

# Language Understanding Systems

*Evaluation in NLP*

Evgeny A. Stepanov

SISL, DISI, UniTN  
`evgeny.stepanov@unitn.it`

# Outline

1 Basic Concepts

2 Evaluation Metrics

# Section 1

## Basic Concepts

# Evaluation of the NLP System

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

# 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.*

# 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.*

# 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**)



# 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)

# 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

# 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

## *Manual Evaluation*

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

# 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

## *Manual Evaluation*

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

- For some tasks the gold standard might be unobtainable

# 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

## *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

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

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

# 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.

# 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?

# Gold Standard / References

- Where Gold Standard comes from?
- *Annotation by experts (human judges)*

# Gold Standard / References

- Where Gold Standard comes from?
- *Annotation by experts (human judges)*
- How do we know that Gold Standard is good?

# 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*

# 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:

- *chance*: random decision
- *majority*: assign everything to the largest category
- etc.



# Lower & Upper Bounds of the Performance

## *Lower Bound*

**Baseline** – trivial solution to the problem:

- *chance*: random decision
- *majority*: assign everything to the largest category
- etc.

## *Upper Bound*

**Inter-rater agreement** – human performance.

# Lower & Upper Bounds of the Performance

## *Lower Bound*

**Baseline** – trivial solution to the problem:

- *chance*: random decision
- *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

<i>Training</i>	for training / extracting rules / etc.
<i>Development</i>	for optimization / intermediate evaluation
<i>Testing</i>	for final evaluation

## Section 2

# Evaluation Metrics

# The Simplest Case

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

# The Simplest Case

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

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

# Contingency Table

		REF	
		<i>POS</i>	<i>NEG</i>
HYP	<i>POS</i>	<b>TP</b>	<b>FP</b>
	<i>NEG</i>	<b>FN</b>	<b>TN</b>

# Contingency Table

		REF	
		<i>POS</i>	<i>NEG</i>
HYP	<i>POS</i>	<b>TP</b>	<b>FP</b>
	<i>NEG</i>	<b>FN</b>	<b>TN</b>

<b>TP</b>	<i>True Positive</i>	a
<b>FP</b>	<i>False Positive</i>	b
<b>FN</b>	<i>False Negative</i>	c
<b>TN</b>	<i>True Negative</i>	d



# Accuracy

		REF	
		<i>POS</i>	<i>NEG</i>
HYP	<i>POS</i>	<b>TP</b>	<b>FP</b>
	<i>NEG</i>	<b>FN</b>	<b>TN</b>

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN} \quad (2)$$

# Accuracy

		REF	
		POS	NEG
HYP	POS	TP	FP
	NEG	FN	TN

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN} \quad (2)$$

- What if TN is infinite or unknown?

# Accuracy

		REF	
		POS	NEG
HYP	POS	TP	FP
	NEG	FN	TN

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN} \quad (2)$$

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

# Precision & Recall

		REF		
		<i>POS</i>	<i>NEG</i>	
HYP	<i>POS</i>	<b>TP</b>	<b>FP</b>	<i>Precision</i>
	<i>NEG</i>	<b>FN</b>	<b>TN</b>	
		<i>Recall</i>		

$$Precision = \frac{TP}{TP + FP} \quad (3)$$

# Precision & Recall

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

$$Precision = \frac{TP}{TP + FP} \quad (3)$$

$$Recall = \frac{TP}{TP + FN} \quad (4)$$

# Precision & Recall

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

$$Precision = \frac{TP}{TP + FP} \quad (3)$$

$$Recall = \frac{TP}{TP + FN} \quad (4)$$

- 2 Values: Precision-Recall Trade-Off

# F-Measure

- Harmonic Mean of Precision & Recall

# F-Measure

- Harmonic Mean of Precision & Recall
- Usually evenly weighted

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

$$F_1 = \frac{2 * Precision * Recall}{Precision + Recall} \quad (6)$$



# Edit Distance

- Hypotheses and Reference might differ not only on instance labels, but also on number of instance
- Number of concepts

# Edit Distance

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

# Edit Distance

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

$$*ER = \frac{I + D + S}{N} \quad (7)$$

# 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.