

Natural Language Processing

CSE 325/425



Sihong Xie

Lecture 1:

- Introduction to NLP
- Course outline
- Language models

NLP Applications

- Spell checking

The screenshot shows a spell checker application interface. The main area displays a document titled "Buying a car". The text contains several errors highlighted by blue circles and a red X:

- "For years I have been driving an old used car with a lot of mileage and I hate it. It gets me where I need to go, but I'm tired of fixing leaks and broken parts all the time. Its annoying every times I need to take it to the mechanic. Even when they take care of everything, I know I'll just end up going back there in a few weeks."
- "I have finally decided that I am not going to do it anymore. I have decided to buy a new car! Unfortunately, I have a problem. I have no idea what car to get. Do I want something fast? Do I want something big? Do I want something stylish? Something economical? I have so many choices that I don't even know where to begin. I am not sure if I will be able to make the decision on my own. I don't have not a lot of money, either, so I probably don't have many options."
- "After I did some research, I knew that I would need some expert advice. Eventually, I went to a local dealership to check out some new models. I

The sidebar on the right provides suggestions and notes:

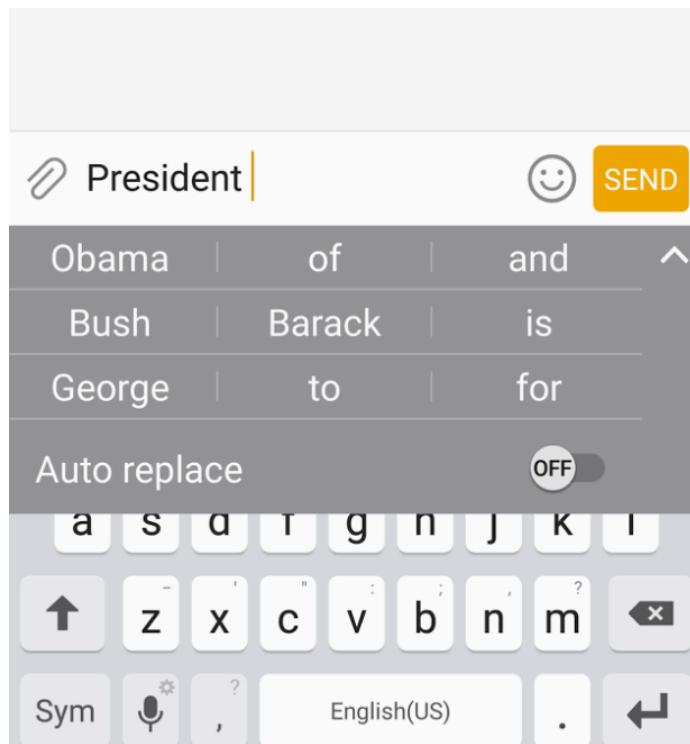
- Possibly confused word: "Its"
- every times → every time
- did → had done

A handwritten note "Discourse" is written in red at the bottom right of the sidebar.

Bottom status bar: GENERAL (DEFAULT), 566 WORDS, 20 CRITICAL ISSUES, SCORE: 1

NLP Applications

- Auto completion.



NLP Applications

- Sentiment analysis

By 10152297075776231 30th November 2011

place as interior is very nice ,staffs seated us nicely and helped explained menu for the first impression was woowww , menu was long and bit confused but food wasent authentic at all totally different than what we had in dubai wine list was ok and we enjoyed our drink but food wasent good soup come quickly but other place was taste very bad we had chicken curry ,duck breast and it was discusting thai pad was good rice was to much sticky ,price wise was bit expensive for what we had . It was first and last time we will go their .



- Negative
- Positive
- Neutral
- Conflict

NLP Applications

- Machine translation

| <i>Input sentence:</i> | <i>Translation (PBMT):</i> | <i>Translation (GNMT):</i> | <i>Translation (human):</i> |
|---|---|--|---|
| 李克強此行將啟動中加總理年度對話機制，與加拿大總理杜魯多舉行兩國總理首次年度對話。 | Li Keqiang premier added this line to start the annual dialogue mechanism with the Canadian Prime Minister Trudeau two prime ministers held its first annual session. | Li Keqiang will start the annual dialogue mechanism with Prime Minister Trudeau of Canada and hold the first annual dialogue between the two premiers. | Li Keqiang will initiate the annual dialogue mechanism between premiers of China and Canada during this visit, and hold the first annual dialogue with Premier Trudeau of Canada. |

NLP Applications

- Chatbots



Image captioning



"man in black shirt is playing guitar."



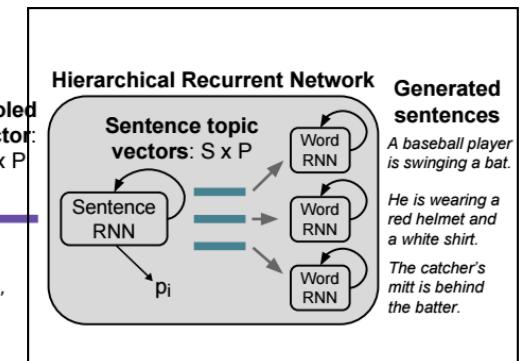
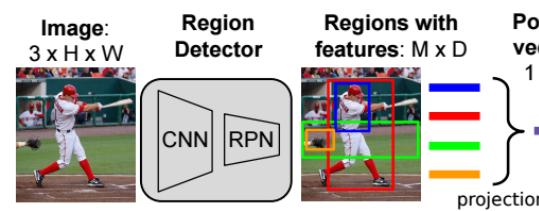
"construction worker in orange safety vest is working on road."



