Natural Language Processing

Lê Hồng Phương
Hanoi University of Science
Vietnam National University
phuonglh@gmail.com
Outline

- What is NLP?
- Basic NLP problems
- Why is NLP hard?
- Overview of the course
What is NLP?

- Computers using natural languages as input and/or output.

```
language → computer → language
```

```
understanding (NLU)  generation (NLG)
```
Question Answering

• IBM's Watson
• Watson won Jeopardy on February 16, 2011.

WILLIAM WILKINSON’S “AN ACCOUNT OF THE PRINCIPALITIES OF WALLACHIA AND MOLDOVIA” INSPIRED THIS AUTHOR’S MOST FAMOUS NOVEL

Bram Stoker
Hi Dan, we’ve now scheduled the curriculum meeting. It will be in Gates 159 tomorrow from 10:00-11:30.

Chris

The fourth Wells account moving to another agency is the packaged paper-products division of Georgia-Pacific Corp., which arrived at Wells only last fall. Like Hertz and the History Channel, it is also leaving for an Omnicom-owned agency, the BBDO South unit of BBDO Worldwide. BBDO South in Atlanta, which handles corporate advertising for Georgia-Pacific, will assume additional duties for brands like Angel Soft toilet tissue and Sparkle paper towels, said Ken Haldin, a spokesman for Georgia-Pacific in Atlanta.
10TH DEGREE is a full service advertising agency specializing in direct and interactive marketing. Located in Irvine CA, 10TH DEGREE is looking for an Assistant Account Manager to help manage and coordinate interactive marketing initiatives for a marquee automotive account. Experience in online marketing, automotive and/or the advertising field is a plus. Assistant Account Manager Responsibilities Ensures smooth implementation of programs and initiatives. Helps manage the delivery of projects and key client deliverables... Compensation: $50,000-$80,000. Hiring Organization: 10TH DEGREE

Industry: Advertising
Position: Assistant Account Manager
Location: Irvine, CA
Company: 10TH DEGREE
Salary: $50,000-$80,000
Information Extraction

• Motivation:
  – Complex search:
    • “Find me all the jobs in advertising paying at least $50,000 in Boston”
  – Statistical queries:
    • “How has the number of jobs in accounting changed over the years?”
Machine Translation

 Vietnamese: Ông già đi nhanh quá!

 English: He was too old to go fast!
Machine Translation

Google thinks the French are filthy.

Google pense que les Français sont sales.

Ukraine : que doivent faire les Européens et les Américains

Ukraine: What should Europeans and Americans
Text Summarization

The Down John is up.
The S&P500 jumped.
Housing price rose.

Economy is good.
Text Summarization

U.S. bombs Islamic State after President Obama call to prevent Iraq 'genocide'
(World, 23 articles) [UPDATE]

Islamic State militants have killed hundreds of Iraq's minority Yazidis, burying some alive and taking women as slaves, an Iraqi government minister said on Sunday, as U.S. warplanes again bombed the insurgents. U.S. warplanes bomb[ed] Islamist fighters marching on Iraq's Kurdish capital on Friday after ROUGH CUT (NO REPORTER NARRATION) U.S. President Barack Obama said Washington must act to prevent "genocide". American warplanes and drones struck Islamist militants near this northern Iraqi city Friday, putting the U.S. military back in action in the skies over Iraq less than three years after the troops withdrew and Obama declared the war over. The president said Iraqi security forces need to revamp to effectively mount an offensive, which requires a government in Baghdad that the Iraqi military and people have confidence in. The United States launched a series of airstrikes against Sunni militants in northern Iraq on Friday, using Predator drones and Navy F/A-18 fighter jets to destroy rebel positions near city of Irbil, the U.S. military said Friday. A Pentagon spokesman said two F/A-18 aircraft from an aircraft carrier in the Gulf had dropped laser-guided 500-pound bombs on a mobile artillery piece used by the fighters to shell Kurdish forces defending Arbil.

Other stories about Iraq, Iraqi and Kurdish:

- U.S. launches more air strikes against ISIS in Iraq (6 articles) [UPDATE]
- Islamic State militants seize Iraq's largest dam (4 articles) [UPDATE]
**User:** I need a flight from Boston to Washington, arriving by 10 pm.

**System:** What day are you flying on?

**User:** Tomorrow

**System:** Returns a list of flights

- Tôi nay rạp Lotte ra mắt bộ phim Transformers đây!
- Phim đó có diễn viên nào nổi tiếng đóng không?
- ABC, XYZ
- Okie.
- Máy giờ thì chiều?
Sentiment Analysis

The final sentiment is determined by looking at the classification probabilities below.

**Subjectivity**
- neutral: 0.0
- polar: 1.0

**Polarity**
- pos: 0.7
- neg: 0.3
Sentiment Analysis

The camera feels flimsy, is plastic and very light in weight. You have to be very delicate in the handling of this camera.

