CS310 Natural Language Processing
自然语言处理
Course Introduction

Instructor: Yang Xu

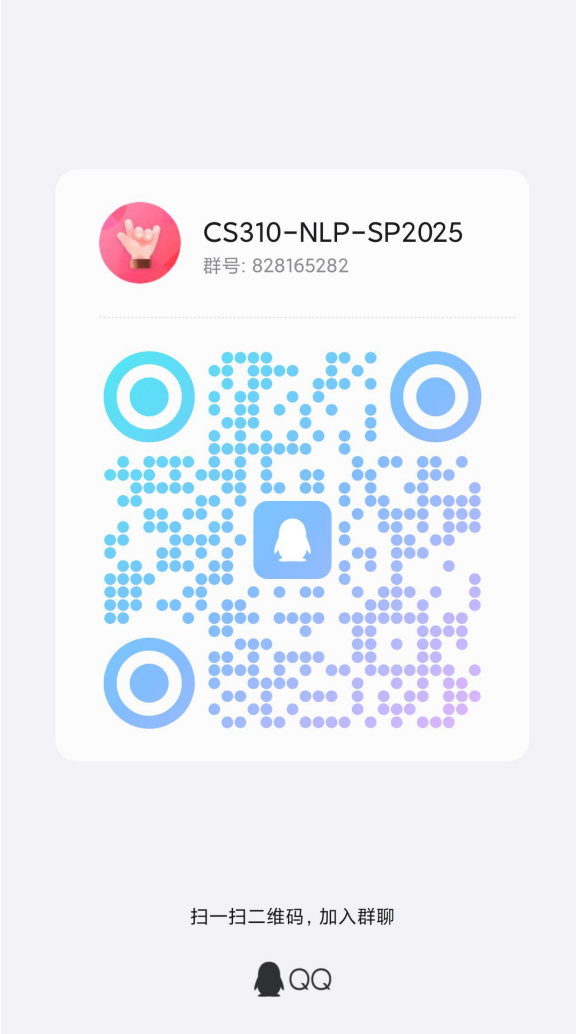
主讲人：徐炀

xuyang@sustech.edu.cn

Content

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

Join the QQ group:



回答问题
课程号？
CS310

What is NLP?

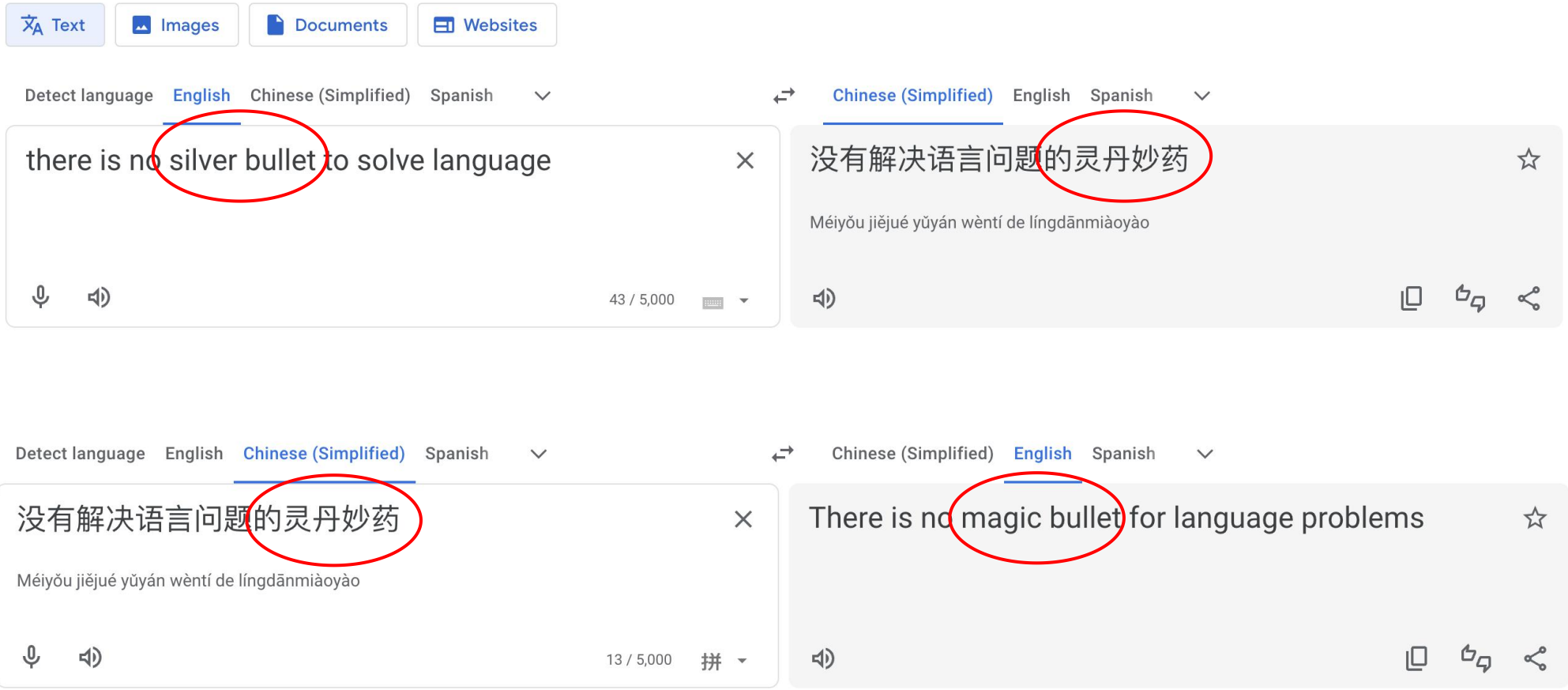
- "*NLP is a branch of artificial intelligence (AI) that enables computers to **comprehend**, **generate**, and **manipulate** human language.*"
  -- oracle.com

