自然语言处理（CS310）课程大纲

课程详述

COURSE SPECIFICATION

以下课程信息可能根据实际授课需要或在课程检讨之后产生变动。如对课程有任何疑问，请联系授课教师。

The course information as follows may be subject to change, either during the session because of unforeseen circumstances, or following review of the course at the end of the session. Queries about the course should be directed to the course instructor.

1.	**课程名称 Course Title**	自然语言处理 Natural Language Processing
2.	**授课院系 Originating Department**	计算机科学与工程系 Department of Computer Science and Engineering
3.	**课程编号 Course Code**	CS310
4.	**课程学分 Credit Value**	3
5.	**课程类别 Course Type**	专业选修课 Major Elective Courses
6.	**授课学期 Semester**	春季 Spring
7.	**授课语言 Teaching Language**	英文 English
8.	**授课教师、所属学系、联系方式（如属团队授课，请列明其他授课教师） Instructor(s), Affiliation& Contact （For team teaching, please list all instructors）**	徐炀，副教授，计算机科学与工程系，xuyang@sustech.edu.cn Yang Xu, Associate Professor, Department of Computer Science and Technology, xuyang@sustech.edu.cn

9.	实验员/助教、所属学系、联系方式 **Tutor/TA(s), Contact**	待公布 To be announced
10.	选课人数限额(可不填) **Maximum Enrolment（Optional）**	

11.	授课方式 **Delivery Method**	讲授 **Lectures**	习题/辅导/讨论 **Tutorials**	实验/实习 **Lab/Practical**	其它(请具体注明) **Other（Please specify）**	总学时 **Total**
	学时数 **Credit Hours**	32	0	32	0	64

12.	先修课程、其它学习要求 **Pre-requisites or Other Academic Requirements**	CS303 人工智能 Artificial Intelligence OR an equivalent course in another university
13.	后续课程、其它学习规划 **Courses for which this course is a pre-requisite**	无
14.	其它要求修读本课程的学系 **Cross-listing Dept.**	无。不接受跨系选课。 None. Not applicable for other departments other than CSE.

教学大纲及教学日历 SYLLABUS

15. 教学目标 **Course Objectives**

This course provides a thorough introduction to the technology of natural language processing (NLP) and the research field of computational linguistics (CL). The two terms, NLP and CL, while mostly used interchangeably, carry slightly different meanings in this course: NLP refers to the technology that enable computers with the capability of processing, understanding, and producing human language in intellectual ways, and thus, is more **application-oriented**. CL refers to the inter-disciplinary research field that study human languages with computational methods, from multiple perspectives, such as physiology, psychology, linguistics, cognitive sciences, culture and humanity etc., and thus is more **theory-flavored**. This course covers a weighted mixture of NLP (70%) and CL (30%), including topics such as the computational word semantics, the extraction of structural text information, automatic translation, question answering, dialogue with user inputs, psycholinguistics, cognitive sciences, and many more. This course focus on teaching the fundamentals, with self-contained lectures and associated readings and implementation practices.

本课程涵盖自然语言处理（natural language processing, NLP）技术和计算语言学（computational linguistics, CL）这两大紧密联系又可加以区分的领域。本课程对 NLP 和 CL 之定义有如下区别：NLP 指使计算机能够以智能方式处理、理解和生成人类语言的技术，主要面向可用系统的搭建。 CL 指用计算机方法研究人类语言的跨学科研究，涵盖语言学、心理学、认知科学、文化和人文等多个学科，理论意味更强。就大体内容而言，本课程是 NLP (70%) 和 CL (30%) 的加权混合，具体主题包括词汇语义、结构文本信息提取、自动翻译、问答、与用户输入对话、心理语言学、认知科学等等。本课程重视理论与实践的结合。

16. 预达学习成果 Learning Outcomes

After taking this course, the students should be able to:

1. Understand the fundamental concepts, technological philosophy, common problems and open tasks in the field of natural language processing.
2. Understand the basic concepts in linguistics and the underlying human language phenomena that can be studied with computational methods.
3. Use the proper technology and algorithms to solve common computational tasks related to language data.
4. Build NLP pipelines that solve and evaluate common NLP tasks using programming frameworks.

