Computational Linguistics: Evaluation Methods

Raffaella Bernardi

University of Trento

1. Admin

Perusall sends email reminders to students 3, 2, and 1 day before the deadline of an assignment. Only students that have not started an assignment will be sent a reminder.

Reminders are enabled by default, but you can disable reminders for your course by unchecking Enable assignment reminders under Settings > Advanced.

Students can also individually opt out of receiving such reminders by clicking Notifications > Manage notifications, and then unchecking Notify me when an assignment that I haven't yet completed is about to be due.

http://disi.unitn.it/~bernardi/Courses/CL/20-21.html

2. Standard practice used in NLP experiments

A typical NLP experiment is based on:

- an annotated dataset (e.g., a collection of image caption pairs (data points).)
- a task defined over the dataset (generation of IC, retrieval of IC)
- a comparison of models' performance on the task

2.1. Evaluation methods

- intrinsic evaluations: model predictions are compared to manually produced “gold-standard” output (e.g. word analogies) ;
- extrinsic evaluations: models are evaluated on a downstream task;
- benchmarks: competitions are organized to compare models, (the “leader-board” approach);
- adversial evaluation: inputs are transformed by perturbations;
- probing/auxiliary (or decoding) tasks: the encoded representations of one system to train another classifier on some other (probing) task of interest. The probing task is designed in such a way to isolate some linguistic phenomena and if the probing classifier performs well on the probing task we infer that the system has encoded the linguistic phenomena in question.

2.2. Dataset, Annotation, Task

- The annotated dataset is collected automatically (e.g. from the web) or
- some part of the datapoints (e.g. the images) are collected automatically and then humans are asked to annotate them or to perforn the task it self.
- Human annotation is obtained via crowdsourcing (uncontrolled dataset) (to simulate a more "naturalistic" collection of data) or
- Synthetic data are produced (eg., Filler in the gap paper) (controlled/diagnostic dataset).
- The dataset is then randomly split into training (e.g. 60%), validation (e.g. 20%) and testing (eg. %20) sets or
- for small datasets several random splits are performed (cross-validation)
- making sure that the test set contains unseen data (the training/validation/test sets do not overlap).

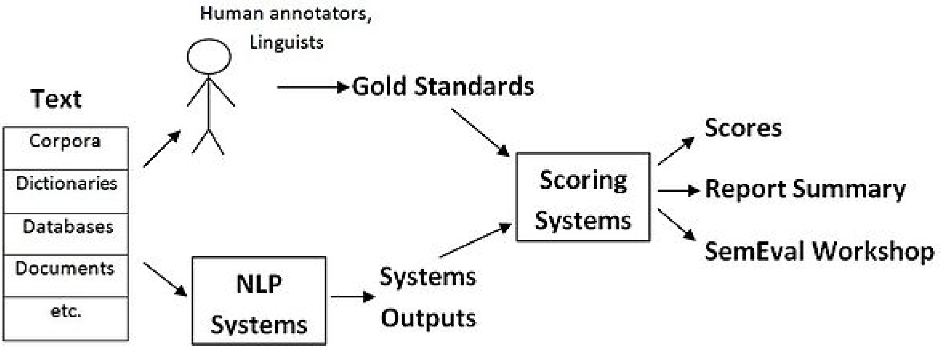
2.3. Examples of tasks/benchmarks

- NL understanding: GLUE https://gluebenchmark.com/, Winograd schema https://cs.nyu.edu/faculty/davise/papers/WinogradSchemas/WS.html
- QA: SQuAT https://rajpurkar.github.io/SQuAD-explorer/
- NL entailment: RTE, SNLI, SICK
- NL Dialogue: BaBi,
- Language and Vision: MS-COCO, FOIL, Visual Genome, VisDial, Guess-What?!

List of NLP Datasets https://github.com/niderhoff/nlp-datasets

2.4. Evaluation campaign

Eg., SemEval: An ongoing series of evaluations of computational semantic analysis



3. Behind the scene

The whole enterprise is based on the idea that:

- "if we take a **random sample** of the "population" (data) the results we obtain can be generalized to the whole "population"."
- **Independent observation assumption**: "observations (data points) in your sample are independent from each other, meaning that the measurements for each sample subject are in no way influenced by or related to the measurements of other subjects." Dependence in the data can turn into **biased** results.
- "the **null hypothesis** (H_0) states that there is no relationship between the measured quantities in the population, while its "rival", the "alternative hypothesis" assumes that there is a relationship."
- "**Statistical tests** tells us whether the differences obtained are statististically significant – they calculate the probability of observing a relationship in a sample even though the relationship does not exist in the population of interest."