"girl in pink dress is jumping in air."



"black and white dog jumps over bar."



Essentially a language model

*predict the next word
based on previous words*

(Hard) NLP Tasks

Grammar

- Tasks on 4 Levels
 - Lexicon: segment "New York-New Haven railroad".
 - Syntax: structure of sentences (finding subject, object, and verb)
 - Semantics: I am a huge *fan*.
 - Discourse:
 - John went on a trip to NYC
 - He left on an early morning flight.
- Need knowledge of
 - the world,
 - the language,
 - how to combine them.



*"One morning I
shot an elephant in
my pajamas. How
he got into my
pajamas I'll never
know."*

*~Groucho Marx
American comedian
1890-1977*



Current state (-of-the-art, deep learning!)

making good progress

mostly solved

Spam detection

| | |
|-------------------|---|
| Let's go to Agra! | ✓ |
| Buy V1AGRA ... | ✗ |

Part-of-speech (POS) tagging

| | | | | |
|-----------|-------|-------|-------|------------|
| ADJ | ADJ | NOUN | VERB | ADV |
| Colorless | green | ideas | sleep | furiously. |

Named entity recognition (NER)

| | | |
|----------|----------|---------------------------|
| PERSON | ORG | LOC |
| Einstein | met with | UN officials in Princeton |

Sentiment analysis

| | |
|---------------------------------------|---------|
| Best roast chicken in San Francisco! | Like |
| The waiter ignored us for 20 minutes. | Dislike |

Coreference resolution

Carter told Mubarak he shouldn't run again.

Word sense disambiguation

I need new batteries for my **mouse**.

Parsing

I can see Alcatraz from the window!

Machine translation (MT)

第13届上海国际电影节开幕...

The 13th Shanghai International Film Festival...

Information extraction (IE)

You're invited to our dinner party, Friday May 27 at 8:30

Party
May 27
add

still really hard

Question answering (QA)

Q. How effective is ibuprofen in reducing fever in patients with acute febrile illness?

Paraphrase

XYZ acquired ABC yesterday

ABC has been taken over by XYZ

Summarization

The Dow Jones is up

The S&P500 jumped

Housing prices rose

Economy is good

Dialog

Where is Citizen Kane playing in SF?

Castro Theatre at 7:30. Do you want a ticket?

Slides from Dan Jurafsky @ Stanford

Course outline (26 lectures)

1. Language models (4 lectures)

- N-gram and Laplacian smoothing.
- Word vectors.

2. Modeling sequences (10 lectures)

- HMM and CRF (5 lectures)
- RNN and LSTM (2 lecture)

3. Syntactic Parsing (10 lectures)

- Context-free grammar
- CYK algorithm
- PCFG and learning PCFG.
- Recurrent neural networks.

4. More recent deep learning

- Attention models (2 lecture)
- seq2seq models (2 lecture)

1.1 Machine learning basics (1 lecture)

- Data, loss, optimization (backprop and matrix calculus).

1.2 PyTorch basics (1 lecture)

- Computation graphs.
- Tensors, nn, functions.

5. Assignments

3 coding projects:

- Programming in Python
- All students are required to submit.
6 homework assignments:
 - Written
 - All students are required to submit.

Grading

- 6 homework (30%)
- No exam.
- 3 individual programming assignments (60%).
- Final online interview (10%).
- Active plagiarism detection: severe penalty (from 0 credit for a detected submission to University level hearing committee).
 - see the [Dean of Student office website](#) for the details.

Programming environments

- Done with Python; no other languages will be accepted.
- Sketch of codes will be provided.
- If you need GPU for training neural network (first project):
 - Google Colab (<https://colab.research.google.com/>).
 - GPU-equipped, free of cost.
 - CUDA, [Anaconda](#), and many machine learning packages pre-installed.

What this course is not about

- Data science (preliminaries).
- Data mining (hands-on).
- Machine learning (statistical and theoretical).
- Reinforcement learning.
- AI (traditional symbolic AI).
- Deep learning (we cover a rather good basic).

Sample space, events, random variable

- NLP is about using probability to model languages, and using optimization to find the right model.
- The universal set Ω contains all objects that can occur in a specific context.
 - Discrete: $\Omega = \text{All words in a corpus (vocabulary)}$ $\Omega = \{\text{the, a, curry, boring}\}$
 - Continuous: $\Omega = \text{Sentimental score in } [-1, 1]$ $X = \begin{matrix} & \downarrow & \downarrow & \downarrow & \downarrow \\ 0 & 1 & 2 & 3 \end{matrix}$
- Event: subsets of Ω ($A \subseteq \Omega$)
 - A document: bag of words (subset of the vocabulary)
 - Positive sentimental score $[0, 1]$
- Random variable $X : \Omega \rightarrow \text{set of numbers}$
 - e.g., map a word to an integer index.
 - The sentimental score is a number already.
- Probability distribution of a random variable $P(X = 1) = P(\omega \in \Omega : X(\omega) = 1)$
- Probability of an event $A \subseteq \Omega$:

$$P(X = 1) = P(\text{seeing 'a' in the corpus}) = 5\%$$

decomposition

$$\begin{aligned} p(a) p(\text{boring} | a) &\xleftarrow{\text{n-gram}} P(A) \rightarrow [0, 1] \\ p(\text{curry} | \text{boring}, a) \end{aligned}$$

Joint Prob.

$$\begin{aligned} P(\text{a boring curry}) &\xleftarrow{\text{Independence}} \\ &= p(a) P(\text{boring}) P(\text{curry}) \end{aligned}$$