European Summer School in Language, Logic and Information
ESSLLI 2007

Evaluation in natural language processing

Diana Santos
Linguateca - www.linguateca.pt

Dublin, 6-10 August 2007

Goals of this course

- Motivate evaluation
- Present basic tools and concepts
- Illustrate common pitfalls and inaccuracies in evaluation
- Provide concrete examples and name famous initiatives

Plus

- provide some history
- challenge some received views
- encourage critical perspective (to NLP and evaluation)

Messages I want to convey

- Evaluation at several levels
- Be careful to understand what is more important, and what it is all about. Names of disciplines, or subareas are often tricky
- Take a closer look at the relationship between people and machines
- Help appreciate the many subtle choices and decisions involved in any practical evaluation task

- **Before** doing anything, think hard on how to evaluate what you will be doing

Course assessment

- Main topics discussed
- Fundamental literature mentioned
- Wide range of examples considered
- Pointers to further sources provided
- Basic message(s) clear
- Others?
  - Enjoyable, reliable, extensible, simple?

Evaluation

- E**valu**ation= assign value to
- Values can be assigned to
  - the purpose/motivation
  - the ideas
  - the results
- Evaluation depends on whose values we are taking into account
  - the stakeholders
  - the community
  - the developer's
  - the user's
  - the customer's

What is your quest?

- Why are you doing this (your R&D work)?
- What are the expected benefits to science (or to mankind)?

- a practical system you want to improve
- a practical community you want to give better tools (better life)

OR

- a given problem you want to solve
- a given research question you are passionate (or just curious) about

Different approaches to research

Those based on an originally practical problem
- find something to research upon

Those based on an originally theoretical problem
- find some practical question to help disentangle it

But NLP has always a practical and a theoretical side, and, for both, **evaluation is relevant**

House (1980) on kinds of evaluation schools

- Systems analysis
- Behavioral objectives
- Decision-making
- Goal-free
  - don't look at what they wanted to do, consider everything as side effects
- Art criticism
- Professional review
- Quasi-legal
- Case study

Attitudes to numbers

*but where do all these numbers **come from**?* (John McCarthy)

***Pseudo-science**: because we're measuring something it must be science* (Gaizauskas 2003)

*I often say that when you can measure what you are speaking about, and express it in numbers, you know something about it; but when you cannot measure it and cannot express it in numbers your knowledge is of a meager and unsatisfactory kind; it may be the beginning of knowledge, **but you have scarcely, in your thoughts, advanced to the stage of science**, whatever the matter may be.*

Lord Kelvin, *Popular Lectures and Addresses*, (1889), vol 1. p. 73.

Qualitative vs. quantitative

- are not in opposition
- both are often required for a satisfactory evaluation
- there has to be some relation between the two
  - partial order or ranking in qualitative appraisals
  - regions of the real line assigned labels

- often one has many qualitative (binary) assessments that are counted over (TREC)
- one can also have many quantitative data that are related into a qualitative interpretation (Biber)

Qualitative evaluation of measures

- Evert, Stefan & Brigitte Krenn. " Methods for the qualitative evaluation of Lexical Association Measures", *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics* (Toulouse, 9-11 July 2001), pp. 188-195.
- Sampson, Geoffrey & Anna Babarczy. "A test of the leaf–ancestor metric for parse accuracy", *Journal of Natural Language Engineering* **9**, 2003, pp. 365–80.

Lexical association measures

$$t \approx \frac{\frac{n_{11}}{n_{++}} - \frac{n_{1+}n_{+1}}{n_{++}^2}}{\sqrt{\frac{n_{11}}{n_{++}^2}}} = \frac{n_{11} - m_{11}}{\sqrt{n_{11}}}$$

- several methods

(frequentist, information-theoretic and statistical significance)
- the problem: measure strength of association between words (Adj, N) and (PrepNoun, Verb)
- standard procedure: **manual judgement of the *n*-best candidates** (for example, corrects among the 50 or 100 first)
  - can be due to chance
  - no way to do evaluation per frequency strata
  - comparison of different lists (for two different measures)

$$G^2 = 2\sum_{i,j} n_{ij} \log \frac{n_{ij}}{m_{ij}} \qquad X^2 = \sum_{i,j} \frac{(n_{ij} - m_{ij})^2}{m_{ij}}$$

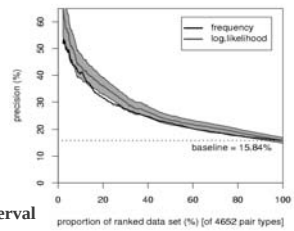
<word>	n_{13}	m_{11}	exact	rank	G^2-χ^2	rank	X^2-χ^2	rank	T	rank
and	22	21.14	.8255	1	.8512	1	.8503	1	.9958	1
services	1	0.50	.3910	2	.5293	2	.4735	2	.9840	2
financial	2	0.78	.1842	3	.2493	3	.1673	3	.9692	3
domestic	1	0.20	.1817	4	.2033	4	.0740	4	.9601	4
motor	1	0.15	.1363	5	.1435	5	.0256	6	.9592	6
recent	2	0.56	.1073	6	.1343	6	.0523	5	.9567	5
instance	1	0.10	.0982	7	.0971	7	.0053	7	.9378	7
engineering	1	0.09	.0847	8	.0815	8	.0022	8	.9317	8
utility	1	0.08	.0724	9	.0678	9	.0007	10	.9248	10
sugar	2	0.06	.0916	10	.0013	10	.0000	16	.8214	16
glass	2	0.04	.0006	11	.0005	11	.0000	18	.7746	18
in	7	21.11	.0006	12	.0003	12	.0019	9	.9315	9
food	4	0.29	.0002	13	.0002	13	.0000	15	.8475	15
newspaper	3	0.10	.0002	14	.0001	14	.0000	17	.8099	17
drug	5	0.47	.0001	15	.0001	15	.0000	13	.8531	13
to	9	27.88	.0000	16	.0000	16	.0003	11	.9203	11
appliance	2	0.00	.0000	17	.0000	17	.0000	23	.4153	23
movie	4	0.09	.0000	18	.0000	18	.0000	19	.7116	19
of	5	29.22	.0000	19	.0000	19	.0000	12	.9002	12
futures	8	0.29	.0000	20	.0000	21	.0000	20	.6872	20
the	110	58.95	.0000	21	.0000	20	.0000	14	.8524	14
oil	11	0.41	.0000	22	.0000	22	.0000	21	.6416	21
airline	17	0.29	.0000	23	.0000	24	.0000	24	.2947	24
an	42	4.03	.0000	24	.0000	23	.0000	22	.5965	22

Figure 8: test for association : <**word**> **industry**

From Pedersen (1996)

Precision per rank

Source: *The significance of result differences*
ESSLLI 2003, Stefan Evert & Brigitte Krenn



95% Confidence Interval

Parser evaluation

GEIG (Grammar Evaluation Interest Group) standard procedure, used in Parseval (Black et al., 1991), for phrase-structure grammars, comparing the candidate **C** with the key in the treebank **T**

- first, removing auxiliaries, null categories, etc...
- **cross-parentheses score**: the number of cases where a bracketed sequence from the standard overlaps a bracketed sequence from the system output, but neither sequence is properly contained in the other.
- **precision and recall**: the number of parenthesis pairs in C∩T divided by the number of parenthesis in C, and in T

- labelled version (the label of the parenthesis must be the same)

The leaf ancestor measure

golden (key): [S [N1 *two* [N1 *tax revision*] *bills*] *were passed*]
candidate: [S [NP *two tax revision bills*] *were passed*]
lineage - sequence of node labels to the root, golden:candidate

- *two* N1 [S: NP [S
- *tax* [N1 N1 S: NP S
- *revision* N1] N1 S : NP S
- *bills* N1] S : NP] S
- *were* S : S
- *passed* S] : S]

Computing the measure

- Lineage similarity: sequence of node labels to the root
- uses (Levenshtein's) editing distance Lv (1 for each operation `Insert`, `Delete`, `Replace`)
- 1-Lv(cand,golden)/(size(cand)+size(golden))
- `Replace` =f with values in {0,2}
  - If the category is related (shares the same first letter, in their coding), f=0.5, otherwise f=2 (partial credit for partly-correct labelling)
- Similarity for a sentence is given by averaging similarities for each word

Application of the leaf ancestor measure

two N1 [S: NP [S	0.917
tax [N1 N1 S: NP S	0.583
revision N1] N1 S : NP S	0.583
bills N1] S : NP] S	0.917
were S : S	1.000
passed S] : S]	1.000

- LAM (average of the values above): 0.833
- GEIG unlabelled F-score: 0.800
- GEIG labelled F-score: 0.400

Evaluation/comparison of the measure

- Setup
  - Picked 500 random chosen sentences from SUSANNE (the golden standard)
  - Applied two measures: GEIG (from Parseval) and LAM to the output of a parser
- Ranking plots
  - Different ranking
  - no correlation between GEIG labelled and unlabelled ranking!
- Concrete examples of extreme differences (favouring the new metric)
- Intuitively satisfying property: since there are measures per words, it is possible to pinpoint the problems, while GEIG is only global
- *which departures from perfect matching ought to be penalized heavily can only be decided in terms of "educated intuition"*

Modelling probability in grammar (Halliday)

- The grammar of a natural language is characterized by overall quantitative tendencies (two kinds of systems)
  - equiprobable: 0.5-0.5
  - skewed: 0.1-0.9 (0.5 redundancy) – unmarked categories
- In any given context, ... global probabilities may be significantly perturbed. ... the local probabilities, for a given situation type, may differ significantly from the global ones. *"resetting" of probabilities ... characterizes functional (register) variation in language*. This is how people recognize the "context of situation" in text. (pp. 236-8)
- "probability" as a theoretical construct is just the technicalising of "modality" from everyday grammar

There is more to evaluation on heaven and earth...

