NLP evaluation

C. Grivaz, J.-C. Chappelier

Laboratoire d'Intelligence Artificielle
Faculté I&C

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

NLP evaluation motivations

- Evaluate the improvement of the technology on a specific task
- Provide gold standard and objective comparison methods
- Develop research and technology in NLP

C. Grivaz, J.-C. Chappelier
ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

NLP evaluation protocols

1. Define a control task
2. Get/regroup large amount of *typical* data (for the task)
3. Assess the quality of the data
4. Test and compare NLP systems on similar data
5. Publish and discuss results

An example: Discourse relation classification

Task: find Linked propositions:

- *Jane has short hair, but Charles has long hair*
  ☞ contrast
- *Marc fell, John pushed him*
  ☞ explanation
- *Everybody got angry and began throwing rotten tomatoes. In short, it was a complete disaster.*
  ☞ result
  ☞ summary
- ...

C. Grivaz, J.-C. Chappelier
ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

Automatic discourse relation classification: motivations

- Question answering, automatic summary
- Deeper understanding

- But difficult: unmarked relations make up a lot of the total discourse relations.

Discourse relations as a classification problem

We are coming back from shopping. I bought aubergines.
John is happy, he has bought a toaster

Items pairs of clauses.

A class per discourse relation explanation, result, contrast, summary, narration, ...

Methodology train a classifier on an annotated corpus.

C. Grivaz, J.-C. Chappelier
ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

Not an easy task

- Some relations are *marked*:

  Jane has short hair, but Charles has long hair

  Everybody got angry and began throwing rotten tomatoes. In short, it was a complete disaster

- But the *marker* is often *ambiguous* :

  Everybody got angry and began throwing rotten tomatoes

- And most of them are *not* marked:

  Marc fell, John pushed him

Need for a set of correct answers

Contrary to some other tasks, there is generally no simple way to know if a NLP system gives correct results

especially when the goal of an NLP task is to mimic something that a human can do

☞ gold standard : set of correct answers to a task, for a *sample* of correct inputs

Evaluation methodology:

the sample of input is then given to the automatic system and its output is compared to the gold standard

Manually annotated corpora

In the case of NLP, the gold standard often takes the form of an annotated corpus.

Example (The Penn Discourse Treebank)

Intelogic holds 27.5% of Datapoint's common shares outstanding.

```
( (S
    (NP-SBJ (NNP Intelogic) )
    (VP (VBZ holds)
      (NP
        (NP (CD 27.5) (NN %) )
        (PP (IN of)
          (NP
            (NP
              (NP (NNP Datapoint) (POS 's) )
              (JJ common) (NNS shares) )
            (ADJP (JJ outstanding) )))))
    (. .) ))
```

The Penn Discourse Treebank

- Discourse annotations over a part of the Penn Treebank
- Claims to be theory neutral

C. Grivaz, J.-C. Chappelier
ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

The Penn Discourse Treebank: example

Although preliminary findings were reported more than a year ago, the latest results appear in today's New England Journal of Medicine [...]

```
  ____Explicit____
534..542 [...] Although
[...]
although, Comparison.Contrast
____Arg1____
600..722 [...] the latest results appear in today's
New England Journal of Medicine
____Arg2____
543..598 [...] preliminary findings were reported
more than a year ago
```

Gold standard impact

- Gold standard creation is extremely expensive
- However, Evaluation size is cheaper than training size
- Amortization (but biais): if a gold standard exists, the whole field is likely to use it for comparison and evaluation

C. Grivaz, J.-C. Chappelier

Gold standard creation process

- Properly define the task in an annotator manual
- Select the corpus to annotate
- Train annotators:
  - annotation instructions
  - assess annotation quality: inter-annotator agreement (or other appropriate measures)
- Annotate

C. Grivaz, J.-C. Chappelier
ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

Humans do not always agree on NLP tasks

- Despite the annotator manual, divergences always exist
- These divergences highly depend on the subjectivity of the task
- A resource is considered good only if the divergences are low

☞ measure Inter-annotator agreement

Disagreement example: word sense disambiguation

Task: Word Sense Disambiguation (WSD):

label each word of a text (= within context) to its corresponding sense (typically from an ontology)

Example (easy):

> *I can hear bass sounds.*
> *They like grilled bass.* `[fish, named "bar" in French]`

Example (not so easy):
disambiguate usage of `national` with an ontology where:

> *1) limited to or in the interest of a particular nation*
> *2) concerned with or applicable to or belonging to an entire nation or country*

`[from WordNet 3.1]`

C. Grivaz, J.-C. Chappelier

Even relatively objective task lead to disagreement: syntax example

Put the block in the box on the table.