3.1. Current debate on evaluation

- Sampling: no attention about sampling. WEIRD (Western, Educated, Industrialized, Rich and Developed) population.;
- Sampling: the indipendent observation assumption is often violated (e.g., text from the same author);
- Test set same distribtuion of the training set
- It would be good to evaluate systems using a stratified/controlled test set;
- More attention should be given to the baseline and the models compared.
- When dealing with NN, the avarage of the results obtained using different seeds should be reported
- Evaluation metrics: more attention should be given to the metric used in the evaluation and (the right) statistical test should be reported;
- Qualitative evaluation and error analysis should complement the automatic metric evaluation.

3.2. Further wishes

- Fair comparison: e.g. same pre-training corpus (see Baroni et al 2014)
- Test-only benchmarks
- evaluation against controlled data sets, with breakdown evaluation.
- replicability
- Open science: all code, material should be well documented and made available to the community.

3.3. Interesting readings

Dror et al ACL 2018: **The Hitchhikers Guide to Testing Statistical Significance in Natural Language Processing**

Alexander Koplenig **Against Statistical significance testing in corpus linguistics** Follow up on Stefan Th. Gries, who follow up on Kilgarriff

van der Lee, C; Gatt, A; van Miltenburg, E and Krahmer, E, **Human evaluation of automatically generated text: Current trends and best practice guidelines** Computer Speech and Language, in press.

Tal Linzen How **Can We Accelerate Progress Towards Human-like Linguistic Generalization?**. Next Reading Group.

4. Dataset annotation: Kappa agreement

- Kappa is a measure of how much judges agree or disagree.
- Designed for categorical judgments
- Corrects for chance agreement
- $P(A)$ = proportion of time judges agree
- $P(E)$ = what agreement would we get by chance

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

Values of κ in the interval

- $[0.8 - 1]$ (good agreement),
- $[0.67 - 0.8]$ (fair agreement),
- $[\cdot - 0.67]$ (dubious basis for an evaluation).

4.1. Calculating the kappa statistic

		Judge 2 Relevance		
		Yes	No	Total
Judge 1	Yes	300	20	320
Relevance	No	10	70	80
	Total	310	90	400

Observed proportion of the times the judges agreed
$P(A) = (300 + 70)/400 = 370/400 = 0.925$

Pooled marginals $P(nonrelevant) = (80 + 90)/(400 + 400) = 170/800 = 0.2125$

$P(relevant) = (320 + 310)/(400 + 400) = 630/800 = 0.7878$

Probability that the two judges agreed by chance
$P(E) = P(nonrelevant)^2 + P(relevant)^2 = 0.2125^2 + 0.7878^2 = 0.665$

Kappa statistic
$\kappa = (P(A) - P(E))/(1 - P(E)) = (0.925 - 0.665)/(1 - 0.665) = 0.776$ (still in acceptable range)

5. Quantitative Evaluation Metrics

- From Information Retrieval: Accuracy, Precision, Recall, F-Measure
- From other disciplines (e.g Psychology and Neuroscience): Pearson Correlation, Spearman Correlation Perplexity, Purity, Representational Similarity Analysis
- Specific of NLP: BLEU and METEOR (machine translation and natural language generation), ROUGE (summarization), USA and LAS (dependency parsing)

6. Evaluation Metrics from IR

Accuracy Percentage of documents correctly classified by the system.

Error Rate Inverse of accuracy. Percentage of documents wrongly classified by the system

Precision percentage of relevant documents correctly retrieved by the system (TP) with respect to all documents **retrieved by the system** (TP + FP). (how many of the retrieved books are relevant?)

Recall percentage of relevant documents correctly retrieved by the system (TP) with respect to all documents **relevant for the human** (TP + FN). (how many of the relevant books have been retrieved?)

	Relevant	Not Relevant
Retrieved	True Positive (TP)	False Positive (FP)
Not retrieved	False Negative (FN)	True Negative (TN)

6.1. Definitions

	Relevant	Not Relevant
Retrieved	True Positive (TP)	False Positive (FP)
Not retrieved	False Negative (FN)	True Negative (TN)

Accuracy	$\frac{\mathbf{TP\ +\ TN}}{TP+TN+FP+FN}$
Error Rate	$\frac{\mathbf{FP+FN}}{TP+TN+FP+FN}$
Precision	$\frac{TP}{TP+\mathbf{FP}}$
Recall	$\frac{TP}{TP+\mathbf{FN}}$

6.2. Exercise

a) In a collaction of 100 documents, 40 documents are relavant for a given search. Two IR systems (System I on the left and System II on the right) behave as following w.r.t. the given search and collection. Calculate the above measures.

