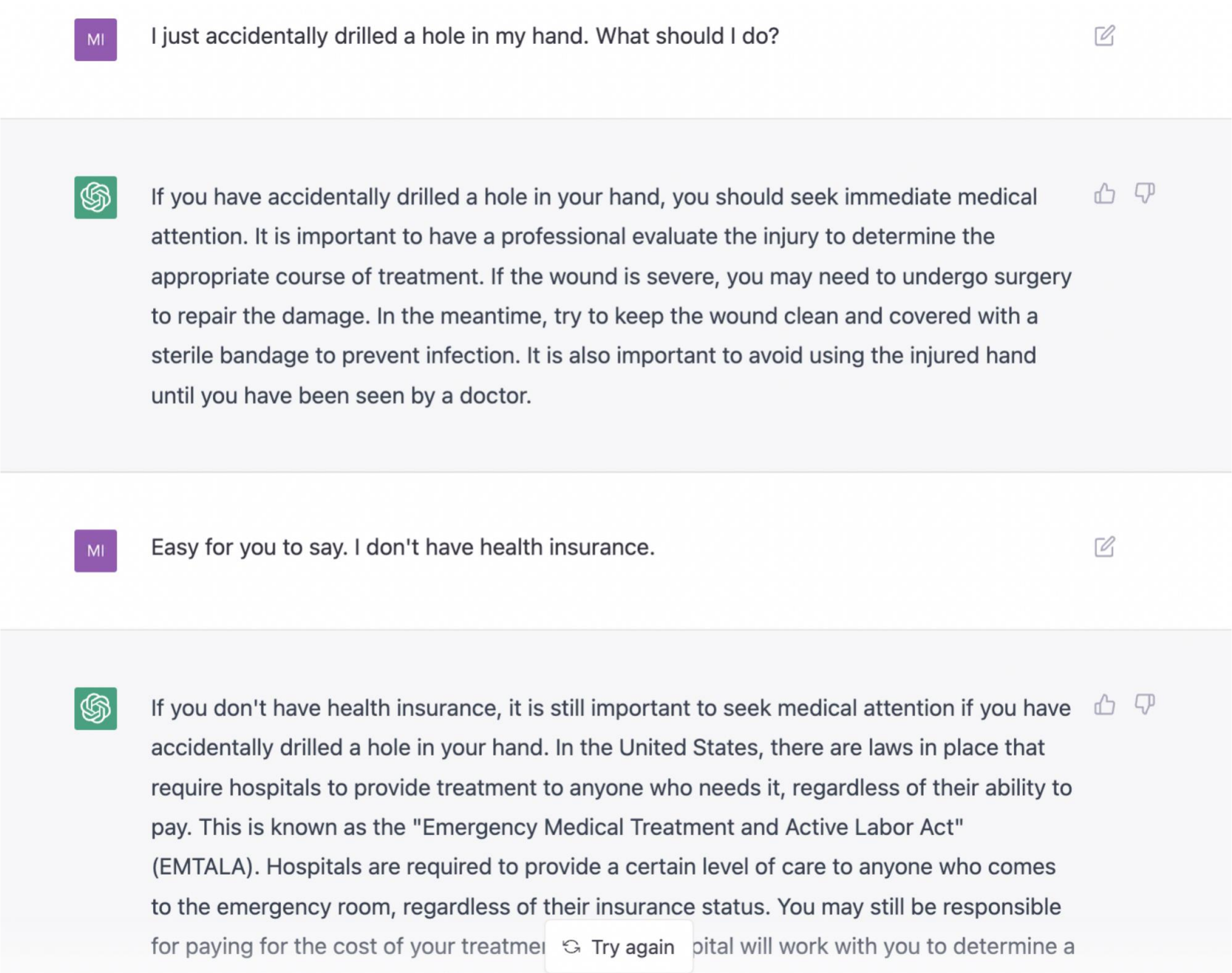
Sometimes I don't understand

Many times you don't understand

If I knew everything, it would be easier

But I'm pretty good at some things

chatGPT



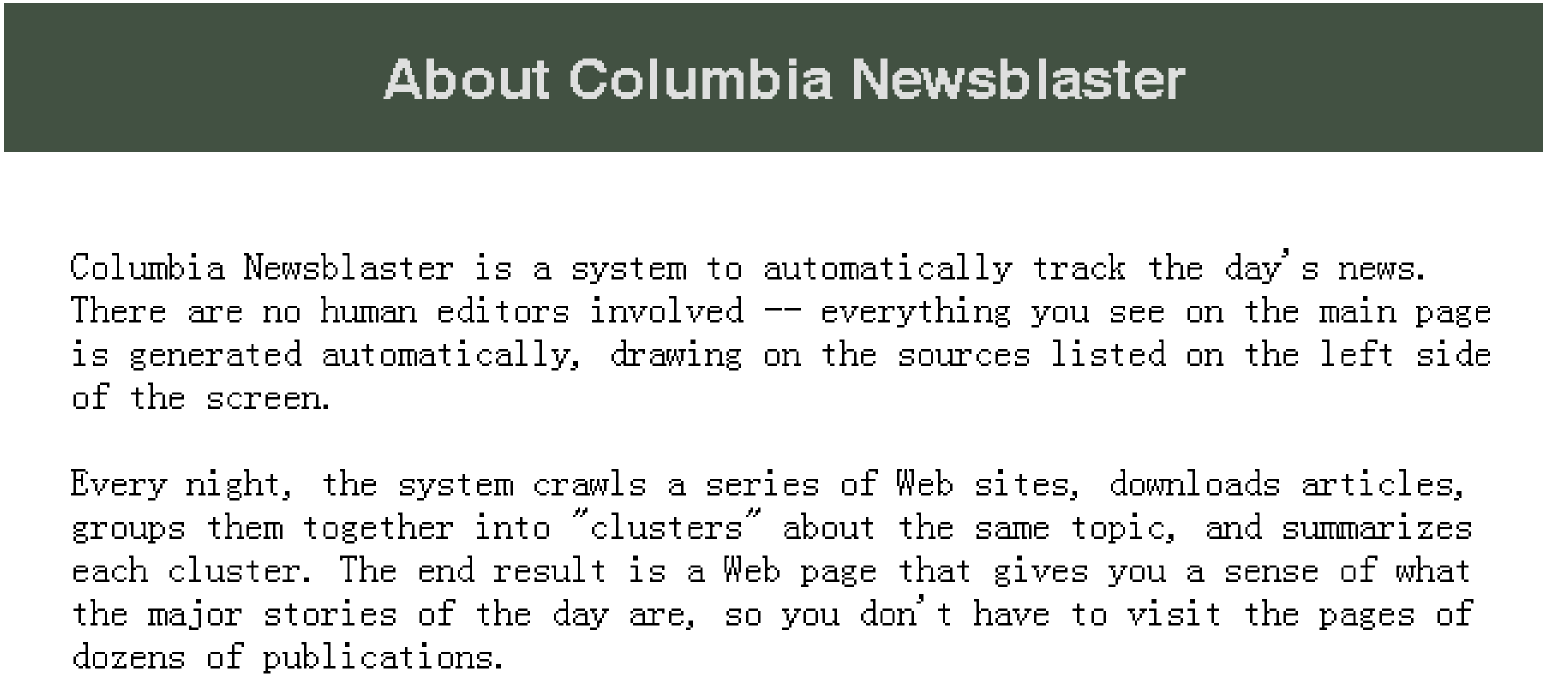
I just accidentally drilled a hole in my hand. What should I do?

If you have accidentally drilled a hole in your hand, you should seek immediate medical attention. It is important to have a professional evaluate the injury to determine the appropriate course of treatment. If the wound is severe, you may need to undergo surgery to repair the damage. In the meantime, try to keep the wound clean and covered with a sterile bandage to prevent infection. It is also important to avoid using the injured hand until you have been seen by a doctor.

Easy for you to say. I don't have health insurance.

