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)

Aspects of Evaluation

• General aspects
  • To measure progress

• Commercial aspects
  • To ensure consumer satisfaction
  • Edge against competitors / PR

• Scientific aspects
  • Good science

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

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

Stages of Development

• Early
  • Intrinsic evaluation on component level

• Mid
  • Intrinsic evaluation on system level

• Late
  • Extrinsic evaluation on system level
Manual Evaluation

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

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”

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

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

Corpora

• A body of data considered to represent ”reality” in a balanced way
  • Sampling

• Raw format vs. annotated data

Ethics

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

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 (NNP Grace) (NNP Energy) (POS 's)) (CD seven) (NN board) (NNS seats)))))

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")

Corpora can be…

- a collection of sound samples

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”

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
Example of a Gold Standard

Gold standard for tagging, shallow parsing and clause boudering

<table>
<thead>
<tr>
<th>Han</th>
<th>NPB</th>
<th>CLB</th>
</tr>
</thead>
<tbody>
<tr>
<td>är</td>
<td>VCB</td>
<td>CLI</td>
</tr>
<tr>
<td>mest</td>
<td>ADV/PB/APMINB</td>
<td>CLI</td>
</tr>
<tr>
<td>road</td>
<td>APMINB/APMINI</td>
<td>CLI</td>
</tr>
<tr>
<td>av</td>
<td>PPR</td>
<td>CLI</td>
</tr>
<tr>
<td>äldre</td>
<td>APMINB/NPB</td>
<td>CLB</td>
</tr>
<tr>
<td>sorter</td>
<td>NPE/PP</td>
<td>CLI</td>
</tr>
</tbody>
</table>

Example of a Gold Standard

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 ambiguous 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!

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
  - \( F_\alpha = \frac{1}{(\alpha \times \frac{P}{R}) + (1-\alpha) \times \frac{R}{P})} \)
  - \( F_5 = 2 \times P \times R \times (P+R) \)

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.

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)

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)
Upper Bound

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

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: \[
\frac{(B-C)^2}{B+C}
\]
- Look up the Chi-square distribution if B+C is large
- Otherwise calculate exact value using binomial distribution

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!
  • Misspeli

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
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

Concrete Examples

• Tagging
  • 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?

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)

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.

Concrete Examples

• Text categorizing
  • How many documents were correctly classified?

• Clustering
  • How pure where the clusters?
  • Entropy, distance measures etc.

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…