What is the attachment site of *on the table* ?

C. Grivaz, J.-C. Chappelier

Measuring inter annotator agreement

- "Inter annotator agreement" (IAA) is considered a measure of the quality of gold standards
- It is also a measure of the subjectivity of a task
- It must be objectively measured and reported

Raw agreement

Simplest measure of agreement:

$$\text{raw agreement} = \frac{\text{nb items agreed}}{\text{total nb of items}}$$

Raw agreement drawback

Raw agreement doesn't take *by-chance agreement* into account

Example

Two annotators annotate items having only one class in 80% of the time, systematically disagreeing about ambiguous items (2 classes)

	yes	no
yes	0	7
no	13	80

$$\text{raw agreement} = \frac{80}{100}$$

They get a 80% raw agreement despite their complete disagreement

Dealing with chance agreement

Taking chance agreement into account:

- Idea: substract chance agreement

$$\frac{\text{observed\_agreement} - \text{chance\_agreement}}{1 - \text{chance\_agreement}}$$

- Several measures exist
- Measures differ in the way they represent chance agreement

Cohen's kappa

Cohen's κ ("kappa") is the most famous inter annotator agreement coefficient

for 2 graders only (generalization: Fleiss' kappa)

It takes each annotator into account (independently)

Example

	yes	no
yes	0	10
no	20	70

- chance of saying yes: A: 0.2, B: 0.1
- chance of saying no: A: 0.8 B: 0.9
- Both yes if independant: $0.2 * 0.1 = 0.02$
- Both no if independant: $0.8 * 0.9 = 0.72$
- chance of independant agreement = 0.72+0.02=0.74

$$\kappa = \frac{\text{observed\_agreement} - \text{chance\_agreement}}{1 - \text{chance\_agreement}} = \frac{0.7 - 0.74}{1 - 0.74}$$
$$= -0.15$$

Interpretation of Cohen's kappa

- Positive: better than chance
- Negative: worse than chance (correlated disagreement)
- 1: perfect agreement
- 0 statistical independence
- more than 0.6 is usually considered ok, and more than 0.8 considered good

C. Grivaz, J.-C. Chappelier
ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

Practices

- IAA measures are almost always reported,
  but often only the raw agreement is given
- IAA is often only measured on a sample,
  sometimes on the whole corpus
- Each rest of the corpus is often annotated by only one person
- Only one annotation set is given at the end.
  When several annotations exist, they are merged

Annotation framework examples

OntoNotes Several cycles of redefinition to increase IAA

OntoNotes release Annotations done in parallel and independently by two annotators and then adjudicated by one

TimeBank IAA on a sample, release: only one annotator per item

Penn Treebank IAA on a sample, release: automatic annotation corrected twice by different annotators

Importance of separating the data

Comparing the programme output to a gold standard

Methodological issue: clearly separate the data:

- Separate training (and validation) from testing
  Do it fully honestly blindly randomly!! ;-)
- Validation set: allows to estimate overfitting or meta-parameters.
  Not to be confused with test set![1]

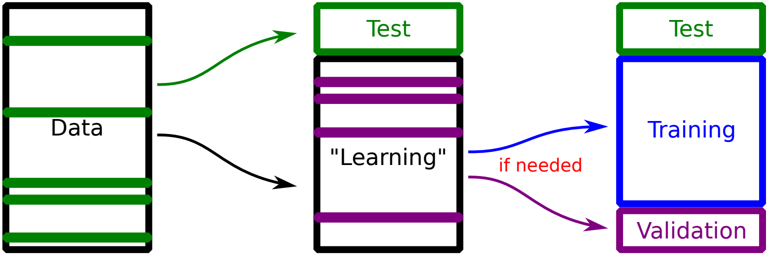
  ☞ clearly separated from test set (validation set is indeed a kind of training set):

  - Train on the training set
  - Test and adjust meta parameters on validation set
  - Reduce overfitting using the validation set
  - Final testing on the testing set (don't even look at it before!)
- Repeat *all* this several times (to estimate variance)

[1] The more so as so-called "*cross-validation*" is an evaluation method, done on the test set, which has *nothing* to do with the validation set!!