- evaluation of a system
- evaluation of measures
- hypotheses testing
- evaluation of tools
- evaluation of a task
- evaluation of a theory
- field evaluations
- evaluation of test collections
- evaluation of a research discipline
- evaluation of evaluation setups

Sparck Jones & Galliers (1993/1996)

- The first and possibly only book devoted to NLP evaluation in general
- written by primarily IR people, from an initial report
- a particular view (quite critical!) of the field
- *In evaluation, what matters is the setup*. [system + operational context]
- *clarity of goals are essential to an evaluation, but unless these goals conform to something 'real' in the world, this can only be a first stage evaluation. At some point the* ***utility*** *of a system has to be a consideration, and for that one must know what it is to be used for and for whom, and testing must be with these considerations in mind* (p. 122)

Sparck Jones & Galliers (1993/1996) contd.

Comments on actual evaluations in NLP (p. 190):

- evaluation is strongly task oriented, either explicitly or implicitly
- evaluation is focussed on systems without sufficient regard for their environments
- evaluation is not pushed hard enough for factor decomposition

Proposals

- mega-evaluation structure: 'braided chain': *The braid model starts from the observation that tasks of any substantial complexity can be decomposed into a number of linked sub-tasks.*
- four evaluations of a fictitious PlanS system

Divide and conquer? Or lose sight?

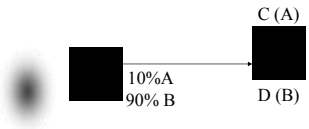
- **blackbox**: description of what the system should do
- **glassbox**: know which sub-systems there are, evaluate them separately as well

BUT

- some of the sub-systems are **user-transparent** (what should they do?) as opposed to **user-significant**
- the dependence of the several evaluations is often neglected!

- Evaluation in series: task A followed by task B (Setzer & Gaizauskas, 2001): If 6 out of 10 entities in task A, then maximum 36 out of 100 relations in task B

The influence of the performance of prior tasks



Even if C(A) is 100% accurate, the output of the whole system is not signifi-cantly affected

- A word of caution about the relevance of the independent evaluation of components in a larger system

Dealing with human performance

- developing prototypes, iteratively evaluated and improved
  but, *as was pointed out by Tennant (1979), people always adapt to the limitations of an existing system* (p. 164)
- doing Wizard-of-Oz (WOZ) experiments
  not easy to deceive subjects, difficult to the wizard, a costly business
- *to judge system performance by assuming that perfect performance is achievable is a fairly serious mistake* (p. 148)

Jarke et al. (1985): setup

- Alumni administration: demographic and gift history data of school alumni, foundations, other organizations and individuals
- Questions about the schools alumni and their donations are submitted to the Assoc. Dir. for EA from faculty, the Deans, student groups, etc.
- Task example:
  `A list of alumni in the state of California has been requested. The request applies to those alumni whose last name starts with an "S". Obtain such a list containing last names and first names.`
- Compare the performance of 8 people using NLS to those using SQL
- 3 phases: 1. group 1: NLS, group 2: SQL; 2. vice versa; 3. subjects could choose

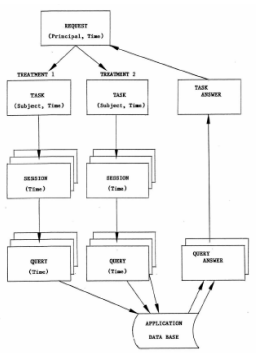
Hypotheses and data

H1 There will be no difference between using NLS or SQL
H2 People using NLS will be more efficient
H3 Performance will be neg. related to the task difficulty
H4 Performance will be neg. related to perception of difficulty and to pos. related to their understanding of a solution strategy

- Forms filled by the subjects
- Computer logs

- 39 different requests (87 tasks, 138 sessions, 1081 queries)

Jarke et al. (contd.)



1. Measure the success of subjects in performing their task or sub-task. Success is based on both the syntactical correctness of queries submitted as well as the contribution of the answer towards accomplishing the overall task.
2. Measure the effort involved in accomplishing the task or sub-task.
3. Measure the factors that are likely to influence success and/or effort.
4. Capture subjects' perceptions about a treatment.

Coding scheme

- Eight kinds of situations that must be differentiated
  - 3. a syntactically correct query produces no (or unusable) output because of a semantic problem – it is the wrong question to ask
  - 5. a syntactically and semantically correct query whose output does not substantially contribute to task accomplishment (e.g. test a language feature)
  - 7. a syntactically and semantically correct query cancelled by a subject before it has completed execution

Results and their interpretation

- Task level
  - Task performance summary disappointing: 51.2% NLS and 67.9% SQL
  - Number of queries per task: 15.6 NLS, 10.0 SQL
- Query level
  - partially correct output from a query: 21.3% SQL, 8.1% NLS (3:1!)
  - query length: 34.2 tokens in SQL vs 10.6 in NLS
  - typing errors: 31% in SQL, 10% NLS
- Individual differences; order effect; validity (several methods all indicated the same outcome)
- H1 is rejected, H2 is conditionally accepted (on token length, not time), H3 is accepted, the first part of H4 as well

Outcome regarding the hypotheses

H1 There will be no difference between using NLS or SQL

Rejected!

H2 People using NLS will be more efficient

Conditionally accepted *(on token length, not time)!*

H3 Performance will be neg. related to the task difficulty

Accepted!

H4 Performance will be neg. related to perception of difficulty and to pos. related to their understanding of a solution strategy

First part accepted!

Jarke et al. (1985): a field evaluation

- Compared database access in SQL and in NL

Results

- no superiority of NL systems could be demonstrated in terms of either query correctness or task solution performance
- NL queries are more concise and require less formulation time

Things they learned

- importance of feedback
- disadvantage of impredictability
- importance of the total operating environment
- restricted NL systems require training...

User-centred evaluation

- 9 in 10 users happy? or all users 90% happy?
- Perform a task with the system
  - before
  - after
- Time/pleasure to learn
- Time to start being productive
- Empathy

- Costs much higher than technical evaluations
- Most often than not, what to improve is not under your control...

Three kinds of system evaluation

- **Ablation**: destroy to rebuild
- **Golden collection**: create solutions before evaluating
- **Assess after running**: based on cooperative pooling
  - Include in a larger task, in the real world

Problems with each

- Difficult to create a realistic point of departure (noise)
- A lot of work, not always all solutions to all problems... difficult to generalize
- Too dependent on the systems' actual performance, too difficult to agree on beforehand criteria

Evaluation resources

- 3 kinds of test materials (evaluation resources) (*SPG*)
  - coverage corpora (examples of all phenomena)
  - distribution corpora (maintaining relative frequency)
  - test collections (texts, topics, and relevance judgements)
- test suites (coverage corpora + negative instances)
- corrupt/manipulated corpora

- a corpus/collection of what? unitizing!!
  - A corpus is a **classified** collection of **linguistic objects** to **use** in NLP/CL

Unitizing

- Krippendorff (2004)

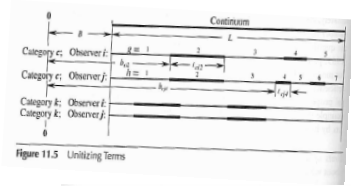
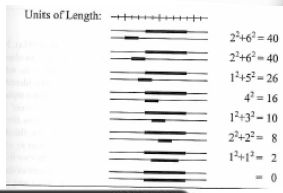


Figure 11.5 Unitizing Terms

- Computing differences in units



A digression on frequency, and on units

What is more important: the most frequent of the least frequent?

- stopwords in IR
- content words of middle frequency in indexing
- rare words in author studies, plagiarism detection

- What is a word?
  - Spelling correction assessment: `correctionassessement`
  - *Morfolimpíadas* and the tokenization quagmire (disagreement on 15.9% of the tokens and 9.5% types, Santos et al. (2003))
  - Sinclair's quote on the defence of multiwords: **p** followed by **aw** means **paw**, followed by **ea** means **pea**, followed by **ie** means **pie** ... is nonsensical!
  - Does punctuation count for parse similarity?

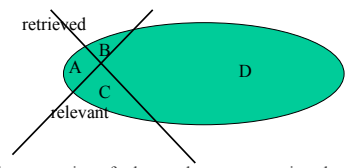
Day 2

The basic model for precision and recall

P=A/(A+B)

R=A/(A+C)

C: missing
B: in excess



- **precision** measures the proportion of relevant documents retrieved out of the retrieved ones
- **recall** measures the proportion of relevant documents retrieved out of the relevant ones
- if a system retrieves all documents, recall is always one, and precision is **accuracy**

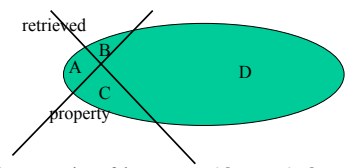
Some technical details and comments

- From two to one: F-measure
- $F_\beta = (\beta^2+1)*precision*recall/(\beta^2*precision+recall)$ $\quad \frac{2*P*R}{P+R}$
- A feeling for common values of precision, recall and F-measure?
- Different tasks from a user point of view
  - High recall: to do a state of the art
  - High precision: few but good (enough)
- Similar to a contingency table

Extending the precision and recall model

P=A/(A+B)

R=A/(A+C)



- **precision** measures the proportion of documents **with a particular property** retrieved out of the retrieved ones
- **recall** measures the proportion of documents retrieved **with a particular property** out of the relevant ones
- correct, useful, similar to X, displaying novelty, ...

Examples of current and common extensions

- given a candidate and a key (golden resource)

Each decision by the system can be classified as

- correct
- **partially correct**
- missing
- in excess

- instead of binary relevance, one could have different scores for each decision
  - graded relevance (very relevant, little relevant, ...)

