NLP For Economists

Corpus Collection

Sowmya Vajjala

Munich Graduate School of Economics - LMU Munich

Guest Course, October 2020

Goals

- Discuss about different means of obtaining/creating corpus for NLP research
- source: Chapter 2 in practicalnlp.ai

What is corpus collection?

- Any NLP system relies on the presence of some corpus for building and evaluating it.
- When we are working on developing new methods/algorithms, we often work with standard datasets to compare among different approaches.
- When we are working on using NLP for domain specific problems, we won't often have such datasets.
- Data can come in various forms, but often, we need some form of "labeled" data.

- where do we get that?

Use available data

Freely accessible:

- scraping websites
- collecting tweets
- newspapers, wikipedia etc.
- public archives of research papers

.... or use licensed data (e.g., NLP researchers use LDC corpora)

Setup data collection experiments

- crowd sourcing
- user studies
- online surveys

....

Automated means

- We can take a small dataset and use some tricks to "generate" more data with similar linguistic characteristics
  - synonym replacement
  - back translation
  - bigram flipping, changing named entities etc
  - using heuristics, regular expressions etc to "create" labeled training data (e.g., Snorkel)

....

Labeling data with snorkel

https://www.snorkel.org/use-cases/

- using labeling functions (LFs) in Snorkel: noisy, programmatic rules and heuristics that assign labels to unlabeled training data. (e.g., using specific wordlists, regular expressions, other heuristics etc.)

```
from snorkel.labeling import labeling_function

@labeling_function()
def lf_contains_link(x):
    # Return a label of SPAM if "http" in comment text, otherwise ABSTAIN
    return SPAM if "http" in x.text.lower() else ABSTAIN

@labeling_function()
def check_out(x):
    return SPAM if "check out" in x.text.lower() else ABSTAIN
```

....

Issues to keep in mind

- you may be collecting the data without asking for permissions (ask: is everything freely visible on the web essentially free for such use?)
- you may potentially be storing personally identifyable information
- training data can bias the predictions of an NLP system (among others)
- ask yourself: is the corpus i am collecting serving my purpose, or is something missing?
- ask yourself: how am i storing/retrieving/sharing this corpus?
- ask yourself: is it static? or constantly changing/updated data? (eg., tweets, news articles etc)

Conclusion

- Many ways to get data, and label it.
- There are some concerns around it too. Worth checking: Diesner, J., & Chin, C. L. (2016, May). Gratis, libre, or something else? Regulations and misassumptions related to working with publicly available text data. In Actes du Workshop on Ethics In Corpus Collection, Annotation & Application (ETHI-CA2), LREC, Portoroz, Slovénie. `http://cpanel.ischool.illinois.edu/~jdiesner/lab/publications/LREC_DiesnerChin.pdf`

ToDo for you

- Think about what kind of data you want to look for and how to obtain it
- Take a relook at Joco corpus in terms of the issues I mentioned 2 slides ago, and about how it was created.
- Do you need annotated data? what sort of annotations are needed? How do you get them? - think about these questions.