The text is **neg.**

The final sentiment is determined by looking at the classification probabilities below.

**Subjectivity**
- neutral: 0.0
- polar: 1.0

**Polarity**
- pos: 0.4
- neg: 0.6
Outline

- What is NLP?
- **Basic NLP problems**
- Why is NLP hard?
- Overview of the course
Basic NLP Problems

- Segmentation
  - Sentence segmentation
  - Word segmentation

- Tagging
  - Part-of-speech tagging (POS tagging)
  - Named entity recognition (NER)

- Parsing
  - Syntactic parsing
  - Semantic parsing
Segmentation

• Every NLP task needs to do text normalization
  – Segmenting/tokenizing words in running text
  – Normalizing word formats
  – Segmenting sentences in running text
Issues in Tokenization

• Finland's capital → Finland Finlands Finland's ?
• what're, I'm, isn't → what are, I am, is not
• Hewlett-Packard → Hewlett Packard ?
• state-of-the-art → state of the art ?
• Lowercase → lower-case, lowercase, lower case?
• San Francisco → one token or two ?
• m.p.h., Ph.D → ??
Tokenization: Language Issues

• French:
  – *L'ensemble* → one token or two?
  – *L ? L' ? Le ?*
  – Want *l'ensemble* to match with *un ensemble*.
  – *Pomme de terre* → 3 tokens or one?

• German noun compound are not segmented:
  – Lebensversicherungsgesellschaftsangestellter
  – “life insurance company employee”
  – German information retrieval needs **compound splitter**.
Tokenization: Language Issues

- Chinese and Japanese: no spaces between words.
- Vietnamese: spaces between syllables but not words:
  - Học sinh học sinh học
  - Thuộc địa bàn thành phố
  - Chiếc áo bà ba / bà ba đi chợ chua về
  - Ăn quả chanh chua / cô gái chanh chua
Tagging

• Strings to tagged sequences.

• Examples:
  – POS Tagging:

    • *Profits/N soared/V at/P Boeing/N Co./N ,/, easily/ADV topping/V forecasts/N on/P Wall/N Street/N ./.

  – NER:

    • *Profits/O soared/O at/O Boeing/ORG Co./ORG ,/O easily/O topping/O forecasts/O on/O Wall/LOC Street/LOC ./O
Syntactic Parsing

• Input: *Boeing is located in Seattle.*

• Output:
Outline

- What is NLP?
- Basic NLP problems
- Why is NLP hard?
- Overview of the course
Why is NLP Hard?

“At last, a computer that understands you like your mother”
Ambiguity

• “At last, a computer that understands you like your mother”
  – (*) It understands you as well as your mother understands you.
  – It understands (that) you like your mother.
  – It understands you as well as it understands your mother.

• What interpretation do you select?
Ambiguity at Many Levels

- At the acoustic level (speech recognition)
  - “... a computer that understands you like your mother”
  - “... a computer that understands you like cured mother”
Ambiguity at Many Levels

- At the **syntactic** level:

```
understands you
like your mother [does]
```
Ambiguity at Many Levels

• At the semantic (meaning) level:
  – Two definitions of mother:
    • mother, female parent -- (a woman who has given birth to a child (also used as a term of address to your mother); "the mother of three children")
    • mother -- (a stringy slimy substance consisting of yeast cells and bacteria; forms during fermentation and is added to cider or wine to produce vinegar)
  • This is an instance of word sense ambiguity.
More Word Sense Ambiguity

- At the **semantic** (meaning) level:
  - They put money to the bank.
    - = buried in mud?
  - I saw a man with a telescope.
Ambiguity at Many Levels

• At the **discourse** (multi-clause) level:
  
  – *Alice says* they've built a computer that understands you like your mother.
  
  – *But she*...
    
    • ...doesn't know any details.
    • ...doesn't understand me at all.

• This is an instance of **anaphora**, where *she* co-referees to some other discourse entity.
Language Technology

• Mostly solved:
  – Spam detection, POS tagging, NER

• Making good progress:
  – Sentiment analysis, co-reference resolution, word sense disambiguation, machine translation, information extraction

• Still really hard:
  – Paraphrase, summarization, QA, dialogue
Outline

• What is NLP?
• Basic NLP problems
• Why is NLP hard?
• Overview of the course
Objectives

- Teaches key theory and methods for rule-based and statistical NLP.
- Practical, robust real-world applications
Overview of the Course

- Lexical analysis:
  - word segmentation, sentence segmentation, stemming
- Statistical NLP:
  - Naive Bayes models
  - Log-linear models, sequence models (HMMs, MEMMs, CRFs)
- Tagging problems: POS tagging, NER
- Syntactic parsing
  - Constituency parsing, dependency parsing
- Question answering system
Prerequisites

- Basic linear algebra, probability and statistics, algorithms
- Programming skills
References

• Books:

• Lecture notes: