Lec 9 Experimental Design

Research Direction

Application - driven — how to make this system better
- └ new task
- └ improve accuracy
- └ improve efficiency

Curiosity - driven — answer some question
- └ linguistic analysis
- └ whether all langs are equally hard to model

▷ Getting research idea
- Bottom-up — incremental improvement identify problem → try fix them
- Top-down — high-level ideas to concrete testing

Survey

→ ACL Anthology
→ Google Scholar

Research Question

Q: Are all langs equally hard to model?
Hypothesis: Unlikely to have architecture good for all langs

For application based:
Q: Does X make Y better? — a natural question
→ What underlying assumption makes Y better?
→ Does it work on toy dataset isolating certain data?
└ e.g. context helps translation, but probably more so for conversation dataset
└ is context even necessary? Test human to find out?
→ If it's not better is it a dataset issue or model issue?

Run Exps

1. Find data (reuse, repurpose, create)
   → Hugging face → ELRA → LDC → Paper with code
   - Annotate data,
     - with annotation guideline!
     - quality assessment — make multiple human annotate
     → Kappa statistics
   - How much data — need sufficient data for stat significance for training, it depends
   - Try document the data
2. Run experiments
   - Modularize into directory & automate each step
3. Evaluation

Stat Significance

paired	vs	unpaired
two models on same dataset		diff in mean for two models on unrelated groups

→ Bootstrap: re-sample many times
→ t-test: only works on additive measures!

▷ Power analysis to estimate test data needed

Ex.
baseline acc	90%
new acc	93%
want	$p < 0.05$
dataset size needed	?

See reference on how to compute this

Report Results

Try plan in advance.
Assume best case with experiments

Analysis, Conclusion

See future lecture