Review for Midterm

Natural Language Processing

1. Overview of NLP

- What is NLP?
  - General definition
  - Natural vs. artificial languages
    - Why artificial languages are convenient
    - Why natural languages are challenging
- Two sides of NLP
  - NLU vs. NLG

1. Overview of NLP

- Applications of NLP
  - Text annotation
    - Tagging
    - Metadata extraction/generation
    - Classification
    - Document summarization
  - Corpus analytics
    - Theme extraction
    - Clustering
    - Taxonomy mapping
    - Sentiment analysis

1. Overview of NLP

- Applications of NLP
  - Search applications
    - Query repair
    - Query refinement
    - Results postprocessing (ranking, clustering, synopsizing)
  - Speech applications
    - Speech recognition (dictation, command-and-control)
    - Speech synthesis (standard, emotive)

1. Overview of NLP

Advanced applications

- Machine translation
- Knowledge discovery
- Question handling
  - Question typing/matching
  - Question answering
- Inference engines
  - Textual entailment
  - Knowledge discovery

2. Levels of Analysis

Levels of analysis in NLP

- Lexical
- Syntactic
- Semantic
- Discourse

2. Levels of Analysis

Lexical analysis

- Morphology and stemming
- Attachment of corpus-based metadata (frequency, collocations, co-occurrences, etc.)
- Enumeration of word senses
- Domain relations
- Terminology extraction—basic
  - Internal weighting (tf) vs. external weighting (encyclopedic)
  - Weighted collocations
  - Weighted NER
- Spell correction

2. Levels of Analysis

Syntactic analysis

- POS tagging
- Sentence boundary detection
- Parsing
  - Deep
  - Light (chunking)
- Revisiting tagging
  - Weighted noun phrases
- Discrete text field analysis (extracting, unitizing, normalizing)
- Lemmatization

2. Levels of Analysis

Semantic analysis

- NER
- Relationship extraction (between NEs)
- WSD
- Classification
- Tagging
- Topic segmentation
- Sentiment analysis
- Question typing/matching

2. Levels of Analysis

Discourse analysis

- Anaphora resolution
- Discourse modeling
- Textual entailment
- Pragmatic analysis
- Question answering
- Speech interpretation

3. Trade-Offs in NLP

Understand differences between:

- Shallow vs. deep NLP
- Statistical vs. symbolic
- Feature engineering vs. feature learning
- Top-down vs. bottom-up
- Transparent vs. opaque (AI vs. XAI)

4. Working in NLP

NLP and data science

- Using ML in NLP
- NLU for feature learning in ML
- NLG for presentation layers of results in data science
- NLU to validate discrete data

4. Working in NLP

Job roles that utilize NLP

- Software engineer
- Knowledge engineer
- Data scientist
- Applied linguistics researcher
- Cognitive scientist
- Marketing technologist
- DBA

4. Working in NLP

Sectors that utilize NLP

- Information retrieval and presentation
- eCommerce
- Customer service desks
- Law enforcement or military
- Legal
- Business intelligence
- Consumer devices
- Embedded technologies
- Publishing

4. Working in NLP

Organizations that relate to NLP

- ACM (SIGKDD and SIGAI)
- IEEE
- AAAI
- IJCAI
- AAAL
- ICLA

Pick one of these organizations, and be prepared to describe its orientation or focus, and how it relates to NLP.

5. Low-Level Analysis

Text preprocessing

- Sentence segmentation
  - Sentence boundary detection—obvious approach
  - Counterexamples to the obvious approach, how to address
  - Use: Combine with higher-level analysis for extraction of key sentences (use in highlighting/summarizing)
- Lexical analysis—word tokenization

5. Low-Level Analysis

Text normalization

- Contractions
- Content words vs. function words vs. stop words
- Fuzzy spelling
- Morphology, stemming

5. Low-Level Analysis

Low-level document feature extraction

- Terminology extraction
  - Frequency based
  - Differential frequency analysis (TF-IDF)
  - Collocations (*n*-grams)

5. Low-Level Analysis

Low-level document feature extraction

- Secondary features and applications
  - Comparative lexical diversity
  - Reading level assessment
  - Example: Dale-Chall formula

6. Lexical Knowledge Bases

Lexical knowledge bases

- Difference between lexicon and dictionary
- Distinction of lexical knowledge base
- Introduction to WordNet
- Improvements to low-level analysis utilizing a lexical KB

6. Lexical Knowledge Bases

Resources for creating or extending lexical knowledge bases

- Vocabulary resources
- Examples: Princeton WordNet, Getty TGN, Wiktionary, Urban Dictionary Encyclopedic Resources
  - Examples: Wikipedia, IMDB, DotDash (formerly About.com)
- Taxonomical/topical resources on websites
  - Sitemaps for CNN Money, LA Times, Vogue, etc.
  - RSS page for SFGate, etc.

6. Lexical Knowledge Bases

Applications of lexical knowledge bases

- NER—recognition, weighting, filtering, and domain assignment using the appropriate KB
- Improved reading level assessment (using a baseline vocabulary)

7. POS Tagging

The basics of POS tagging

- POS ambiguity
- POS tagsets
  - Penn Treebank tagset
  - Why most people use it

Be prepared to take a random sentence and manually tag it with Penn Treebank POS tags (you'll be able to look at a reference table of the tagset while doing so).

7. POS Tagging

How POS taggers work

- Sliding window
- Transformation rules
- Statistical approaches and HMMs

7. POS Tagging

Types of POS taggers

- Rule-based tagger
- Statistical tagger
- ML-classifier tagger

7. Using POS Tags

Potential benefit of using POS tags

- WordNet synset lookups
- NER
- Word clouds
- Sentiment analysis

8. Shallow Parsing

Reasons to choose shallow parsing

- POS tagging vs. shallow parsing vs. deep parsing

How to create chunks

- RegEx approach
- ML-based shallow parsers

Implementing chunks

- IOB annotation
- Chunks and chinks

8. Full Grammar Parsing

Understanding full parsers

- Two types of parsers: dependency and constituency
- Understanding parse trees
- Using a toolkit to create parse trees
- Pros and cons of full-parse trees
- Ambiguity of full parsing

8. Using Parse Trees

Uses for full-parse trees

- Validation of candidate answer for question answering
- Attachment of sentiment to particular entities (verb-to-noun)

8. Using Parse Trees: Combining Lexical and Syntactic Analyses

Topic definition and processing

- Creating a word class for a theme using a lexical KB (WordNet)
- Applying syntax parsers to a word class
  - Objects for verbs, modifiers for nouns (from dependency parse)
  - NPs and VPs containing sentiment trigger (from constituency parse)

8. Using Parser Trees: Combining Lexical and Syntactic Analyses

Sentiment detection

- Why simple valence scoring is not good enough
  - Mixed signals cancel out within a sentence
  - Can’t tell what each signal attaches to
- Improvement by adding a parser
  - Extract NPs, VPs that contain unidirectional sentiment
  - Makes for a more informative synopsis

DataScience@SMU