EMORY UNIVERSITY SCHOOL OF MEDICINE

Department of Biomedical Informatics

BMI 500: https://tinyurl.com/bmi500

Introduction to Biomedical Informatics

5. Natural Language Processing

21st Sept, 2022

Abeed Sarker

Department of Biomedical Informatics, Emory University, Atlanta, GA USA

Expectations: Deliverables

- Participation in class
- Exploring and comparing texts using NLP
- Understanding the creation of end-to-end (*full stack*) NLP pipelines

Overview questions

- What is Natural Language Processing (NLP)?
- Why is NLP important?
- What are some of the basic challenges for NLP?
- What has NLP accomplished so far, particularly in biomedical informatics?
- What is the future of NLP?

Overview questions

- What is Natural Language Processing (NLP)?
- Why is NLP important?
- What are some of the basic challenges and approaches for NLP?
- What has NLP accomplished so far, particularly in biomedical informatics?
- How can we get started with NLP?

What is natural language processing?

- Natural language = Human language
- Natural language != formal/programming language
- Overarching objectives:
  - Understand meanings
    - How do we understand language anyways?
  - Curate information/knowledge
    - How do we pass language through the years?
  - Automate language-related tasks
    - Natural language processing + information retrieval + machine learning – changed the world as we used to know it

Open domain *vs.* restricted domain NLP

- Open domains *vs.* restricted domains
  - *e.g.,* news *vs.* medical publication
- NLP in restricted domains is more complicated
  - Implementing systems often requires domain knowledge (*e.g.,* medical knowledge about diseases, symptoms etc.)
  - Domain specific terminologies

Natural language processing tasks

- Parsing
- Part of Speech Tagging
- Named Entity Recognition
- Natural Language Generation
- Speech Recognition
- Summarization
- Question Answering
- Machine Translation
- Some intersection between NLP and IR

Overview questions

- What is Natural Language Processing (NLP)?
- Why is NLP important?
- What are some of the basic challenges and approaches for NLP?
- What has NLP accomplished so far, particularly in biomedical informatics?
- How can we get started with NLP?

