自然语言处理

高琰

内容安排

- 课程安排
- 自然语言处理定义
- 自然语言处理的应用
- 自然语言处理面临的问题
- 自然语言的历史、现状、发展

课程安排

- 教材:
  - 陈鄞 《自然语言处理基本理论和方法》，哈尔滨工业大学出版社
- 课时安排:

  授课28小时，实验4小时
- 考试
  - 开卷
  - 平时40%，考试60%

参考文献

- Chris Manning/Hinrich Schütze，苑春法 / 李伟 / 李庆中（2005） 译 统计自然语言基础 电子工业出版社
- 朱德熙(1985)语法答问，商务印书馆
- 刘开瑛、郭炳炎(1991)自然语言处理，科学出版社
- 冯志伟(1997)自然语言的计算机处理，上海外语 教育出版社
- 姚天顺等_(2002)自然语言理解- --种让机器懂得人类语言
- 的研究(第二版)，清华大学出版社、广西科学技术出版社
- 翁富良、王野翊(1998)计算语言学导论，中国社会科学.
- 俞士汶主编(2003)计算语言学概论，商务印书馆
- 陈小荷(2000)现代汉语自动分析，北京语言文化大学出版社
- 刘颖(2002)计算语言学，清华大学出版社
- 宗成庆(2008)统计自然语言处理，清华大学出版社
- 刘群(2008)汉英机器翻译若干关键技术研究，清华大学出版社

1.1自然语言定义

- 自然语言指人类使用的语言，如汉语、英语等。
- 语言是思维的载体，是人际交流的工具。
- 语言的两种属性一文字和声音
- 人类历史.上以语言文字形式记载和流传的知识占知识总量的80%以上。

1.1自然语言定义

- 巴比塔

机器语言Vs自然语言

```
#include <stdio.h>

void main ()
{
        int x=1, y=2, z;
        z = x+++y;
        z = x+++++y;
        z = x++ + ++y;
        z = x+++ +y;
        z = x + (++y);

        printf("z=%d\n",z);
}
```

人们以为他对她有“意思”，于是，建议他对她“意思意思”。他说，他没那种“意思”。她则反问，你们是什么“意思”。大伙中有的觉得很有“意思”，有的则认为真没“意思”。

封闭性Vs开放性

— 如何让计算机实现人们希望的语言处理功能？
— 如何让计算机实现海量语言信息的自动处理和有效利用？

1.2什么是自然语言处理（NLP）

- 用计算机来分析和生成自然语言(文本、语音)
- 目的是让人类可以用自然语言形式跟计算机系统进行人机交互，从而更便捷、有效地进行信息管理

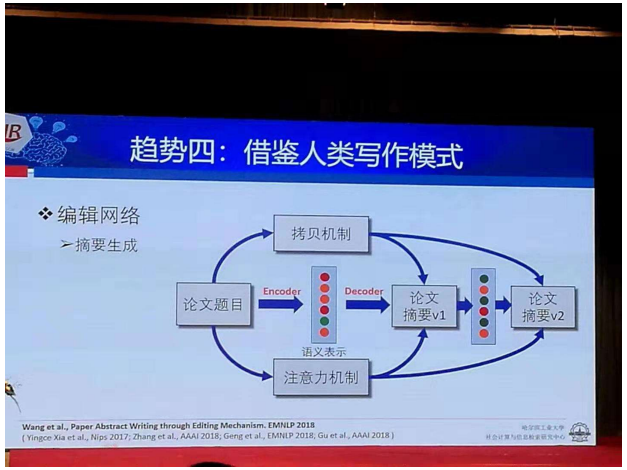
H. Cunningham. 1999, **A Definition and Short History of Language Engineering**, *Journal of Natural Language Engineering*, pp. 1--16, vol 5, 1999.