学习目标

1. 了解自然语言处理领域的基本概念、技术理念、常见问题和开放任务。
2. 了解语言学的基本概念以及计算方法，用以研究广泛的人类语言现象。
3. 能够使用合适的技术和算法，以解决常见的与语言数据相关的计算任务。
4. 掌握搭建 NLP 工作流程的能力，能够使用编程框架解决和评估常见的 NLP 任务。

17. 课程内容及教学日历 （如授课语言以英文为主，则课程内容介绍可以用英文；如团队教学或模块教学，教学日历须注明主讲人）

Course Contents (in Parts/Chapters/Sections/Weeks. Please notify name of instructor for course section(s), if this is a team teaching or module course.)

64 hours in total. 2 hours lecture and 2 hours lab for each week.

Grading percentage

Assignments: 55% (A1: 5%; A2-A6: 10% each)

Final Project: 25% (Report 15%; Presentation 10%)

Lab: 15% (Attendance 5%; Practice 10%)

Attendance to lecture: 5%

Schedule

Week 1: Introduction

- Introduction to NLP and CL
- Python basics and basic text processing

[Lab] Setup Python and packages; Practice text processing; Text frequency information

Week 2: Word Vectors and Neural Networks

- Logistic regression; Bag of Words
- Neural network; computational graph; backprop; loss function; training/testing workflow

[Lab] PyTorch tutorial for building neural network models; Training/testing workflow go-through

[A1] Neural-network-based text classifier

Week 3: Recurrent Neural Network

[Lab] PyTorch implementation of RNN (LSTM); RNN for classification task; Explore hidden layers in LSTM

Week 4: Language Models

- N-gram models; Neural language models

[Lab]: Data preparation for implementing neural LMs

[A2] Neural language models: Word2vec (skip-gram) and causal LM (BiLSTM)

Week 5: Sequence Labeling (Part-of-speech tags and named entities)

[Lab] Data preparation for sequence labeling tasks

[A3] Named entity recognition (NER) task

Week 6: Context-Free Grammar and Parsing

[Lab] Processing treebank data; Syntax tree visualization

Week 7: Dependency Parsing

[Lab] Data preparation and processing for dependency parsing task

[A4] Neural dependency parsing

Week 8: Attention and Transformer

[Lab] Manual implementation of attention; understanding position embedding

Week 9: Sequence to sequence and translation

- Seq2seq architecture: source and target; encoder and decoder

[A5] Seq-seq translation with transformers

Week 10: Pretraining Transformer-based Models

[Lab] HuggingFace transformers library

[Custom project out]

Week 11: Large Language Models and Prompting

- History of GPTs
- Few-shots; one-shot; zero-shot

[Lab] Deployment LLMs locally (GPU or CPU); Play with prompts

[A6] Fine tuning an LLM (BERT-like)

Week 12: Natural Language Generation

- Curious cases of NLG; Sampling methods: Greedy, beam search, temperature etc.
- Human judgements

[Lab] Experiment with different sampling methods; Evaluate with human judgements

Week 13: Reinforcement Learning with Human Feedback and Computational Ethics

- Instruction fine-tuning; alignment with human intention

[Lab] Project update and assistance

Week 14: Limits and Future of LLMs and NLP

[Lab] Experiment with BIG-Bench using locally deployed LLMs.

Week 15: Cognitive Science and Language

[Lab] Project update and assistance

Week 16: Project Report and Presentation

[Lab] Project report and presentation

18. 教材及其它参考资料 Textbook and Supplementary Readings

Speech and Language Processing (3rd ed. draft), Dan Jurafsky and James H. Martin

Free online version available: https://web.stanford.edu/~jurafsky/slp3/

课程评估 ASSESSMENT

19.

