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

南京大学
NANJING UNIVERSITY

机器翻译的研究历程
--基于规则的机器翻译

黄书剑

西班牙语是西班牙的官方语言，也是拉丁美洲大多数国家的官方语言，同时还是联合国的工作语言之一。下面展示的是 10 个西班牙语语句以及它们的英语译文，请据此按要求回答问题。

西班牙语	英语
Garcia y una pistola.	Garcia and a pistol.
Carlos Garcia tiene tres asociados.	Carlos Garcia has three associates.
Sus asociados no son fuertes.	His associates are not strong.
Garcia tambien tiene empresas en Europa.	Garcia has companies in Europe too.
Sus clientes están enfadados.	His clients are angry.
Los asociados tambien están enfadados.	The associates are also angry.
Los clients y los asociados son enemigos.	The clients and the associates are enemies.
Sus empresa no venden pistola.	His company do not sell pistol.
Los grupos pequeños no son modernos.	The small groups are not modern.
Los asociados están sorprendidos.	The associates are surprised.

问题一 请将下面（1）中的西班牙语句子翻译成英语，并按要求回答（2）中的问题。

（1）Los asociados enfadados y Garcia no son enemigos.

（2）简单而言，“语序”指的就是不同成分在语句中的排列顺序。请回答，从前面这句话中可看出西班牙语和英语的语序有哪两点不同？

问题二 请将下面的英语句子翻译成西班牙语：

（3）Three small clients sell pistols in Europe too.

问题三 请把下面（4）、（5）两个英语句子翻译成西班牙语（已知 hungry 和 poor 在翻译出来的西班牙语句子中都是 hambrientos），并按要求回答（6）中的问题。

（4）The associates are hungry.

机器翻译的起源

- Warren Weaver, “Translation”, 1949

Indeed, what seems to W.W. to be the most promising approach of all is one based on the ideas expressed in Section 2 above - that is to say, an approach that goes so deeply into the structure of languages as to come down to the level where they exhibit common traits.