- "*NLP is an interdisciplinary subfield of computer science and linguistics … giving computers the ability to **support** and **manipulate** human language*"
  -- wikipedia.org

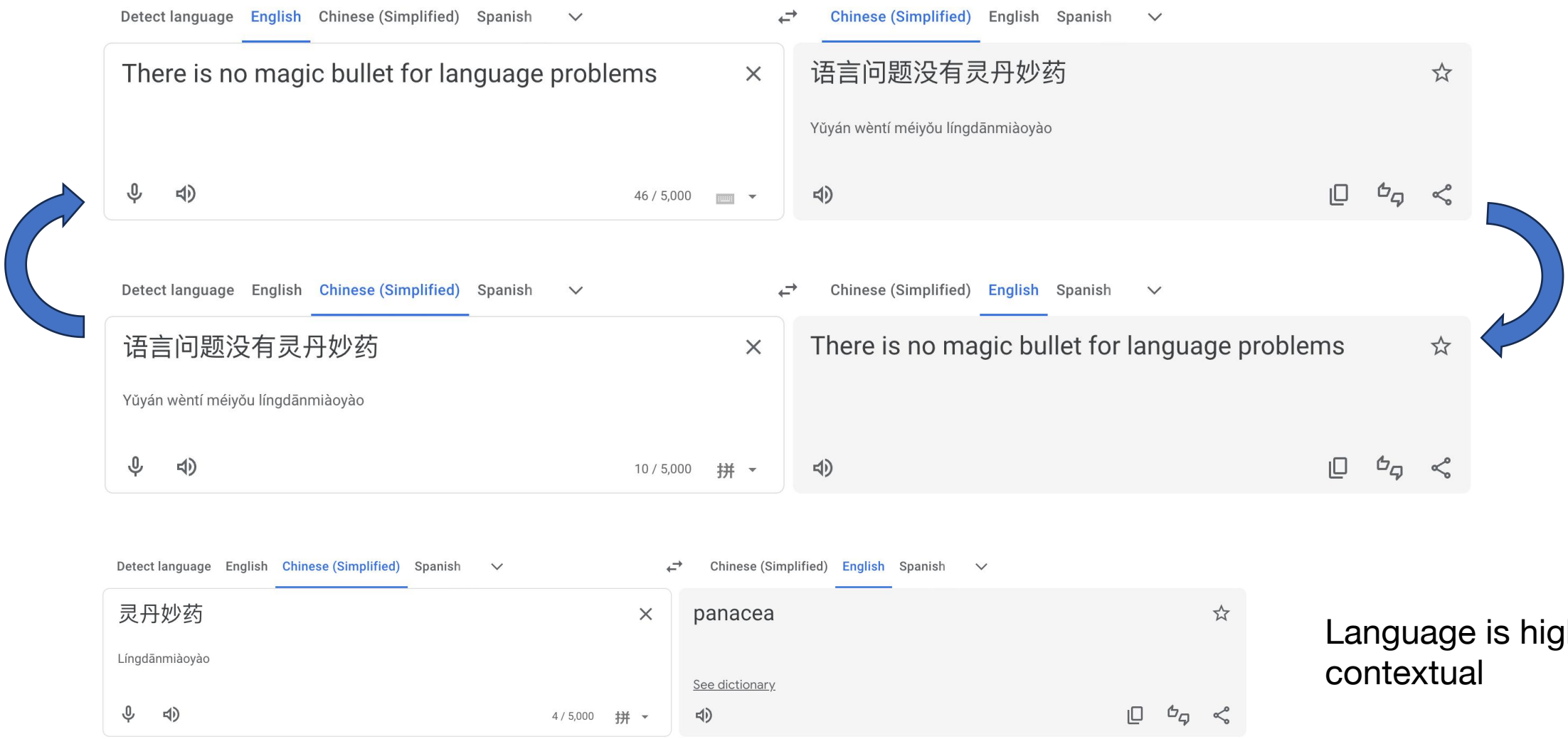
- "*Computational linguistics is the scientific study of language from a computational perspective. Computational linguists are interested in **providing** computational models of various kinds of linguistic **phenomena**.*"
  -- aclweb.org (Association for Computational Linguistics)