评估形式 Type of Assessment	评估时间 Time	占考试总成绩百分比 % of final score	违纪处罚 Penalty	备注 Notes
出勤 Attendance				
课堂表现 Class Performance		5%		Attendance in lecture
小测验 Quiz				
课程项目 Projects				
平时作业 Assignments		55%		6 Programming assignments A1: 5%; A2-A6: 10% each
期中考试 Mid-Term Test				
期末考试 Final Exam				
期末报告 Final Presentation		25%		Report 15%; Presentation 10%

其它（可根据需要改写以上评估方式） **Others (The above may be modified as necessary)**		15%		Attendance 5%; Practice 10%

20. 记分方式 **GRADING SYSTEM**

☑ A. 十三级等级制 **Letter Grading**
☐ B. 二级记分制（通过/不通过） **Pass/Fail Grading**

课程审批 REVIEW AND APPROVAL

21. **本课程设置已经过以下责任人/委员会审议通过**
This Course has been approved by the following person or committee of authority

课程详述

COURSE SPECIFICATION

以下课程信息可能根据实际授课需要或在课程检讨之后产生变动。如对课程有任何疑问，请联系授课教师。

The course information as follows may be subject to change, either during the session because of unforeseen circumstances, or following review of the course at the end of the session. Queries about the course should be directed to the course instructor.

1.	**课程名称 Course Title**	自然语言处理 Natural Language Processing
2.	**授课院系 Originating Department**	计算机科学与工程系 Department of Computer Science and Engineering
3.	**课程编号 Course Code**	CS310
4.	**课程学分 Credit Value**	3
5.	**课程类别 Course Type**	专业选修课 Major Elective Courses
6.	**授课学期 Semester**	春季 Spring
7.	**授课语言 Teaching Language**	英文 English
8.	**授课教师、所属学系、联系方式（如属团队授课，请列明其他授课教师） Instructor(s), Affiliation& Contact （For team teaching, please list all instructors）**	高汝霆，助理教授，计算机科学与工程系，kot@sustech.edu.cn TOM Ko Yu Ting, Assistant Professor, Department of Computer Science and Technology, kot@sustech.edu.cn

9.	实验员/助教、所属学系、联系方式 **Tutor/TA(s), Contact**	待公布 To be announced
10.	选课人数限额(可不填) **Maximum Enrolment（Optional）**	

11.	授课方式 **Delivery Method**	讲授 **Lectures**	习题/辅导/讨论 **Tutorials**	实验/实习 **Lab/Practical**	其它(请具体注明) **Other（Please specify）**	总学时 **Total**
	学时数 **Credit Hours**	32	0	32	0	64

12.	先修课程、其它学习要求 **Pre-requisites or Other Academic Requirements**	CS303 人工智能 Artificial Intelligence OR an equivalent course in another university
13.	后续课程、其它学习规划 **Courses for which this course is a pre-requisite**	无
14.	其它要求修读本课程的学系 **Cross-listing Dept.**	无。不接受跨系选课。 None. Not applicable for other departments other than CSE.

教学大纲及教学日历 SYLLABUS

15. 教学目标 **Course Objectives**

This course covers all relevant knowledge in automatic speech recognition and partially introduces natural language processing. Basic concepts in machine learning, e.g. Bayesian decision theory, maximum-likelihood training, pattern classification will also be covered. The most important components in speech recognition, e.g. acoustic model, language model and pronunciation dictionary will be discussed separately. Some very popular language tasks, e.g. machine translation, information extraction, text classification will also be discussed. At the end of the course, the students should know how to design a real-life application based on the components they learned.

16. 预达学习成果 **Learning Outcomes**

After taking this course, the students should be able to:

1. Understand the basic concepts in pattern classification and machine learning, as well as all the relevant components in automatic speech recognition.
2. Build a speech recognition system from scratch. If time is allowed, they can also learn how to implement the system in a mobile device.
3. Understand some speech-related techniques like speaker verification, language identification and text-to-speech.
4. Understand the basic in natural language processing like machine translation, information extraction and text classification.
5. Understand how deep learning is applied in speech recognition and NLP tasks.

