Seminar für Sprachwissenschaft, University of Tübingen

NLP without a readymade annotated dataset

WS 2020/21 - Course Description Version 3

Instructor: Sowmya Vajjala

- National Research Council, Canada
- *Email:* sowmya.vajjala@nrc-cnrc.gc.ca

Course Overview: Natural Language Processing (NLP) is a part of many day to day applications we use, such as search engines, virtual assistants on your smartphones and various functionalities in your email provider. When we think of NLP, we think of the various algorithms, neural network architectures, and so on. However, what drives all of them are large collections of annotated corpora. However, in many research and real-world scenarios, when we encounter a new problem which can be solved using NLP, we don't have such ready made datasets. This course gives you an overview of how to approach such scenarios - how to collect and cleanup the textual data, how to develop initial labeled datasets, and how to build first solutions from them. The course will cover both research and practical aspects.

Learning Outcomes: Upon successful completion of this course, students are expected to be able to know the following:

- 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

Pre-requisites:

1. Intermediate proficiency in any programming language (Python preferred)
2. Comfortable installing libraries etc on their laptops
3. Knowledge of the usage of virtual environments (venv, anaconda) is useful

Meeting time: January 8 2021-January 29, 2021, M W F, 17:00 s.t. - 18:30 (Central European Time).

Dates: (10 sessions in total)

1. 8th Jan 2021 (Friday)
2. 11th, 13th, 15th Jan 2021 (Mon, Wed, Fri)
3. 18th, 20th, 22nd Jan 2021 (Mon, Wed, Fri)

4. 25th, 27th, 29th Jan 2021 (Mon, Wed, Fri)

Credits: 3 CP + 3 CP for the term paper

Course Format We will meet 3 times in a week (1.5 hours/session). Primary mode of instruction is through online video lecture + discussion. Lecture slides for each session will be uploaded in advance and you are expected to read the recommended readings before listening to the lecture. Assignments will all be uploaded before the start of the course and are due a week after the last class (06th February 2021). Term paper (if you write) is due 2 weeks after the last class, i.e., 13th Feb 2021.

Resources/Reading Materials Books: There is no single textbook. We will try to rely on publicly accessible resources for as much as possible.

1. "Speech and Language Processing" by Jurafsky and Martin (2nd Edition: https://github.com/rain1024/slp2-pdf. 3rd Edition: https://web.stanford.edu/~jurafsky/slp3/)
2. "Python for Everybody" Charles Severance https://www.py4e.com/html3/ (For Python)
3. "Practical Natural Language Processing" by Sowmya Vajjala, Bodhisattwa Majumder, Anuj Gupta and Harshit Surana. https://www.amazon.de/Practical-Natural-Language-P dp/1492054054/. The book is also available for free on O'Reilly's online learning platform (if your university is subscribed). A 30 day trial code to read the book online is here: https://learning.oreilly.com/get-learning/?code=PNLP20.
4. NLTK book -https://nltk.org/book

Course Website: on Moodle, more details are announced later.

List of Topics (tentative)

Note: The following syllabus is tentative, and more detailed information along with relevant readings will be added by the time the course starts.

1. **Introduction (1 session)**
   - Course overview
   - Introduction to NLP
   - NLP system development pipeline

optional Python Overview

Readings: (Note that you are not obligated to read everything thoroughly).

- Chapter 1 from "Speech and Language Processing" by Jurafsky and Martin, 2nd edition (available online) (SLP book from now on)
- "Python for Everybody" by Charles Severence. https://www.py4e.com/html3/ (Py4e from now on)

- Chapter 2 from "Practical Natural Language Processing" by Vajjala et.al. (PNLP book from now on)

2. **NLP Pipeline (1 Session)**

   - Various steps in NLP system development process
   - An example demonstrating the process

   Readings: Chapter 2 in "Practical Natural Language Processing"

3. **Corpus collection, extraction, exploration (1 Session)**

   - Collecting textual data from various sources (e.g., social media text, ethical issues etc)
   - Reading text in different formats (e.g., pdf, html, text, doc etc)
   - Corpus analysis (basic analysis - e.g., frequent words/phrases etc)
   - Probing the corpus for linguistic phenomenon coverage
   - Understanding bias in the corpus and other issues
   - Basic visualization