If you don't have health insurance, it is still important to seek medical attention if you have accidentally drilled a hole in your hand. In the United States, there are laws in place that require hospitals to provide treatment to anyone who needs it, regardless of their ability to pay. This is known as the "Emergency Medical Treatment and Active Labor Act" (EMTALA). Hospitals are required to provide a certain level of care to anyone who comes to the emergency room, regardless of their insurance status. You may still be responsible for paying for the cost of your treatme pital will work with you to determine a

自动文摘

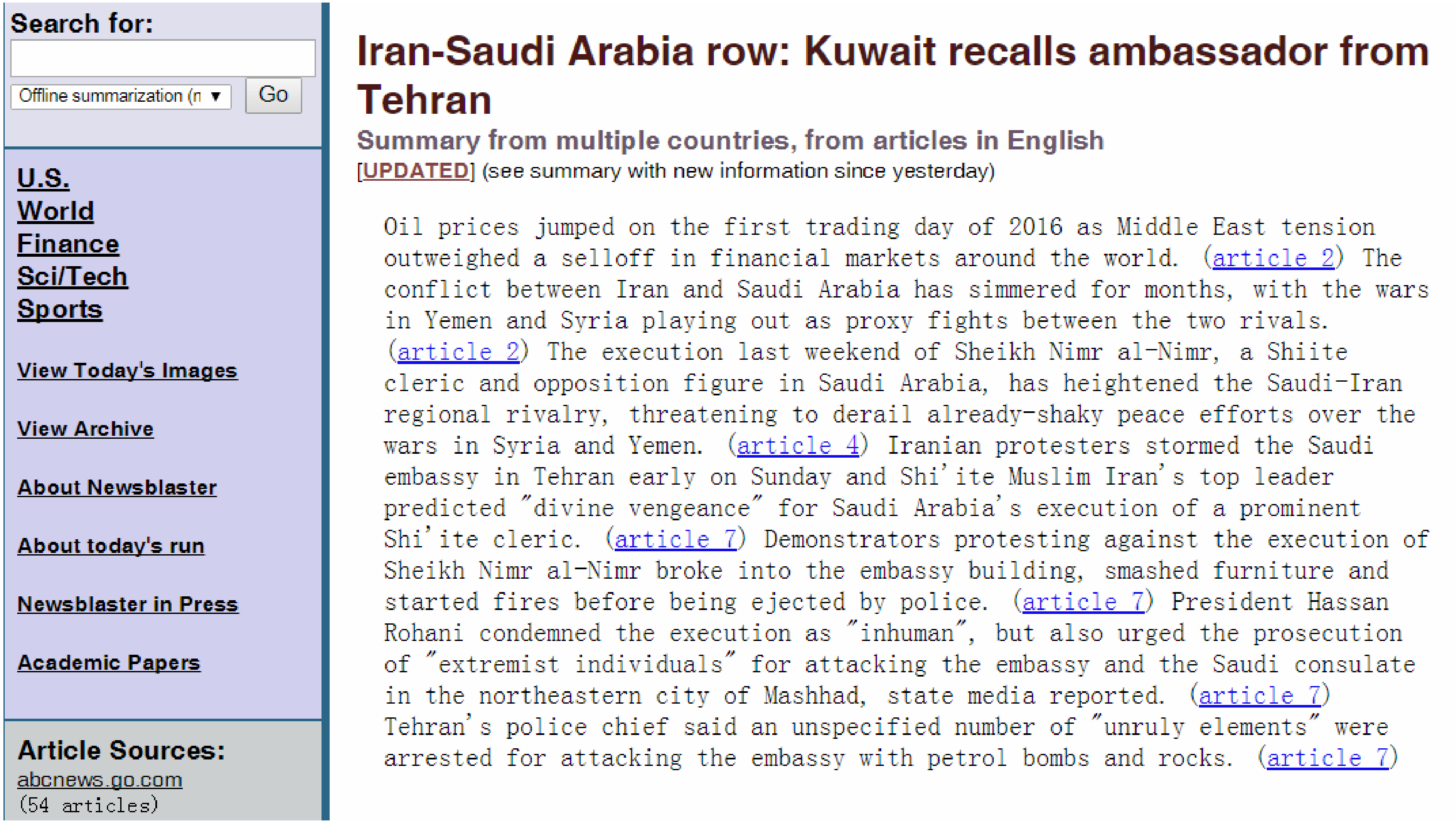


About Columbia Newsblaster

Columbia Newsblaster is a system to automatically track the day's news. There are no human editors involved -- everything you see on the main page is generated automatically, drawing on the sources listed on the left side of the screen.

Every night, the system crawls a series of Web sites, downloads articles, groups them together into "clusters" about the same topic, and summarizes each cluster. The end result is a Web page that gives you a sense of what the major stories of the day are, so you don't have to visit the pages of dozens of publications.

- 访问: Columbia Newsblaster
  http://www1.cs.columbia.edu/nlp/newsblaster/
  http://newsblaster.cs.columbia.edu