17. **课程内容及教学日历 （如授课语言以英文为主，则课程内容介绍可以用英文；如团队教学或模块教学，教学日历须注明主讲人）**

Course Contents (in Parts/Chapters/Sections/Weeks. Please notify name of instructor for course section(s), if this is a team teaching or module course.)

64 hours in total. 2 hours lecture and 2 hours lab for each week.

Week 1: Introduction

- Introduction to course
- Introduction to automatic speech recognition and natural language processing

[Lab] Introduction to linux server, shell script.

Week 2: Bayesian decision theory

- Introduction
- The Normal Density
- Maximum likelihood estimation
- Bayesian parameter estimation

[Lab] Introduction to python

Week 3: Feature Extraction

- Speech production

- Speech perception
- MFCC

[Lab] Installation of Kaldi (speech tool) on linux server.

Week 4: Acoustic Modeling I

- Hidden Markov Models
- Context dependent modeling units
- Decision Tree based clustering

[Lab] Run a complete ASR system and read the result .

Week 5: Decoding, Alignment, and WFSTs

- Alignment generation
- ASR decoding
- Weighted finite state transducers (WFST)
- N-gram language model

[Lab] Implement a language model and convert it to a WFST

Week 6: Acoustic Modeling II

- Feed forward neural network
- Recurrent neural network

[Lab] Train acoustic model with GPU

Week 7: Speaker Adaptation

- Introduction to Speaker Adaptation
- Speaker Adaptation in GMM
- Speaker Adaptation in DNN

[Lab] Paper reading and presentation

Week 8: Data Augmentation

- Speed perturbation
- Reverberation
- Spectrum augmentation

[Lab] Implement the spectrum augmentation on Kaldi

Week 9: Speaker Verification

- I-vector
- Speaker embeddings

[Lab] Implement the speaker verification example on Kaldi

Week 10: Regular Expressions and Text Normalization

- What is a language?
- Edit distance

[Lab] Download and execute an open-source text classification tool

Week 11: Text classification

- One hot encoding
- Word2Vec

[Lab] Convert the existing text classification tool to English

Week 12: Vector Semantics

- Embeddings

[Lab] Understand the importance of text normalization.

Week 13: Sentimental Classification

- The basic framework

[Lab] Learn BERT from Google

Week 14: Machine Translation

- Self-attention

[Lab] Implement a machine translation system

Week 15: Information Extraction

- Slot fitting

[Lab] Paper reading and presentation

Week 16: Summary & Revision

[Lab] Revision, Q&A.

18. **教材及其它参考资料 Textbook and Supplementary Readings**

Automatic Speech Recognition: A Deep Learning Approach 2015th edition by Dong Yu and Li Deng

Various articles in journals and conference proceedings given during the lectures.

课程评估 ASSESSMENT

19.

评估形式 **Type of Assessment**	评估时间 **Time**	占考试总成绩百分比 **% of final score**	违纪处罚 **Penalty**	备注 **Notes**
出勤 **Attendance**				
课堂表现 **Class Performance**		20%		Attendance in lecture and lab

小测验 **Quiz**				
课程项目 **Projects**				
平时作业 **Assignments**		50%		Project, programs and reports.
期中考试 **Mid-Term Test**				
期末考试 **Final Exam**		30%		Unseen exam
期末报告 **Final Presentation**				
其它（可根据需要改写以上评估方式） **Others (The above may be modified as necessary)**				

20. 记分方式 **GRADING SYSTEM**

☑ A. 十三级等级制 **Letter Grading**
☐ B. 二级记分制（通过/不通过） **Pass/Fail Grading**

课程审批 REVIEW AND APPROVAL

21. 本课程设置已经过以下责任人/委员会审议通过

This Course has been approved by the following person or committee of authority