

LING 334 - Introduction to Computational Linguistics

Week 10

—

Applied NLP

Goals This Week

- Survey the landscape of contemporary CL/NLP
 - Research structure
 - Publicly available data and software
- Point you at resources for learning more after this class
- Hopefully useful for final projects and beyond!

This Class vs. Real World

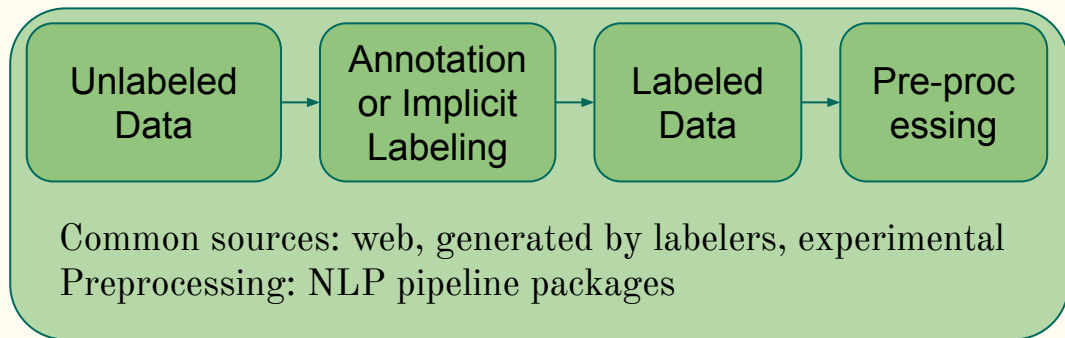
- You will generally not need to implement these algorithms from scratch!
 - Especially now that you know how they work.
 - We are beneficiaries of a huge history of publicly available data and code.
 - On a day-to-day basis, just use what's out there!
- However the goals will be less clear-cut, and more defined by you.
 - What questions are you interested in? What data? Why CL/NLP?

Existing Resources

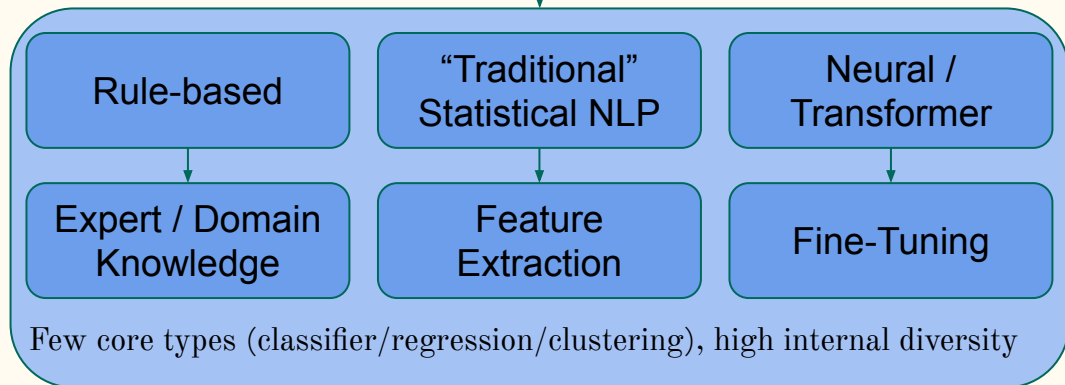
- Tutorials, books, course materials
 - CL/NLP is well-resourced in free materials!
 - I'll give some pointers today, but always look around
- Code Documentation
 - Do not be afraid to dig in and read open-source code!
- People of course!
 - For instance, I'll be here and am interested in helping you continue your CL/NLP journey if you do!

Trajectories of CL/NLP Projects (non-exhaustive)

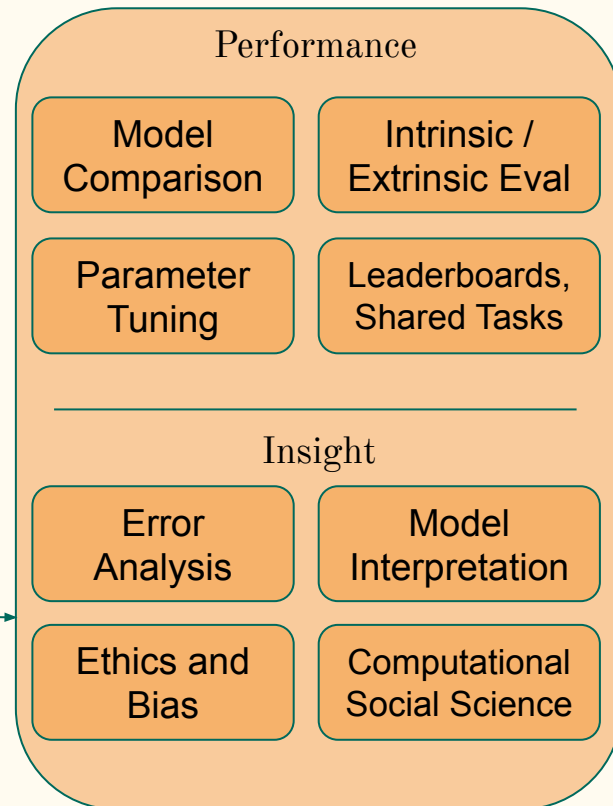
1. Data



2. Model



3. Analysis



Data

- As we've seen, crucially important!
- Generally defines what is possible to do.
- Often collecting your own dataset is productive!
 - You know what you're getting, you define the protocol.
 - E.g., almost anything on the internet can be scraped into a dataset.
- But also time-consuming.
- Frequently, better to use existing resources to start.