C. Grivaz, J.-C. Chappelier
ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

Training, validation and test sets



C. Grivaz, J.-C. Chappelier

The confusion matrix

The confusion matrix is not a measure itself,
but it gives complete information about the success and errors.

All the evaluation measures are summaries of the confusion matrix in one way or another.

The confusion matrix represents, for each reference class, how the system classifies its corresponding items.

Example

		reference		
		A	B	C
system	A	35	2	10
	B	3	46	1
	C	5	6	12

Evaluation measures

- Standard/Usual (not specific to NLP):
  - Accuracy
  - Precision, Recall (and F-score)
  - ROC curve
- Dedicated ones

C. Grivaz, J.-C. Chappelier
ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

Accuracy

$$\text{accuracy} = \frac{\text{number of correctly classified items}}{\text{total number of items}}$$
$$= \text{(normalized) trace of the confusion matrix}$$

- Can be used with any number of classes
- Used for classification tasks where all class have the same importance
- Accuracy does not take the difference between two classes into account:
  - asymmetry can result from classes of different importance (e.g. diagnostic)
  - or a class containing much more items than another

A task with asymmetrical classes: information retrieval

IR seen as a binary classification task

- a document is *relevant* or *irrelevant* to a query

Example of assymetry:

- Take a query to which 20 out of 100'000 documents are relevant
- The perfect classifier has the following accuracy

$$\frac{100'000}{100'000} = 100\%$$

- The uninteresting *all documents are irrelevant* classifier gets

$$\frac{99'980}{100'000} = 99.98\%$$

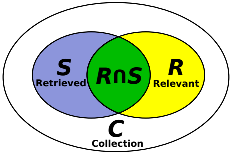
☞ For uneven classes, accuracy may not distinguish excellent from very poor systems

Two types of error for information retrieval and similar tasks

- False positives:
  documents retrieved that should not have been
- False negatives:
  document not retrieved that should have been

A specific confusion matrix:

	relevant	irrelevant
retrieved	true positives	false positives
not retrieved	false negatives	true negatives



Precision, Recall and F-score

Deal with unbalanced classes:

- Use two measures instead of one:
  Precision and Recall (to be defined in next slides)

F-score is a summary of the two measures

Precision

$$\text{precision} = \frac{\text{correctly retrieved documents}}{\text{total number of retrieved documents}}$$
$$= \frac{\text{true positives}}{\text{true positives} + \text{false positives}}$$

- Estimates the likelihood that a retrieved document is indeed relevant to the query
- Ignores false negatives. Take only false positives into account
- Ignores non-retrieved documents. Takes only retrieved documents into account
- Can be biaised by retrieving no documents: gives a perfect score to the system that retrieves no document

Recall (a.k.a. “true positive rate”)

$$\text{recall} = \frac{\text{correctly retrieved documents}}{\text{total number of relevant documents}}$$

$$= \frac{\text{true positives}}{\text{true positives} + \text{false negatives}}$$

- Estimates (one minus) the probabilty to miss relevant documents
- Ignores false positives. Take only false negatives into account
- Ignores irrelevant documents. Takes only relevant documents into account
- Can be biaised by retrieving all documents: gives a perfect score to the system that retrieves all documents

Precision & Recall: example

Spam filtering example:

	System	Reference
email0	OK	OK
email1	OK	Spam
email2	OK	OK
email3	Spam	OK
email4	OK	OK
email5	OK	OK
email6	OK	OK
email7	Spam	Spam
email8	OK	OK
email9	OK	OK
emailA	OK	Spam
emailB	Spam	Spam
emailC	OK	OK
emailD	OK	OK
emailE	OK	OK
emailF	Spam	Spam

