

Biomedical Natural Language Processing (NLP): Methods and Applications HI6330/HI6330W – Syllabus

Tentative schedule:

- NLP essentials
 - Week 1 – Introduction to NLP
 - Week 2 - 3 – Dealing with words
 - Week 3 - 4–Processing sentences and corpus
- NLP applications
 - Week 5– Text classification
 - Week 6– Information extraction
 - Week 7– Information retrieval
 - Week 8– Question answering
 - Week 9– Sentimental analysis
 - Week 10–Mid-term exam
- Biomedical NLP tasks
 - Week 11 – Deep learning
 - Week 12 – Biomedical NLP resources and systems
 - Week 13 – no class (AMIA Annual Symposium)
 - Week 14– Biomedical NLP tasks: i2b2 and BioCreative Challenges
 - Week 15-16– Final Project (presentation date: TBD)

Contents:

1. NLP Essentials (week 1-4)

1.1 Introduction to NLP (week-1)

1.1.1 Origin

Artificial intelligence, Turing Test, Rationalist vs empiricist, statistical NLP (Frederik Jelinek)

1.1.2 Application (week-1)

IR-google, Machine Translation-google, QA-watson, Information Extraction-MS Outlook, spam-detection- Email, spelling error checking-MS word

1.1.3 Challenge (week-1)

Ambiguity

1.2 Linguistic Essentials (week 2-4)

1.2.1 Word (week-2)

Tokenization (Regular Expression)

Normalization (stemmer, lemmatizer)

Part of Speech, Tagging

Word Sense Disambiguation (probabilistic theory essential, naïve bayes)

1.2.2 Sentence (week-3)

Sentence boundary (Rule-based)

Syntactic structure

Parsing, Chunking

1.2.3 Corpus-based Work (week-4)

Annotation, Tagging, Agreement

Popular corpora

Corpus-based tasks, mutual information, entropy

N-gram, Smoothing, Language Model

2. NLP Applications and Techniques (week 5-9)

2.1 Text classification (week-5)

Decision Trees, Random Forests, Support Vector Machines,
Bayesian classification

2.2 Information Extraction

2.2.1 Named Entity Recognition (week-6)

Markov Models (HMM), Conditional Random Fields (CRFs)
Knowledge based, Machine Learning based

2.2.2 Mining Relations (week-6)

Rule based, Machine Learning based

2.3 Information Retrieval (IR) (week 7)

2.3.1 IR models (week-7)

Set-theoretic, Vector Space Model, Probabilistic Model

2.3.2 IR techniques (week-7)

Document model, term weighting, smoothing

2.3.3 IR system (week-7)

Web crawler, Redundancy, Indexing, Ranking

2.4 Question Answering (week-8)

2.4.1 Question Analysis

Question classification, Information role Recognition, Query
generation

2.4.2 Answer Retrieval

Sentence/passage retrieval

2.4.3 Answer Generation

Answer extraction, Answer summarization, Answer re-ranking

2.5 Sentimental Analysis (week-9)

----- **Mid-term Exam (week-10)** -----

2.6 Deep learning (week-11)

3. Biomedical NLP Applications (week 12-15)

3.1 Biomedical Resources (week-12)

Literature: Ontology (Gene ontology), Network, Knowledgebase (PharmGKB)

Clinical: ICD codes, UMLS, RxNorm, SnoMed, Clinical Trial, I2B2

3.2 Biomedical NLP systems (week-12)

MedLEE, MetaMap, cTAKES, KnowledgeMap

3.3 Biomedical NLP Applications (week-12)

Biomedical Entity Recognition & Normalization

Biomedical Entity Relation Mining

----- **NO class (AMIA) (week-13)** -----

3.4 Biomedical NLP Case Study (week 14)

3.4.1 Bio creative (week-14)

3.4.2 I2B2 (week-14)

----- **Final Project (week-15-16)** -----