NLP without Annotated Dataset

Course Review

Sowmya Vajjala

Seminar für Sprachwissenschaft, University of Tübingen, Germany

29 January 2021

Topics we covered

1. NLP Overview
2. NLP system development pipeline
3. Corpus collection, extraction, exploration
4. Automatically labeling data
5. Snorkel: Spam Classification without annotated data, Data augmentation with annotated data
6. Text Embeddings and Transfer Learning: An overview
7. Other shorter topics: NLP for Endangered Languages, NLP for Language Learning
8. Group Discussions on various topics.

NLP Overview

- Different faces of NLP: Research, Industry, Other disciplines
- Various day to day applications
- Challenges with NLP
- Some common tasks and
- Degrees of language processing

NLP Pipeline

- How does one build NLP based software?
- How do we represent text for NLP?
- A real world case study
- A code walk through of the pipeline

Corpus collection, extraction, exploration

- How can we get data for NLP?
- How do we extract text from various file formats?
- What is inside our corpus?
- What are some issues to consider while developing/using a corpus?

Automatic Data Creation

- Different ways of labeling data
- Weak supervision
- Data augmentation
- Other topics:
  - Text anonymization
  - Text classification

Snorkel

- Generating training data for spam classification
- Using text augmentation for improving spam classification

Text Embeddings and Transfer Learning

- Bag of words to BERT
- What happens during fine-tuning?

Other Topics

- NLP for Endangered Languages
- NLP for Language Learning
- NLP careers
- Lot of interesting discussion papers

Graded Part

- Assignments
- Group discussion (DONE!)
- Term paper
- Classroom participation (DONE!?)

Some Ideas for Termpaper

I am taking text classification as an example a I covered it in class

- Compare different data augmentation methods for a simple text classification task
- Explore different *BERT models for a simple task
- Generating a labeled dataset for some new classification/IE task with Snorkel
- Surveying existing resources and software for a new language, and listing a few directions on how to extend them
- Exploring already existing datasets for its coverage, identifying potential issues with using it etc (e.g., following datasheets paper)
- Evaluating an existing tool (e.g., Spacy NER) in terms of how it does for various categories of text, using standard evaluation sets.

....

Initial ideas: Course Objectives

- Provide an overview of NLP system development pipeline
- Discuss some common approaches for collecting, cleaning and exploring text data
- Introduce some methods to develop labeled data for NLP

Initial ideas: Learning Outcomes

Students should be able to:

- Understand the end to end NLP system development pipeline
- Compile and explore labeled/annotated corpora for NLP
- Build some basic text classification and information extraction systems

... upon successful completion of the course..

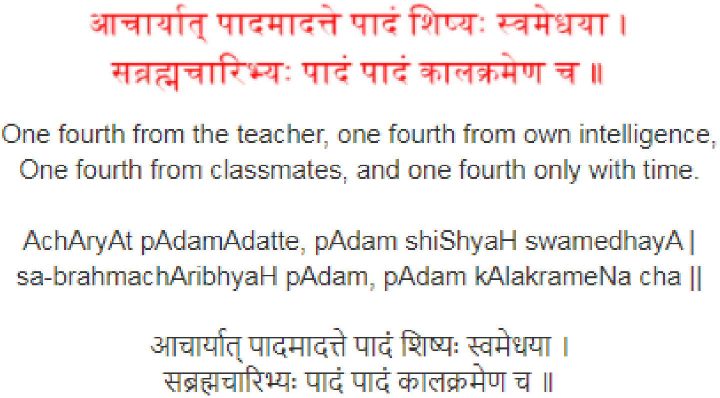
Initial ideas: What the course can't do

- Don't expect to become an NLP expert with one compact course.
- Contents may not always meet your own expectations, but there is a term paper and a group discussion, which gives you opportunities to explore your specific interests related to this topic.
- The course won't teach you programming.

Upcoming Deadlines

- March 1st: Assignments, Term paper
- Today: Make a decision on whether or not you want to submit a term paper, and let me know! (thanks for those who already informed!)
- Soon: Start working on assignments and term paper!

How we learn and grow



आचार्यात् पादमादत्ते पादं शिष्यः स्वमेधया ।
सब्रह्मचारिभ्यः पादं पादं कालक्रमेण च ॥

One fourth from the teacher, one fourth from own intelligence,
One fourth from classmates, and one fourth only with time.

AchAryAt pAdamAdatte, pAdam shiShyaH swamedhayA |
sa-brahmachAribhyaH pAdam, pAdam kAlakrameNa cha ||

आचार्यात् पादमादत्ते पादं शिष्यः स्वमेधया ।
सब्रह्मचारिभ्यः पादं पादं कालक्रमेण च ॥

Source

Thank you!

contact: sowmya.vajjala @ nrc-cnrc.gc.ca
Feel free to connect on social networks (Linkedin, twitter etc) if you want to (finding me is up to you)

Wish you all good luck!!