Evaluation of NLP Systems

Martin Hassel
KTH CSC
School of Computer Science and Communication
xmartin@kth.se

Why Evaluate?

- Otherwise you won't know if what you're doing is any good!
- Human languages are very loosely defined
- This makes it hard to prove that something works (as you do in mathematics or logic)

Martin Hassel

Aspects of Evaluation

- General aspects
  - To measure progress
- Commercial aspects
  - To ensure consumer satisfaction
  - Edge against competitors / PR
- Scientific aspects
  - Good science

Martin Hassel

What Is Good Science?

- Induction
  - Testing against a data subset considered fairly representing the complete possible data set
- Popper's theory of falsifiability
  - For an assertion to be falsifiable, in principle it must be possible to make an observation or do a physical experiment that would show the assertion to be false

Martin Hassel

Evaluation Schemes

- Intrinsic
  - Measures the system in of itself
- Extrinsic
  - Measures the efficiency and acceptability of the system output in some task
  - Usually requires ”user” interaction

Martin Hassel

Stages of Development

- Early
  - Intrinsic evaluation on component level
- Mid
  - Intrinsic evaluation on system level
- Late
  - Extrinsic evaluation on system level

Martin Hassel

Manual Evaluation

- Human judges (intrinsic/extrinsic)
  + Semantically based assessment
  – Subjective
  – Time consuming
  – Expensive

Martin Hassel

Semi-Automatic Evaluation

- Task based evaluation (extrinsic)
  + Measures the system's utility
  – Subjective interpretation of questions and answers
- Keyword association (intrinsic)
  + No annotation required
  – Shallow, allows for "good guesses"

Martin Hassel

Automatic Evaluation

- Sentence Recall (intrinsic)
  + Cheap and repeatable
  – Does not distinguish between different summaries
- Vocabulary Test (intrinsic)
  + Useful for key phrase summaries
  – Sensitive to word order differences and negation

Martin Hassel

Why Automatic Evaluation?

- Manual labor is expensive and takes time
- It's practical to be able to evaluate often
  – does this parameter lead to improvements?
- It's tedious to evaluate manually
- Human factor
  – People tend to tire and make mistakes

Martin Hassel

Corpora

- A body of data considered to represent "reality" in a balanced way
  - Sampling
- Raw format vs. annotated data

Martin Hassel

Ethics

- Informants
  - Must be informed
  - Should be anonymous
    – but save demographics!
  - Data should be preserved for ten years

Martin Hassel

Corpora can be…

- a Part-of-Speech tagged data collection

```
Arrangör        nn.utr.sin.ind.nom
var             vb.prt.akt.kop
Järfälla        pm.gen
naturförening   nn.utr.sin.ind.nom
där             ha
Margareta       pm.nom
är              vb.prs.akt.kop
medlem          nn.utr.sin.ind.nom
.               mad
```

Martin Hassel

Corpora can be…

- a parse tree data collection

```
(S
 (NP-SBJ (NNP W.R.) (NNP Grace) )
  (VP (VBZ holds)
   (NP
    (NP (CD three) )
    (PP (IN of)
     (NP
      (NP (NNP Grace) (NNP Energy) (POS 's) )
      (CD seven) (NN board) (NNS seats) ) ) ) )
(. .) )
```

Martin Hassel

Corpora can be…

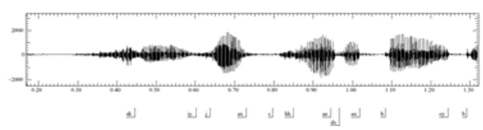
- a RST tree data collection

```
(SATELLITE(SPAN|4||19|)(REL2PAR ELABORATION-
  ADDITIONAL)
(SATELLITE(SPAN|4||7|)(REL2PAR CIRCUMSTANCE)
(NUCLEUS(LEAF|4|)(REL2PAR CONTRAST)
(TEXT _!THE PACKAGE WAS TERMED EXCESSIVE BY
  THE BUSH |ADMINISTRATION,_!|))
(NUCLEUS(SPAN|5||7|)(REL2PAR CONTRAST)
(NUCLEUS(LEAF|5|)(REL2PAR SPAN)
(TEXT _!BUT IT ALSO PROVOKED A STRUGGLE WITH
  INFLUENTIAL CALIFORNIA LAWMAKERS_!))
```

Martin Hassel

Corpora can be…

- a collection of sound samples



Martin Hassel

Widely Accepted Corpora

- Pros
  - Well-defined origin and context
  - Well-established evaluation schemes
  - Inter-system comparability
- Cons
  - Optimizing for a specific data set
  - May establish a common “truth”

Martin Hassel

Gold Standard

- ”Correct guesses” demand knowing what the result should be
- This ”optimal” result is often called a *gold standard*
- How the gold standard looks and how you count can differ a lot between tasks
- The basic idea is however the same

Martin Hassel

Example of a Gold Standard

Gold standard for tagging, shallow parsing and clause boundering

Han	*pn.utr.sin.def.sub*	NPB	*CLB*
är	*vb.prs.akt.kop*	VCB	*CLI*
mest	*ab.suv*	ADVPB|APMINB	*CLI*
road	*jj.pos.utr.sin.ind.nom*	APMINB|APMINI	*CLI*
av	*pp*	PPB	*CLI*
äldre	*jj.kom.utr/neu.sin/plu.ind/def.nom*	APMINB|NPB|PPI	*CLI*
sorter	*nn.utr.plu.ind.nom*	NPI|PPI	*CLI*
.	*Mad*	0	*CLI*

Martin Hassel

Gold Standard or Gold Standards?