"Same" measures do not necessarily mean the same

- *though 'recall' and 'precision' were imported from IR into the DARPA evaluations, they have been given distinctive and distinct meanings, and it is not clear how generally applicable they could be across NLP tasks* (p. 150)
- in addition, using the same measures does not mean the **same** task
  - named entity recognition: MUC, CoNLL and HAREM
  - word alignment: Melamed, Véronis, Moore and Simard
- different understandings of the "same" task require different measures
  - question answering (QA)
  - word sense disambiguation (WSD)

NER: 1st pass...

Eça de Queirós nasceu na Póvoa de Varzim em 1845, e faleceu 1900, em Paris. Estudou na Universidade de Coimbra.

Eça de Queirós nasceu na ***Póvoa de Varzim*** em *1845*, e faleceu *1900*, em ***Paris***. Estudou na ***Universidade de Coimbra***.

Semantic categories I: ***City***, *Year*, ***Person***, ***University***

Semantic categories II: ***Place***, *Time*, ***Person***, ***Organization***

Semantic categories III: ***Geoadmin location***, *Date*, ***Famous writer***, ***Cultural premise/facility***

Evaluation pitfalls because of "same" measure

- the best system in MUC attained F-measure greater than 95%
  -> so, if best scores in HAREM had F-measure of 70%, Portuguese lags behind...

Wrong!

- Several problems:
  - the evaluation measures
  - the task definition

Study at the <**ENAMEX TYPE="ORGANIZATION"**>Temple University</**ENAMEX**>'s <**ENAMEX TYPE="ORGANIZATION"**>Graduate School of Business</**ENAMEX**>

MUC-7, Chinchor (1997)

CONLL, Sang (2002)

Wolff	B-PER
,	O
currently	O
a	O
journalist	O
in	O
Argentina	B-LOC
,	O
played	O
with	O
Del	B-PER
Bosque	I-PER

Evaluation measures used in MUC and CoNLL

MUC: Given a set of semantically defined categories expressed as proper names in English

- universe is: number of correct NEs in the collection
- recall: number of correct NEs returned by the system/number of correct NEs

CoNLL^{fict}: Given a set of words, marked as initiating or continuing a NE of three kinds (+MISC)

- universe: number of words belonging to NEs
- recall: number of words correctly marked by the system/number of words

Detailed example, MUC vs. CoNLL vs. HAREM

U.N. official Ekeus heads for Baghdad 1:30 pm Chicago time.

- [ORG U.N.] official [PER Ekeus] heads for [LOC Baghdad] 1:30 p.m. [LOC Chicago] time. (**CoNLL** 2003: 4)
- [ORG U.N.] official [PER Ekeus] heads for [LOC Baghdad] [TIME 1:30 p.m. [LOC Chicago] time]. (**MUC**)
- [PER U.N. official Ekeus] heads for [LOC Baghdad] [TIME 1:30 p.m. Chicago time]. (**HAREM**)

Detailed example, MUC vs. CoNLL vs. HAREM

He gave Mary Jane Eyre last Christmas at the Kennedys.

- He gave [PER Mary] [MISC Jane Eyre] last [MISC Christmas] at the [PER Kennedys]. (**CoNLL**)
- He gave [PER Mary] Jane Eyre last Christmas at the [PER Kennedys]. (**MUC**)
- He gave [PER Mary] [OBRA Jane Eyre] last [TIME Christmas] at the [LOC Kennedys]. (**HAREM**)

Task definition

MUC: Given a set of semantically defined categories expressed as proper names (in English) (or number or temporal expressions), mark their occurrence in text

- correct or incorrect

HAREM: Given all proper names (in Portuguese) (or numerical expressions), assign their correct semantic interpretation in context

- partially correct
- alternative interpretations

Summing up

- There are several choices and decisions when defining precisely a task for which an evaluation is conducted
- Even if, for the final ranking of systems, the same kind of measures are used, one **cannot** compare results of distinct evaluations
  - if basic assumptions are different
  - if the concrete way of measuring is different

Plus: different languages!

- *handling multi-lingual evaluation: data has to be collected for different languages, and the data has to be comparable: however, if data is functionally comparable it is not necessarily descriptively comparable (or vice versa), since languages are intrinsically different* (p.144)
- while there are proper names in different languages, the difficulty of identifying them and/or classifying them is to a large extent language-dependent
  - *Thursday* vs. *quinta*
  - *John* vs. *O João*
  - *United Nations* vs. *De forente nasjonene*
  - German noun capitalization

Have we gone too far? PR for everything?

- Sentence alignment (Simard et al., 2000)
  P: given the pairings produced by an aligner, how many are right
  R: how many sentences are aligned with their translations
- Anaphora resolution (Mitkov, 2000)
  P: correctly resolved anaphors / anaphors attempted to be resolved
  R: correctly resolved anaphors / all anaphors
- Parsing: 100% recall in CG parsers ...
  (all units receive a parse... so it should be parse accuracy instead)
- Using precision and recall to create one global measure for information-theoretic inspired measures
  P: value / maximum value given output; R: value / maximum value in golden res.

Sentence alignment (Simard et al., 2000)

- Two texts S and T viewed as unordered sets of sentences $s_1\ s_2\ ...\ t_1\ t_2$
- An alignment of the two texts is a subset of SxT
  $A=\{ (s_1, t_1), (s_2, t_2), (s_2, t_3), ... (s_n, t_m)\}$
  A_R - reference alignment
- Precision: $|A \cap A_R|/|A|$
- Recall: $|A \cap A_R|/|A_R|$
- measured in terms of characters instead of sentences, *because most alignment errors occurred on small sentences*
  - weighted sum of pairs source sentence x target sentence (s_1, t_1), weighted by character size of both sentences $|s_1|+|t_1|$

Anaphora resolution (Mitkov, 2000)

- Mitkov claims against indiscriminate use of precision and recall

$$Recall = \frac{Number\ of\ correctly\ resolved\ anaphors}{Number\ of\ all\ anaphors}$$

$$Precision = \frac{Number\ of\ correctly\ resolved\ anaphors}{Number\ of\ anaphors\ attempted\ to\ be\ resolved}$$

- suggesting instead the success rate of an algorithm (or system)

$$Success\ rate_{Anaphora\ resolution\ algorithm} = \frac{Number\ of\ successfully\ resolved\ anaphors}{Number\ of\ all\ anaphors}$$

- and **non-trivial sucess rate** (more than one candidate) and **critical success rate** (even tougher: no choice in terms of gender or number)

Some more distinctions made by Mitkov

- It is different to evaluate
  - an algorithm: based on ideal categories
  - a system: in practice, it may not have succeeded to identify the categories
- Co-reference is different (a particular case) of anaphor resolution
- One must include also possible anaphoric expressions which are not anaphors in the evaluation (false positives)
  - in that case one would have to use another additional measure...

MT evaluation for IE (Babych et al., 2003)

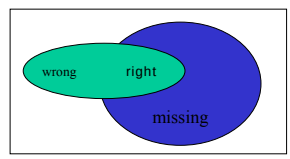
3 measures that characterise differences in statistical models for MT and human translation of each text:

- a measure of "avoiding overgeneration" (which is *linked* to the standard "precision" measure)
- a measure of "avoiding under-generation" (*linked* to "recall")
- a combined score (calculated similarly to the F-measure)

Note however, that the proposed scores could go beyond the range [0,1], which makes them different from precision/recall scores

Evaluation of reference extraction (Cabral 2007)

- Manually analysed texts with the references identified
- A list of candidate references
- Each candidate is marked as
  - correct
  - with excess info
  - missing info
  - is missing
  - wrong
- Precision, recall
- overgeneration, etc



The evaluation contest paradigm

- A given **task**, with success measures and evaluation resources/setup agreed upon
- Several systems attempt to perform the particular task
- **Comparative** evaluation, measuring state of the art
- Unbiased compared to self-evaluation (most assumptions are never put into question)
- Paradigmatic examples:
  - TREC
  - MUC

MUC: Message Understanding Conferences

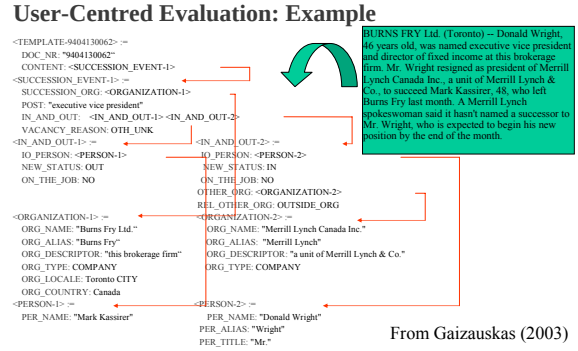
- 1st MUCK (1987)
  - common corpus with real message traffic
- MUCK-II (1989)
  - introduction of a template
  - training data annotated with templates
- MUC-3 (1991) and MUC-4 (1992)
  - newswire text on terrorism
  - semiautomatic scoring mechanism
  - collective creation of a large training corpus
- MUC-5 (1993) (with TIPSTER)
  - two domains: microelectronics and joint ventures
  - two languages: English and Japanese

From Hirschman (1998)
and Grishman & Sundheim (1996)

MUC (ctd.)

- MUC-6 (1995) and MUC-7 (1998): management succession events of high level officers joining or leaving companies
  - domain independent metrics
  - introduction of tracks
    - named entity
    - co-reference
    - template elements: NEs with alias and short descriptive phrases
    - template relation: properties or relations among template elements (*employee-of*, ...)
  - emphasis on portability
- Related, according to H98, because adopting IE measures
  - MET (Multilingual Entity Task) (1996, 1998)
  - Broadcast News (1996, 1998)

Application Task Technology Evaluation vs User-Centred Evaluation: Example



From Gaizauskas (2003)

Comparing the relative difficulty of MUCK2 and MUC-3 (Hirschman 91)

- Complexity of data
  - telegraphic syntax, 4 types of messages vs. 16 types from newswire reports
- Corpus dimensions
  - 105 messages (3,000 words) vs. 1300 messages (400,000 words)
  - test set: 5 messages (158 words) vs. 100 messages (30,000 words)
- Nature of the task
  - template fill vs. relevance assessment plus template fill (only 50% of the messages were relevant)
- Difficulty of the task
  - 6 types of events, 10 slots vs. 10 types of events and 17 slots
- Scoring of results (70-80% vs 45-65%)

Aligning the answer with the key...

