CS310 Natural Language Processing
自然语言处理
Lecture 00 - Course Introduction

Instructor: Yang Xu

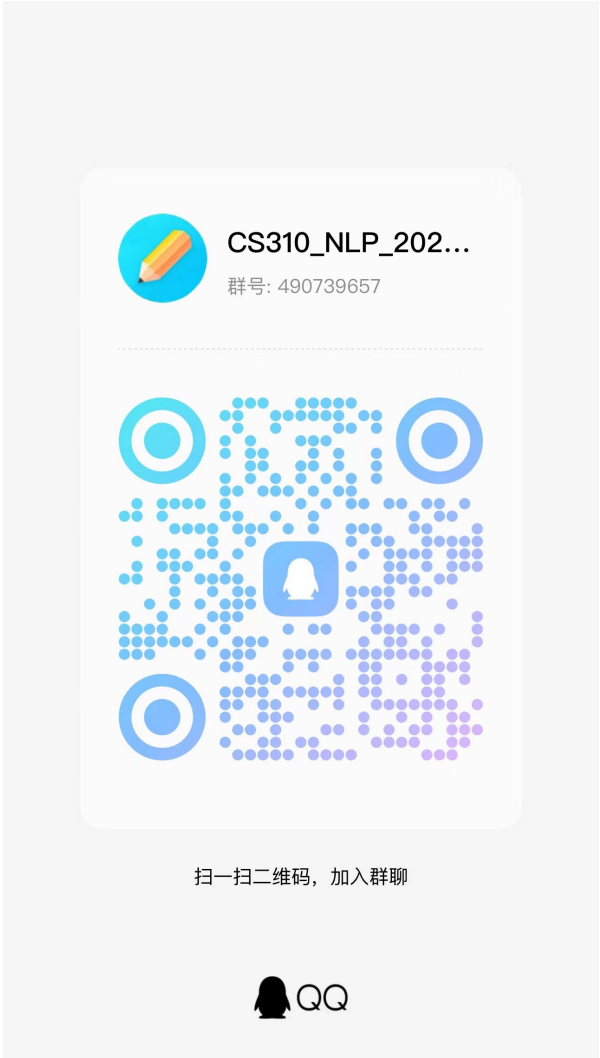
主讲人：徐炀

xuyang@sustech.edu.cn

Table of Content

- What is NLP ?
- What will be taught in this course?
- Programming tools needed
- Mathematical knowledge needed

Join the QQ group:



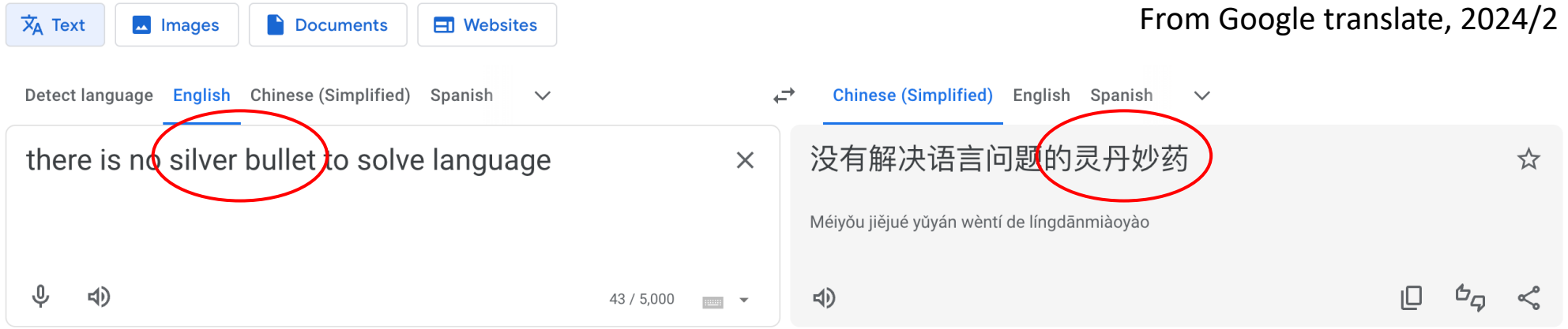
What is NLP?