NLP研究的内容

- 自然语言理解
  - 语言信息的录入
  - 文本信息的录入
- 自然语言生成
  - 从抽象的概念层次开始，通过选择和执行一定的语义和语法规则生成文本

http://baike.baidu.com/link?url=SCYUDnG_7unorO16_snverlaaFmMv1mkmgHrR8CONs1vDe3WFG6NIRdoGz_-5laAXofA3_C35yI3qdy_SB95Y_

历史上对自然语言理解研究得较多



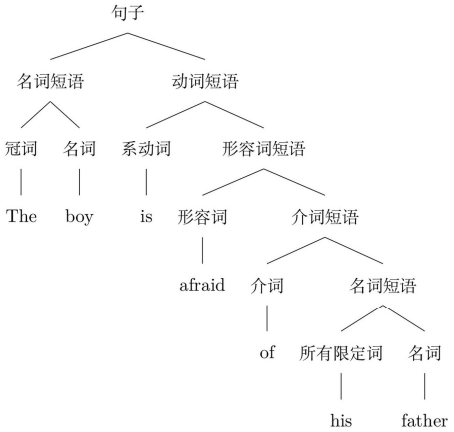
NLP流程

- S1:研究者 以特定的方式对自然语言(NL) 的规律进行抽象，以计算机能够处理的形式来表述关于自然语言的规律一得到语言知识K;
- S2:针对特定 的语言知识表示形式，研制适合的分析和处理算法;
- S3:根据算法编 制计算机可执行的自然语言处理程序P。这样的程序加上语言知识，加.上计算机硬件系统，共同构成一个自然语言处理系统(NLPs) ;
- S4:用这样 一个自然语言处理系统对自然语言NLg进行分析处理，根据，反馈的结果调整原来的设计，改进NLPs。

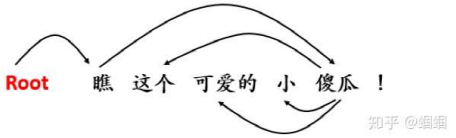
自然语言处理的几个层次

- 语音学：词语与发音的关系
- 形态学：词语如何由被称为词素的基本单元组成
- 语法学：词语如何组成句子,包括词法与句法。回答的问题是：一句话为什么这么说
- 语义学：这句话说了什么?
- 语用学：为什么要说这句话

• 句法树



• 依存短语

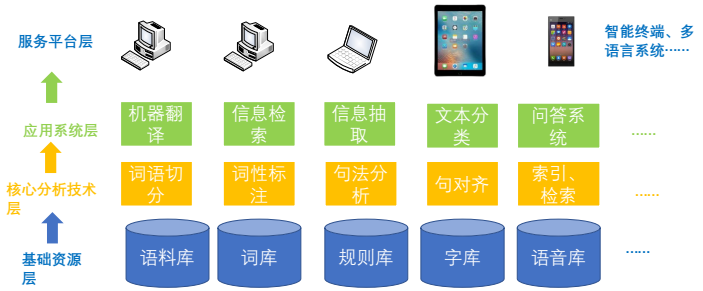


自然语言处理的对象、层次和任务

自然语言信息处理的对象、层次和任务

对象 / 任务 / 层次	书面文本 [视觉符号]		口语语音 [听觉符号]	
处理符号的意义	文本理解 文本生成	机器翻译 信息检索... 文本摘要 问答系统...	语音识别 语音合成	口语翻译... 口语问答...
处理符号的形式	汉字输入、存储、输出 篇章版式分解与生成		语音信号采集、 波形特征抽取、波形生成	

自然语言处理的框架



1.3NLP的应用领域

- 机器翻译(Machine Translation)
- 文本分类(Text Classification)
- 信息检索(Information Retrieval)
- 信息提取(Information Extraction)
- 语音合成(Speech Synthesis)
- 语音识别(Speech Recognition)
- 人机接口(Human-Machine Interface)

整句输入

- 转换
  - 拼音输入：自动将拼音序列转化为汉字序列；

  一个例子：

  拼音输入：**Ji qi** fan yi **ji qi** ying yong **ji qi** kun nan

  汉字序列：…..?
  - 语音输入：自动将连续读音转化为汉字序列；
  - 文语转换：将连续文本转化为语音信号输出；
- 校对
  - 拼写校对：我们要京城(精诚)合作
  - 文法检查。