Figure 3
Target and hypothetical outputs for the example text.

From Kehler et al. (2001)

Scoring the tasks

- MUCK-II
  - 0 – wrong; 1 – missing; 2 – right
- MUC-3
  - 0 – wrong or missing; 1 – right
- Since 100% is the upper bound, it is actually more meaningful to compare the "shortfall" from the upper bound
  - 20-30% to 35-55%
  - MUC-3 performance is half as good as (has twice the shortfall of) MUCK-2
- the relation between difficulty and precision/recall figures is certainly not linear (the last 10-20% is always much harder to get than the first 80%)

What we learned about evaluation in MUC

Chinchor et al. (2003) conclude that evaluation contests are

- good to get a snapshot of the field
- not good as a predictor of future performance
- not effective to determine which techniques are responsible for good performance across systems
- system convergence (Hirschmann, 1991): two test sets, do changes in one and check whether changes made to fix problem s in one test set actually helped in another test set
- costly:
  - investment of substantial resources
  - port the systems to the chosen application

Day 3

The human factor

- Especially relevant in NLP!
- All NLP systems are ultimately to satisfy people (otherwise no need for NLP in the first place)
- Ultimately the final judges of a NLP system will always be people
- To err is human (*errare humanum est*): important to deal with error
- To judge is human – and judges have different opinions ☺
- People change... : important to deal with that, too

To err is human

- Programs need to be robust
  - expect typos, syntactic, semantic, logical, translation mistakes etc.
  - help detect and correct errors
  - let users persist in errors
- Programs cannot be misled by errors
  - while generalizing
  - while keeping stock
  - while reasoning/translating
- Programs cannot be blindly compared with human performance

To judge is human

- Atitudes, opinions, states of mind, feelings
- There is no point in computers being right if this is not acknowledged by the users
- It is important to be able to compare opinions (of different people)
  - inter-anotator agreement
  - agreement by class
- Interannotator agreement is not always necessary/relevant!
  - personalized systems should disagree as much as people they personalized to ...

Measuring agreement...

- agreement with an expert coder (separately for each coder)
- pairwise agreement figures among all coders
- the proportion of pairwise agreements relative to the number of pairwise comparisons
- majority voting (expert coder by the back door): ratio of observed agreements with the majority opinion

- pairwise agreement – or agreement only if all coders agree ?
- pool of coders or one distinguished coder + many helpers

Motivation for the Kappa statistic

- need to discount the amount of agreement if they coded by chance (which is inversely proportional to the number of categories)
- when one category of a set predominates, artificially high agreement figures arise
- when using majority voting, 50% agreement is already guaranteed by the measure (only pairs off coders agains the majority)
- measures are not comparable when the number of categories is different

- need to compare K across studies

The Kappa statistic (Carletta, 1996)

- for pairwise agreement among a set of coders

K=(P(A)-P(E))/(1-P(E))

- P(A): proportion of agreement
- P(E): proportion of agreement by chance

1: total agreement 0: totally by chance

- in order to compare different studies, the units over which coding is done have to be chosen sensibly and **comparably**
- when no sensible choice of unit is available pretheoretically, simple pairwise agreement may be preferable

Per-class agreement

- Where do annotators agree (or disagree) most?

1. The proportion of pairwise agreements relative to the number of pairwise comparisons for each class
   - If all three subjects ascribe a description to the same class,
     - 3 assignments, 6 pairwise comparisons, 6 pairwise agreements: 100% agreement
   - If two subjects ascribe a description to C1 and the other subject to C2
     - two assignments, four comparisons and two agreements for C1: 50% agreement
     - one assignment, two comparisons and no agreement for C2: 0% agreement
2. Take each class and eliminate items classified as such by any coder, then see which of the classes when eliminated causes the Kappa statistic to increase most. (similar to "odd-man-out")

Measuring agreement (Craggs & Wood, 2006)

- Assessing **reliability** of a coding scheme based on agreement between annotators
- *there is frequently a lack of understanding of what the figures actually **mean***
- **Reliability**: degree to which the data generated by coders applying a scheme can be relied upon:
  - categories are not idiosyncratic
  - there is a shared understanding
- the statistic to measure reliability must be a function of the coding process, and not of the coders, data, or categories

Evaluating coding schemes (Craggs & Wood, 2006)

- the purpose of assessing the reliability of coding schemes is not to judge the performance of the small number of individuals participating in the trial, but rather to predict the performance of the scheme in general
- the solution is not to apply a test that panders to individual differences, but rather to increase the number of coders so that the influence of any individual on the final result becomes less pronounced
- if there is a single correct label, training coders may mitigate coder preference

Objectivity... House (1980:86ff)

- confusing objectivity with procedures for determining intersubjectivity
- two different senses for objectivity:
  - quantitative: objectivity is achieved through the experiences of **a number of** subjects or observers – a sampling problem (intersubjectivism)
  - qualitative: factual instead of biased
- it is possible to be quantitatively subjective (one man's opinion) but qualitatitively objective (unbiased and true)
- different individual and group biases...

Validity vs. reliability (House, 1980)

- Substitution of reliability for validity: a common error of evaluation
  - one thing is that you can **rely** on the measures a given tool gives
  - another is that those measures are **valid** to represent what you want
- *there is no virtue in a metric that is easy to calculate, if it measures the wrong thing* (Sampson & Babarczy, 2003: 379)
- Positivism-dangers
  - use highly reliable instruments the validity of which is questionable
  - believe in science as objective and independent of the values of the researchers

Example: the meaning of OK (Craggs & Wood)

Confusion matrix

Coder 1 \ Coder 2	Accept	Acknowledge	
Accept	**90**	**5**	95
Acknowledge	**5**	**0**	5
	95	5	100

- **prevalence "problem"**: when there is an unequal distribution of label use by coders, skew in the categories increases agreement by chance (Di Eugenio & Glass, 2004)
- percentage of agreement: 90%; kappa: small (0.47)
- reliable agreement? **NO!**

3 agreement measures and reliability inference

- percentage agreement – does not correct for chance
- chance-corrected agreement without assuming an equal distribution of categories between coders **Cohen's kappa**
- chance-corrected agreement assuming equal distribution of categories between coders **Krippendorff's alpha 1-D_0/D_e**

- depending on the use/purpose of that annotation...
- are we willing/unwilling to rely on imperfect data?
  - training of automatic systems
  - corpus analysis: study tendencies
- there are no magic thresholds/recipes

Krippendorff's (1980/2004) content analysis

Agreement = 1 – Observed / Expected Disagreement

%-agreement	$A_o = 1 -$	$(b + c) /$
Bennett et al. (1954)	$S = 1 -$	$(b + c) / 2 \cdot ½ \cdot ½,$
Scott (1955)	$\pi = 1 -$	$(b + c) / 2\bar{p}\,\bar{q},$
Krippendorff (1970a)	$\alpha = 1 - \frac{n-1}{n}$	$(b + c) / 2\bar{p}\,\bar{q},$
Cohen (1960)	$\kappa = 1 -$	$(b + c) / p_A q_B + p_B q_A.$

where ½ is the logical probability of 0 or 1; $\bar{p}$ and $\bar{q} = (1-\bar{p})$ are population estimates; $n = 2r$ = the total number of values used jointly by both observers; and $(n-1)/n$ corrects α for small sample sizes.

p. 248

A \ B	a	b	p_B
	c	d	q_B
	p_A	q_A	1

Reliability vs agreement (Tinsley & Weiss, 2000)

- when rating scales are an issue
- interrater reliability – indication of the extent to which the variance in the ratings is attributable to differences among the objects rated
- interrater reliability is sensitive only to the relative ordering of the rated objects
- one must decide (4 different versions)
  - whether differences in the level (mean) or scatter (variance) in the ratings of judges represent error or inconsequential differences
  - whether we want the average reliability of the individual judge or the reliability of the composite rating of the panel of judges

Example (Tinsley & Weiss)

Rater

Candidate	Z	W	Y	V	X	M
A	1	3	6	5	6	5
B	1	3	6	5	4	4
C	2	4	7	6	4	6
D	2	4	7	4	5	6
E	3	5	8	5	4	4
F	3	5	8	6	6	5
G	4	6	9	4	4	5
H	4	6	9	5	5	4
I	5	7	10	4	5	3
J	5	7	10	6	6	6
Mean	3.0	5.0	8.0	5.0	5.1	4.7
SD	1.5	1.5	1.5	.8	.9	.8

Example (Tinsley & Weiss) ctd.

Reliability: average of a single, composite

- K number of judges rating each person
- MS – mean square for
  - persons
  - judges
  - error

$R_i = (MS_p - MS_e)/(MS_p + MS_e(K-1))$

$R_c = (MS_p - MS_e)/MS_p$

Agreement:

- T_n agreement defined as n=0,1,2 points discrepancy

$T_n = (N_a - N_{pc})/(N - N_{pc})$

R_{di}	1.0	.38
R_{gi}	.02	.40
R_{dc}	1.0	.65
R_{gc}	.06	.67
r_i	.23	.94
r_c	.47	.98
T_0	0.0	0.0
T_1	0.0	.67
T_2	0.0	1.0

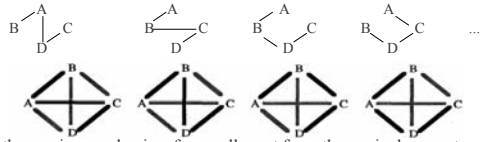
And if we know more?