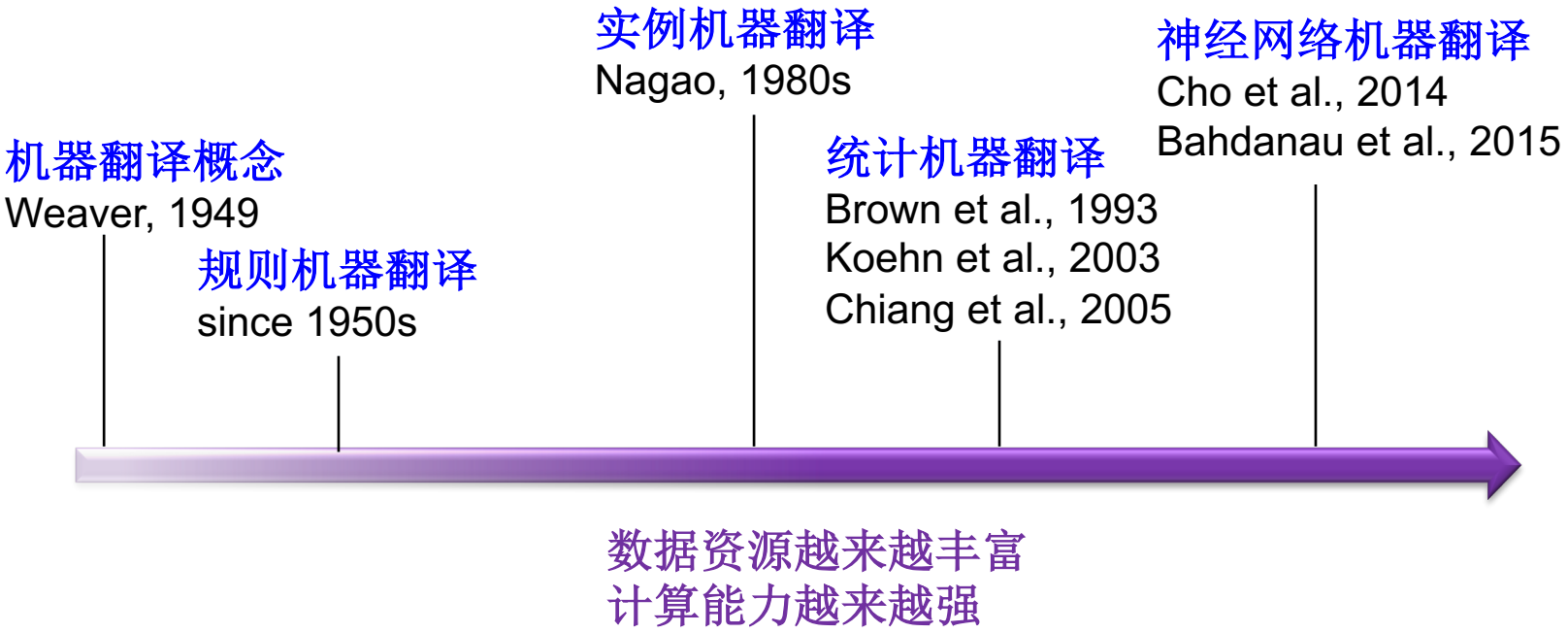
—— Warren Weaver

如何分析？如何更深地分析？如何发现共同特性？
如何以较低的代价具备上述能力？

提纲

- 发展历程简介
  - 基于规则的机器翻译
  - 基于实例的机器翻译
  - 统计机器翻译
  - 神经网络机器翻译

机器翻译的发展


机器翻译概念
Weaver, 1949
规则机器翻译
since 1950s
实例机器翻译
Nagao, 1980s
统计机器翻译
Brown et al., 1993
Koehn et al., 2003
Chiang et al., 2005
神经网络机器翻译
Cho et al., 2014
Bahdanau et al., 2015
数据资源越来越丰富
计算能力越来越强

基于规则的机器翻译（since 1950s）

- **由语言学方面的专家进行规则的制订**
  - 一般包含词典和用法等组成部分
- **例如：**
  - “我” -> (宾语) “me”; (主语) “I”
  - “来自” -> (一般现代时) come from; (第三人称单数) comes from; (过去时) came from
  - “南京大学” -> Nanjing University
  - 汉语主谓宾结构 -> 英语主谓宾结构
  - 我来自南京大学 -> I come from Nanjing University

Georgetown-IBM experiment (since 1952)

Six rules [edit]

- Operation 0 — An exact equivalent for an translated item exists. Any further steps needed.[5]
- Operation 1 — Rearrangement of the position of the words. AB > BA
- Operation 2 — The several choices problem. The result is based on the consecutive words (maximum of three).
- Operation 3 — Also several problems. But the result depends on the previous words (maximum of three).
- Operation 4 — Omissions of the lexical (morphological) item. The source item would be redundant.
- Operation 5 — Insertion of the lexical (morphological) item. The item is not present in the output language.

https://en.wikipedia.org/wiki/Georgetown%E2%80%93IBM_experiment

系统实例

- **南京大学日汉规则翻译系统**
  - http://nlp.nju.edu.cn/homepage/introduction.html
  - 翻译系统包括6万余词的基本词典，10万余词的领域词典，10万余人名地名词典等资源；
  - 句法规则1000余条，动词格框架1800余条，通用格框架60余条，
  - 转换/生成规则大约有800余条;
  - 以及40万句对翻译记忆体

  - 国家七·五科技攻关重大成果奖
  - 江苏省科技进步三等奖

词典示例

日汉翻译引擎词典说明

目　录

日汉翻译引擎词典说明

1、词典构成

翻译词典由以下部分构成：

1） 基本词典（jc_basic.dic）
2） 外来语词典（jc_foreign.dic）
3） 用户词典（jc_custom.dic）
4） 附属词词典（jc_auxi.dic）
5） 活用词尾词典（jc_suffix.dic）
6） 动词格框架词典（jc_case.dic,jc_comcase.dic）
7） 人名词典（jc_person.dic）
8） 地名词典（jc_place.dic）
9） 专业词典（jc_special.dic）

规则示例

日汉翻译引擎规则说明

分词规则

- 解决源语言（日语）的词语切分问题

(1)邻接表

邻接表在分词时用于判断二个词是否可以出现在一个日文句子中相邻位置上，邻接表的形式为二维表，表的纵向为后接信息（参见词典说明 8），表的横向为前接信息（参见词典说明 8），表中元素[i,j]可取 0，1 或 2，其含义为：

① 0：表示后接信息为 i 的词与前接信息为 j 的词不能邻接。

② 1：表示后接信息为 i 的词与前接信息为 j 的词可以邻接，但必须合并为一个词。

③ 2：表示后接信息为 i 的词与前接信息为 j 的词可以作为单独的词邻接。

邻接表以文本方式存贮在磁盘文件 jc_lex.rul 中。

(2)分词规则

分词时用到的个性规则。规则给出了句子片断，以及要切分的位置。例如：“今北京,2”表示在句子中出现串“今北京”时，需要把“今”和“北京”切开，2 表示切分的位置（西文字符单位）。

