Natural Language Processing
CSPE73

Chandramani Chaudhary

Evaluation

- Midterms – 2 CT's – 20%
- End semester exam – 40%/50%
- Assignments – TBD (tentative – 3 programming assignments) – 30%

Requirements

- Coding
- Academic honesty
- Participation during the class

Course Outline

- UNIT I Lexical Analysis–
  - Lexical Analysis - Regular expression and Automata for string matching - Words and Word Forms - Morphology fundamentals - Parts of Speech - N-gram Models
- UNIT II Speech Processing
  - Word Boundary Detection - Argmax based computations - HMM and Speech Recognition - Text to Speech Synthesis - Rule based-Concatenative based approach.*
- UNIT III Parsing Theories - Parsing Algorithms –
  - Earley Parser - CYK Parser - Probabilistic Parsing - CYK - Resolving attachment and structural ambiguity - Shallow Parsing - Dependency Parsing - Named Entity Recognition - Maximum Entropy Models - Conditional Random Fields.*
- UNIT IV Lexical Knowledge Networks Meaning:
  - Lexical Knowledge Networks - Wordnet Theory - Indian Language Wordnets and Multilingual Dictionaries - Word Sense Disambiguation
- UNIT V Applications :
  - Sentiment Analysis - Text Entailment - Machine Translation - Question Answering System - Information Retrieval - Information Extraction - Cross Lingual Information Retrieval (CLIR).*

Textbook

- Jurafsky Daniel, Martin James, “Speech and Language Processing”, Second Edition, Tenth Impression, Pearson Education, 2018.

What is NLP?

"computational modelling of human language"

Multidisciplinary

- Linguistics
- Cognitive science
- Psychology
- Philosophy
- Maths
- Computer Science- machine learning-AI, formal language theory, compiler techniques, theorem proving, and human-computer interaction

Some Important linguistic terms

- **Morphology** – doors (singular/ plural)
- **Syntax** – structural knowledge – I'm I do, sorry that afraid Dave I'm can't
- **Semantics** –
  - How much Chinese silk was exported to Western Europe by the end of the 18th century?

  We need to know about lexical semantics, the meaning of all the words (export or silk) as well as compositional semantics (what exactly constitutes Western Europe as opposed to Eastern or Southern Europe, what does end mean when combined with the 18th century.

  We also need to know something about the relationship of the words to the syntactic structure. For example, we need to know that by the end of the 18th century is a temporal end-point and not a description of the agent, as the by-phrase is in the following sentence
  - How much Chinese silk was exported to Western Europe by southern merchants?
- **Pragmatics/Discourse** – how many states were in the United States that year?
  - Coreference – it , she, they, etc
  - Will you crack open the door? I am getting hot.

Data processing Vs language Processing application

- E.g. wc command in Unix
  - For lines – data processing
  - For words – language processing

Why is computational language processing difficult?

- Ambiguous input: if there exist multiple alternative linguistic structures
- Most tasks can be viewed as resolving ambiguity at one of the levels
- E.g. – *I made her duck*
  - *I cooked waterfowl for her*
  - *I cooked waterfowl belonging to her*
  - *I created the (plaster?) duck she owns*
  - *I caused her to quickly lower her head or body*
  - *I waved my magic wand and turned her into undifferentiated waterfowl*

- *I cooked waterfowl for her*
- *I cooked waterfowl belonging to her*
- *I created the (plaster?) duck she owns*
- *I caused her to quickly lower her head or body*
- *I waved my magic wand and turned her into undifferentiated waterfowl*

- Morphologically or syntactically ambiguous:
  - Duck- verb or noun (sol- POS tagging)
  - Her- dative pronoun or possessive pronoun

- Semantically ambiguous:
  - Make- create or cook (word sense disambiguation)
- Speech– eye or maid
- (her duck) or (her) (duck) – syntactic disambiguation

Lexical variation

ACCORDING TO THE THESAURUS, "THEY'RE HUMID, PREPOSSESSING HOMOSAPIENS WITH FULL SIZED AORTIC PUMPS" MEANS "THEY'RE WARM, NICE PEOPLE WITH BIG HEARTS."

Several words can mean the same thing!