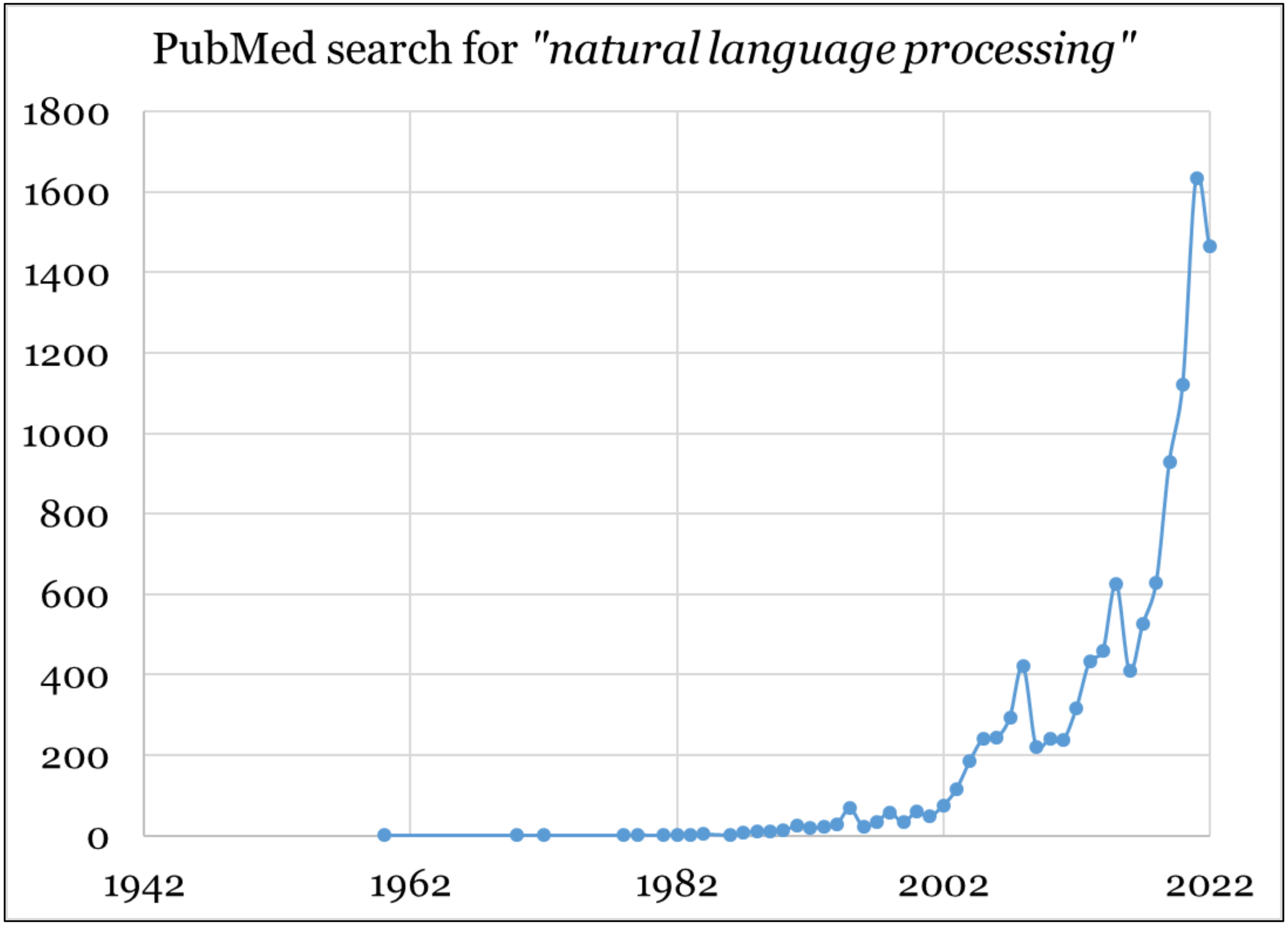
The importance of NLP

- Large volumes of knowledge encapsulated in text
- Internet:
  - Large volumes of information are being generated every day/minute/second
- There is too much information available to process manually
- Information increasing at an exponential rate
- Sources
  - Published science, social media, electronic health records, news papers, emails …

Why should we process language

- Language is how we communicate knowledge
  - The origin of species
  - A brief history of time
- Language is culture; language is experience
  - Poetry
  - Music
- Language is fascinating
  - Sarcasm typically does not translate

NLP Growth



PubMed search for *"natural language processing"*
1800
1600
1400
1200
1000
800
600
400
200
0
1942
1962
1982
2002
2022

Language technology: current state

- Early progress
  - Email spam detection
  - POS tagging
  - Some NER
- Recent developments
  - Sentiment analysis
  - WSD
  - Misc. information extraction
- Difficult problems
  - Summarization
  - Question-answering
  - Language generation
  - Language understanding

Overview questions

- What is Natural Language Processing (NLP)?
- Why is NLP important?
- What are some of the basic challenges and approaches for NLP?
- What has NLP accomplished so far, particularly in biomedical informatics?
- How can we get started with NLP?

Hierarchy of language processing

- The analysis of natural language is not done at a single step
- Instead, language is typically *dealt with* at several layers of abstraction:
  - Lexical level (words or terms)
  - Syntactic level (organization of groups of words in sentences or clauses)
  - Semantic level (meanings of words/phrases)
  - Discourse level (across sentences and documents)

Morphology and morphological analyses

- Morphology concerns the structure of words
  - Words are made up of morphemes
  - The minimal information carrying units
  - Words are made up of a stem and zero or more affixes
- English only has suffixes and prefixes. Examples:
  - Box -> boxes, boxed; Car -> cars; party-> parties; walk-> walked, walking
  - legal->illegal
  - Vast majority of English terms have regular morphology

Stemming

- Most information retrieval and natural language processing applications benefit from reducing all morphological variants into a canonical form
- Stemming is the common approach to removing suffixes
- Porter stemmer
  - Uses a series of simple rules to strip endings
  - Stemming, stemmer, stemmed -> stem
  - Argued, arguing, argues, argue -> argu (the stem itself is not a word or the root)
- Many problems treat words with the same stem as synonyms

Porter stemmer

- Full algorithm (5-6 pages) available at:
  - http://people.scs.carleton.ca/~armyunis/projects/KAPI/porter.pdf
- Many implementations available, including in *nltk*