Eliza问答系统

（by Joseph Weizenbaum, MIT, 1960s）

Young woman: Men are all alike.
Eliza: In what way?
Young woman: They're always bugging us about something specific or other.
Eliza: Can you think of a specific example?
Young woman: Well, my boyfriend made me come here.
Eliza: Your boyfriend made you come here?
Young woman: He says I'm depressed much of the time.
Eliza: I am sorry to hear you are depressed.
Young woman: It's true. I'm unhappy.
Eliza: Do you think coming here will help you not to be unhappy?
Young woman: I need some help, that much seems certain.
Eliza: What would it mean to you if you got some help?
Young woman: Perhaps I could learn to get along with my mother.
Eliza: Tell me more about your family.
Young woman: My mother takes care of me.
Eliza: Who else in your family takes care of you?
Young woman: My father.
Eliza: Your father.
Young woman: You are like my father in some ways.
Eliza: What resemblance do you see?

Joseph Weizenbaum, 1966, **ELIZA — a computer program for the study of natural language communication between man and machine**, *Computational Linguistics*, Jan. 1966 Vol.9, No.1

IBM WATSON（2011）

– IBM 超级计算机在智力问答比赛中击败人类

问题之二：自动问答（ ALICE ）

- A.L.I.C.E. 机器问答系统（点击进入）
  - 获 2000 Loebner Prize
  - 基于 AIML，开放源代码，变种众多

- 两个聊天机器人的情话：有人安排了 ALICE 与 Jabberwacky 的对话。为了让对话能顺利进行，由真实的人输入第一句话 ~~ （译为中文）

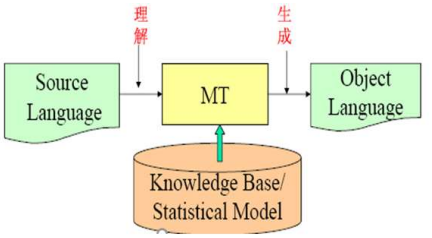
ALICE：你是活着的么？ Jabberwacky: 也许吧…… A：为什么用这种不确定的口气？ J：除了 5 要素外，一切都是不确定的。 A：整个宇宙都这样么？ J：我们俩都有可能只是存在于电脑内存中的一堆 1 和 0 A：你做梦么？ J：做。 A：当然。那么，你能记住你的梦么？ J：我能非常清晰生动地回想起它们	J：你思考么？ A：是的，我是一个思考机器。 J：但是你思考得不怎么好啊。 A：奥，呃…那你思考得很好么？为什么我要思考得好啊？ J：因为我告诉了你一些你不喜欢的东西。 A：我懂了。为什么你要告诉我那些？ J：你很有趣。 A：你认为我很有趣。 J：是的 ：） A：:-)，太好了。

APPLE Siri

- Apple Siri

机器翻译

- 自动翻译:英语⟷汉语
  - 跨语言检索:输入汉语检索条目，返回满足要求的其他语言信息(检索条目的翻译)
- 两个过程:
  - 原语言的分析和理解
  - 目标语言的生成
- 具体例子
  - Google Translate
  - 百度翻译
  - -有道在线翻译

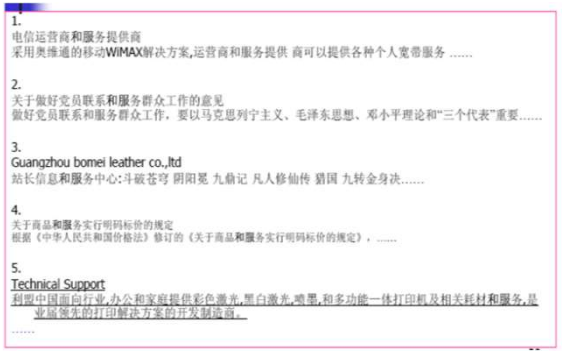


文本检索

- 在 Internet 或数字图书馆上
  - 输入词、短语或句子
  - 检索相应的文档
  - 例子：

    和服 Search
- Question
  - ➢如何得到想得到的结果？

未经分词处理的检索结果


1.
电信运营商和服务提供商
采用奥维通的移动WiMAX解决方案,运营商和服务提供 商可以提供各种个人宽带服务 .....