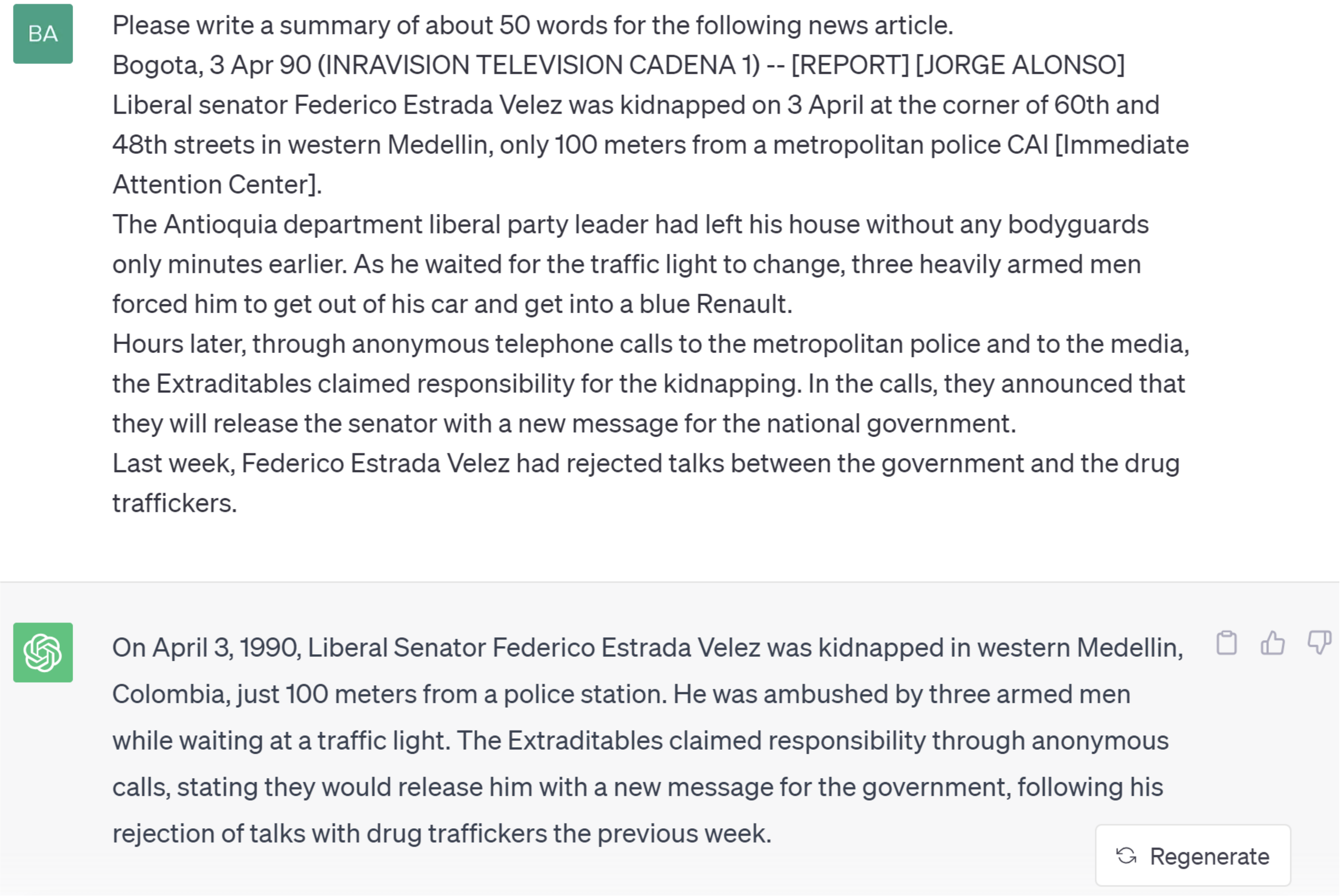


Iran-Saudi Arabia row: Kuwait recalls ambassador from Tehran

Summary from multiple countries, from articles in English

[UPDATED] (see summary with new information since yesterday)

Oil prices jumped on the first trading day of 2016 as Middle East tension outweighed a selloff in financial markets around the world. (article 2) The conflict between Iran and Saudi Arabia has simmered for months, with the wars in Yemen and Syria playing out as proxy fights between the two rivals. (article 2) The execution last weekend of Sheikh Nimr al-Nimr, a Shiite cleric and opposition figure in Saudi Arabia, has heightened the Saudi-Iran regional rivalry, threatening to derail already-shaky peace efforts over the wars in Syria and Yemen. (article 4) Iranian protesters stormed the Saudi embassy in Tehran early on Sunday and Shi'ite Muslim Iran's top leader predicted "divine vengeance" for Saudi Arabia's execution of a prominent Shi'ite cleric. (article 7) Demonstrators protesting against the execution of Sheikh Nimr al-Nimr broke into the embassy building, smashed furniture and started fires before being ejected by police. (article 7) President Hassan Rohani condemned the execution as "inhuman", but also urged the prosecution of "extremist individuals" for attacking the embassy and the Saudi consulate in the northeastern city of Mashhad, state media reported. (article 7) Tehran's police chief said an unspecified number of "unruly elements" were arrested for attacking the embassy with petrol bombs and rocks. (article 7)



Please write a summary of about 50 words for the following news article.