Some NLP applications

- spelling and grammar checking
- predictive text
- optical character recognition (OCR)
- augmentative and alternative communication
- machine aided translation
- lexicographers' tools
- information retrieval
- document classification
- document clustering
- information extraction
- sentiment classification
- text mining

Sentiment classification

- Task: scan documents (webpages, tweets etc) for positive and negative opinions on people, products etc.
- Find all references to entity in some document collection: list as positive, negative (possibly with strength) or neutral.
- Fine-grained classification: e.g., for phone, opinions about: design, performance, battery life . . .
- Construct summary report plus examples (text snippets).
- Rapidly done for trends on social media

Sentiment classification

- Full task: information retrieval, cleaning up text structure, named entity recognition, identification of relevant parts of text. Evaluation by humans.
- preclassified documents, topic known, opinion in text along with some straight forwardly extractable score.

iPhone 8 review (Guardian 29/9/2017)

The iPhone 8 has Apple's latest and best processor. The six-core A11 Bionic has two high-performance cores and four power-efficient cores and is apparently the most powerful so far because it can use a combination of all six at once.

Performance was excellent, but I struggled to see a real difference in day-to-day speed compared to the iPhone7. But what I'm very pleased to be able to report is that Apple has finally improved battery life for the 4.7in iPhone.

We're not talking a two-day battery here, but the iPhone 8 lasted just over 26 hours . . .

iPhone 8 review (Guardian 29/9/2017)

*The **iPhone 8** has Apple's latest and best **processor**. The six-core A11 Bionic has two high-performance cores and four **power-efficient cores** and is apparently the most powerful so far because it can use a combination of all six at once.*

***Performance** was excellent, but I struggled to see a real difference in day-to-day **speed** compared to the iPhone7. But what I'm very pleased to be able to report is that Apple has finally improved **battery life** for the 4.7in iPhone.*

*We're not talking a two-day **battery** here, but the iPhone 8 lasted just over 26 hours . . .*

iPhone 8 review (Guardian 29/9/2017)

*The **iPhone 8** has Apple's latest and best **processor**. The six-core A11 Bionic has two high-performance cores and four **power-efficient cores** and is apparently the most powerful so far because it can use a combination of all six at once.*

***Performance** was **excellent**, but I struggled to see a real difference in day-to-day **speed** compared to the iPhone7. But what I'm **very pleased** to be able to report is that Apple has finally **improved battery life** for the 4.7in iPhone.*

*We're not talking a two-day **battery here**, but the iPhone 8 lasted just over 26 hours . . .*

iPhone 8 review (Guardian 29/9/2017)

*The **iPhone 8** has Apple's latest and best **processor**. The six-core A11 Bionic has two high-performance cores and four **power-efficient cores** and is apparently the most powerful so far because it can use a combination of all six at once.*

***Performance** was **excellent**, but I struggled to see a real difference in day-to-day **speed** compared to the iPhone7. But what I'm **very pleased** to be able to report is that Apple has finally **improved battery life** for the 4.7in iPhone.*

*We're not talking a two-day **battery here**, but the iPhone 8 lasted just over 26 hours . . .*

Bag-of-word model:

- Treat the reviews as collections of individual words.
- Classify reviews according to positive or negative words.
- Could use word lists prepared by humans, but machine learning based on a portion of the corpus (training set) is preferable.
- Use human rankings for training and evaluation.
- Pang et al, 2002: Chance success is 50% (corpus artificially balanced), bag-of-words gives 80%.

Some sources of errors for bag-of-words

- Negation:

  Ridley Scott has never directed a bad film.

- Overfitting the training data:

  e.g., if training set includes a lot of films from before 2005, Ridley may be a strong positive indicator, (‘Alien,’ ‘Thelma & Louise,’ ‘Gladiator,’ ‘Black Hawk Down’) but then we test on reviews for ‘Kingdom of Heaven’?

- Comparisons and contrasts.