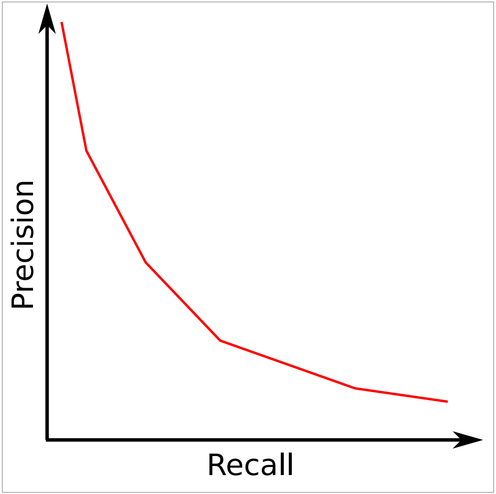
Confusion matrix:

$P =$ $R =$

Note:

- accuracy =
- always-ok system: accuracy= , $R =$, P

Precision vs Recall plots



☞ More in the “Information Retrieval” lecture

C. Grivaz, J.-C. Chappelier

F-score

- Harmonic mean of precision and recall
- The harmonic mean penalises large divergence between numbers, contrary to other means

$$\text{F-score} = 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}}$$

More generally (for given different emphasis to precision and recall):

$$F_\beta = (1+\beta^2) \times \frac{\text{precision} \times \text{recall}}{(\beta^2 \times \text{precision}) + \text{recall}}$$

C. Grivaz, J.-C. Chappelier
ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

Area under ROC curve

ROC curve (ROC = Receiver Operating Characteristic) :

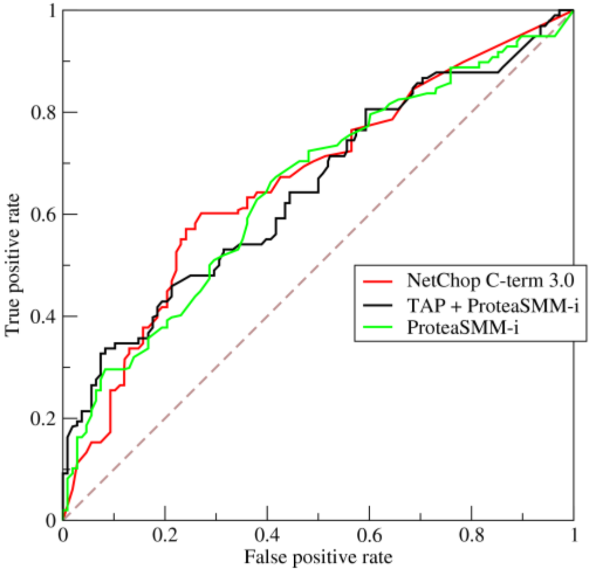
- Plot true positive rate vs false positive rate
  (using a meta parameter; typically, some threshold)

$$\text{true positive rate} = \text{recall}$$

$$\text{false positive rate} = \frac{\text{false positives}}{\text{true negatives + false positives}}$$

- The higher the curve, the better the system
- The area under the curve is sometimes used as an evaluation score

ROC curve



[from Wikipedia, User:BOR]

Example of a non-classification task evaluated as binary classification: PARSEVAL

- A parser output is a syntactical tree
- But parsers are often evaluated as a binary classification task
- Items: constituents
- Classes: exists/does not exist
- Precision: nb of correctly annotated constituent/constituents in parser's output
- Recall: nb of correctly annotated constituent/constituents in gold standard
- Can be computed taking account of labels or not

C. Grivaz, J.-C. Chappelier
ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

Other NLP measures

For some specific NLP tasks, ad-hoc measures have been defined:

- **BLEU** (bilingual evaluation understudy) measure:
  n-gram precision-like measure for machine translation
- **METEOR** (Metric for Evaluation of Translation with Explicit ORdering) measure:
  unigram F-score-like measure for machine translation
- **ROUGE** (Recall-Oriented Understudy for Gisting Evaluation) measures:
  n-gram recall-like measures for automated summarization

C. Grivaz, J.-C. Chappelier
ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

Variability of the results

Whatever evaluation metric you use, measuring it only once on one single test set is <u>**not**</u> appropriate.

You shall estimate its variabilty (e.g. variance) as well!

☞ This means having several different test sets...

How to?

One common way is to use so-called "*cross-validation*".

C. Grivaz, J.-C. Chappelier
ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

Cross-validation

- Idea: using several *test*/*learning* sets splitings to get a more accurate estimation of the results

  (Notice: not necessarily any *validation* set here, despite the name!)