In the rules below, examples of their application, successful or otherwise, are given on the right in l
case. The algorithm now follows:

```
Step 1a

    SSES -> SS                         caresses  ->  caress
    IES  -> I                          ponies    ->  poni
                                       ties      ->  ti
    SS   -> SS                         caress    ->  caress
    S    ->                            cats      ->  cat

Step 1b

    (m>0) EED -> EE                    feed      ->  feed
                                       agreed    ->  agree
    (*v*) ED  ->                       plastered ->  plaster
                                       bled      ->  bled
    (*v*) ING ->                       motoring  ->  motor
                                       sing      ->  sing
```

Initial steps to processing language

- Text 1: ‘*I like to paint*’
- Text 2: ‘*I Like painting’*
- Text 3: *‘I like to play’*
- Which of these two texts are similar?
- Word to word comparison:
  - Text 1 and 2: 1 word in common
  - Text 2 and 3: 1 word in common
  - Text 1 and 3: 3 words in common

Token		**Texts**
I	->	1, 2, 3
Like	->	2
like	->	1, 3
to	->	1, 3
paint	->	1
painting	->	2
play	->	3

After tokenization and stemming

- Text 1: [*I, like, to, paint*]
- Text 2: [*I, Like, paint~~ing~~*]
- Text 3: [*I, like, to, play*]
- Which of these two texts are similar?
- Word to word comparison:
  - Text 1 and 2: 2 words in common
  - Text 2 and 3: 1 word in common
  - Text 1 and 3: 3 words in common

Token		Texts
I	->	1, 2, 3
Like	->	2
like	->	1, 3
to	->	1, 3
paint	->	1, 2
~~painting~~	~~->~~	~~2~~
play	->	3

Lowercasing

- For many NLP tasks, cases of terms are very important
- For example, named entity recognition
- Cases often give us clues about what a word represents
  - Names of people, cities, countries are typically in uppercase (Yahoo! vs. yahoo!)
  - Abbreviations are typically in uppercase
  - Sometimes also helpful for sentence tokenization
- However, in many cases, such as comparing content, case is not important
- Text normalizing/preprocessing commonly involves lowercasing of all texts

Texts after lowercasing

- Text 1: [*I, like, to, paint*]
- Text 2: [*I, like, paint~~ing~~*]
- Text 3: [*I, like, to, play*]
- Which of these two texts are similar?
- Word to word comparison:
  - Text 1 and 2: 3 words in common
  - Text 2 and 3: 2 words in common
  - Text 1 and 3: 3 words in common

Token		**Texts**
I	->	1, 2, 3
~~Like~~	~~->~~	~~2~~
like	->	1, 2, 3
to	->	1, 3
paint	->	1, 2
~~painting~~	~~->~~	~~2~~
play	->	3

Stopword removal

- Stopwords
  - Commonly used words that are typically not important for NLP and information retrieval tasks
- Common stopwords
  - ’and’, ‘but’, ’how’, ’or’…
- These words may be useful in semantic language representation, in sequential models, and in deep language analysis
- Not useful in content-oriented NLP
  - Does it matter how many times the word ‘to’ occurs in a text?

nltk and stopwords

- nltk provides its own list of English stopwords

```
from nltk.corpus import stopwords
print set(stopwords.words('english'))
```

- *to, from, over, being, both, and, are .....*

Texts after stopword removal

- Text 1: [*I, like, ~~to,~~ paint*]
- Text 2: [*I, like, paint~~ing~~*]
- Text 3: [*I, like, ~~to,~~ play*]
- Which of these two texts are similar?
- Word to word comparison:
  - Text 1 and 2: 3 words in common
  - Text 2 and 3: 2 words in common
  - Text 1 and 3: 2 words in common

Token		Texts
I	->	1, 2, 3
~~Like~~	~~->~~	~~2~~
like	->	1, 2, 3
~~to~~	~~->~~	~~1, 3~~
paint	->	1, 2
~~painting~~	~~->~~	~~2~~
play	->	3

Ambiguity

- Sentences are complex; large documents contain many complex sentences
- Ambiguity is one of the many challenges to NLP
- Example 1:
    - `I saw the man on the hill with a telescope`
- So, who had the telescope?

Interpretation of natural language

- I saw the man on the hill with a telescope
    - I saw the man. The man was on the hill. I was using a telescope.
    - I saw the man. I was on the hill. I was using a telescope.
    - I saw the man. The man was on the hill. The hill had a telescope.
    - I saw the man. I was on the hill. The hill had a telescope.
    - I saw the man. The man was on the hill. I saw him using a telescope.

