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

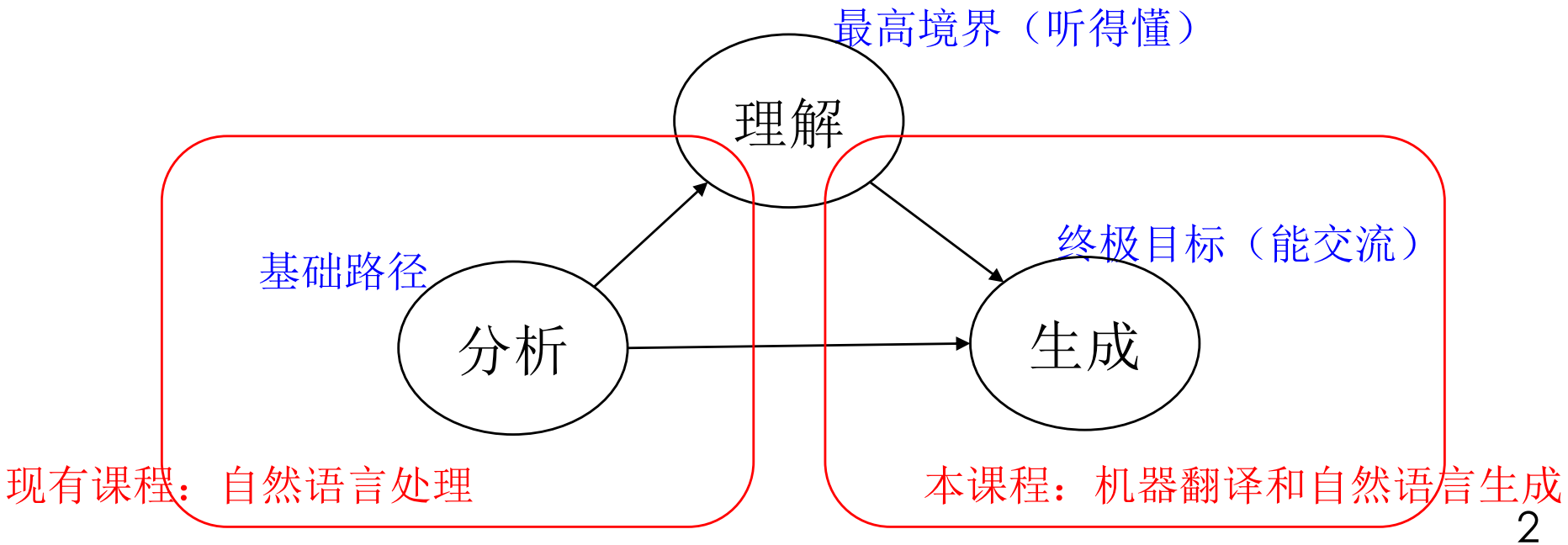
南京大学 NANJING UNIVERSITY

《机器翻译和自然语言生成》课程介绍

黄书剑

自然语言生成问题

- **自然语言处理：分析、理解、生成自然语言**
  - 分析、理解：
    - 难点：表达自由，存在潜在歧义
  - 生成：
    - 难点：信、达、雅



课程意义

- **理论方面**
  - 自然语言研究的重要环节
  - 语言智能研究的前沿
- **应用方面**
  - 自然交互的客观要求
  - 蕴含巨大价值
    - 机器翻译、自动摘要、辅助写作
- **科研训练**
  - 了解基本的自然语言处理和生成方法
  - 尝试寻找、分析、解决问题的过程

预备知识

- **数学基础**
  - 线性代数、微积分
- **机器学习基础**
  - 有监督学习
- **编程实践技能**
  - C++，Python，Deep Learning
- ***自然语言处理基础**

课程内容体系

- 自然语言处理和生成基础
- 语言模型（侧重基本模型和大规模预训练）
- 机器翻译（侧重跨语言的等价性和转换）
- 自动摘要（侧重内容的筛选和重要性评估）
- 复述（侧重同语言的表达多样性和一致性）
- 风格迁移（侧重生成过程中的控制）
- 多模态生成（侧重跨模态信息交互）

教学方式

- **课堂讲授（24学时）**
  - 讲授科学问题、主流技术、语言学分析等
- **技术研讨（8学时）**
  - 针对每个问题进行专题探讨
    - 问题建模、前沿技术、解决方案、实际应用
  - 结合技术报告和研讨

课时安排

- **自然语言分析和生成概述（4学时）**
  - 介绍课程的主要目标和思路，回顾自然语言处理分析和生成的基本方法，包括
  - 课程介绍
  - 自然语言分析和结构化预测
  - 语言模型
  - 生成的基本方法
  - 生成结果的评估

课时安排

- **机器翻译（8学时）**
  - 介绍机器翻译的发展历史和最新研究方向和进展，包括：
    - 传统机器翻译方法
    - 神经网络机器翻译
    - 领域自适应、低资源、无监督机器翻译
    - 非自回归机器翻译
    - 人机交互机器翻译
    - 机器翻译质量评估
  - 课堂技术报告和研讨

课时安排

- **自动摘要（4学时）**
  - 介绍自动摘要的发展历史和最新研究进展，包括：
    - 抽取式和生成式摘要模型
    - 层次注意力机制
    - 篇章上下文建模
  - 课堂技术报告和研讨
- **复述判别和生成（4学时）**
  - 介绍自动摘要的发展历史和最新研究进展，包括：
    - 句子表示方法
    - 语义一致性
  - 课堂技术报告和研讨

课时安排

- **风格迁移（2学时）**
  - 介绍风格迁移相关的技术和最新研究进展，包括:
    - 属性识别和解耦
    - 受控生成等
  - 课堂技术报告和研讨
- **多模态生成（2学时）**
  - 介绍多模态生成的发展历史和最新研究进展，包括:
    - 跨模态的数据处理和知识表示
    - 跨模态表示一致性等
  - 课堂技术报告和研讨

教学资料

- **教材**
  - 无成书教材，以教师自制讲义为主
- **参考书籍**
  - 《机器翻译-基础与模型》，肖桐、朱靖波著，电子工业出版社 （with online version）
  - 《现代自然语言生成》，黄民烈、黄斐、朱小燕著，电子工业出版社
  - 《机器翻译》，李沐 刘树杰 张冬冬 周明著，高等教育出版社
  - 《统计机器翻译》，(德)Philipp Koehn著，宗成庆，张霄军译，电子工业出版社
  - 相关方向前沿学术论文

考核方式

- **技术研讨（60%）**
  - 对前沿问题和文献的阅读、报告和讨论
    - 现场报告和讨论
    - 综述和分析（技术报告）
    - 3人小组
- **课程项目（40%）**
  - 利用前沿技术进行某专项问题的实践
    - 提高实践能力（demo）
    - 在实践中尝试和探索新的解决方案（项目报告）
    - 独立完成（分组须提出申请）

ACL Student Research Workshop

- Selected papers in 2019 [https://sites.google.com/view/acl19studentresearchworkshop/accepted-papers]

- Paraphrases as Foreign Languages in Multilingual Neural Machine Translation
- Improving Mongolian-Chinese Neural Machine Translation with Morphological Noise
- Unsupervised Pretraining for Neural Machine Translation Using Elastic Weight Consolidation
- From Bilingual to Multilingual Neural Machine Translation by Incremental Training
- Normalizing Non-canonical Turkish Texts Using Machine Translation Approaches
- English-Indonesian Neural Machine Translation for Spoken Language Domains
- Automatic Generation of Personalized Comment Based on User Profile
- Using Semantic Similarity as Reward for Reinforcement Learning in Sentence Generation
- Natural Language Generation from Abstract Semantic Representation for Brazilian Portuguese

学习材料

- Python编程
  - cs224n: Week 3: Python Review http://web.stanford.edu/class/cs224n
- 线性代数、微积分
  - Notes from cs231n http://cs231n.stanford.edu/handouts/derivatives.pdf

学习材料

- **机器学习基础**
  - 机器学习 周志华 2-3 章
  - 神经网络与深度学习 邱锡鹏 2-3 章
    https://nndl.github.io/
  - Machine Learning Andrew Ng Week1-3
    https://www.coursera.org/learn/machine-learning

学习材料

- **神经网络**
  - 机器学习 周志华 6 章
  - 神经网络与深度学习 邱锡鹏 4-6 章
    https://nndl.github.io/
  - CS224n: Natural Language Processing with Deep Learning Stanford Week 4-5
    http://web.stanford.edu/class/cs224n/
  - Machine Learning Andrew Ng Week 4-5
    https://www.coursera.org/learn/machine-learning

***在线课程中均包含与编程相关的实践环节，可同步使用，作为实践练习**

联系我们

- **黄书剑**
  - huangsj@nju.edu.cn
- **朱文昊**
  - zhuwh@smail.nju.edu.cn
- **课程网站：**
  - 用于登记课程报告、讨论交流信息
  - https://cslab-cms.nju.edu.cn/
  - 课程邀请码： 2BZ6O
- **QQ群：**
  - 431234713