What is NLP? -- A translation tool

- No silver bullet to “solve” language



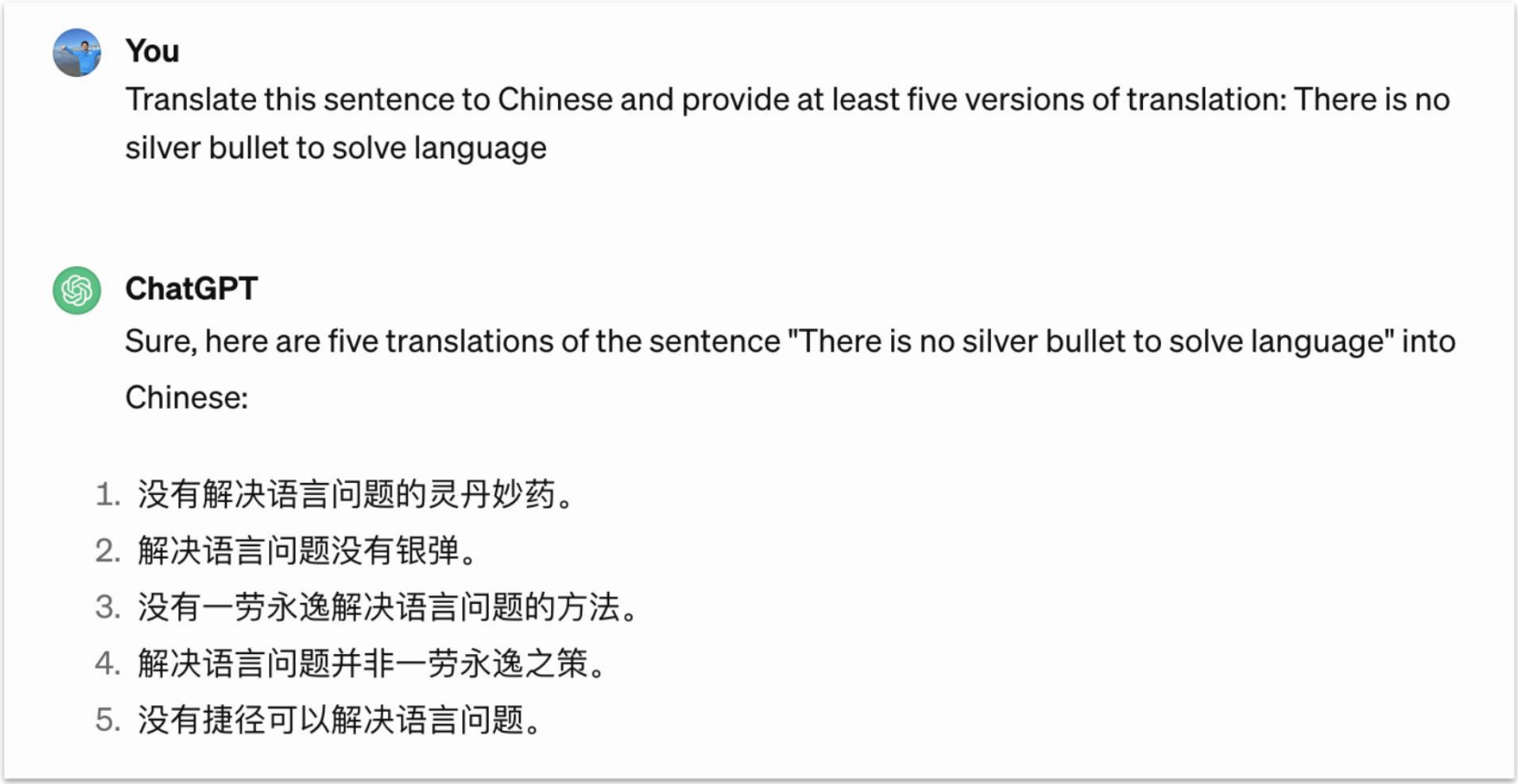
What is NLP? -- A translation tool



Language is highly contextual

What is NLP? -- A technology that understands your intention

