Introduction to Natural Language Processing

Course Wrap Up

Outline of the Course

- This is a *introduction* class to natural language processing
- Focus will be on (1) *introducing* you to key concepts from NLP used to describe and analyze languages (2) *writing* programs in Python to manipulate and analyze language data

What You have Learned

- **Collect texts**: scraping Twitter, scraping news, scraping web pages, scraping …
- **Clean and Process texts**: sentence segmentation, tokenization, normalization: lemmatization, lower casing, stop words, ...
- **Language Model**: N-gram LM, Neural LM, Vision-Language Model
- **Lexical and Vector semantics**: WordNet, word senses, word-doc matrix, word-word matrix, TFIDF, word2vec, Glove, cosine similarity, bias, Transformers, BERT, XLM-Roberta, GPTs, T5, …
- **Machine Learning models**: NB, LR, NN, RNN, LSTM, GRU, Transformers, seq2seq, domain adaptation, vision transformers, contrastive learning, ...
- **NLP tasks**: POS tagging and other structured prediction task (NER), sentiment analysis, morphological modification, machine translation, information extraction, coreference resolution, entity linking, question answering
- **Libraries**: scrapy, requests, newspaper, BeautifulSoup, newspaper, wordcloud, NLTK, spacy, sklearn, gensim, pytorch, transformers, …
- **But there is still much more**: dialogue systems, textual entailment, summarization, language identification, text simplification, …

Your Assignments Did This:

- Getting your basics down on the mathematics behind NLP models (objective function, MLE, perplexities, attention, …), getting your hands dirty on various NLP tasks: language modeling, multilingual text classification, vector space models, textual similarities (doc to doc, word to word), …

- You learned how to formulate the problem, convert text to features, implement models to solve the problem using the features

- You tried a baseline algorithm, cross validation, to see how you can improve

- You learned to implement different features, approaches to solve the problem

An **Exciting** Time for NLP!

- Increase in **computing resources**
- Increase in the **amount of data** and information available in digital form
- Development of highly successful **machine learning methods** and **competitive evaluations** (SemEval, NIST, CoNLL shared tasks, Kaggle)
- Richer understanding of the structure of human language and its **deployment in social contexts**

An **Exciting** Time for NLP!

- Real world NLP systems
  - Conversational agents for making reservations
  - Voice assistant systems
  - Machine translation systems, speech-to-speech translation
  - Automated grading systems
  - Virtual tutors
  - Text analysis companies for marketing intelligence
  - …

State of the Art

- Simple methods often work very well when trained on **large quantities of data**
  - e.g., many text and sentiment classifiers still rely on different sets of words (“bag of words”) without regard to sentence and discourse structure or meaning

State of the Art

- However, most NLP resources and systems are **available only for high resource languages**
  - Many low resource languages are spoken by millions of people e.g., Bengali, Indonesian, Swahili, Javanese, Igbo, …
  - The challenge is how to develop resources and tools for thousands of languages, not just a few

Lessons so far

- More data -> better performance
- Less data ?
  - new approach?
  - look for more data?

Toward More Human-like Learning and Thinking Machines

- Seeing objects and agents rather than features
- Building causal models and not just recognizing patterns
- Recombining representations without needing to retrain
- Learning-to-learn rather than starting from scratch

Building machines that learn and think like people, Lake et al.,
Behavioral and Brain Sciences, 2017

What you can do to stay involved

- Do a research project with me
- If you're graduating, then stay in touch!
  - Email: wijaya@bu.edu

Research Projects

- News Framing
- Machine Translation for Low Resource Languages
    - More generally, learning under low-resource:
        - Few-shot learning
        - Data augmentation
- NLP Applications for Public Health

Thanks!!!

- Course feedback: bu.campuslabs.com/courseeval