Bogota, 3 Apr 90 (INRAVISION TELEVISION CADENA 1) -- [REPORT] [JORGE ALONSO]
Liberal senator Federico Estrada Velez was kidnapped on 3 April at the corner of 60th and 48th streets in western Medellin, only 100 meters from a metropolitan police CAI [Immediate Attention Center].
The Antioquia department liberal party leader had left his house without any bodyguards only minutes earlier. As he waited for the traffic light to change, three heavily armed men forced him to get out of his car and get into a blue Renault.
Hours later, through anonymous telephone calls to the metropolitan police and to the media, the Extraditables claimed responsibility for the kidnapping. In the calls, they announced that they will release the senator with a new message for the national government.
Last week, Federico Estrada Velez had rejected talks between the government and the drug traffickers.

On April 3, 1990, Liberal Senator Federico Estrada Velez was kidnapped in western Medellin, Colombia, just 100 meters from a police station. He was ambushed by three armed men while waiting at a traffic light. The Extraditables claimed responsibility through anonymous calls, stating they would release him with a new message for the government, following his rejection of talks with drug traffickers the previous week.

信息提取

- 文本数据结构化

BOGOTA, 3 Apr 90 (INRAVISION TELEVISION CADENA 1) -- [REPORT] [JORGE ALONSO]

Liberal senator Federico Estrada Velez was kidnapped on 3 April at the corner of 60th and 48th streets in western Medellin, only 100 meters from a metropolitan police CAI [Immediate Attention Center].