- LLM can act as a translation tool



You

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

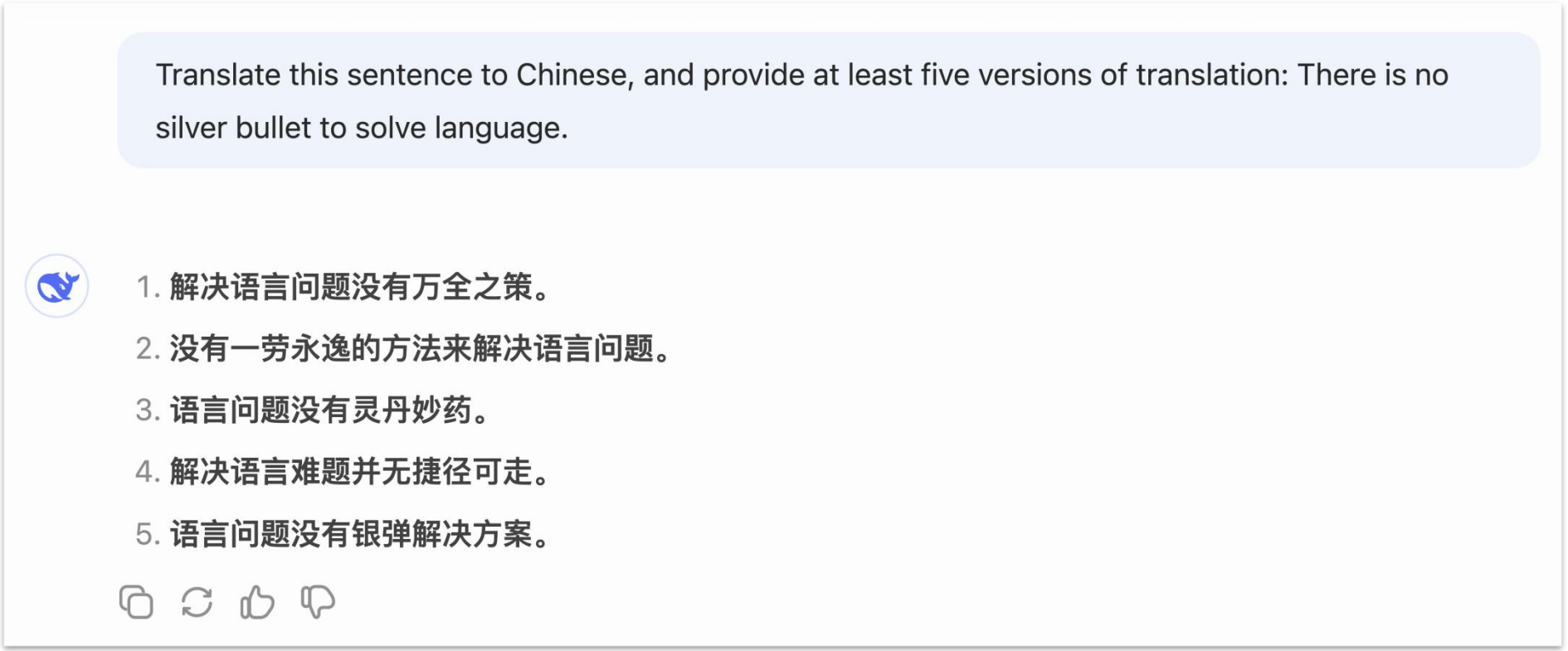
ChatGPT

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。

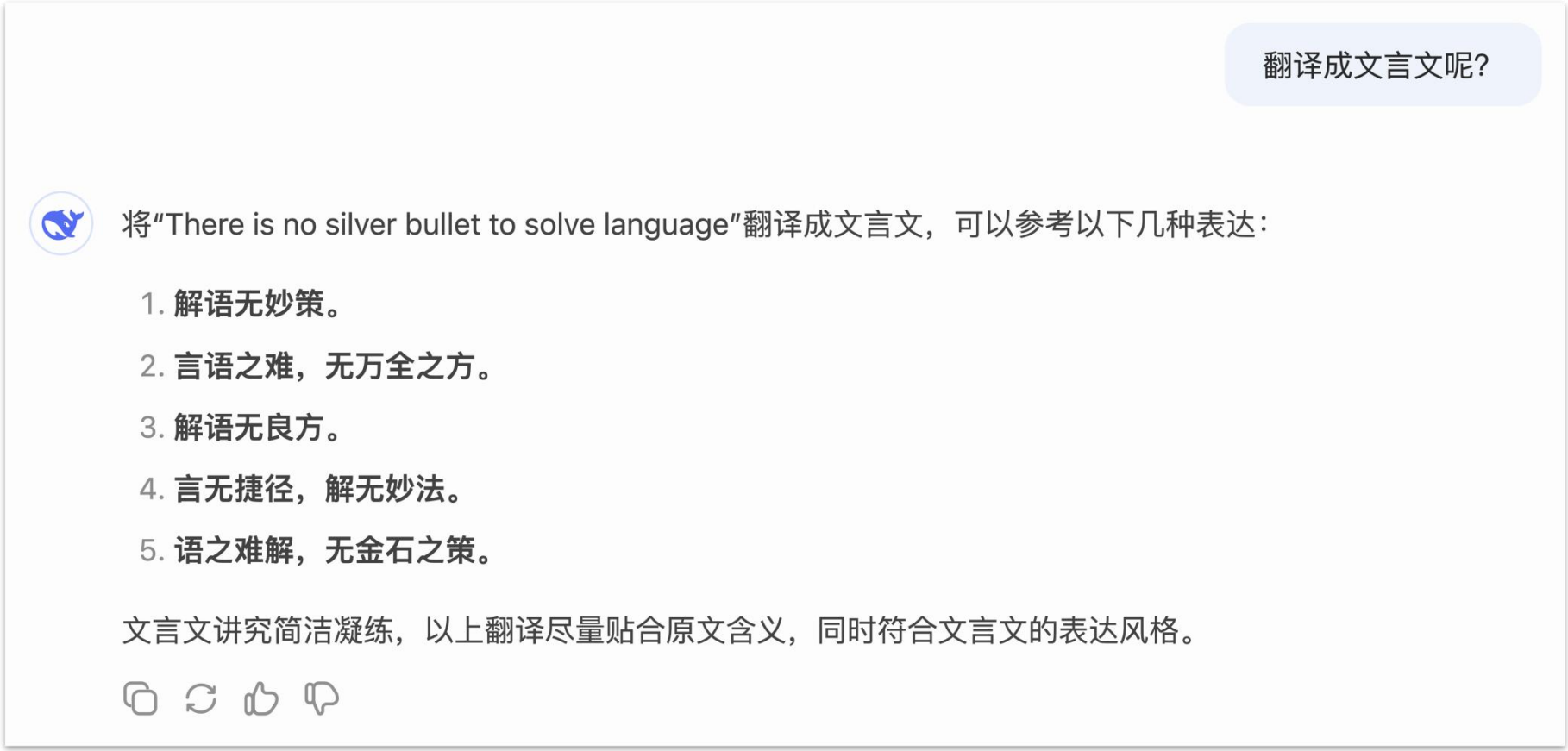
What is NLP? -- A technology that understands your intention

- LLM can act as a translation tool



What is NLP? -- A technology that understands your intention

- LLM can be very intelligent and elegant



What will be taught in CS301?

