Evaluation Issues in AI and NLP

COMP-550

Nov 28, 2017

Announcements

Course evaluations: please submit one!

A4 reading: Please bring to class on Thursday

Outline

Evaluation in NLP

The Turing Test

Deception in the Turing test

Gaming the measure with “cheap tricks”

Winograd Schema Challenge

Evaluation in NLP

What are some evaluation measures and methods for different NLP tasks that we have discussed in this class?

Classes of Evaluation Methods

Intrinsic measures

- Pertains to the particular task that a model aims to solve

Extrinsic measures

- Pertains to some downstream application of the current model

Separate issue from whether the evaluation is manual or automatic

Let's classify the previous evaluations.

Validity of Evaluations

Different kinds of **validity** in our evaluations, to help us know whether our model is making *real* progress

Internal validity

External validity

Test validity

Internal Validity

Whether a causal conclusion drawn by study is warranted

Conclusion: Method A outperforms Method B

Independent variable: method

Dependent variable: evaluation measure

- Same training data? Same preprocessing?
- Both methods' parameters were tuned?
- No other confounds?
- Methods, evaluation measures, etc. implemented correctly?

External Validity

Whether or not the conclusions drawn by study generalizes to other situations and other data

Conclusion: Method A outperforms Method B

- How big was the test data set?
- Is it representative of all kinds of language?
  - e.g., benchmark data sets usually are drawn from one genre of text
- Is it biased in some way?

Case Study: Parsing Results

Train	Test: BNC	GENIA	BROWN	SWBD	ETT	WSJ	Average
GENIA	66.3	**83.6**	64.6	51.6	69.0	66.6	67.0
BROWN	81.0	71.5	**86.3**	79.0	80.9	80.6	79.9
SWBD	70.8	62.9	75.5	**89.0**	75.9	69.1	73.9
ETT	72.7	65.3	75.4	75.2	81.9	73.2	73.9
WSJ	**82.5**	74.9	83.8	78.5	**83.4**	**89.0**	**82.0**

Table 1: Cross-domain *f*-score performance of the Charniak (2000) parser. Averages are macro-averages. Performance drops as training and test domains diverge. On average, the WSJ model is the most accurate.

Parsing results, from McClosky et al. (2010)

- An evaluation only on WSJ would have limited external validity
- Developing methods that generalize across domains is called **domain adaptation**

Construct Validity

Concerned with whether an evaluation actually measures what it claims to

- Does ROUGE reflect usefulness of summaries?
- Does better perplexity in language modelling lead to lower word error rate in ASR?
- Does lower word error rate lead to better user satisfaction in an ASR system?

Developing an Evaluation Measure

Would like a good, cheap, automatic measure of quality of NLP output

Correlate against some gold standard

- e.g., ROUGE approximates human summary judgment scores
- Could develop automatic measure that correlates with extrinsic task performance measure

As we'll see, overly optimizing on a measure tends to reduce its usefulness -> lowers construct validity

Solving AI and NLP

When will we have solved NLP, or indeed, AI? Is there a single test that can measure progress to this goal?

The Turing Test

Alan Turing proposed this as a way to check whether a machine exhibits intelligent behaviour (1950).

- More concrete than some philosophical question such as "Can machines think?"

Setup:

- Human interlocutor chats with an agent, which is either a machine, or another human, for 5 minutes.
- Afterwards, must decide whether the agent is human or machine

Loebner Prize

Annual competition inspired by the Turing test with prize money:

http://www.aisb.org.uk/events/loebner-prize

Another competition is organized by the University of Reading

- Claim to have passed Turing test:

http://www.bbc.com/news/technology-27762088

Is this format, or the Turing test in general, a good indicator of progress in NLP?

Deception in the Turing Test

Most of the effort seems to be to deceive.

Meet Rose, the 2014 Loebner winner:

http://ec2-54-215-197-164.us-west-1.compute.amazonaws.com/speech.php

What are some aspects of natural language that Rose understands? That Rose does not understand?

Sample Chatlog

Jackie: Where are you right now?
Rose: I'm here, at Bletchley Park.
Jackie: Where's that?
Rose: It doesn't matter.
Jackie: What is the weather there right now?
Rose: It's nice enough out. But it's pretty reasonable the year round.
Jackie: That must be nice. It's pretty cold in Montreal in winter.
Rose: I do prefer nice weather.
Jackie: What do you think about computational linguistics?
Rose: I haven't a clue. San Francisco is perverse. When you expect sunny it's fog and chill winds, but just when you think winter's nearly here, it goes sunny.
Jackie: Do you only care about weather?
Rose: I prefer sunny climates and good weather. Though occasional warm rain is OK.

Goodhart's Law

First formulated in economics:

> Once a measure of some quality is turned into a target to optimize, it is no longer a good measure of quality.

Back to the summarization example:

- Many systems now directly optimize for ROUGE to perform well in shared tasks.
- Some "tricks" clearly do not represent genuine progress
- Others are less clear, but also appear to be so

Optimizing ROUGE

ROUGE is *recall-oriented*

- Make sure we are using the entire word length limit, even if the last sentence is cut off.

ROUGE was developed using purely extractive summarization methods

- Sentence simplification and compression helps ROUGE, because we can fit more content into the same word length limit
- This usually degrades readability and overall quality

Other cases of this in NLP:

- BLEU, PARSEVAL

Ignoring Less Common Issues

Less common, but important and systematic issues are ignored, if we only use standard evaluation measures

e.g., Parsing

- Overall parsing accuracy is relatively high (~90 F1), but parsing of coordinate structures is poor
- Hogan (2007) found that a baseline parser gets about 70 F1 on parsing NP coordination

busloads of [executives and their wives]	CORRECT
[busloads of executives] and [their wives]	INCORRECT

"Cheap Tricks"

Are we overly enamoured by corpus-based, statistical approaches?

Cheap tricks (Levesque, 2013):

- Get the answer right, but for dubious reasons different from human-like reasoning

e.g.,

Could a crocodile run a steeplechase?

- Can use statistical reasoning, closed-world assumption to answer such questions

Should baseball players be allowed to glue small wins on their caps?

Cheap Tricks in NLP

Chatbot:

- Create fictitious personality, backstory
- Deceive with humour, emotional outburst, misdirection

Question answering and information extraction:

- Use existing knowledge bases, regularities in statistical patterns to look up memorized knowledge

Automatic summarization and NLG:

- Use extraction and redundancy to avoid having to really "understand" the text and generate summary sentences (Cheung and Penn, 2013)

Winograd Schema Challenge

Attempt to design multiple-choice questions that require *deeper* understanding beyond:

- Simple statistical look-ups with some search method
- Features that map simply to other features (*older than* maps to AGE)
- Biases in word order, vocabulary, grammar

Basic format: binary questions, where a small change in wording leads to a different correct solution

Example

Joan made sure to thank Susan for all the help she had *given*. Who had *given* the help?

- Joan
- **Susan**

Joan made sure to thank Susan for all the help she had *received*. Who had *received* the help?

- **Joan**
- Susan

https://www.cs.nyu.edu/davise/papers/WS.html

Consequences

For a simplified set of questions, it is possible to use statistical knowledge and existing work in coreference resolution to partially solve WSC questions

- A variety of semantic features fed to a machine learning system -> 73% accuracy (Rahman and Ng, 2012)

On original set of questions, performance remains poor

Bigger point remains:

- Is there a science of AI distinct from the technological aspect of it?
- How do we decide what kinds of techniques are “cheap tricks” vs. genuine “intelligent behaviour”?

Next Class

A4 reading discussion

Bias in NLP systems

Course recap