Data - Public Sources

- Kaggle - <https://www.kaggle.com/datasets>
- HuggingFace Datasets - <https://huggingface.co/datasets>
- PapersWithCode - <https://paperswithcode.com/datasets>
- ACL Resources Wiki -
[https://aclweb.org/aclwiki/List of resources by language](https://aclweb.org/aclwiki/List_of_resources_by_language)
- DatasetList - <https://www.datasetlist.com/>
- AWS Open Data - <https://registry.opendata.aws/>
- OpenSLR - <http://www.openslr.org/resources.php>
- And many more!!! Seek and ye shall find.

Data - More Closed Sources

- Linguistic Data Consortium - catalog ldc upenn edu
 - At NU we have access to all of these corpora, and I manage them, so you can email me if you find something you're interested in!
- BYU Corpora - <https://www.corpusdata.org/corpora.asp>
 - Several widely-used and well-curated corpora; as with LDC we have access to these and I manage them.

Data - Public APIs

- Twitter, Reddit, other large portals provide “Application Programming Interfaces” which can be used to collect data.
- Varying rules and processes, commonly have to apply for developer access.
- Always be aware of licensing!
Especially for commercial projects.

Preprocessing and NLP Pipelines

Many high-quality libraries for basic preprocessing and NLP pipeline functions:

Tokenization, lemmatization, part-of-speech tagging, syntactic parsing, named entity recognition, etc.

Here's a few popular examples.

NLP Pipelines - NLTK

Natural Language Toolkit - <https://www.nltk.org/>

Earlier, classic library. Still great! Free and comprehensive
corresponding book: <https://www.nltk.org/book/>

Particularly useful for access to lexical resources and corpora.

Drawbacks:

- no concept of distributional semantics
- Largely focused on English

NLP Pipelines - CoreNLP and Stanza

Stanford CoreNLP - <https://stanfordnlp.github.io/CoreNLP/>

Also classic, highly used in research and beyond.

Built in Java, but interfaces to Python and other languages.

Stanza - <https://stanfordnlp.github.io/stanza/>

More modern neural network pipeline, written in Python.

Support for 70+ natural languages, built to interface with:

Universal Dependencies - <https://universaldependencies.org/>

Attempt at universal cross-linguistic syntax representation.

NLP Pipelines - spaCy

spaCy - <https://spacy.io/>

Fast, modern, deep functionality.

Natively incorporates static and contextual word vectors.

Dependency Visualizer demo:

<https://explosion.ai/demos/displacy>

Models - Statistical NLP

NLTK, spaCy, others include classifier models etc.

I recommend however using them as preprocessing / feature extraction for dedicated machine learning libraries.

ML Libraries - Scikit-Learn

Scikit-Learn - <https://scikit-learn.org/stable/>

Many well-implemented models for statistical ML,
excellent documentation and tutorials on the web.

E.g.,

[https://scikit-learn.org/stable/tutorial/text_analytics/working
with_text_data.html](https://scikit-learn.org/stable/tutorial/text_analytics/working_with_text_data.html)

ML Libraries - Neural / Transformer

Keras - <https://keras.io/>

Good abstractions and plug-and-play layers; recent accompanying book which looks great

TensorFlow - <https://tensorflow.org/>

Google product, institutional support there

PyTorch - <https://pytorch.org/>

De facto standard, backbone for many applications

Modeling Packages - Gensim

Gensim - <https://radimrehurek.com/gensim/>

Focus on topic modeling, but support for other sorts of semantic clustering algorithms.

Modeling Packages - AllenNLP

AllenNLP - <https://allennlp.org/>

Incorporates many cool task-specific models.

Check out the demos:

<https://demo.allennlp.org/>

Modeling Packages - HuggingFace

HuggingFace - <https://huggingface.co/>

Frankly pretty amazing coalescence of open-source effort.

Large library of easily usable pretrained models, especially LLMs / transformers: <https://huggingface.co/models>

They've single-handedly accelerated research in our field.

Practical Considerations for Modeling

Speed and memory can be issues, especially with LLMs

GPU acceleration is sometimes essential depending on the task

Graphics cards are built to do lots of big matrix multiplications really fast, turns out this is what we need for neural ML; NVIDIA huge player here

Analysis

As many possibilities as stars in the sky!

Often, matplotlib for visualization in Python
or ggplot2 for visualization in R.

Great Free Courses on Neural Stuff

Stanford CS224n:

<https://www.youtube.com/playlist?list=PLoROMvodv4rOhcuXMZkNm7j3fVwBBY42z>

CMU CS 11-747:

<https://www.youtube.com/playlist?list=PL8PYTP1V4I8AkaHEJ7l0Orlex-pcxS-XV>

Great Free Course Materials

Berkeley Applied NLP:

<https://people.ischool.berkeley.edu/~dbamman/info256.html>

<https://github.com/dbamman/anlp21>

(great Jupyter walkthroughs of many of these libraries)