

Course Recap

CMSC 473/673

Frank Ferraro

Course Goals

- Be introduced to some of the core problems and solutions of NLP (big picture)
- Learn different ways that success and progress can be measured in NLP
- Relate to statistics, machine learning, and linguistics
- Implement NLP programs
- Read and analyze research papers
- Practice your (written) communication skills

NLP ↔ Machine Learning

Goal: Learn parameters
(weights) θ to develop a
scoring function that
says how “good” some
provided text is

$$S = p_{\theta}(s, h)$$

s: Michael Jordan, coach Phil Jackson and the star cast, including Scottie Pippen, took the Chicago Bulls to six National Basketball Association championships.

h: The Bulls basketball team is based in Chicago.

Goal: Learn parameters (weights) θ to develop a scoring function that says how “good” some provided text is

Text Annotation Tasks

1. Classify the entire document
2. Classify word tokens individually
3. Classify word tokens in a sequence
4. Identify phrases (“chunking”)
5. Syntactic annotation (parsing)
6. Semantic annotation

Course Recap

Basics of Probability

Requirements to be a distribution
("proportional to", \propto)

Definitions of conditional probability, joint probability, and independence

Bayes rule, (probability) chain rule

Tasks and Classification (use Bayes rule!)

Posterior decoding vs. Bayes decomposition model

Evaluations: accuracy, precision, recall, and F_β (F_1) scores

Basics of ML

Course Recap

Basics of Probability

Requirements to be a distribution (“proportional to”, \propto)

Definitions of conditional probability, joint probability, and independence

Bayes rule, (probability) chain rule

Tasks and Classification (use Bayes rule!)

Posterior decoding vs. Bayes decomposition model

Evaluations: accuracy, precision, recall, and F_β (F_1) scores

Basics of language modeling

Goal: model (be able to predict) and give a score to *language* (whole sequences of characters or words)

Simple count-based model

Smoothing (and why we need it): Laplace (add- λ)

Evaluation: perplexity

Word Modeling

Course Recap

Basics of Probability

- Requirements to be a distribution (“proportional to”, \propto)

- Definitions of conditional probability, joint probability, and independence

- Bayes rule, (probability) chain rule

Tasks and Classification (use Bayes rule!)

- Posterior decoding vs. Bayes decomposition model

- Evaluations: accuracy, precision, recall, and F_β (F_1) scores

Basics of language modeling

- Goal: model (be able to predict) and give a score to *language* (whole sequences of characters or words)

- Simple count-based model

- Smoothing (and why we need it): Laplace (add- λ)

- Evaluation: perplexity

Maximum Entropy Models

- Meanings of feature functions and weights

- Use for language modeling or conditional classification (“posterior in one go”)

- How to learn the weights: gradient descent

Distributed Representations & Neural Language Models

- What embeddings are and what their motivation is

- A common way to evaluate: cosine similarity

Course Recap

Basics of Probability

Requirements to be a distribution (“proportional to”, \propto)

Definitions of conditional probability, joint probability, and independence

Bayes rule, (probability) chain rule

Tasks and Classification (use Bayes rule!)

Posterior decoding vs. Bayes decomposition model

Evaluations: accuracy, precision, recall, and F_β (F_1) scores

Basics of language modeling

Goal: model (be able to predict) and give a score to *language* (whole sequences of characters or words)

Simple count-based model

Smoothing (and why we need it): Laplace (add- λ)

Evaluation: perplexity

Maximum Entropy Models

Meanings of feature functions and weights

Use for language modeling or conditional classification (“posterior in one go”)

How to learn the weights: gradient descent

Distributed Representations & Neural Language Models

What embeddings are and what their motivation is

A common way to evaluate: cosine similarity

Recurrent Sequence Modeling

Neural/recurrent cell

Training setup

Language modeling, sequence tagging, sequence-to-sequence

Examples: LM, POS/NER tagging, machine translation, summarization

Pytorch

Word Modeling →
Sequences/Structures

Word Modeling

Course Recap

Basics of Probability

Requirements to be a distribution (“proportional to”, \propto)

Definitions of conditional probability, joint probability, and independence

Bayes rule, (probability) chain rule

Tasks and Classification (use Bayes rule!)