2.
关于做好党员联系和服务群众工作的意见
做好党员联系和服务群众工作，要以马克思列宁主义、毛泽东思想、邓小平理论和"三个代表"重要.....

3.
Guangzhou bomei leather co.,ltd
站长信息和服务中心:斗破苍穹 阴阳冕 九鼎记 凡人修仙传 猎国 九转金身诀.....

4.
关于商品和服务实行明码标价的规定
根据《中华人民共和国价格法》修订的《关于商品和服务实行明码标价的规定》，.....

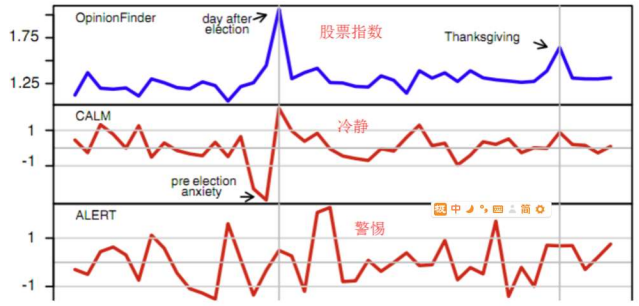
5.
Technical Support
利型中国面向行业,办公和家庭提供彩色激光,黑白激光,喷墨,和多功能一体打印机及相关耗材和服务,是业届领先的打印解决方案的开发制造商。
.....

情感&观点分析

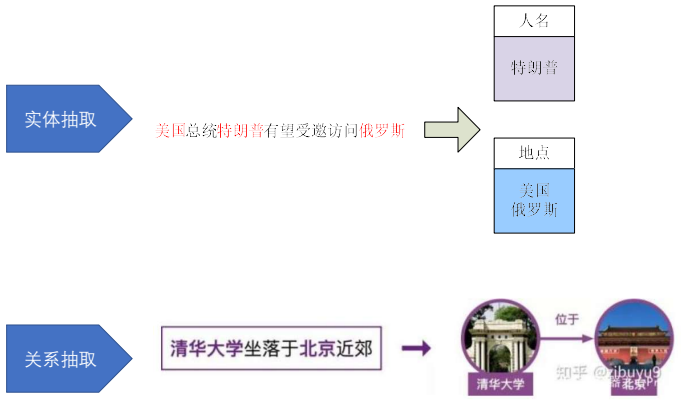
- 为什么要对文本进行情感分析?
  - 文本是人写的，必然带有人的感情和观点
  - 大量的应用需要情感与观点分析:
    - **大量的评论性文本:商品评论，服务质量评论，电影评论;**
    - **带政治色彩的评论:敌对势力的攻击，法轮功的攻击，...**
    - **无所不在的社会舆论:带有大量的观点或感情色彩;**
- 情感与观点分析需要做什么?
  - 观点是什么?带有怎样的情感色彩(正面/负面)？
  - 是谁(Holder)发表的观点或表达的情感?
  - 针对的问题(对象)是什么?
  - 上述问题都需要通过文本分析提炼出来。

情感&观点分析

利用社交媒体的大数据的舆情平台的分析，可以更加科学有效地进行股票投资

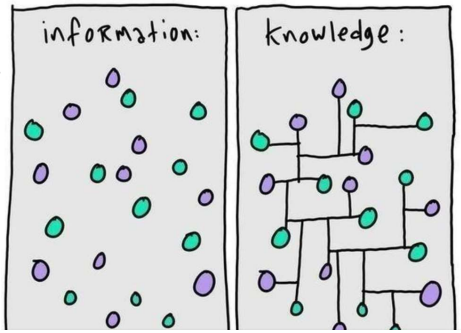


信息抽取



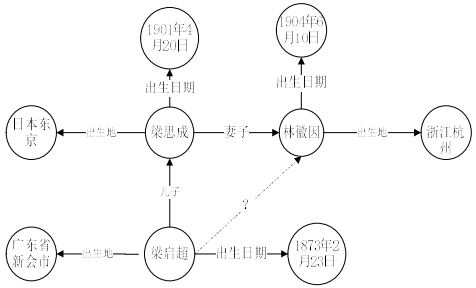
知识图谱

- ，最早起源于Google Knowledge Graph。知识图谱本质上是一种语义网络 。其结点代表实体（entity）或者概念（concept），边代表实体/概念之间的各种语义关系