OK, that may be enough for content analysis, where a pool of independent observers are classifying using mutually exclusive labels

- But what if we know about (data) dependencies in our material?
- Is it fair to consider everything either equal or disagreeing?
- If there is structure among the classes, one should take it into account
- **Semantic consistency** instead of **annotation equivalence**

Comparing the annotation of co-reference

- Vilain et al. 95 discuss a model-theoretic coreference scoring scheme
- key links: <A-B B-C B-D>; response: <A-B, C-D>



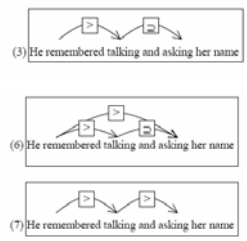
- the scoring mechanism for recall must form the equivalence sets generated by the key, and then determine, for each such key set, how many subsets the response partitions the key set into.

Vilain et al. (1995) ctd

- let S be an equivalence set generated by the key, and let $R_1 \ldots R_m$ be equivalent classes generated by the response.
- For example, say the key generates the equivalence class S= { A B C D} and the response is simply <A-B> . The relative partition p(S) is then {A B} {C} and {D} . |p(S)|=3
- c(S) is the minimal number of "correct" links necessary to generate the equivalence class S. c(S) = (|S| -1) c({A B C D})=3
- m(S) is the number of "missing" links in the response relative to the key set S. m(S) = (|p(S)| – 1) m({A B C D})=2
- **recall** = (c(S) – m(S))/ c(S) 1/3
- switching figure and ground, **precision** = (c'(S') – m'(S'))/ c'(S') (partitioning the key according to the response)

Katz & Arosio (2001) on temporal annotation

- Annotation A and B are **equivalent** if all models satisfying A satisfy B and all models satisfying B satisfy A.
- Annotation A **subsumes** annotation B iff all models satisfying B satisfy A.
- Annotations A and B are **consistent** iff there are models satisfying both A and B.
- Annotations A and B are **inconsistent** if ther are no models satisfying both A and B.
- the **distance** is the number of relation pairs that are not shared by the annotations normalized by the number that they do share



Not all annotation disagreements are equal

- Diferent weights for different mistakes/disagreements
- Compute the cost for particular disagreements
- Different fundamental opinions
- Mistakes that can be recovered, after you are made aware of them
- Fundamental indeterminacy, vagueness, polisemy, where any choice is wrong

Comparison window (lower and upper bounds)

- One has to have some idea of what are the meaningful limits for the performance of a system before measuring it
- Gale et al. (1992b) discuss word sense tagging as having a very narrow evaluation window: 75% to 96%?
- And mention that part of speech has a 90-95% window
- Such window(s) should be expanded so that evaluation can be made more precise
  - more difficult task
  - only count verbs?

Baseline and ceiling

- If a system does not go over the baseline, it is not useful
  - PoS tagger that assigns every word the tag N
  - WSD system that assigns every word its most common sense
- There is a ceiling one cannot measure over, because there is no consensus: Ceiling as human performance
- *Given that human annotators do not perform to the 100% level (measured by interannotator comparisons) NE recognition can now be said to function to human performance levels* (Cunningham, 2006)

Wrong! confusing possibility to evaluate with performance:

- Only 95% consensus implies that only 95% can be evaluated; it does **not** mean that the automatic program reached human level...

NLP vs. IR baseline's

- In NLP: The easiest possible working system
  - systems are **not** expected to perform better than people
  - NLP: systems that do human tasks
- In IR: what people can do
  - systems **do** expect to perform better than people
  - IR: systems that do inhuman tasks

Keen (1992) speaks of benchmark performances in IR: important to test approaches at high, medium and low recall situations

Paul Cohen (1995): kinds of empirical studies

- empirical = exploratory + experimental

- exploratory studies – yield causal hypotheses
- assessment studies – establish baselines and ranges
- manipulation experiments – test hypotheses by manipulating factors
- observation experiments – disclose effects by observing associations

- experiments are **confirmatory**
- **exploratory** studies are the informal prelude to experiments

Experiments

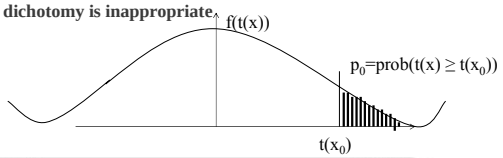
- Are often expected to have a yes/no outcome
- Are often rendered as the opposite hypothesis to reject with a particular confidence
- The opposite of order is random, so: often, the hypothesis to reject, standardly called H_0, is that some thing is due to chance alone
- There is a lot of statistical lore for hypotheses testing, which I won't cover here
  - often they make assumptions about population distributions or sampling properties that are hard to confirm or are at odds with our understanding of linguistic phenomena
  - apparently there is a lot of disagreement among language statisticians

Noreen (1989) on computer-intensive tests

Techniques with a minimum of assumptions - and easy to grasp.

Simon: "resampling methods can fill all statistical needs"

- computer-intensive methods estimate the probability $\mathbf{p_0}$ that a given result is due to chance
- there is not necessarily any particular p_0 value that would cause [the researcher] to switch to a complete disbelief, and so **the accept-reject dichotomy is inappropriate**



Testing hypotheses (Noreen, 1989)

- **Randomization** is used to test that one variable (or group) is unrelated to another (or group), shuffling the first relative to the other.
- If the variables are related, then the value of the test statistic for the original unshuffled data should be unusual relative to the values obtained after shuffling.
- exact randomization tests: all permutations; approximate rand. tests: a sample of all (assuming all are equally possible)
  1. select a test statistic **that is sensitive to the veracity of the theory**
  2. shuffle the data N times and count when it is greater than the original (nge)
  3. if (nge+1)/(NS+1) < x, reject the hypothesis (of independence)
  4. x (lim NS->∞) at confidence levels (.10, .05, .01) (see Tables)

Testing hypotheses (Noreen, 1989) contd

- **Monte Carlo Sampling** tests the hypothesis that a sample was randomly drawn from a specified population, by drawing random samples and comparing with it
- if the value of the test statistic for the real sample is unusual relative to the values for the simulated random samples, then the hypothesis that it is randomly drawn is rejected

1. define the population
2. compute the test statistic for the original sample
3. draw a simulated sample, compute the pseudostatistic
4. compute the significance level (nge+1)/(NS+1) < p0
5. reject the hypothesis that it is random if p0 < rejection level

Testing hypotheses (Noreen, 1989) contd

- **Bootstrap resampling** aims to draw a conclusion about a population based on a random sample, by drawing artificial samples (with replacement) from the sample itself.
- are primarily used to estimate the significance level of a test statistic, i.e., the probability that a random sample drawn from the hypothetical null hypothesis population would yield a value of the test statistic at least as large as for the real sample
- several bootstrap methods: the shift, the normal, etc.
- must be used in situations in which the conventional parametric sampling distribution of the test statistic is not known (e.g. median)
- unreliable and to be used with extra care...

Examples from Noreen (1989)

Hyp: citizens will be most inclined to vote in close elections

- Data: Voter turnout in the 1844 US presidential election (decision by electoral college): per U.S. state, participation (% of voters who voted); spread (diff of votes obtained by the two candidates)
- Test statistic: - correlation coefficient beween participation and spread
- Null hypothesis: all shuffling is equally likely
- Results: only in 35 of the 999 shuffles was the negative correlation higher -> the significance level (nge+1/NS+1) is 0.036
- p(exact signif. level < 0.01; 0.05; 0.10) = 0; .986; 1)

Examples from Noreen (1989)

Hyp: the higher the relative slave holdings, the more likely a county voted for secession (in 1861 US), and vice-versa

- Data: actual vote by county (secession vs. union) in three categories of relative slave holdings (high, medium, low)
- Statistic: absolute difference from total distribution (55%-45% secession-union) for high and low counties, and deviations for medium counties
- 148 of the 537 counties deviated from the expectation that distribution was independent of slave holdings
- Results: After 999 shuffles (of the 537 rows) there was no shuffle on which the test statistic was greater than the original unshuffled data

Noreen: stratified shuffling

- Control for other variables

... is appropriate when there is reason to believe that the value of the dependent variable depends on the value of a categorical variable that is not of primary interest in the hypothesis test.

- for example, study grades of transfer/non-transfer students
  - control for different grading practices of different instructors
  - shuffling only within each instructor's class
- Note that several "nuisance categorical variables" can be controlled simultaneously, like instructor and gender

Examples from Noreen (1989)

- High-fidelity speakers (set of 1,000) claimed to be 98% defect-free
- a random sample of 100 was tested and 4 were defective (4%)
- should we reject the set?
- statistic: number of defective in randomly chosen sets of 100
- by Monte Carlo sampling, we see that the probability of a set with 980 good and 20 defective provide 4 defects in a 100 sample is 0.119 (there were 4 or more defects in 118 of the 999 tested examples)

assess how significant/decisive is one random sample

Examples from Noreen (1989)

- Investment analyst's advice on the ten best stock prices
- Is the rate of return better than if it had been chosen at random?
- Test statistic: rate of return of the ten
- Out of 999 randomly formed portfolios by selecting 10 stocks listed on the NYSE, 26% are better than the analyst's

assess how random is a significant/decisive sample

NLP examples of computer intensive tests

Chinchor (1992) in MUC

- Hypothesis: systems X and Y do not differ in recall
- statistic: absolute value of difference in recall; null hypothesis: none
- approximate randomization test – per message – 9,999 shuffles
- for each 105 pairs of MUC systems...