Posterior decoding vs. Bayes decomposition model

Evaluations: accuracy, precision, recall, and F_β (F_1) scores

Basics of language modeling

Goal: model (be able to predict) and give a score to *language* (whole sequences of characters or words)

Simple count-based model

Smoothing (and why we need it): Laplace (add- λ), interpolation, backoff

Evaluation: perplexity

Maximum Entropy Models

Meanings of feature functions and weights

Use for language modeling or conditional classification (“posterior in one go”)

How to learn the weights: gradient descent

Distributed Representations & Neural Language Models

What embeddings are and what their motivation is

A common way to evaluate: cosine similarity

Recurrent Sequence Modeling

Neural/recurrent cell

Training setup

Language modeling, sequence tagging, sequence-to-sequence

Examples: LM, POS/NER tagging, machine translation, summarization

Pytorch

Building Blocks of Universal Language Models

Attention

Transformers

Examples: BERT, GPT-2, and T5

Huggingface transformers

Word Modeling →
Sequences/Structures

Word Modeling

Course Recap

Basics of Probability

Requirements to be a distribution (“proportional to”, \propto)

Definitions of conditional probability, joint probability, and independence

Bayes rule, (probability) chain rule

Tasks and Classification (use Bayes rule!)

Posterior decoding vs. Bayes decomposition model

Evaluations: accuracy, precision, recall, and F_β (F_1) scores

Basics of language modeling

Goal: model (be able to predict) and give a score to *language* (whole sequences of characters or words)

Simple count-based model

Smoothing (and why we need it): Laplace (add- λ), interpolation, backoff

Evaluation: perplexity

Maximum Entropy Models

Meanings of feature functions and weights

Use for language modeling or conditional classification (“posterior in one go”)

How to learn the weights: gradient descent

Distributed Representations & Neural Language Models

What embeddings are and what their motivation is

A common way to evaluate: cosine similarity

Recurrent Sequence Modeling

Neural/recurrent cell

Training setup

Language modeling, sequence tagging, sequence-to-sequence

Examples: LM, POS/NER tagging, machine translation, summarization

Pytorch

Building Blocks of Universal Language Models

Attention

Transformers

Examples: BERT, GPT-2, and T5

Huggingface transformers

Ethical Considerations of NLP

Data bias

Model and featurization bias

LLMs

Word Modeling

Sequences/Structures

Pick Your Toolkit

PyTorch
Huggingface Transformers
Deeplearning4j
TensorFlow
DyNet
Caffe

Keras
MxNet
Gluon
CNTK
...

Comparisons:

https://en.wikipedia.org/wiki/Comparison_of_deep_learning_software

<https://deeplearning4j.org/compare-dl4j-tensorflow-pytorch>

build a system to translate
create a QA system

engineering focus

Natural Language Processing

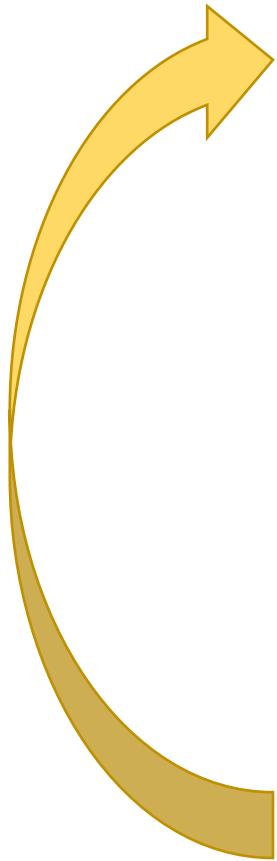
≈

Computational Linguistics

science focus

computational bio
computational chemistry
computational X

these views can co-exist peacefully



Natural Language Processing \approx Computational Linguistics

Both have impact in/contribute to/draw from:

Machine learning

Linguistics

Information Theory

Cognitive Science

Data Science

Psychology

Systems Engineering

Political Science

Logic

Digital Humanities

Theory of Computation

Education

The NLP Research Community

Papers

- ACL Anthology (<http://aclweb.org/anthology>) has nearly everything, free! As of late 2023:
 - Over 92,000 papers!
 - Free-text searchable
 - Great way to learn about current research on a topic
 - New search interfaces currently available in beta
 - Find recent or highly cited work; follow citations
 - Used as a dataset by various projects
 - Analyzing the text of the papers (e.g., parsing it)
 - Extracting a graph of papers, authors, and institutions (Who wrote what? Who works where? What cites what?)

Conferences

- Most work in NLP is published as 8-page conference papers with 3 double-blind reviewers.
- Main annual conferences: ACL, EMNLP, NAACL
 - Also EACL, IJCNLP, COLING
 - + journals (TACL, *Computational Linguistics* [CL])
 - + various specialized conferences and workshops
- Big events, and growing fast!

Thank you for a great semester!

Natural language processing

Semantics

ITE 358

ferraro@umbc.edu

Vision & language processing

Learning with low-to-no supervision