分词规则放在文件 jc_seg.rul 中。

(3)特殊邻接规则

分析规则

- 分析源语言（日语）的词语之间的关系（句子结构）
  - 形式：〈条件部〉[〈语义检查〉]〈动作部〉

彼 彼 12xxxx NYH,NAR 他
は は 7030xx
バス bus 11xxxx NSM 公共汽车
に に 7010xx
乗っ 乗る 550103 VYMW 乗坐,登上,上当
て て 7020xx
東京 東京 1102xx NTP,NOF 东京
へ へ 7010xx
行っ 行く 550103 VYMW 去
た た 801305
。 。 9901xx 。

```
[s]彼はバスに乗って東京へ行った。(分类=000003,语义码=,深层信息=PAST.VP,表层信息=(null),译词=(null))
  ├ [cvp2]彼はバスに乗って(分类=060303,语义码=,深层信息=ZD.VP,表层信息=て,译词=(null))
  |   ├ [pp]彼は(分类=020002,语义码=,深层信息=TOP,表层信息=は,译词=(null))
  |   |   ├ [np]彼(分类=010001,语义码=,深层信息=(null),表层信息=(null),译词=(null))
  |   |   |   └ 彼(分类=12xxxx,语义码=NYH,深层信息=彼,表层信息=(null),译词=他)
  |   |   └ は(分类=7030xx,语义码=,深层信息=は,表层信息=(null),译词=)
  |   ├ [pp]バスに(分类=020002,语义码=,深层信息=OBJ,表层信息=に,译词=(null))
  |   |   ├ [np]バス(分类=010001,语义码=NSM,深层信息=(null),表层信息=(null),译词=(null))
  |   |   |   └ バス(分类=11xxxx,语义码=NSM,深层信息=bus,表层信息=(null),译词=公共汽车)
  |   |   └ に(分类=7010xx,语义码=,深层信息=に,表层信息=(null),译词=)
  |   └ [vp2]乗って(分类=050302,语义码=,深层信息=ZD.VP,表层信息=て,译词=(null))
  |       ├ 乗っ(分类=550103,语义码=VYMW,深层信息=乗る,表层信息=(null),译词=乘坐)
  |       └ て(分类=7020xx,语义码=,深层信息=て,表层信息=(null),译词=)
  ├ [cvp3]東京へ行った(分类=060002,语义码=,深层信息=PAST.VP,表层信息=た,译词=(null))
  |   ├ [pp]東京へ(分类=020002,语义码=,深层信息=STO,表层信息=へ,译词=(null))
  |   |   ├ [np]東京(分类=010001,语义码=NTP,深层信息=(null),表层信息=(null),译词=(null))
  |   |   |   └ 東京(分类=1102xx,语义码=NTP,深层信息=東京,表层信息=(null),译词=东京)
  |   |   └ へ(分类=7010xx,语义码=,深层信息=へ,表层信息=(null),译词=)
  |   └ [vp3]行った(分类=050002,语义码=,深层信息=PAST.VP,表层信息=た,译词=(null))
  |       ├ 行っ(分类=550103,语义码=VYMW,深层信息=行く,表层信息=(null),译词=去)
  |       └ た(分类=801305,语义码=,深层信息=た,表层信息=(null),译词=)
```

生成规则

- **根据分析结果生成目标语言（汉语）**
  - 形式：

〈条件〉>〈动作〉
〈条件〉::=〈条件项〉{ ＋〈条件项〉}
〈条件项〉::= 〈条件函数〉{ * 〈条件函数〉}
〈动作〉::=〈动作项〉{ *〈动作项〉}

  - 动作定义：

add(W,P)：在位置 P 添加译词 W
del(P)：删除位置 P 的译词
del(W,P)：删除位置 P 包含的译词 W
sit(N [,P])：设置对象 P 或当前结点在当前 NP / CVP 中的位置 N（结点的位置默认值为“-1”）
lian()：取当前词（名词）的字典量词放在该词的译词之前
shock([p])：终止当前规则库的检索（空操作），或在结点 p 禁用对规则库的检索。适用于在格助词的规则操作中禁用格短语的规则，或在助动词的的规则操作中禁用动词短语的规则
alter(P): 把当前结点的译词移至位置 P。
chgcase(C,P)：把位置 P 的深层格改成 C。
copy(P1,P2)：把位置 P1 的译词复制到位置 P2 处，位置 P2 可以带_f 和_b，但 P1 不允许。

基于规则的模型更新

- **扩大词典**
  - “翻译系统包括6万余词的基本词典，10万余词的领域词典，10万余人名地名词典等资源；”
- **制定新的规则**
  - 保证与原先规则兼容
  - 不引入新的错误
  - “句法规则1000余条，动词格框架1800余条，通用格框架60余条，转换/生成规则大约有800余条;”
- **模型更新需要专家进行**

基于规则的机器翻译回顾

- 需要语言学家大量的工作；维护难度大；翻译规则容易发生冲突

“Every time I fire a linguist, the performance of the speech recognizer goes up”

late 1980s – early 1990s

or " Anytime a linguist leaves the group the recognition rate goes up "

Frederick Jelinek (1932-2010)
Researcher in Information Theory, Speech Recognition, and Natural Language Processing
Ph.D. EE MIT 1962
Professor at Cornell 1962-1974
Head of Continuous Speech Recognition group, IBM T.J. Watson 1972-1993
Head of Center for Language and Speech Processing, JHU 1993-2010