- Sometimes many ”answers” are (potentially) equally correct!
  - Machine Translation
  - Text Summarization
- If possible:
  - List all correct answers (all tags for ambigous words)
  - Compare answers to (several) examples of correct answers
  - Translate data to a simpler (less detailed?) format (IOB-parsing)
  - Solve some other problem which is more easily evaluated, and that builds on the problem we really want to evaluate (synonyms in ORD or TOEFL)
  - Evaluate manually!

Martin Hassel

Some Common Measures

- Precision = correct guesses / all guesses
- Recall = correct guesses / correct answers
- Precision and recall often are mutually dependant
  - higher recall → lower precision
  - higher precision → lower recall
- F-score: combines precision and recall
  - $F\alpha = 1 / ((\alpha*(1/P)) + (1-\alpha)*(1/R))$
    α = weighting factor
  - $F.5 = 2*P*R/(P+R)$

Martin Hassel

More Evaluation Terminology

- True positive
  – Alarm given at correct point
- False negative
  – No alarm when one should be given
- False positive
  – Alarm given when none should be given
- (True negative)
  – The algorithm is quiet on uninteresting data
- In e.g. spell checking the above could correspond to detected errors, missed errors, false alarms and correct words without warning.

Martin Hassel

How Good Is 95%?

- It depends on what problem you are solving!
- Try to determine expected upper and lower bounds for performance (of a specific task)
- A baseline tells you the performance of a naïve approach (lower bound)

Martin Hassel

Lower Bound

- Baselines
  - Serve as lower limit of acceptability
  - Common to have several baselines
- Common baselines
  - Random
  - Most common choice/answer (e.g. in tagging)
  - Linear selection (e.g. in summarization)

Martin Hassel

Upper Bound

- Sometimes there is an upper bound lower than 100%
- Example 1:
  3% of all answers in the evaluation corpus are (randomly) errornous
  - Impossible to learn where random errors occur

Martin Hassel

Upper Bound

- Example 2:
  In 10% of all cases experts disagree on the correct answer
  - Human ceiling (inter-assessor agreement)
  - Low inter-assessor agreement can sometimes be countered with comparison against several ”sources”

Martin Hassel

Is 95.3% Better Than 94.8%?

- It depends, have you tested against 212 examples or 10 millions examples?
- Statistical significance testing tells us how often chance would give us this difference if both methods perform on par
- If you evaluate many methods on the same data (or the same method with many different parameter settings) you must take this into consideration

Martin Hassel

Example of a Significance Test

McNemar's Test

- Null hypothesis: both methods are equally good
- Example: If we toss a coin, what is the probability that we get *B* heads and *C* tails?
- If the probability is low, reject the null hypothesis (i.e. the difference between the methods is significant)
- In practice: ((*B*-*C*)^2)/(*B*+*C*)
- Look up the Chi-square distribution if B+C is large
- Otherwise calculate exact value using binomial distribution

Martin Hassel

Limited Data

- Limited data is often a problem, especially in machine learning
- We want lots of data for training
  - Better results
- We want lots of data for evaluation
  - More reliable numbers
- If possible, create your own data!
  - Missplel

Martin Hassel

Limited Data

***N*-fold Cross Validation**

- Idea:
  1. Set 5% of the data aside for evaluation and train on 95%
  2. Set another 5% aside for evaluation and repeat training on 95%
  3. … and again (repeat in total 20 times)
- Take the mean of the evaluation results to be the final result

Martin Hassel

Considerations

- How you evaluate affects the direction of the research
  - Information retrieval
  - Text summarization
- Evaluation data:
  - The training data or the trimming data (does not reflect reality)
  - The same data is used over and over again (significance)
- Evaluation cycles: slow / fast
- Hardware demanding: memory / drive space
- Resource demanding: lots of data

Martin Hassel

Concrete Examples

- Taggning
  - Force the tagger to assign exactly one tag to each token – precision?
- Parsing
  - What happens when almost correct?
  - Crossing-brackets, partial trees, how many sentences got full trees?
- Spell checking
  - Recall / precision for alarms
  - How far down in the suggestion list is the correct suggestion?

Martin Hassel

Concrete Examples

- Grammar checking
  - How many are false alarms (precision)?
  - How many errors are detected (recall)?
  - How many of these have the correct diagnosis?
- Machine translation
  - How many n-grams overlap with gold standard(s)?
  - BLEU scores
- Text Summarization
  - How many n-grams overlap with gold standard(s)?
  - ROUGE scores (premiers short summaries)

Martin Hassel

Concrete Examples

- Synonyms
  - How many questions in the TOEFL test can the program answer correct?
- Information retrieval
  - What is the precision of the first X hits? At Y% recall?
  - Mean precision, precision-recall graphs.

Martin Hassel

Concrete Examples

- Text categorizing
  - How many documents were correctly classified?
- Clustering
  - How pure where the clusters?
  - Entropy, distance measures etc.

Martin Hassel

Conferences & Campaigns

- TREC – Text REtrieval Conferences
  - Information Retrieval/Extraction and TDT
- CLEF – Cross-Language Evaluation Forum
  - Information Retrieval on texts in European languages
- DUC – Document Understanding Conference
  - Automatic Text Summarization
- SENSEVAL
  - Word Sense Disambiguation
- ATIS – Air Travel Information System
  - DARPA Spoken Language Systems

and few more…

Martin Hassel