- for the sample of (100) test messages used, ... indicates that the results of MUC-3 are statistically different enough to distinguish the performance of most of the participating systems
- caveats: some templates were repeated (same event in different messages), so the assumption of independence may be violated

From Chinchor (1992)

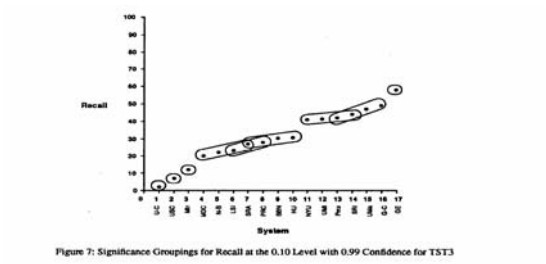


Figure 7: Significance Groupings for Recall at the 0.10 Level with 0.99 Confidence for TST3

Day 4

TREC: the Text REtrieval Conference

- Follows the Cranfield tradition
- Assumptions:
  - Relevance of documents independent of each other
  - User information need does nc change
  - All relevant documents equall desirable
- Single set of judgements representative of a user population
- Recall is knowable

- From Voorhees (2001):

Cranfield Tradition

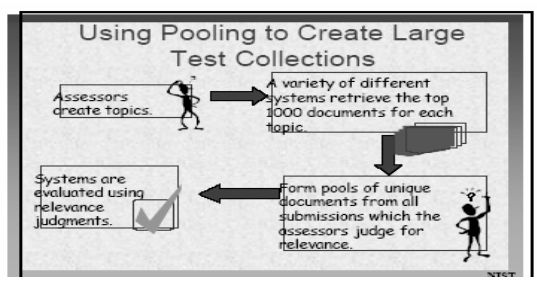
- Laboratory testing of retrieval systems first done in Cranfield II experiment (1963)
  - fixed document and query sets
  - evaluation based on relevance judgments
  - relevance abstracted to topical similarity
- Test collections
  - set of documents
  - set of questions
  - relevance judgments

NIST

Pooling in TREC

Dealing with unknowable recall

From Voorhees (2001)



History of TREC (Voorhees & Harman 2003)

- Yearly workshops following evaluations in information retrieval from 1992 on
- TREC-6 (1997) had a cross-language CLIR track (jointly funded by Swiss ETH and US NIST), later transformed into CLEF
- from 2000 on TREC started to be named with the year... so TREC 2001, ... TREC 2007
- A large number of participants world-wide (industry and academia)
- Several tracks: streamed, human, beyond text, Web, QA, domain, novelty, blogs, etc.

Use of precision and recall in IR - TREC

Precision and recall are set based measures... what about ranking?

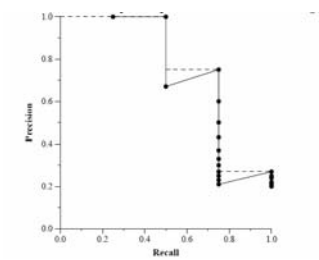
- **Interpolated precision at 11 standard recall levels**: compute precision against recall after each retrieved document, at levels 0.0, 0.1, 0.2 ... 1.0 of recall, average over all topics
- **Average precision**, not interpolated: the average of precision obtained after each relevant document is retrieved
- **Precision at X document** cutoff values (after X documents have been seen): 5, 10, 15, 20, 30, 100, 200, 500, 1000 docs
- **R-precision**: precision after R (all relevant documents) documents have been retrieved

Example of TREC measures

- Out of 20 documents, 4 are relevant to topic t. The system ranks them as 1st, 2nd, 4th and 15th.

Average precision:

- 1,1,0.75,0.266 =.754

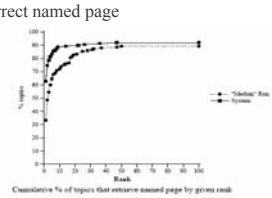


From http://trec.nist.gov/pubs/trec11/appendices/MEASURES.pdf

More examples of TREC measures

Named page: known item

- (inverse of the) rank of the first correct named page
- MRR: mean reciprocal rank



Cumulative % of topics that retrieve named page by given rank

Novelty track

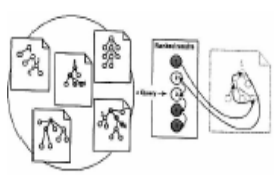
- Product of precision and recall

(because set precision and recall do not average well)

- Median graphs

INEX: when overlaps are possible

- the task of an XML IR system is to identify the most appropriate granularity XML elements to return to the user and to list these in decreasing order of relevance
- components that are most specific, while being exhaustive with respect to the topic
- probability that a comp. is relevant
- P(rel|retr)(x)= xn/(xn+esl$_{x,n}$)
  - esl: expected source length
  - x: document component
  - n total number of relevant components



From Kazai & Lalmas (2006)

The TREC QA Track: Metrics and Scoring

From Gaizauskas (2003)

- Principal metric for TREC8-10 was **Mean Reciprocal Rank (MRR)**
  - Correct answer at rank 1 scores 1
  - Correct answer at rank 2 scores 1/2
  - Correct answer at rank 3 scores 1/3
  - …

  Sum over all questions and divide by number of questions
- More formally:

$$MRR = \frac{\sum_{i=1}^{N} r_i}{N}$$

N = # questions

r_i = reciprocal of best (lowest) rank assigned by system at which a correct answer is found for question i, or 0 if no correct answer found

- Judgements made by human judges based on answer string alone (lenient evaluation) and by reference to documents (strict evaluation)

The TREC QA Track: Metrics and Scoring

- For list questions
  - each list judged as a unit
  - evaluation measure is accuracy:
    # distinct instances returned / target # instances
- The principal metric for TREC2002 was Confidence Weighted Score

$$\text{confidence weighted score} = \frac{\sum_{i=1}^{Q} \#\text{correct in first } i \text{ positions}/i}{Q}$$

where Q is number of questions

From Gaizauskas (2003)

The TREC QA Track: Metrics and Scoring

- A systems overall score will be:
  1/2*factoid-score + 1/4*list-score + 1/4*definition-score
- A factoid answer is one of: correct, non-exact, unsupported, incorrect.
  Factoid-score is % factoid answers judged correct
- List answers are treated as sets of factoid answers or "instances"
  Instance recall + precision are defined as:
  IR = # instances judged correct & distinct/|final answer set|
  IP = # instances judged correct & distinct/# instances returned
  Overall factoid score is then the F1 measure:
  F = (2*IP*IR)/(IP+IR)
- Definition answers are scored based on the number of "essential" and "acceptable" information "nuggets" they contain – see track definition for details

From Gaizauskas (2003)

Lack of agreement on the purpose of a discipline: what is QA?

- Wilks (2005:277):
- *providing ranked answers [...] is quite counterintuitive to anyone taking a common view of questions and answers. "Who composed Eugene Onegin? and the expected answer was Tchaikowsky [...] listing Gorbatchev, Glazunov etc. is no help*

- Karen Sparck-Jones (2003):
- *Who wrote "The antiquary?"*
- *The author of Waverley*
- *Walter Scott*
- *Sir Walter Scott*

- *Who is John Sulston?*
- *Former director of the Sanger Institute*
- *Nobel laureate for medicine 2002*
- *Nematode genome man*
- There are no context-independent grounds for choosing any one of these

Two views of QA

- IR: passage extraction before IE
- but "what colour is the sky?" passages with *colour* and *sky* may not have *blue* (Roberts & Gaizauskas, 2003)

- AI: deep understanding
- but "where is the Taj Mahal?" (Voorhees & Tice, 2000) how can you know what the user has in mind/knows, and therefore wants to know?

- Basically, what is difficult for one approach may be easy for the other

QA evaluation (Wilks, 2005)

- Methodological difference between AI/NLP and IR evaluation:
  AI, linguistics and IR were respectively seeking propositions, sentences and byte-strings and there is no clear commensurability between the criteria for determining the three kinds of entities
- Evaluation setup has a lot to say
  *If your text-derived answer was **A**, but wanting to submit 250 bytes of answer meant that you, inadvertently, could lenghten that answer rightwards in the text to include the form **A AND B***
  ... your answer would become wrong in the very act of conforming to format

Evaluating test collections for IR

Creating evaluation resources is time consuming

- Cormack et al. (1998) compared three strategies for creating IR collections:
  - Pooling – the standard
  - Interactive Search and Judging (ISJ)
  - Move-to-Front Judging
- Sanderson & Joho (2004) advocate no system pooling
  - Relevance feedback
  - Manual queries for ISJ (intersected with TREC judgements)
  - The set of automatic queries (intersected with TREC judgements)
- Vorhees & Buckley (2002) suggest reporting error rate for a collection (how likely the outcome of one experiment leads to the wrong conclusion)

For any technique there is a collection where it will help
(Bruce Croft)

Evaluating results of IR evaluation

- The effect of topic set size on Retrieval Experiment Error (Voorhees & Buckley, 2002)
  - How many topics and how much specific topics influence comparative results?
- Zobel (1998)
  - How reliable are the different measures used in TREC?
- Raghavan et al. (1989)
  - The effects on non-linear ordering of retrieval results
  - Precision of interpolation given a set of queries
  - Stopping criteria (fixed number of relevant documents vs. fixed number of documents retrieved)

Evaluation of "one sense per discourse"

- Yarowsky (1992, 1995)/Gale et al. (1992) published their famous OSPD
- while Krovetz & Croft (1992), unaware of that, published an extensive study on lexical ambiguity in IR
- Krovetz (1997, 2000) investigated the question further
- Wilks (2005) supports Krovetz, but thousands of researchers keep citing Yarowsky/Gale/Church as established truth...

Who is right?

The "one sense per discourse" hypothesis

- *if a polysemous word such as **sentence** appears two or more times in a well-written discourse, it is extremely likely that they all share the same sense* (Gale et al., 1992:233)
- the tendency to share sense in the same discourse is extremely strong, roughly 98%

Empirical evidence:

- a random sample of 108 nouns, reading the articles in **Grolier's encyclopedia**, was judged by 3 judges (only 6 in 300 had > one sense)
- 102 of 106 pairs of 5 words in one sense only **in the Brown corpus**

Why bother?