Word sense disambiguation

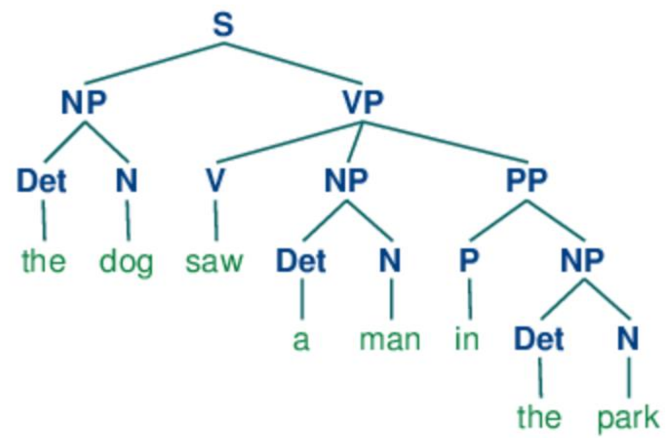
- Resolve the meaning of a term in a text segment
- The word ***bank*** has multiple meanings:
- Did you put your money in the ***bank***?
  - (noun) An institution for receiving and lending money
- We sat and chatted by the river ***bank***
  - (noun) the land alongside or sloping down to a river or lake
- Context is everything!

Denotation *vs.* connotation

- Denotation
  - Original (dictionary) meaning of a text
- Connotation
  - Implied meanings of texts that are not literal

Other common preprocessing methods

- Non-alphanumeric character removal
- Encoding conversion
- Parsing
  - Can be computationally expensive
  - Need training on domain-specific texts
- Vectorization
- Pre-training



Overview questions

- What is Natural Language Processing (NLP)?
- Why is NLP important?
- What are some of the basic challenges and approaches for NLP?
- What has NLP accomplished so far, particularly in biomedical informatics?
- How can we get started with NLP?

Biomedical NLP

- Rapid growth in biomedical literature
  - MEDLINE (25+ million articles)
- An overwhelming amount of (very valuable for discovery) information is “hiding” in biomedical text
- Information overload
- Types of biomedical data
  - Published literature
  - Electronic health records/clinical notes
  - Social media health data (very recent; very exciting)

Challenges to biomedical NLP

- NLP is more challenging compared to non-medical text
  - Lexical level challenges
    - Identifying words (tokenization)
    - Identifying lexical variants (due to inflection and derivation)
    - Disambiguation and normalization (especially for unstructured texts)
    - Identification of multi-token terms
  - Complex domain-specific terminologies
  - Complex associations (*e.g.*, between medications and treatments)

Resources for Biomedical NLP

- Vocabularies/ontologies/knowledge bases
- For example, the Unified Medical Language System (UMLS)
  - A collection of many health and biomedical vocabularies
  - Three tools:
    - Metathesaurus
    - Terms and codes from many vocabularies
    - Semantic network
    - Broad categories and semantic types
    - Relationships between semantic types
- More next week…

Overview questions

- What is Natural Language Processing (NLP)?
- Why is NLP important?
- What are some of the basic challenges and approaches for NLP?
- What has NLP accomplished so far, particularly in biomedical informatics?
- How can we get started with NLP?

Python and nltk

- For this week's lab work, we will do some basic NLP using python and nltk
- Python
  - Very popular for NLP and data science in general
  - Version 3.* is currently supported, although many people still use 2.*
  - Distributions available *(e.g.*, Anaconda): https://www.anaconda.com/products/individual
- nltk – **N**atural **L**anguage **T**ool**k**it
  - Has been popular for a while
  - Available: https://www.nltk.org/

Pre-requisites

- Python 3.* distribution
  - Anaconda is great
- nltk
  - It's a good idea to run `nltk.download()`
- A good IDE can help
  - My personal preference is PyCharm: https://www.jetbrains.com/pycharm/
  - Many other IDEs available
- Now to the lab work!

NLP lab work (week 5)

- Tasks:
  - NLP basics
- Homework: https://drive.google.com/file/d/1AQq9r1JR022ubdS-BF2MGp6y8hgvcVK7/view?usp=sharing
- Practice homework (optional): https://drive.google.com/file/d/1S7N_Fwn9tCDmcGPNkhhqAcyM6TqNltqR/view?usp=sharing
- Solutions will be posted next week