   Readings: Chapter 1-4 from "NLTK book" (https://nltk.org/book) [Assignment 1 on this topic]

4. **Automatically labeling data (2 sessions)**

   1.1 Regular expressions: Overview

   1.2 Automatic labeling of data (with Snorkel)

   1.3 Data augmentation - an overview (with snorkel)

   1.4 Illustration of some NLP applications with such data (text classification, information extraction)

   Readings: Py4e, Chapter 11; Snorkel usecases (https://www.snorkel.org/use-cases/), A visual survey of data augmentation for NLP (https://amitness.com/2020/05/data-augmentation-for-nlp/) [Assignment 2 on this topic]

5. **Working with small datasets: transfer learning (1 session)**

   (a) Different forms of neural embeddings

   (b) Using neural embeddings in NLP

   (c) Transfer learning with BERT

6. **Student presentations (3 sessions)** Students can work in teams of 2-4 people and present one of the research papers related to course topics, from a given list of papers. Papers are listed at the end of this document. If you want to present a different paper, talk to me first.

7. Recap (1 session)

   - Discussion on topics covered

- Review of exercises
- Resources for the future

8. Student term papers
    - Work on a short project involving NLP and write a report describing your work (6-8 pages long in single column, latex formatted document)
    - Some ideas are listed towards the end of this document. If you want to work on something else, talk to me first.

Assignments/Grading (for 6 CP)

1. 2 Assignments (30% of the grade)
2. 1 presentation (30% of the grade)
3. 1 term paper (30% of the grade)
4. classroom participation (10% of the grade)

Important Deadlines

1. Decide on a team for group discussion (13th Jan 2021)
2. Decide on a paper for group discussion (15th Jan 2021)
3. Group Discussions (22nd-27th Jan 2021)
4. Assignments 1 and 2 Submission (6th Feb 2021)
5. Decide on term paper topic (29th Jan 2021)
6. Term paper submission (13th Feb 2021)

Note: All Assignments are due by 6th Feb 2021. Group Discussions are evaluated during class sessions. Term paper is due 2 weeks after the last class, i.e., 13th Feb 2021.

Papers for Group Discussion: (Note: There is a lot of work in all these sub-divisions I made below. I just chose a few I know of. If you want to choose something else, let me know in advance!)

- **NLP pipeline**
    1. Smith, N. A. (2020). Contextual word representations: putting words into computers. Communications of the ACM, 63(6), 66-74.
    2. Bernier-Colborne, G., & Langlais, P. (2020, May). HardEval: Focusing on Challenging Tokens to Assess Robustness of NER. In Proceedings of The 12th Language Resources and Evaluation Conference (pp. 1704-1711).

3. Srivastava, A., Makhija, P., & Gupta, A. (2020, November). Noisy Text Data: Achilles' Heel of BERT. In Proceedings of the Sixth Workshop on Noisy User-generated Text (W-NUT 2020) (pp. 16-21).
4. Discussion of an existing software/casestudy: https://eng.uber.com/cota/
5. Ribeiro, M. T., Wu, T., Guestrin, C., & Singh, S. (2020). Beyond Accuracy: Behavioral Testing of NLP Models with CheckList. arXiv preprint arXiv:2005.04118.

- **Corpus collection, Exploration**

1. Marcus, M., Santorini, B., & Marcinkiewicz, M. A. (1993). Building a large annotated corpus of English: The Penn Treebank.
2. Keung, P., Lu, Y., Szarvas, G., & Smith, N. A. (2020, November). The Multilingual Amazon Reviews Corpus. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP) (pp. 4563-4568).
3. Bustamante, G., Oncevay, A., & Zariquiey, R. (2020, May). No data to crawl? monolingual corpus creation from PDF files of truly low-resource languages in Peru. In Proceedings of The 12th Language Resources and Evaluation Conference (pp. 2914-2923).
4. Bender, E. M., & Friedman, B. (2018). Data statements for natural language processing: Toward mitigating system bias and enabling better science. Transactions of the Association for Computational Linguistics, 6, 587-604.
5. Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Daumé III, H., & Crawford, K. (2018). Datasheets for datasets. arXiv preprint arXiv:1803.09010.