	Relevant	Not Relevant
Retrieved	30	0
Not retrieved	10	60

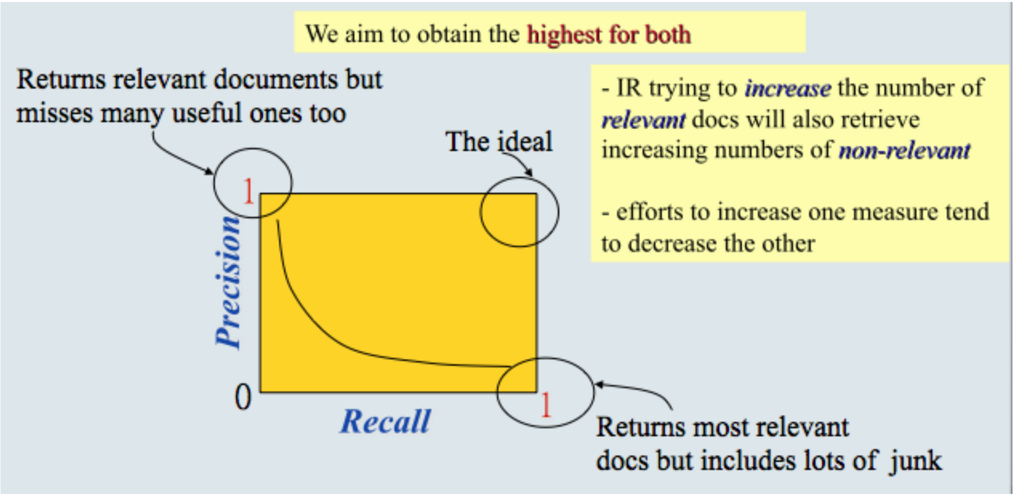
	Relevant	Not Relevant
Retrieved	40	50
Not retrieved	0	10

Which system is better?

Solutions

	Acc	ER	P	R
System I	0.90	0.1	1	0.44
System II	0.90	0.5	0.75	1

6.3. Trade off



We aim to obtain the highest for both

Returns relevant documents but misses many useful ones too

- IR trying to ***increase*** the number of ***relevant*** docs will also retrieve increasing numbers of ***non-relevant***

- efforts to increase one measure tend to decrease the other

Returns most relevant docs but includes lots of junk

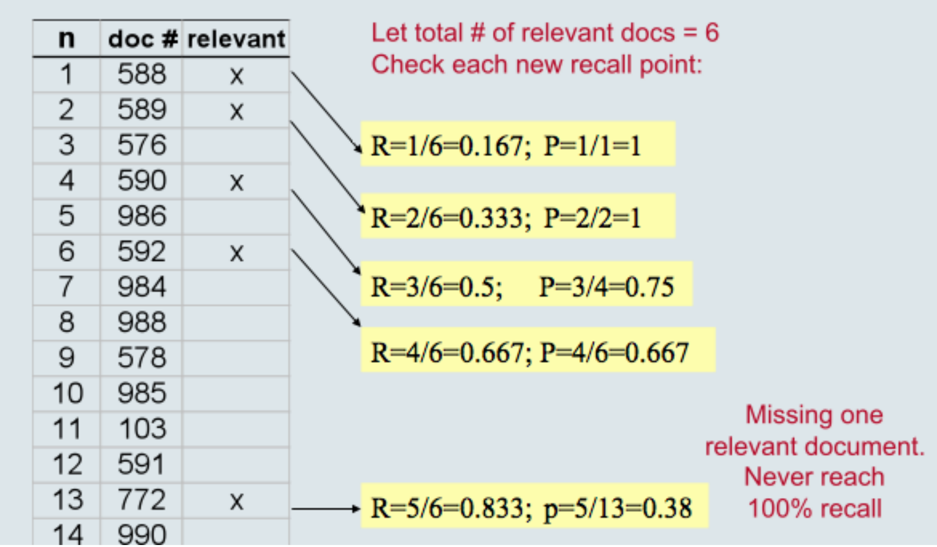
6.4. F-Measure

Combine in a single measure Precision (P) and Recall (R) giving a **global estimation of the performance** of an IR system

$$\mathrm{F} \quad \frac{2PR}{R+P}$$

	Acc	ER	P	R	F
System I	0.90	0.1	1	0.44	**0.85**
System II	0.90	0.5	0.75	1	0.6

6.5. Precision/Recall: at position



n	doc #	relevant
1	588	x
2	589	x
3	576	
4	590	x
5	986	
6	592	x
7	984	
8	988	
9	578	
10	985	
11	103	
12	591	
13	772	x
14	990	