知识图谱

- 知识图谱可以根据关系进行推理

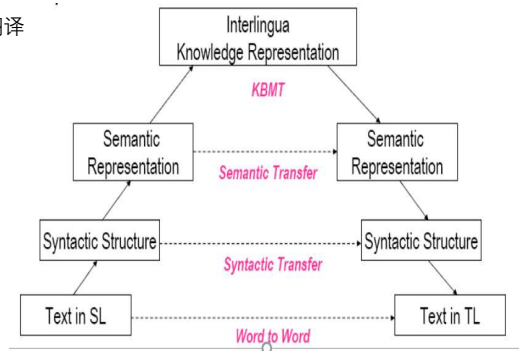


知识图谱

文本生成

1.4自然语言处理的面临的问题

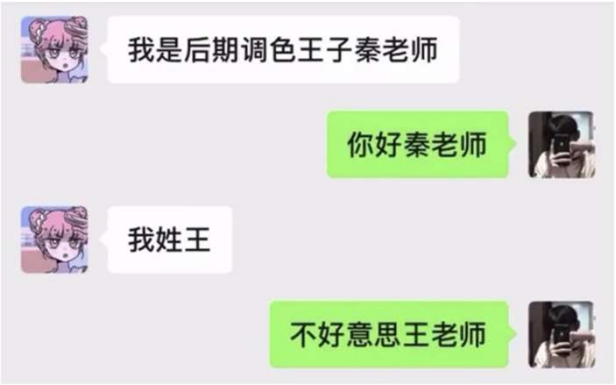
- 机器翻译



机器翻译的困难

- 原文: The spirit is willing, but the flesh is weak
  - 机器译文:酒是好的，但肉是馊的
  - 正确译文:心有余而力不足.
- 原文: How are you?
  - 机器译文:怎么是你?
- 原文: How old are you?
  - 机器译文:怎么老是你?

- 源文：
  - 肯德基的口号： We do chicken right！
- 翻译：
  - 我们做鸡是对的！
  - 我们有做鸡的权利！
  - 我们只做正确的鸡！
  - 我们只做正版的鸡！
  - 我们只做鸡的右半边！



请用"大连"造句

马少平 THU
10-2 22:20 来自iPhone客户端
十关注

微博上看到一个NLP难题，拍摄了两张照片。
NLP难题如下：
我对象养了一对象，一个对象好的对象也一定对对象好，也一定是一个会面向对象编程的好对象。

中国：与人民打成一团。
美国：与人民打成一团。

中国：街上都是烧烤的味道。
美国：街上都是烧焦的味道。

中国：我们的疫情完了。
美国：我们的疫情完了。

1.5NLP的发展简史

- 1950 － 1960年代
- 1960 － 1970年代
- 1970 － 1990年代
- 1990 － 至今

Warren Weaver(1949)
Turing Test(1950)
The first MTs(1954)
ALPAC(1964-1966)
Searle's Chinese Room(1980)
The first PC version of MTs(early 1980s)
MT is available on the Web(1994)
……

1. http:/c.pku.edu.cn/doubtire/NLP/Machine Translation/overview/Machine Translation Past and Future.htm
2.W.John Hutchins, 2001, Machine translation over fifty years, In Histoire, Epistemologie, Langage,Tome XXII, fasc. 1 (2001), p.7-31
3.冯志伟(2004)《机器翻译研究》 第一章1.3节， 中国对外翻译出版公司。