- *"NLP is a branch of artificial intelligence (AI) that enables computers to **comprehend**, **generate**, and **manipulate** human language."*
  -- oracle.com
- *"NLP is an interdisciplinary subfield of computer science and linguistics … giving computers the ability to **support** and **manipulate** human language"*
  -- wikipedia.org
- *"Computational linguistics is the scientific study of language from a computational perspective. Computational linguists are interested in **providing** computational models of various kinds of linguistic **phenomena**."*
  -- aclweb.org (Association for Computational Linguistics)

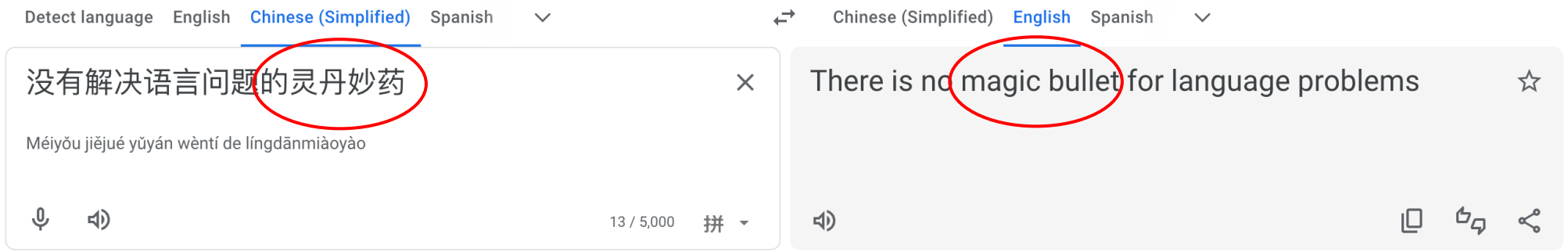
What is NLP? -- An example

- No silver bullet to “solve” language

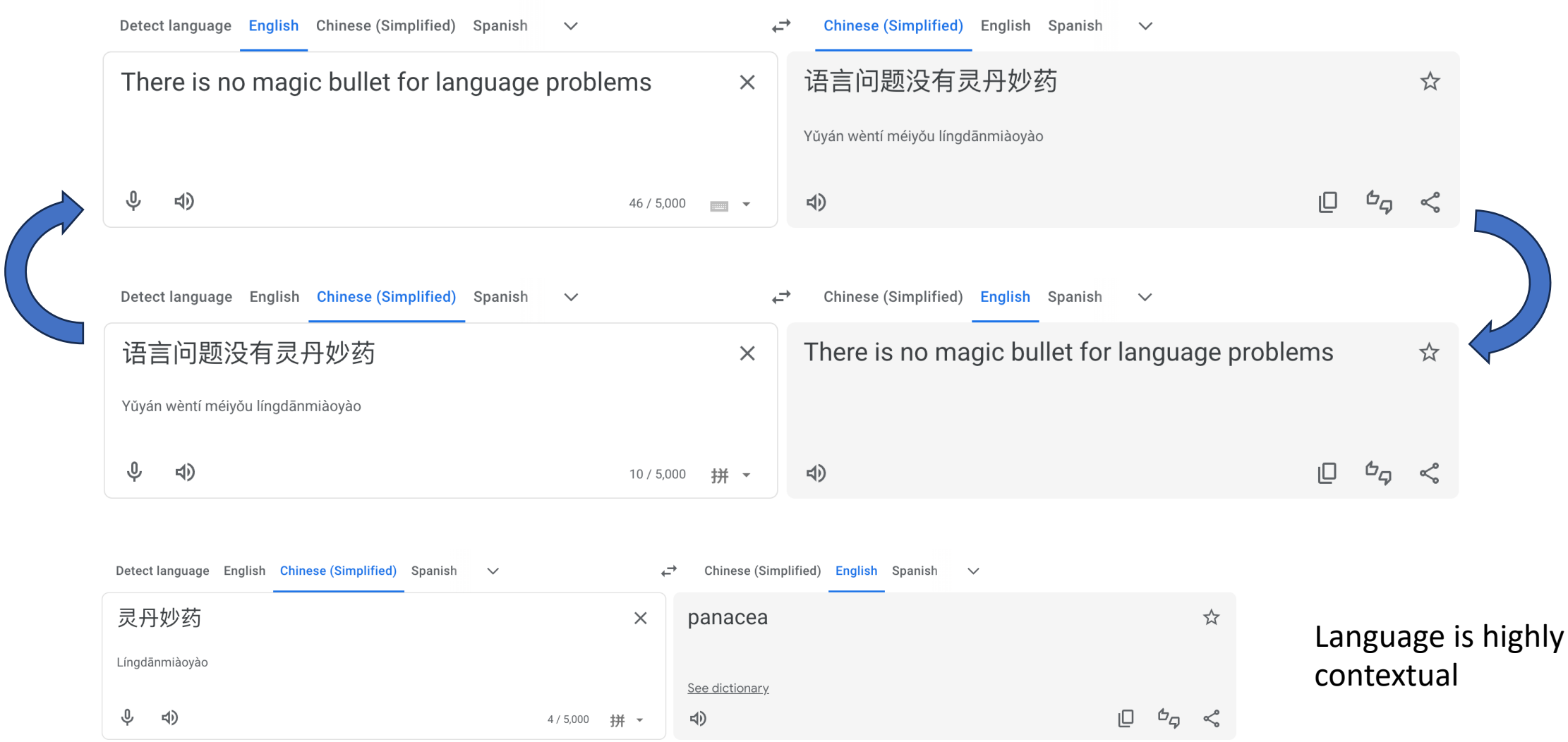
From Google translate, 2024/2



From Google translate, 2024/2



What is NLP? -- An example



Language is highly contextual

What is NLP? -- An example

- Ask GPT to translate

You

Translate this sentence to Chinese and provide at least five versions of translation: There is no silver bullet to solve language

ChatGPT

Sure, here are five translations of the sentence "There is no silver bullet to solve language" into Chinese:

1. 没有解决语言问题的灵丹妙药。
2. 解决语言问题没有银弹。
3. 没有一劳永逸解决语言问题的方法。
4. 解决语言问题并非一劳永逸之策。
5. 没有捷径可以解决语言问题。

What is NLP? -- An example

- Translate to 文言文

You
翻译成文言文呢?

ChatGPT
当然，请看：

1. 蓋語言難解，無一捷徑可赴也。
2. 治語言之難，無以一技所搏也。
3. 語言之難，無以金玉良方可去也。
4. 語言之難，非有一劍可斬也。
5. 語言之難，未有以金錢買之良藥也。

用文言文翻译这句话：there is no silver bullet to solve language

无以银弹解語言之困。

请提供五个版本的翻译

1. 言語之困，非一藥可治。
2. 解語言之謎，無速效良方。
3. 語言之難，非一蹴可幾。
4. 語言之結，無直解之策。
5. 語言之障，無一劑可救。

请提供五个版本的翻译

当然，这里是“没有银弹来解决语言问题”的五种不同文言文风格的翻译：