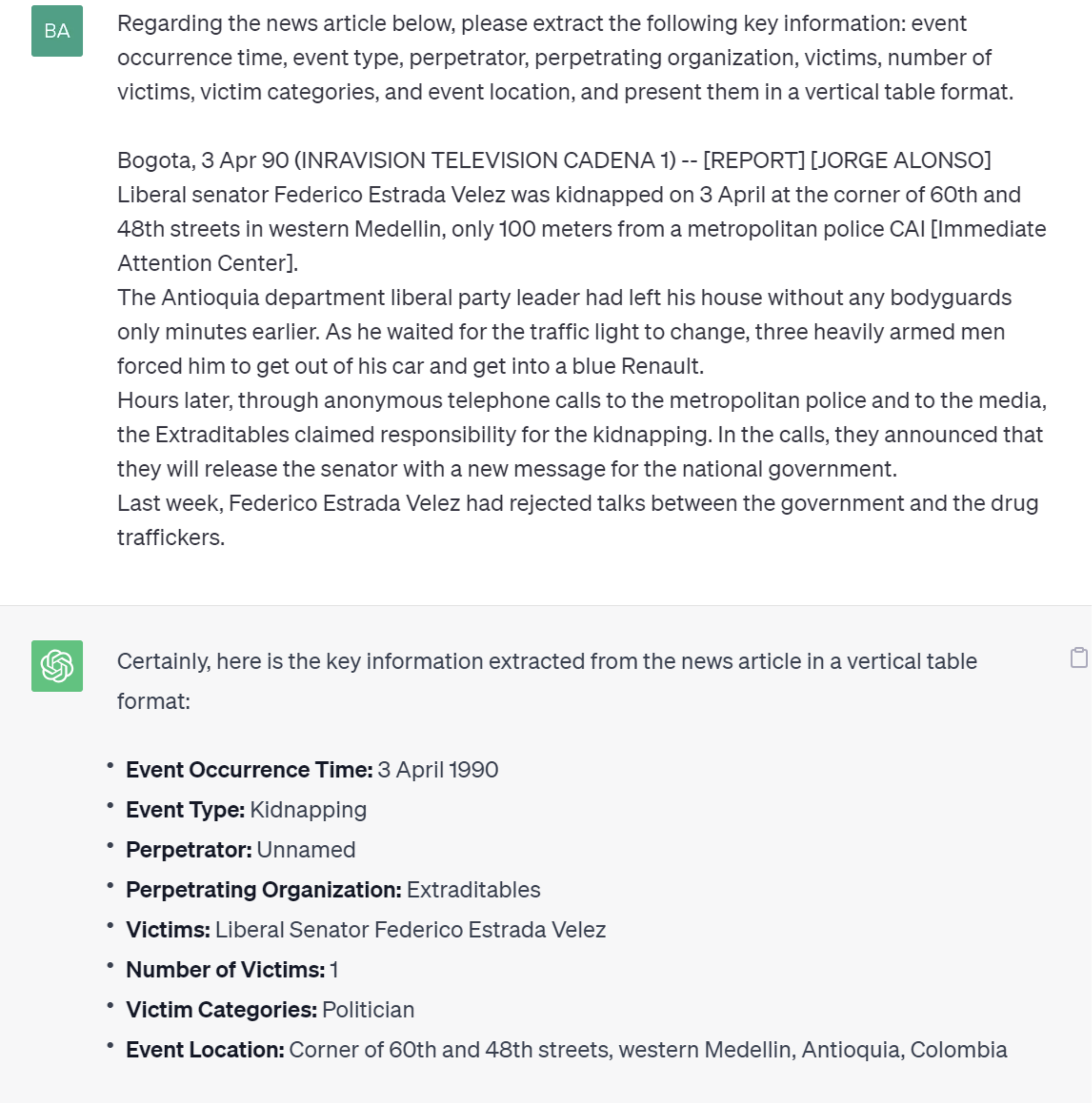
The Antioquia department liberal party leader had left his house without any bodyguards only minutes earlier. As he waited for the traffic light to change, three heavily armed men forced him to get out of his car and get into a blue Renault.