- **Data Labeling and Augmentation**

1. Ratner, A. J., Bach, S. H., Ehrenberg, H. R., & Ré, C. (2017, May). Snorkel: Fast training set generation for information extraction. In Proceedings of the 2017 ACM international conference on management of data (pp. 1683-1686).
2. Wei, J., & Zou, K. (2019, November). EDA: Easy Data Augmentation Techniques for Boosting Performance on Text Classification Tasks. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP) (pp. 6383-6389).
3. Amjad, M., Sidorov, G., & Zhila, A. (2020, May). Data augmentation using machine translation for fake news detection in the Urdu language. In Proceedings of The 12th Language Resources and Evaluation Conference (pp. 2537-2542).
4. Chen, Y., Liu, S., Zhang, X., Liu, K., & Zhao, J. (2017, July). Automatically labeled data generation for large scale event extraction. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 409-419).
5. Tambi, R., Kale, A., & King, T. H. (2020, May). Search Query Language Identification Using Weak Labeling. In Proceedings of The 12th Language Resources and Evaluation Conference (pp. 3520-3527).

- **NLP Research in under resourced scenarios**

    1. Rijhwani, S., Anastasopoulos, A., & Neubig, G. (2020, November). OCR Post-Correction for Endangered Language Texts. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP) (pp. 5931-5942).
    2. Chau, E. C., Lin, L. H., & Smith, N. A. (2020). Parsing with multilingual bert, a small corpus, and a small treebank. arXiv preprint arXiv:2009.14124.
    3. Zhang, B., Lu, D., Pan, X., Lin, Y., Abudukelimu, H., Ji, H., & Knight, K. (2017, November). Embracing non-traditional linguistic resources for low-resource language name tagging. In Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers) (pp. 362-372).
    4. Mehta, D., Santy, S., Mothilal, R. K., Srivastava, B. M. L., Sharma, A., Shukla, A., ... & Bali, K. (2020). Learnings from Technological Interventions in a Low Resource Language: A Case-Study on Gondi. arXiv preprint arXiv:2004.10270.
    5. Kakwani, D., Kunchukuttan, A., Golla, S., Gokul, N. C., Bhattacharyya, A., Khapra, M. M., & Kumar, P. (2020, November). iNLPSuite: Monolingual Corpora, Evaluation Benchmarks and Pre-trained Multilingual Language Models for Indian Languages. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings (pp. 4948-4961).

- **Other Topics in NLP**

    1. Ribeiro, M. T., Singh, S., & Guestrin, C. (2018, February). Anchors: High-Precision Model-Agnostic Explanations. In AAAI (Vol. 18, pp. 1527-1535).
    2. Lau, J. H., & Baldwin, T. (2020, July). Give Me Convenience and Give Her Death: Who Should Decide What Uses of NLP are Appropriate, and on What Basis?. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (pp. 2908-2913).
    3. Loukina, A., Madnani, N., & Zechner, K. (2019, August). The many dimensions of algorithmic fairness in educational applications. In Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications (pp. 1-10).
    4. Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. arXiv preprint arXiv:1910.01108.
    5. Zhong, H., Xiao, C., Tu, C., Zhang, T., Liu, Z., & Sun, M. (2020). How Does NLP Benefit Legal System: A Summary of Legal Artificial Intelligence. arXiv preprint arXiv:2004.12158.

Note: Presentation time is around 15 min per group, with 10-15 min more for discussion.

Term paper ideas: (You can choose to work on any other idea that interests you. Talk to me first to know whether it is a) feasible and b) relevant for this course)

- Pick a demo paper from any of the recent (2017-20) NLP conferences (ACL, EACL, EMNLP, NAACL, COLING) or any open source tool from NLP-OSS workshop series. Use the tool and write a report evaluating it for various use cases.
- Use Snorkel for spam classification, in a language of your choice (e.g., German) and write a report on your observations about working with the tool, and its performance. You have to look for any relevant datasets yourselves.

Note: Any code you write should be submitted with the report.