Evaluation Strategies

Regina Barzilay

April 26, 2004

Evaluation

Challenges:

- Intrinsic subjectivity of some discourse related judgments
- Hard to find corpora for training/testing
  - Lack of standard corpora for most of the tasks
- Different evaluation methodology for different tasks

Evaluation

- Intrinsic:
  - comparison with an "ideal" output
  - subjective quality evaluation
- Extrinsic (task-based):
  - impact of the output quality on the judge's performance in a particular task

Intrinsic Evaluation

Comparison with an "ideal" output:

- Requires a large testing set
- Especially suited for classification tasks
- Typical measures: precision, recall and F-measure
- Confusion metric can help to understand the results
- Statistical significance tests used to validate improvement
- Must include baselines, including a straw baseline (majority class, or random) and comparable methods

Intrinsic Evaluation

Subjective quality evaluation

- Advantages:
  - Doesn't require any testing data
  - Gives an easily understandable performance evaluation
- Disadvantages:
  - Requires several judges and a mechanism for dealing with disagreement
  - Tightly depends on the quality of instructions
  - Hard to isolate different components
  - Hard to reproduce

Task-based Evaluation

Advantages:

- Doesn't require any testing data
- Gives an easily understandable performance evaluation

Disadvantages:

- Hard to find a task with good discriminative power
- Requires multiple judges
- Hard to reproduce

Intrinsic Evaluation

Comparison with an "ideal" output:

- Advantages:
  - Results can be easily reproducible
  - Allows to isolate different factors contributing to system performance
- Disadvantages:
  - In the presence of multiple "ideal" outputs, penalizes alternative solutions
  - Distance between an "ideal" and a machine-generated output may not be proportional to the human perception of quality

Task-based Evaluation

Examples:

- Dialogue systems: Book a flight satisfying some requirements
- Summarization systems: Retrieve a story about X from a collection of summaries
- Summarization systems: Determine if a paper X should be part of related work for a paper on topic Y

Today

- Basic of annotations; agreement computation
- Estimating difference in the distribution of two sets
  - Significance of the method's improvement
  - Impact of a certain change on the system's performance
- Comparing rating schemes

Large Annotation Efforts

- Dialogue acts
- Coreference
- Discourse relations
- Summarization

The first three are available through LDC, the last one is available through DUC

Developing an Annotation Scheme

Main steps:

- Basic scheme
- Preliminary Annotation
- Informal evaluation
- Scheme revision and re-coding
- Coding manual
- Formal evaluation: inter-code reliability

Ready to code real data

Basic Scheme

Preliminary categories that seem to cover the range of phenomena of interest

- Different categories functionally important and/or easy to distinguish

Example: Dialogue Act Classification

Taxonomy principles:

- Activity-specific
  - Must cover activity features
  - Make crucial distinctions
  - Avoid irrelevant distinctions
- General
  - Aim to cover all activities
  - Specific activities work in a sub-space
  - Activity-specific clusters as “macros”

Informal Evaluation and Development

- Analysis of problematic annotations
  - Are some categories missing?
  - Are some categories indistinguishable for some coding decisions?
  - Do categories overlap?
- Meetings between annotators and scheme designers and users
- Revision of annotation guidelines
- More annotations

Result: Annotation manual

Example: Dialogue Act Classification

- Informativeness:
  - Difference in conditions/effects vs. confidence in label
  - Generalization vs. distinctions
    - Example: *state, assert, inform, confess, concede, affirm, claim*
- Granularity:
  - Complex, multi-functional acts vs. simple acts (the latter relies on multi-class classification)

Preliminary Annotation

- Algorithm
  - Automated annotation if possible
    - Semi-automated (partial, supervised decisions)
  - Decision trees for human annotators
- Definitions, guidelines
- Trial run with multiple annotators
  - Ideally following official guidelines or algorithm rather than informally taught

Reliability of Annotations

- The performance of an algorithm has to be evaluated against some kind of correct solution, the *key*
- For most linguistic tasks *correct* can be defined using human performance (not linguistic intuition)
- If different humans get different solutions for the same task, it is questionable which solution is correct and whether the task can be solved by humans at all
- Measures of reliability have to be used to test whether human performance is reliable
- If human performance is indeed reliable, the solution produced by human can be used as a key against which an algorithm can be evaluated

Agreement: Balanced Distribution

	A	B	C
1	2	0	0
2	2	0	0
3	2	0	0
4	0	2	0
5	0	2	0
6	0	2	0
7	0	0	2
8	0	0	2
9	0	0	2
10	1	1	0

$p(A) = 9/10 = 0.9$

Formal Evaluation

- Controlled coding procedures
  - Individuals coding unseen data
  - Coding on the basis of manual
  - No discussion between coders
- Evaluation of inter-code reliability
  - Confusion matrix
  - Statistical measure of agreement

Reliability of Annotations

- Kowtko et al. (1992) and Litman&Hirschberg use pairwise agreement between naive annotators
- Silverman et al. (1992) have two groups of annotators: a small group of experienced annotators and a large group of naive annotators. Assumption: the annotations are reliable, of there is only a small difference between groups.