Let total # of relevant docs = 6
Check each new recall point:

R=1/6=0.167; P=1/1=1

R=2/6=0.333; P=2/2=1

R=3/6=0.5; P=3/4=0.75

R=4/6=0.667; P=4/6=0.667

R=5/6=0.833; p=5/13=0.38

Missing one relevant document. Never reach 100% recall

7. Metrics from other disciplines

Correlation statistical relation between two variables (eg. dependent phenomena include the correlation between the height of parents and their offspring.);

Purity is a measure of the extent to which clusters contain a single class;

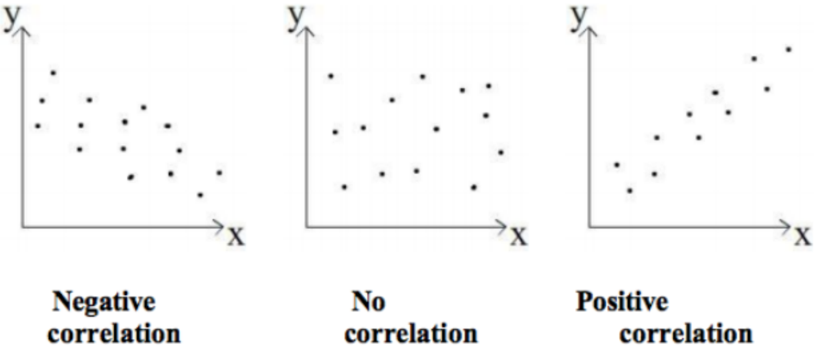
Perplexity is a measurement of how well a probability distribution predicts a sample. A low perplexity indicates the probability distribution is good at predicting the sample;

RSA pairwise comparisons of stimuli to reveal their representation in higher-order space.

7.1. Correlation coefficients

If we have studied the behaviour of some data w.r.t. more task/phenomena (variables), we can take a pair of such variables and see if they are correlated (=tend to change together). In particular, we can check if one variable increases what happens to the other variable:

- the other variable has a tendency to decrease, then there is a negative correlation.
- the other variable does not tend to either increase or decrease, then there is no correlation.
- the other variable has a tendency to also increase, then there is a positive correlation



Taken from:
http://www.statstutor.ac.uk/resources/uploaded/pearsons.pdf and
http://www.statstutor.ac.uk/resources/uploaded/spearmans.pdf

To decide whether two variables are correlated, we can compute standard correlation coefficients.

7.2. Standard correlation Coefficients

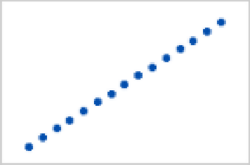
A coefficient describes both the direction and the strength of the relationship.

Pearson's correlation coefficient The Pearson correlation, r, evaluates the **linear** relationship between two continuous variables. A relationship is linear when a change in one variable is associated with a **proportional** change in the other variable.

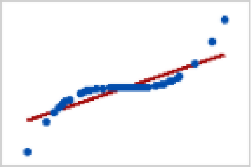
Spearman Correlation coefficient The Spearman correlation, ρ, evaluates the **monotonic** relationship between two continuous or ordinal variables. In a monotonic relationship, the variables tend to change together, but not necessarily at a constant rate. The Spearman correlation coefficient is based on the ranked values for each variable rather than the raw data.

7.3. Comparison

The Pearson and Spearman correlation coefficients can range in value from -1 to $+1$, they are represented in a scatterplot.



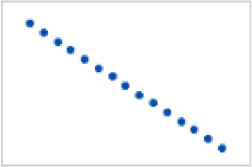
Pearson = +1, Spearman = +1



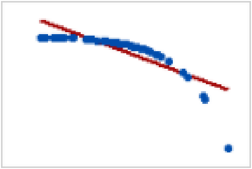
Pearson = +0.851, Spearman = +1

$r = +1, \rho = +1$ For the Pearson correlation coefficient to be +1, when one variable increases then the other variable increases by a consistent amount. This relationship forms a perfect line. The Spearman correlation coefficient is also +1

$r = +0.85, \rho = +1$ If the relationship is that one variable increases when the other increases, but the amount is not consistent, the Pearson correlation coefficient is positive but less than +1. The Spearman coefficient still equals +1



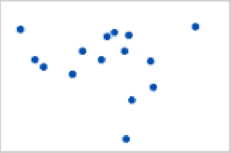
Pearson = −1, Spearman = −1



Pearson = −0.799, Spearman = −1

$r = -1, \rho = -1$ If the relationship is a perfect line for a decreasing relationship, then both correlation coefficients are -1.