- Technology of NLP and research of Computational Linguistics (CL)
- NLP (70%) => application
- CL (30%) => theories
- **Topics**: computational word semantics, the extraction of structural text information, language models, question answering, natural language generation, dialogue with user inputs, psycholinguistics, cognitive sciences (touch a bit)
- 词汇语义、结构文本信息提取、语言模型、问答、自然语言生成、用户输入对话、心理语言学、认知科学 (略微涉及)

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.

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: Neural Networks and Text Classification [A1]

Week 3: Word Vectors [A2]

Week 4: Recurrent Neural Network

Week 5: Language Models [A3]

Week 6: Sequence Labeling [A4]

Week 7: Context-Free Grammar and Parsing

Week 8: Dependency Parsing [A5]

Week 9: Attention and Transformer

Week 10: Pretraining Transformer-based Models [A6]

Week 11: Large Language Models and Prompting

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

Some advices for learning 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
- Be bold and cautious at the same time

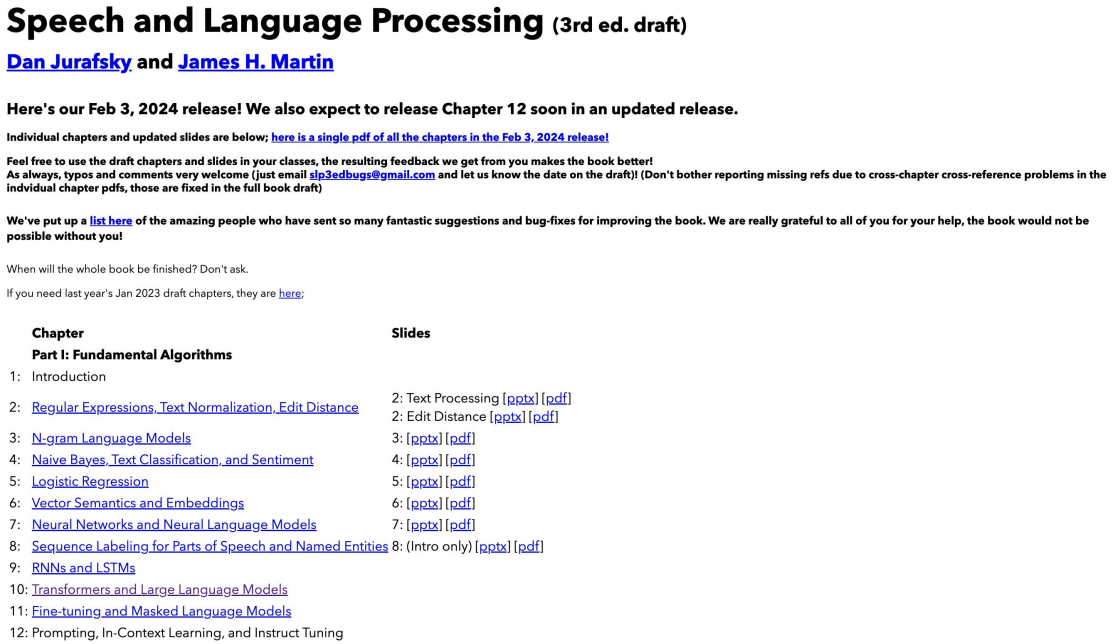
About Instructors

- 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 interests:
- Computational linguistics, psycholinguistics, AI + humanity

- TA：劳乐怡，黎永源，刘之辰，邹欣桐

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

- Programming Assignments: 55% (A1: 5%; A2-A6: 10% each)
- Final Project: 25% (Report 15%; Presentation 10%)
- Lab: 15% (Programming Practice)
- Attendance to lecture: 5%

Assignments

- 55% (A1: 5%; A2-A6: 10% each)
- [A1] Neural network-based text classifier
- [A2] Neural word embedding models (Word2vec)
- [A3] Recurrent neural network-based language models
- [A4] Named entity recognition (NER) task
- [A5] Neural dependency parsing
- [A6] Fine tuning a transformer-based language model

- 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 8

- Topic: Detecting model-generated text
- Details will be released after week 8.

Labs

- 15% (Practice questions)
- Labs are meant to be practices
- Covering some of the pre-processing and preparation steps for the 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

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 Chapter 7 of SLP3 - Neural Networks and Neural Language Models
- Complete A0 and use it to decide if you are ready to take the course