Language Understanding Systems

Evaluation in NLP

Evgeny A. Stepanov

SISL, DISI, UniTN & VUI, Inc.
evgeny.stepanov@unitn.it

Outline

1. Basic Concepts

2. Evaluation Metrics

3. Exercises

Section 1

Basic Concepts

Evaluation of the NLP System

Why do we want to evaluate a system / an algorithm's performance?

- *To measure one or more of its qualities.*
- *Proper evaluation criteria is a way to specify the problem.*

How do we evaluate a system / an algorithm's performance?

Automatic vs. Manual Evaluation

Automatic Evaluation

Compare the system's output with the gold standard (**reference**)

- Cons: An effort to produce the gold standard (manual)
- Pros: Re-usable; no additional cost

OBJECTIVE

Manual Evaluation

Ask **human judges** to estimate the quality w.r.t. certain criteria

- For some tasks the gold standard might be unobtainable
- No agreed automatic evaluation method

SUBJECTIVE

Intrinsic vs. Extrinsic Evaluation

Intrinsic

- in isolation
- w.r.t. gold standard (references)
- e.g. POS-Tagging performance

Extrinsic

- as a part of other system
- usefulness for some other task
- e.g. effect of POS-Tagger on parsing performance

Black-Box vs. Glass-Box

Black-Box

Evaluation of Performance

- speed
- accuracy
- etc.

Glass-Box

Evaluation of Design

- algorithm
- used resources
- etc.

Gold Standard / References

- **Where Gold Standard comes from?**
- *Annotation by experts (human judges)*
- **How do we know that Gold Standard is good?**
- *Evaluate agreement between the annotators/judges*
- Most simple agreement measure: % of agreed instances

Lower & Upper Bounds of the Performance

Lower Bound

Baseline – trivial solution to the problem:

- *random*: random decision
- *chance*: random decision w.r.t. the distribution of categories in the training data
- *majority*: assign everything to the largest category
- etc.

Upper Bound

Inter-rater agreement – human performance.

A system is expected to perform within the lower and upper bounds.

Data Split

Training	for training / extracting rules / etc.
Development	for optimization / intermediate evaluation
Testing	for the final evaluation

Section 2

Evaluation Metrics

The Simplest Case

$$Accuracy = \frac{\text{Num. of Correct Decisions}}{\text{Total Num. of Instances}} \tag{1}$$

- Known number of instances
- Single decision for each instance
- Single correct answer for each instance
- All errors are equal

Contingency Table

		REF	
		POS	*NEG*
HYP	*POS*	**TP**	**FP**
	NEG	**FN**	**TN**

TP	*True Positive*	a
FP	*False Positive*	b
FN	*False Negative*	c
TN	*True Negative*	d

Accuracy

		REF	
		POS	*NEG*
HYP	*POS*	**TP**	**FP**
	NEG	**FN**	**TN**

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN} \tag{2}$$

- **What if TN is infinite or unknown?**
- e.g.: Number of irrelevant queries to a search engine

Precision & Recall

		REF		
		POS	*NEG*	
HYP	*POS*	**TP**	**FP**	*Precision*
	NEG	**FN**	**TN**	
		Recall		

$$Precison = \frac{TP}{TP + FP} \tag{3}$$

$$Recall = \frac{TP}{TP + FN} \tag{4}$$

- **2 Values: Precision-Recall Trade-Off**

F-Measure

- Harmonic Mean of Precision & Recall
- Usually evenly weighted

$$F_\beta = \frac{(1 + \beta^2) * Precision * Recall}{\beta^2 * Precision + Recall} \tag{5}$$

$$F_1 = \frac{2 * Precision * Recall}{Precision + Recall} \tag{6}$$

Edit Distance

- Hypotheses and Reference might differ not only on instance labels, but also on number of instances
- Number of concepts
- WER: Word Error Rate
- CER: Concept Error Rate

$$*ER = \frac{I + D + S}{N} \tag{7}$$

More Advanced Topics

- Cross-Validation
- Significance Tests
- Agreement Measures
- Sampling (random, stratified)
- Binary vs. Multi-class classification
- Multi-label data
- Regression
- Re-ranking
- Ensemble Methods
- etc.

Section 3

Exercises

Exercises

Given the sample data, where Column 1 – References and Column 2 – Hypotheses:

1 Compute raw TP, FP, FN, TN.

2 Compute Accuracy, Precision, Recall, F-Measure

Write scripts...

Synthetic Data

- Generate a Data Set where:
  - 5 classes
  - the distribution is 20%, 20%, 30%, 25%, 5%
- Sampling:
  - Split into training and test sets as 90% & 10%
  - Random vs. Stratified Sampling