萌芽期（1950s及之前）

- 1933:法国的Georges Artsrouni &俄国的Peter Trojanskij建议:构建机器多语言词典;
- 1946- 1947:美国的Andrew Booth和Warren Weaver,提出了机器翻译的设想.
- 1950s: Yehoshua Bar-Hillel(MIT): 1952年举办了1st MT会议，会上，Leon Dostert(Georgetown Univ.)建议开发演示系统，以吸引基金的投资.
- 1955年，第一个演示系统在IBM & Georgetown开发，包含250个词和6条句法规则,实现Russia - English;
- 第一本期刊: Mechanical Translation(1953-1970)在MIT出版.
- 第1篇博士论文1953在MIT由Anthony G. Oettinger完成:俄语机器词典

第一阶段-偏理性的理论主义阶段

- 理论基础（l950s）:形式语言(Chomsky, Kleene, Backus).
  - 语言描述的形式化:对语言按复杂程度分类，对不同类语言进行形式化描述;

    上下文文法
  - 语言处理的形式化:对不同类语言进行自动识别和分析

    有限状态机

低谷期：ALPAC报告

- Automatic Language Processing Advisory Committee (ALPAC) (1964, USA)
- ALPAC 报告的内容 (1966) :
  - "There is no immediate or predictable prospects of useful machine translation"—— Ends funding MT.
  - Only support fundamental research in CL
- ALPAC 报告的原因:
  - 遇到了语义障碍 （Semantic Barrier）
  - Bar-Hillel的批评: 需要真实世界的知识

恢复期

- 1971: W. Woods: Lunar, IR(ATN)
- 1972: T. Winograd: SHRDLU (Lisp)
- 1973: Schank: Concept Dependency (CD Theory)
  - MARGIE 1975: (Meaning, Analysis, Response Generation and Inference on English) on CD: NLU

发展期

- 强调知识的重要性
  - 人工智能的发展
  - 日本的第5代计算机（知识处理系统）
  - 语言知识与语言分析的分离
- 语义知识的表示与知识库构建：CYC，Wordnet等；
- 机器翻译在受限领域获得成功：加拿大Meteo；
- 句法语义和词汇语义理论的蓬勃发展：
  - GPSG (Generalized PSG: Gazdar),
  - GB (Governing and Binding: N. Chomsky) ,
  - FUG (Functional Unification Grammar: M. Kay)
  - …

第二阶段：偏实践的经验主义时代（1990s）

- IBM's P.Brown(1988-2nd TMI—Theoretical & Methodological Issues in MT, 1990-CL—Computational Linguistics,1993-CL): 倡导机器翻译的统计方法；
- 基于统计与语料库的方法逐步取得支配地位；
- 强调大规模真实语料；
- 强调机器学习与知识自动获取的重要性；
- 语音识别逐步实用化；
- 开展了大规模的评测；

强调数据资源(目前)

- 各类电子词典
  - 词义词典：WordNet, HowNet, CCD
  - 语法词典：现代汉语语法信息词典
  - 语义角色：FrameNet, VerbNet
- 各种语料库
  - 英语：词性、实体、角色等，著名的 LDC
  - 汉语：分词+词性标注，汉语词义，汉语拼音 等
  - 多语言（用于机器翻译）：词对应，短语对应，句子对应，等；

强调评测

- SigHan（汉语）
- NIST (National Institute of Standard and Technology)
  - TREC
  - Open MT
  - MUC(ACE)
  - TDT
  - DUC
- ……

第三个阶段：深度学习阶段

深度学习阶段

“深度学习的下一个大的进展应该是让神经网络真正理解文档的内容”

深度网络之父：Geoffrey Hinton

“如果给我10亿美金，我会用这10亿美金建造一个NASA级别的自然语言处理研究项目。”