- to improve the performance of a WSD algorithm
- aid in collecting annotated test material for evaluating WSD
  - tag all instances in one fell swoop
- One sense per collocation:
- *if we actually build a disambiguation procedure using exclusively the content word to the right as information, such a system performs with 97% precision on new data where a content word appears to the right and for which there is information in the model* (Yarowsky 93: 269)
- binary senses of words *plant space tank motion bass palm poach axes duty drug sake crane*

Value of "one sense per discourse"

- Yarowsky (1995) applies the OSPD assumption to his unsupervised WSD method
- in average (12 words): 94.8% accuracy
- plus OSPD: 96.1 or 96.5 [depending on where in the algorithm]
- *yielding improvements of **27% reduction in error rate***

(96.5-94.8=1.7, error rate=100-94.8=5.2, 1.7/5.2=.32=32% reduction)
(96.1-94.8=1.3, error rate=100-94.8=5.2, 1.3/5.2=.25=25% reduction)

Some technical wake-up calls

Keen (1992) in "Presenting results of experimental retrieval comparisons", on the use of improvement percentages:

- is liable to mislead
  - to process ratio results that are already interpretable as percentages by the percentage improvement – a sort of double percentage
- an unhelpful metric
  - a difference on a poorly performing collection shows much greater improvement values than the same difference on a better performing collection
- at odds with statistical significance testing
- differences on the **mean results** vs. differences in **improvements**

Keen's example in detail

- system A: 70.6% precision
- system B: 45.4% precision

- Difference: 25.2%
- Percentage improvement: 55.5%
  - uses a different base figure: 25.2%/45.4%=55.5 % (?)

Krovetz & Croft (1992) study

- analyse 64 queries for a collection of 3204 computer science abstracts
  - 300 query word types corresponding to 35,000 tokens
- analyse 83 queries, in a collection of 423 short articles from the *Time* magazine (still, texts were in average 6 times longer)

- word senses taken from LDOCE
- mean number of senses for the collection and for the queries
  - CACM: 4.7/6.8 – *Time*: 3.7/8.2
- *What is surprising is the large number of words that are of high ambiguity and low frequency* (p. 127)
- In the *Time* collection, 6.5% of the matches has more than one sense

Krovetz (1997): More than one sense per discourse

- Two sense-tagged corpora (based on WordNet)
  - SemCor (all open class words of a Brown corpus subset)
  - DSO (191 highly ambiguous words, 121 nouns, 70 verbs) (Brown+WSJ)
  - "Discourse" is 2000 words in Brown, or article in WSJ
- SemCor: 41% of 47% for potentially ambiguous nouns, 50% of 66% for pot. amb. verbs and 18% of 63% pot. amb. adjectives
- DSO: all appeared in more than one sense. 39% of the files had more than one sense of the **same** word
- Explanation: homonymy vs polysemy

H: easy (Yarowsky), P: complex, real life (Krovetz)

More on homonymy vs. polysemy

- Buitelaar (1998) notes that in WordNet < 5.2% of noun types have non-related senses, while 94.8% have more than one related sense
- Santos (1996) reports on 1,609 categorially (POS) ambiguous types in Portuguese (between different open class N/A, Adv, V, Vpp)
  - common origin: 12%
  - homonymy: 20%
  - derivation-related: 21%
  - other: 47% (more difficult to formalize the relationship; ambiguous intra-categorially, > 2 with different classification outcomes)
- Krovetz: Operational way to distinguish H from P: systematically inconsistent judgements across judges -> polysemy (related senses)

Measure the difficulty of the problem

- One important and often neglected preoccupation!
- Should be done beforehand...

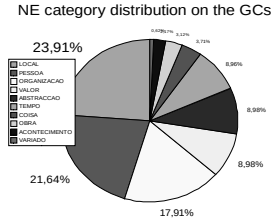
While building the golden collection in HAREM (Santos et al., 2006)

- How many cases are easy?
  - enough to be in the gazetteer
  - that string has only one sense (gets always the same label)
- How many cases require disambiguation?
  - among categories (person or local, etc.)
  - among NEs or non NEs (proper-common noun disambiguation)
- How many cases are vague?
- How many cases are creative? (difficult)

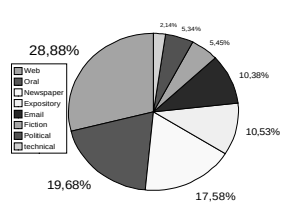
The golden collection of HAREM (PT)

- 257 (127) documents, with 155,291 (68,386) words and 9,128 (4,101) manually annotated NEs
- Places: 2157 (980) tokens, 979 (462) types

NE category distribution on the GCs



Text Genre Distribution on the GCs



From Santos & Cardoso (2006)

Digression: what is hard?

- Hard questions in QA (Voorhees & Tice)
- Automatic evaluation of QA (Breck et al., 2000)
- **WiQA** (Jijkoun & de Rijke, 2007): support, importance, novelty, non-repetition: the task of an automatic system participating in WiQA 2006 is to locate information snippets in Wikipedia which are:
  - outside the given source article,
  - in one of the specified target languages,
  - substantially new w.r.t. the information contained

  in the source article, and important for the topic of the source article, in other words, worth including in the content of (the future editions of) the article. One specific application of the task defined in this way can be a system that helps a Wikipedia editor to update or expand an article using the information available elsewhere.

Brewster et al: Extract explicit knowledge?

Extract ontologies from text... knowledge discovery from static repositories?

- *No matter how large our corpus, if it is domain specific, the major part of the domain ontology will not be specified because it is taken as given, or assumed to be part of the* ***background knowledge*** *the reader brings to the text.* (Brewster et al., 2003)
- *A text is an act of* ***knowledge maintenance.*** (...) *A primary purpose of a text at some level is to* ***change*** *the relationship between existing concepts, or* ***change*** *the instantiations of these concepts [...] or* ***adding*** *new concepts to the existing domain ontology* (Brewster et al., 2005)

Day 5

The world of machine translation

- A long history, with the ALPAC report (1966)
- The most difficult application...
  - two different languages/worlds
  - two different evaluation criteria: fidelity and intelligibility
- ARPA HLT evaluation: (fully automatic) MT in 1993
- FEMTI: framework for the evaluation of MT (ISLE project)
- BLEU

White & O'Connell (1994) on ARPA HLT

- 8 system plus novice translators were evaluated on three criteria
  - Fluency
  - Adequacy
  - Comprehension (Afterwards: did they understood what was conveyed?)
- 22 French, Japanese or Spanish texts translated into English
- Reference translations produced by professional translators; 11 native speakers of English as evaluators
- Fluency: 5 points scale; standardized comprehension test
- Adequacy: linguistic components, average 11-12 words, scale of 1 to 5 whether the meaning was absent or present in the translation, comparing with the reference translation

Hovy et al. (2002): FEMTI

- Quality model: 6 characteristics (ISO/IEC) are functionality, reliability, usability, efficiency, maintainability, portability
- External and internal quality (related to the users' needs vs. software)
- Quality in use: effectiveness, productivity, safety, satisfaction
- *Which parameters to pay attention to, and how much weight to assign each one, remains the prerrogative of the evaluator*
- Suggestion: User profiles determine the weighting of partial scores
- Fluency:
  - readability: sentences read naturally
  - comprehensibility: text is easy to understand
  - coherence: possible to grasp and understand the structure
  - cohesion: text-internal links such as lexical chains are maintained in the translation

BLEU (Bilingual Evaluation Understudy) (Papineni et al, 2001)

- using *n*-gram similarity of a candidate to a set of reference translations (sentence based)
- modified precision of a candidate translation:
  number of clipped words (n-grams) that occur in any reference transl. / number of total words (n-grams) in the candidate
  sum of clipped n-grams in all sentences / sum of candidate n-grams
- *word-weighted average of sentence-level modified precisions, rather than a sentence-weighted average*
- combination of the modified precisions of 1 to 4 grams
- sentence-brevity penalty

Example from Papinemi et al

Candidate 1: obeys .

It is a guide to action that ensures that the military will forever heed Party commands.

Candidate 2: to insure the troops hearing the activity guidebook direct.

It is the guiding principle which guarantees the military forces always being under the command of the Party.

It is the practical guide for the army always to heed the directions of the party.

P1=17/18
P2=5/18

BLEU formulas

$$\text{BLEU} = \text{BP} \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right) \quad \text{BP} = \begin{cases} 1 & \text{if } c > r \\ e^{(1-r/c)} & \text{if } c \leq r \end{cases}$$

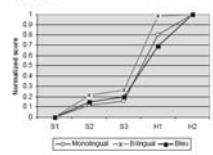
$$\log \text{BLEU} = \min(1 - \frac{r}{c}, 0) + \sum_{n=1}^{N} w_n \log p_n$$

- c, r – length of the candidate or reference translations
- As a baseline, Papineni et al suggest:
  - wn – uniform weights: 1/N
  - N = 4
- Note that the matches are position independent.