$r = 0.799, \rho = -1$ If the relationship is that one variable decreases when the other increases, but the amount is not consistent, then the Pearson correlation coefficient is negative but greater than -1. The Spearman coefficient still equals -1 in this case.



Pearson = −0.093, Spearman = −0.093

$r = -0.093, \rho = -0.093$ When a relationship is random or non-existent, then both correlation coefficients are nearly zero.

Taken from https://support.minitab.com/en-us/minitab-express/

7.4. Purity of clusters

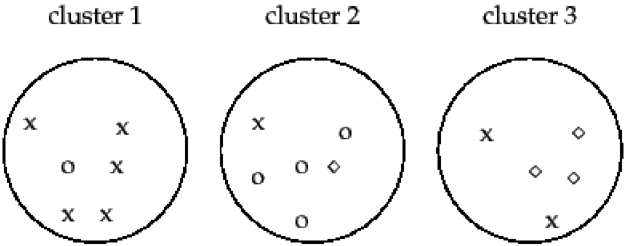
Purity is a measure of the extent to which clusters contain a single class.

Formally, given some set of clusters M and some set of classes D, both partitioning N data points, purity can be defined as:

$$\frac{1}{N}\sum_{m\in M} max_{d\in D}|m \cap d|$$

Its calculation can be thought of as follows: For each cluster, count the number of data points from the most common class in said cluster. Now take the sum over all clusters and divide by the total number of data points.

7.5. Example



► **Figure 16.1** Purity as an external evaluation criterion for cluster quality. Majority class and number of members of the majority class for the three clusters are: x, 5 (cluster 1); o, 4 (cluster 2); and ⋄, 3 (cluster 3). Purity is $(1/17) \times (5+4+3) \approx 0.71$.

Note that this measure doesn't penalise having many clusters. So for example, a purity score of 1 is possible by putting each data point in its own cluster."

8. Main areas of CL

ACL 2020, EMNLP 2020

- Cognitive Modeling and Psycholinguistics
- Computational Social Science and Social Media
- Dialogue and Interactive Systems
- Discourse and Pragmatics
- Ethics and NLP
- Generation
- Information Extraction
- Information Retrieval and Text Mining
- Interpretability and Analysis of Models for NLP
- **Language Grounding to Vision, Robotics and Beyond**

- Theory and Formalism in NLP (Linguistic and Mathematical)
- Machine Learning for NLP
- Machine Translation
- NLP Applications
- Phonology, Morphology and Word Segmentation
- Question Answering
- Resources and Evaluation
- Semantics: Lexical
- Semantics: Sentence Level
- Semantics: Textual Inference and Other Areas of Semantics
- Sentiment Analysis, Stylistic Analysis, and Argument Mining
- Speech and Multimodality
- Summarization

- Syntax: Tagging, Chunking and Parsing

EACL 2021 adds:

- Green and Sustainable NLP
- NLP and Crisis Management

9. Further readings

Patrick Paroubek, Stphane Chaudiron, Lynette Hirschman. Principles of Evaluation in Natural Language Processing. Traitement Automatique des Langues, ATALA, 2007, 48 (1), pp.7-31.

Karen Sparck Jones and Julia R. Galliers, Evaluating Natural Language Processing Systems: An Analysis and Review
https://www.aclweb.org/anthology/J98-2013.pdf

Dodge et al: Show Your Work: Improved Reporting of Experimental Results EMNLP 2019 https://www.aclweb.org/anthology/D19-1224.pdf

Adversarial Evaluation for Models of Natural Language Noah A. Smith
https://arxiv.org/abs/1207.0245

Elephant in the Room: An Evaluation Framework for Assessing Adversarial Examples in NLP Ying Xu, Xu Zhong, Antonio Jose Jimeno Yepes, Jey Han Lau
https://arxiv.org/abs/2001.07820

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, Sameer Singh Beyond Accuracy: Behavioral Testing of NLP models with CheckList

https://arxiv.org/abs/2005.04118

Jurafsky and Martin, Speech and Language Processing Sec. 4.7 and 4.8
https://web.stanford.edu/~jurafsky/slp3/ed3book.pdf