- Repeat k times:
  - split the original data set into n subsets:
  - Repeat n times with a different test (sub)set each time:
    - use $n-1$ subsets for learning and 1 for testing
    - compute evaluation using the (different) test set

- estimate variability of the results

☞ $k \times n$ cross-validation (e.g. 2×5, 1×10): run k times a (different) n-fold cross-validation

Note: why $k \times n$ rather than $1 \times (k\,n)$?

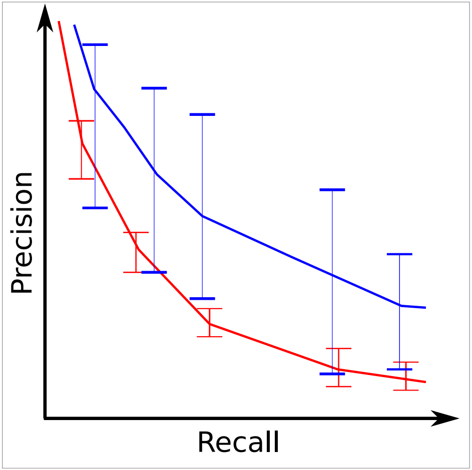
☞ increases variabilty; e.g. chance to have two given samples in the same subset is $\simeq k/n$ versus $1/(k\,n)$.

("$\simeq k/n$" is in fact $1-(n-1)^k/n^k = k/n - \sum_{i=2}^{k}(-1)^i\binom{k}{i}/n^i$)

C. Grivaz, J.-C. Chappelier
ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

Statistically significant evaluation

- Having evaluations allow to compute standard deviations of results
- Which allows to compute confidence intervals or even *confidence boxes*



Comparing two systems in a statistically significant way

Simple example: (paired) Student's t-test: compare two classifiers on the *same* data of T test subsets
(assuming normal distribution and equal variance; generalizations: Welch's t-test, ANOVA)

Δ_i: performance difference between the two classifiers on test subset #i

empirical arithmetic mean: $\mu = \frac{1}{T} \sum_{i=1}^{T} \Delta_i$

empirical unbiaised standard deviation: $s = \sqrt{\frac{1}{T-1} \sum_{i=1}^{T} (\Delta_i - \mu)^2}$

Then $t = \frac{\mu\sqrt{T}}{s}$ is compared to some threshold value for the desired confidence level.

For instance, at 95%, $|t|$ must be bigger than 1.645 (for $T \gg 1$)

To have a result statistically significant at more than 99%, $|t|$ must be bigger than 2.326

C. Grivaz, J.-C. Chappelier

The impact of inter annotator agreement on maximal accuracy

- The best possible result is that of a human
- But diversity exist as long as the IAA is not perfect
- This diversity is not only made of mistakes but of subjectivity as well
- So it would not be good for a computer system to go closer to the gold standard than humans do

Common evaluation protocols

- Allow for objective comparison of systems
- have given rise to a number of hand annotated corpora for specific tasks (e.g. Penn Treebank, many are distributed by the Linguistic Data Consortium (LDC, `http://www.ldc.upenn.edu/`) and the European Language Resources Association (ELRA, `http://www.elra.info/`))
- Evaluation campaigns : specific task, specific evaluation framework, specific time (e.g. conference workshops)
- Example: TREC (information retrieval), ParsEval, SensEval (word sense disambiguation)

C. Grivaz, J.-C. Chappelier
ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

Conclusions

- NLP systems need to be evaluated in order to be objectively compared
- Most NLP task can only be evaluated by being compared to solutions done by humans
- Humans do not always agree and some tasks are subjective
- Several measure exist that need to be computed and which significance need to be statistically measured
- To get clean results, test data should never be used in anyway for development

References

[1] *Consequences of Variability in Classifier Performance Estimates*, by T. Raeder, T. R. Hoens and N. V. Chawla, in 10th IEEE International Conference on Data Mining (ICDM), pp. 421–430, 2010.

[2] *On Comparing Classifiers: Pitfalls to Avoid and a Recommended Approach*, by S. L. Salzberg, in. Data Mining and Knowledge Discovery, 1, pp. 317–327, 1997.