1. 无银丸破言语之困。
2. 语难无神丹妙药。
3. 言事之难，非银弹所能解。
4. 谈言译语，求索无灵丹。
5. 语疾无神医，银弹亦徒劳。

What is NLP? -- An example

You

为什么深圳市的学校开学早一点?

ChatGPT

深圳市学校提前开学可能是因为多种因素的综合考虑：

1. 气候因素：深圳地处南方，气候相对温暖，冬季较短，春季来得早，因此学校可能会提前开学，以适应天气变化。
2. 经济因素：深圳是中国经济特区之一，经济发展较快，社会竞争压力较大，学校可能希望提前开学，以保证教学时间充足，学生能够更好地适应竞争环境。
3. 教学需求：深圳的学校可能有自己的教学计划和安排，提前开学可能有利于完成教学任务，为学生提供更多的学习机会和资源。
4. 政策因素：地方政府可能会根据当地的实际情况和教育需求，制定相关的教育政策，包括开学时间等方面的规定。

综上所述，深圳市学校提前开学可能是出于气候、经济、教学需求和政策等多方面考虑的结果。

What will be taught in CS301?

- Technology of NLP and research of Computational Linguistics (CL)
- NLP (70%) => application oriented
- CL (30%) => theory flavored
- **Topics**: computational word semantics, the extraction of structural text information, automatic translation, question answering, dialogue with user inputs, psycholinguistics, cognitive sciences
- 词汇语义、结构文本信息提取、自动翻译、问答、用户输入对话、心理语言学、认知科学

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

Schedule

Week 2: Word Vectors and Neural Networks [A1]

Week 3: Recurrent Neural Network

Week 4: Language Models [A2]

Week 5: Sequence Labeling [A3]

Week 6: Context-Free Grammar and Parsing

Week 7: Dependency Parsing [A4]

Week 8: Attention and Transformer

Week 9: Sequence to sequence and translation [A5]

Week 10: Pretraining Transformer-based Models

Week 11: Large Language Models and Prompting [A6]

Week 12: Natural Language Generation

Week 13: Reinforcement Learning with Human Feedback and Computational Ethics

Week 14: Limits and Future of LLMs and NLP

Week 15: Cognitive Science and Language

Week 16: Project Report and Presentation

Learning outcomes of CS301

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.