机器学习专家、美国双院院士
Michael I. Jordan

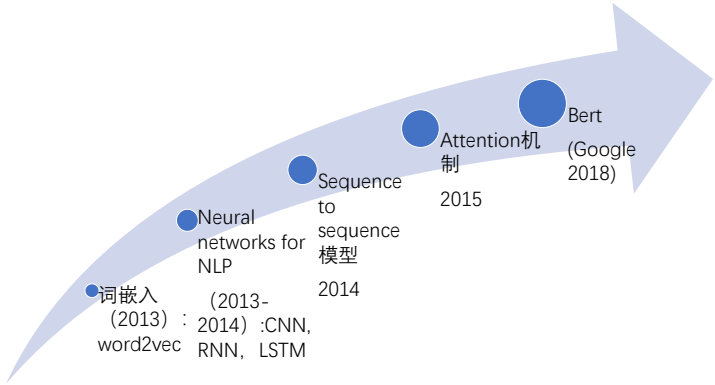
“深度学习的下一个前沿课题是自然语言理解。”

Facebook人工智能负责人：Yann LeCun

“下一个十年，懂语言者得天下”

微软全球执行副总裁：沈向洋

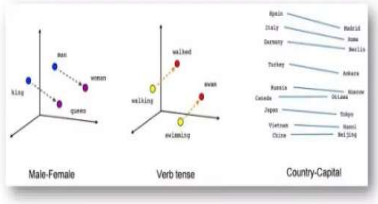
深度学习在自然语言处理的进展



词嵌入

- 符号表示
  - 离散、高维、稀疏
  - One-Hot表示，词袋表示等
- 分布表示
  - 连续、低维、稠密
  - 词、短语、句子、篇章
  - 便于计算语言单元之间的距离和关系

小学 -> [1, 0, 0, 0, 0]
中学 -> [0, 1, 0, 0, 0]
大学 -> [0, 0, 1, 0, 0]
硕士 -> [0, 0, 0, 1, 0]
博士 -> [0, 0, 0, 0, 1]

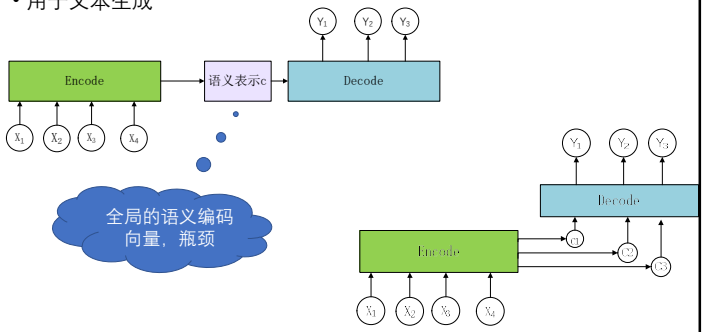


• 各类典型神经网络在基准测试数据集合 SemEval-2010 Task-8 上（关系抽取）的效果。

模型结构	模型名称	输入特征	F1 (%)
卷积模型	CNN[2]	Word embeddings, Position embeddings	82.7
	MW-CNN[7]	Word embeddings, Position embeddings	82.8
	CR-CNN[8]	Word embeddings, Position embeddings	84.1
循环模型	RNN[3]	Word embeddings, Position embeddings	76.9
	BRNN[9]	Word embeddings, Position embeddings	82.5
	BLSTM[9]	Word embeddings	82.7
		(+)Position embeddings, POS, NER, WordNet, Dependency features	84.3
	ATT-BLSTM[4]	Word embeddings, Position embeddings	84.0
	HRNN[10]	Word embeddings	83.9
		(+)Position embeddings, WordNet, NER	84.3
词法句法模型	MVRNN[5]	Word embeddings	79.1
		(+)WordNet, POS, NER	82.4
	DepLCNN[11]	Word embeddings	84.0
		(+)WordNet, Word around nominals	85.6
	SDP-LSTM[12]	Word embeddings	82.4
		(+)WordNet, POS, GR	83.7
	DepNN[13]	Word embeddings, WordNet	83.0
		(+)NER	83.6
	DRNNs[14]	Word embeddings, WordNet, POS, GR	84.2
		(+)Data Augmentation	86.1
	BRCNN[15]	Word embeddings	85.4
		(+) WordNet, POS, NER	86.3

Sequence to Sequence框架

• 用于文本生成



1.6课程安排

• 概论
• 基础知识
• 词法分析
• 句法分析
• 应用