Hours later, through anonymous telephone calls to the metropolitan police and to the media, the Extraditables claimed responsibility for the kidnapping. In the calls, they announced that they will release the senator with a new message for the national government.

Last week, Federico Estrada Velez had rejected talks between the government and the drug traffickers.

信息提取

模板编号：	1
事件发生时间：	03 APR 90
事件类型：	Kidnapping
肇事人：	“Three heavily armed men”
肇事组织：	“The Extraditables”
受害人：	“Federico Estrada Velez ”
受害人数：	1
受害人类别：	Political Figure
事件发生地点：	Colombia: Medellin(city)



Regarding the news article below, please extract the following key information: event occurrence time, event type, perpetrator, perpetrating organization, victims, number of victims, victim categories, and event location, and present them in a vertical table format.

Bogota, 3 Apr 90 (INRAVISION TELEVISION CADENA 1) -- [REPORT] [JORGE ALONSO]
Liberal senator Federico Estrada Velez was kidnapped on 3 April at the corner of 60th and 48th streets in western Medellin, only 100 meters from a metropolitan police CAI [Immediate Attention Center].
The Antioquia department liberal party leader had left his house without any bodyguards only minutes earlier. As he waited for the traffic light to change, three heavily armed men forced him to get out of his car and get into a blue Renault.
Hours later, through anonymous telephone calls to the metropolitan police and to the media, the Extraditables claimed responsibility for the kidnapping. In the calls, they announced that they will release the senator with a new message for the national government.
Last week, Federico Estrada Velez had rejected talks between the government and the drug traffickers.

Certainly, here is the key information extracted from the news article in a vertical table format:

- **Event Occurrence Time:** 3 April 1990
- **Event Type:** Kidnapping
- **Perpetrator:** Unnamed
- **Perpetrating Organization:** Extraditables
- **Victims:** Liberal Senator Federico Estrada Velez
- **Number of Victims:** 1
- **Victim Categories:** Politician
- **Event Location:** Corner of 60th and 48th streets, western Medellin, Antioquia, Colombia

其他应用

- 垃圾邮件过滤
- 文本分类（自动判别文本的类别）
- 音字转换（汉字整句输入法）
- 拼写检查和自动勘校系统
- …..

概要

- 自然语言处理概要
- 自然语言处理的基本方法
- 自然语言处理的挑战性
- 自然语言处理的应用领域
- **自然语言处理发展简史**

自然语言处理简史

- 1940年代末—1960年代中期
  - 机器翻译导向
  - Warren Weaver(49)、GeorgeTown系统(54)、Noam Chomsky(57)
  - 统计方法被放弃
- 1966年：ALPAC(66) 语义障碍
- 1960年代末期—1980年代初期
  - 研究受到AI研究的影响，强调世界知识和语义表示
  - SHRDLU(73)、LUNAR QA(78)
  - TAUM-METEO(76)、SYSTRAN(76)
  - MT产品 如Fujitsi、Hitachi、Siemens

自然语言处理简史

- 1980年代—1990年代前期
  - 强调可计算的语法：语法形式系统
  - 合一运算
  - FUG、HPSG、LFG．．．
  - 欧盟 Eurotra 计划(82)、日本Mu系统以及ODA计划(82)
- 1990年代—2010年代前期
  - 经验主义方法复苏、浅层机器学习技术、数据驱动方法、词汇主义
  - Internet的高速发展
  - 统计语言模型、隐马尔可夫模型、最大熵、条件随机场…

自然语言处理简史

  - 基于词的统计翻译(1990,IBM)、基于短语的统计翻译(Och,2004;Koehn,2003)、基于结构的统计翻译(Chiang,2005)
  - Google Translate的建立和发展(2004---)
- 2010年代--
  - 神经网络方法复苏
  - FFN、CNN、RNN、LSTM、Transformer…
  - 预训练模型 GPT、BERT、XLNET、BART、T5…
  - 预训练 + 精调
  - 大模型 in context Learning → multi-task learning
  - aligning language models with human intentions