Instructor's goal of NLP

- Know about the tools -- (you will not find the tree until you sharpen your ax)
- Be familiar with the fundamentals; start with simple questions
- Find a way between imagination and verification
- Abandon the idea to "master" it, but to "be part of" it

- "大胆假设，小心求证"——胡适
- "得之小心，失之费力"——邵雍

About Instructors

- Instructor: Yang Xu 徐炀, Associate Professor, Dept. of Computer Science and Engineering, SUSTech
- Computational Linguistics and Consciousness Sciences Lab,计算语言学与意識科学实验室, clcs-sustech.github.io
- Research areas:
- 2018-2023： Assistant Professor, San Diego State University, USA
- 2018： Ph.D., The Pennsylvania State University
- 2013： M.A., Psychology, Tsinghua University
- 2010： M.E., Electronic Engineering, Tsinghua University

- TA: Tengfei Liu 刘腾飞, and TBD

Textbook and resources

- Speech and Language Processing (3rd ed. draft), Dan Jurafsky and James H. Martin, **SLP3**
- Free online version available: https://web.stanford.edu/~jurafsky/slp3/

Speech and Language Processing (3rd ed. draft)

Dan Jurafsky and James H. Martin

Here's our Feb 3, 2024 release! We also expect to release Chapter 12 soon in an updated release.

Individual chapters and updated slides are below; here is a single pdf of all the chapters in the Feb 3, 2024 release!

Feel free to use the draft chapters and slides in your classes, the resulting feedback we get from you makes the book better!
As always, typos and comments very welcome (just email slp3edbugs@gmail.com and let us know the date on the draft)! (Don't bother reporting missing refs due to cross-chapter cross-reference problems in the indvidual chapter pdfs, those are fixed in the full book draft)

We've put up a list here of the amazing people who have sent so many fantastic suggestions and bug-fixes for improving the book. We are really grateful to all of you for your help, the book would not be possible without you!

When will the whole book be finished? Don't ask.

If you need last year's Jan 2023 draft chapters, they are here;

	Chapter	Slides
	Part I: Fundamental Algorithms	
1:	Introduction	
2:	Regular Expressions, Text Normalization, Edit Distance	2: Text Processing [pptx] [pdf] 2: Edit Distance [pptx] [pdf]
3:	N-gram Language Models	3: [pptx] [pdf]
4:	Naive Bayes, Text Classification, and Sentiment	4: [pptx] [pdf]
5:	Logistic Regression	5: [pptx] [pdf]
6:	Vector Semantics and Embeddings	6: [pptx] [pdf]
7:	Neural Networks and Neural Language Models	7: [pptx] [pdf]
8:	Sequence Labeling for Parts of Speech and Named Entities	8: (Intro only) [pptx] [pdf]
9:	RNNs and LSTMs	
10:	Transformers and Large Language Models	
11:	Fine-tuning and Masked Language Models	
12:	Prompting, In-Context Learning, and Instruct Tuning	

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%

Assignments

- 55% (A1: 5%; A2-A6: 10% each)
- [A1] Neural-network-based text classifier
- [A2] Neural language models: Word2vec (skip-gram) and causal LM (BiLSTM)
- [A3] Named entity recognition (NER) task
- [A4] Neural dependency parsing
- [A5] Seq-seq translation with transformers
- [A6] Fine tuning an LLM (BERT-like)

- Template Jupyter Notebook file and necessary code/data will be provided
- Finished individually; submit the modified notebook
- Deadline in roughly two weeks

Final Project

- 25% (Report 15%; Presentation 10%)
- Group of 2 - 3 students
  - Find your partners by Week 4
- Default topic: Implement a BERT model and fine-tune it.
- Custom topics: To be discussed with instructor during Week 8 - 10 and determined by Week 10.

Labs

- 15% (Attendance 5%; Practice 10%)
- Labs are meant to be practices, not tests.
- Covering some of the pre-processing and preparation steps for assignments.

Example

Week 2: Word Vectors and Neural Networks

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

[A1] Neural-network-based text classifier

Example

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

[Lab] Data preparation for sequence labeling tasks

[A3] Named entity recognition (NER) task

Table of Content

- What is NLP ?
- What will be taught in this course?
- **Programming tools needed**
- **Mathematical knowledge needed**

Python and Jupyter Notebook

- Install Python, Jupyter locally
- Other packages needed: NLTK (optional), numpy, PyTorch, transformers, …

Lab computers and GPU support

- Lab computers are preinstalled with anaconda
- For GPU-enabled cloud computing support, consult计算机科学与工程系HPC-AI服务平台 http://172.18.34.4/

Math

- Probability
- Calculus: derivative, chain rule, gradient descent
- Linear algebra: vector, matrix
- Algorithms: dynamic programming

- Use A0 to estimate how much effort you need to spend.
- (A0 is not graded)

To-Do List

- Read the syllabus
- Read Chapter 6 of SLP3
- Complete A0 and use it to decide if you are ready to take the course