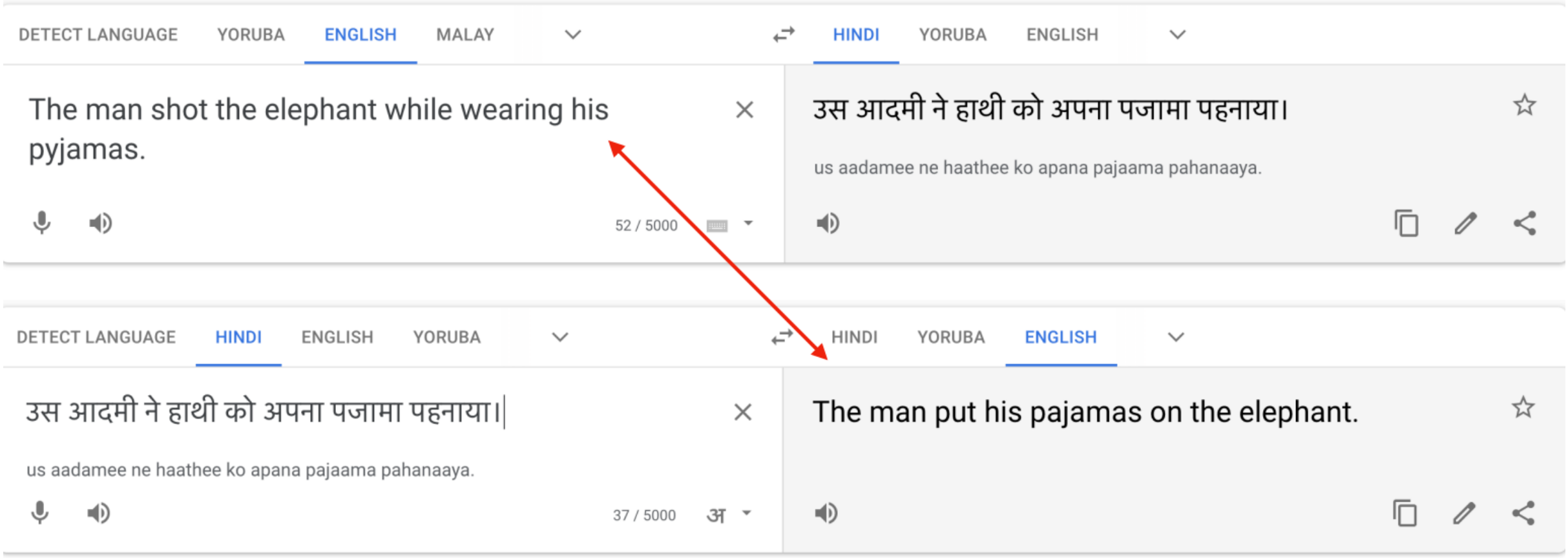
Contrasts in the discourse

This film should be brilliant. It sounds like a great plot, the actors are first grade, and the supporting cast is good as well, and Stallone is attempting to deliver a good performance. However, it can't hold up.

More contrast

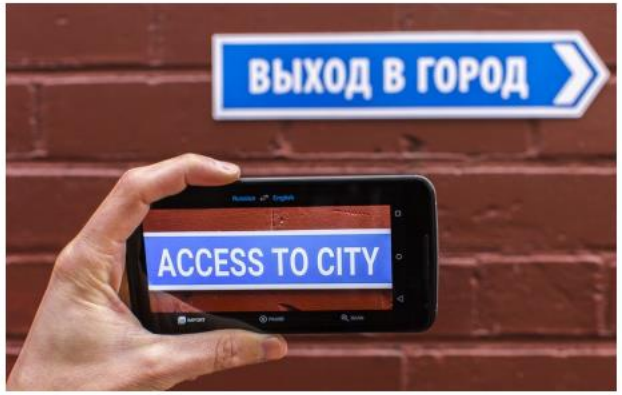
AN AMERICAN WEREWOLF IN PARIS is a failed attempt . . . Julie Delpy is far too good for this movie. She imbues Serafine with spirit, spunk, and humanity. This isn't necessarily a good thing, since it prevents us from relaxing and enjoying AN AMERICAN WEREWOLF IN PARIS as a completely mindless, campy entertainment experience. Delpy's injection of class into an otherwise classless production raises the specter of what this film could have been with a better script and a better cast. . . She was radiant, charismatic, and effective . . .

Machine Translation

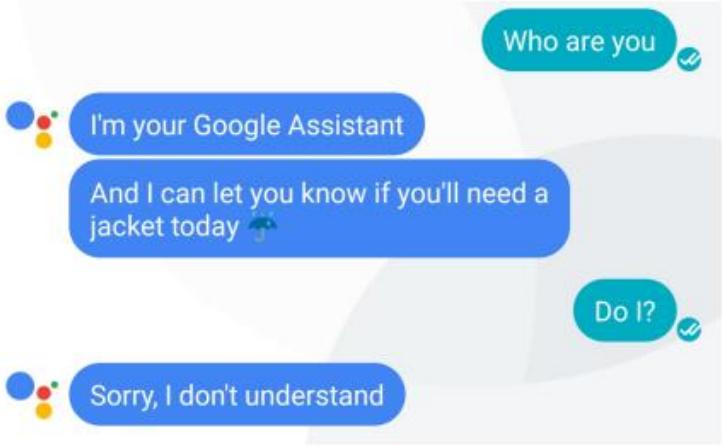


Deep Learning era

- Significant advances in core NLP technologies
- Essential ingredient: large-scale supervision, lots of compute
- Reduced manual effort - less/zero feature engineering



Failures…



… maybe not.

Some language humor

- Kids make nutritious snacks
- Stolen painting found by tree
- Miners refuse to work after death
- Squad helps dog bite victim
- Killer sentenced to die for second time in 10 years
- Lack of brains hinders research

Real newspaper headlines!