However, what does reliability mean in these cases?

Agreement: Balanced Distribution

	A	B	C
1	2	0	0
2	2	0	0
3	2	0	0
4	0	2	0
5	0	2	0
6	0	2	0
7	0	0	2
8	0	0	2
9	0	0	2
10	1	1	0

$p(A) = 9/10 = 0.9$

$p(E) = (A/T)^2 + (B/T)^2 + (C/T)^2 = 0.335$

Kappa

- The kappa statistics can be used when multiple annotators have to assign markables to one of a set of non-ordered classes
- Kappa is defined as:

$$K = \frac{P(A) - P(E)}{1 - p(E)}$$

where P(A) is the actual agreement between annotators, and P(E) is the agreement by chance

Agreement: Skewed Distribution

	A	B	C
1	2	0	0
2	2	0	0
3	2	0	0
4	0	0	0
5	0	0	0
6	0	0	0
7	0	0	0
8	0	0	0
9	0	0	0
10	1	1	0

$p(A) = 9/10 = 0.9$

Agreement: Skewed Distribution

	A	B	C
1	2	0	0
2	2	0	0
3	2	0	0
4	0	0	0
5	0	0	0
6	0	0	0
7	0	0	0
8	0	0	0
9	0	0	0
10	1	1	0

$p(A) = 9/10 = 0.9$

$p(E) = (A/T)^2 + (B/T)^2 + (C/T)^2 = 0.815$

Today

- Basic of annotations; agreement computation
- Estimating difference in the distribution of two sets
  - Significance of the method's improvement
  - Impact of a certain change on the system's performance
- Comparing rating schemes

Kappa Interpretation

- Complete agreement: $K = 1$; random agreement: $K = 0$ random agreement
- In our example: K for balance set is 0.85, and for skewed one is 0.46
- Typically, $K > 0.8$ indicates good reliability

Many statisticians do not like Kappa! (alternative: interclass agreement)

Paired Data

- Goal: determine the impact of a certain fact on a given distribution
- Test is performed on the same sample
- Example scenario: we want to test whether adding parsing information improves performance of a summarization system on a predefined set of texts
- Null hypothesis: the actual mean difference is consistent with zero

Student's t-Test

- Goal: determine whether two distributions are different
- Samples are selected independently
- Example scenario: we want to test whether adding parsing information improves performance of a summarization system
- Null hypothesis: the difference is due to chance
  For $N = 10$, $X_{avg} \pm 2.26 * \sigma / N^{\frac{1}{2}}$ (with 95% confidence)
- "Statistical significance": the probability that the difference is due to chance

Chi-squared test

- Goal: compare expected counts
- Example scenario: we want to test whether number of backchannels in a dialogue predicted by our algorithm is consistent with their distributions in real text
- Assume "normal" distribution with mean μ and standard deviation σ:

$$\chi^2 = \sum_{i=1}^{k} \frac{(x_i - \mu)^2}{\sigma^2}$$

Today

- Basic of annotations; agreement computation
- Estimating difference in the distribution of two sets
- Comparing rating schemes

Anova

- Goal: determine the impact of a certain fact on several distributions (assumes cause-effect relation)
- Samples are selected independently
- Null hypothesis: the difference is due to chance
- Computation:

$$F = \frac{found variation of the group averages}{expected variation of the group averages}$$

- Interpretation: if $F = 1$ null hypothesis is correct, while large values of F confirm the impact

Chi-squared test

- In some cases, we don't know the standard deviation for each count:

$$X^2 = \sum_{i=1}^{k} \frac{(x_i - E_i)^2}{E_i}$$

$$E_i = p_i N$$

- Assume the Poisson distribution (the standard deviation equals the square of the expected counts)
- Restrictions: not applicable for small E_i

Kendall's τ

- Goal: estimate the agreement between two orderings
- Computation:

$$\tau = 1 - 2\frac{I}{N(N-1)/2},$$

where N is a sample size and I is the minimal number of interchanges required to map the first order into the second