More on BLEU

- Proposed for use in the R&D cycle of machine translation technology
- The more reference translations, the higher the precision
- Even a human translator will hardly score 1 (except if s/he produces a translation equal to one of the reference translations)
- experiments to judge 5 "systems":
  - 250 Chinese-English sentence pairs
  = 5 translations of 50 sentences
  - rated by two groups of human judges
  - from 1 (very bad) to 5 (very good)
  - 10 bilinguals and 10 monolinguals
  - linearly normalized by the range

Figure 7: BLEU vs Bilingual and Monolingual Judgments



Technical digression: the 3 steps of evaluation

1. Qualitative evaluation of individual events/pieces
   - unitizing
2. Transformation into a quantitative score
   - number magic?
   - questions of scales
3. Aggregation of that score or a multiplicity of scores
   - many events
   - many factors
   - unitizing again

Human vs. automatic evaluation

- Not really a choice!
- Human values are always there
- BLEU was tested/suggested with 4 reference translations
- Computers help apply evaluations that rely on human values / data / performance
- Automatic evaluations are reliable (can be repeated with the same or consitent outcome) but not necessarily valid

- One of the most important advantages of automatic evaluations is that they rely on what humans do well – not ask humans to do weird things

Human similarity and human acceptance

- *Most approaches to automatic MT evaluation implicitly assume that both criteria should lead to the same results, but this assumption has not been proved empirically or even discussed* (Amigó et al., 2006)
- And: *Human Likeness implies Human Acceptability but the reverse is not true.*
- The authors argue that instead of trying to model human acceptability (which accepts some never-by-humans translations) one should try to model human likeness (which implies human acceptance)

Four works on "word alignment"

- Melamed
- Véronis
- Moore
- Simard

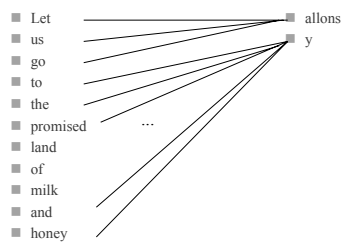
Word alignment (Melamed, 1997)

- A fast method for inducing accurate translation lexicons

! assuming that the words are translated one to one

- *A word-to-word model of translational equivalence can be evaluated either over types or over tokens*
- Link types
  - Recall: fraction of the bitext vocabulary represented in the model
  - Precision: manual evaluation of the links (if they ever occurred in some context, correct). Incomplete may count as correct or not (two values)
- Link tokens
  - 51 sentences observed manually, type of error described
  - → Manual construction of a golden resource

BLinker: Word alignment (tokens)



Alternative

go – allons

Melamed (1998a,b) BLinker's project

Véronis 1998: translation spotting

- *Translation spotting can be seen as a simpler sub-problem of full [word] alignment. Given a particular word or expression in the source text, it consists in detecting its translation in the target text.*
- Simplified problem because it chooses content words (no annotation of pronouns or auxiliaries)
- Comments may be included, plus classification of the following:
  - different coordination level (non-parallel conjuncts),
  - not translated (ommisssion)
  - translated by a referring expression (pronoun...)
  - translated with a spelling error
  - non-parallel conjuncts
  - divergent translations (the translator has completely rephrased the fragment)

Word alignment (tokens): translation spotting of sequences

- Simard (2003): The term *translation spotting* (TS) refers to the task of identifying the target-language (TL) words that correspond to a given set of source-language (SL) words in a pair of text segments known to be mutual translations.

		Sentence Pair	
	Query:	SL (English)	TL (French)
1.	*and a growing gap*	Is this our model of the future, regional disparity ***and a growing gap*** between rich and poor?	Est ce là le modèle que nous visons, soit la disparité régionale **et un fossé de plus en plus large** entre les riches et les pauvres?
2.	*the government's commitment*	***The government's commitment*** was laid out in the 1994 white paper.	**Le gouvernement** a exposé **ses engagements** dans le livre blanc de 1994.
3.	*close to* [...] *years*	I have been fortunate to have been travelling for ***close to*** 40 ***years***.	J'ai eu la chance de voyager pendant **près de** 40 **ans**.
4.	*to the extent that*	***To the extent that*** the Canadian government could be open, it has been so.	Le gouvernement canadien a été aussi ouvert qu'il le pouvait.

Ø

Figure 1: Translation spotting examples

Example of TrSpotSeq (fictive)

- These companies **indicated their support** for the government 's decision. -> Ces compagnies ont déclaré qu' elles appuyaient la décision du gouvernement .
- These companies **indicated their support** for the government 's decision. -> La déclaration de cettes compagnies de qu' elles appuyaient la décision du gouvernement .
- These companies **indicated their support** for the government 's decision. -> Ces compagnies ont démontré un appui déclaré à les mesures décidées hier par le gouvernement .
- These companies **indicated their support** for the government 's decision -> l'appui que ces compagnies ont hier déclaré au décisions de notre gouvernement c'est la preuve...

Simard's evaluation of TrSpotSeq

- All sequences of chunks from [a Hansard] text that contained three or more word tokens were then looked up in the Hansard TM [the SL queries]. Among the sequences that did match sentences in the TM
  - 100 were selected at random. (avg: 41 translations per sequence)
  - and were manually annotated according to Véronis (1998) guidelines
  - reference corpus: 4,100 pairs of sets of words and their translation
- evaluation of automatic TS was then done with exactness, precision, recall, and F-measure
- averaged over all pairs of the test corpus (*and not over SL queries, which means that more "productive" queries weigh more heavily in the reported results*)
- empty Ø -> single "null word"

Word alignment types (Melamed)

Given a set of bilingual texts, find reasonable translation candidates from bilingual evidence: automatic creation of bilingual dictionaries

Results from post-hoc human analysis of 500 entries (if in the dictionary - > correct, otherwise classified manually according to the context)

- dictionary-like (89%)
- change in part of speech (*protection – protégée* from *have protection – être protégée*) (3.4%)
- part of 1-M translations (*immédiatement – right* from *right away* or *right now*) (7.6%)

(due to Melamed's heuristic of 1-1 translations only)

Word alignment types (Moore)

- out of context, evaluated by judges as `correct`, `incorrect` or `not sure`
- type coverage is the proportion of distinct lexical types in the entire training corpus (including both languages) for which there is at least a translation given
- token coverage is the proportion of the total number of occurrences of items in the text represented by the types included within the type coverage

TypeC	TokenC	**Accuracy:** Single-word, MW, Compound
0.040	0.859	**0.902** 0.927 0.900 0.615
0.632	0.989	**0.550** 0.784 0.429 0.232

Word types, Melamed vs. Moore

- Porter stemmer
- skips at most over one or two function words
- single words only
- threshold to compute association scores:
  - discard L(u,v) < 1
- lemma after deep syntactic analysis
- unbounded distances in a sentence
- multiword relationships
- single tokens for compounds
  - ouvrir_session/log_on
  - annuler/roll_back
  - mot_de_passe/password
- multiwords independently defined for each language (lang_dep parsers)
- threshold to compute association scores: pairs that co-occur at least once and each word has frequency > 1 in each language
- compound lemmas sentence dependent

The moral is:

- The "same" task may be radically different
- Even if the final way of presenting results is the same, and the task is presented as the same, comparing outcomes may be misleading
- Exercise your judgement on all assumptions and choices!

Evaluation of tools (Sazedj & Pinto, 2006)

- Propose an evaluation framework for ontology-based annotation tools
- for annotation of Web pages
- 5 tools selected

- Methodology:
  - a set of well-defined criteria
  - with (relative) metrics associated to each criterion
  - distinction between domain-independent vs. domain-specific
  - feature-based (f), set-dependent (d) and set-independent criteria (set-dependent criteria have metrics that depend on the particular set of tools chosen, e.g. interoperability)

Evaluation of tools (Sazedj & Pinto, 2006)

- Start by defining a set of criteria

Interface	*Metadata*	*Procedure*	*General*
Self-documentation	Association (f)	Expressiveness (f)	Documentation
Simplicity	Flexibility (f)	Input (f)	Scalability (f)
Timeliness	Heterogeneity (f)	Output (f)	Stability
Usability (d)	Integrity (f)	Precision*	
	Interoperability (d)	Recall*	
	Scope (f)	Reliability*	
		Speed*	

Table 1. All criteria, classified into four dimensions.

- Explain how to quantify them: **Scope**: 1 point per feature:
  (a) annotate the minimum unit of information
  (b) annotate any multiple of that minimum unit
  (c) annotate the resource as a whole

Evaluation of tools (Sazedj & Pinto, 2006) ctd

- Create special corpora and ontologies to evaluate the specific annotation tools
- Annotate them manually (studying inter-annotator agreement)
- Compare the annotations produced with each tool

... taking into consideration differences in "interpretation"

- does date include numbers?
- should anaphorically-related entities also count for evaluation?

- How to compare semi-automatic and automatic annotation tools?

John M. Ellis (1993): what is basic?

Generalized methodological fallacy:

- start with the simple and then try to accomodate the complex
  - scientific unambiguous terms such as *triangle* instead of *good*
  - factual discourse instead of evaluative
  - truth judgments instead of moral judgements
- one should start with the complex, and provide theories that deal with it. Then "simple cases" can be looked at

- Also a beautiful debate of the "container metaphor", the "communicative purpose" and the "similarity as basic" misconceptions of language!

Messages I want(ed) to convey

- Evaluation at several levels
- Be careful to understand what is more important, and what it is all about.
  - Names of disciplines, or subareas are often tricky
- Take a closer look at the relationship between people and machines
- Help appreciate the many subtle choices and decisions involved in any practical evaluation task
  - one has to **design** an evaluation

Subjects not covered

- numerical error
- traditional statistical analysis
- evaluation in corpus linguistics
- new/exciting evaluation contests
  - ACE
  - CLEANEVAL
  - SemEval

Comments and feedback most welcome!

- Please send **Diana.Santos AT sintef.no**
  - doubts
  - suggestions
  - critical remarks
  - all kinds of assessments
- because I intend to create a feedback-improved version after the course, where I will try to incorporate (and thank) everything you send me
- Thank you for attending this course!

Possible extras

Sparck Jones & Galliers (1996/1993) contd

- 3 classes of criteria
  - effectiveness
  - efficiency
  - acceptability
- EAGLES: 3 kinds of evaluation, and the consumer report paradigm (checklist formation and use)
  - progress
  - adequacy
  - diagnostic
- Church & Hovy (1993): task dependent evaluation criteria, for MT

Sparck Jones & Galliers (1996/1993) contd.

- the agreed answers are honed, consensual ones and are to that extent 'unnatural' (p. 175)
- comparisons with purely human means of getting the info
- the particular danger which arises with diagnostic evaluation is that the attempt to attribute responsibility for performance by even sharper decomposition and focussing can lead to artificiality and distortion (p. 137)
- evaluations have to be designed for the individual case (p. 194)
- the scope of the task evaluation is set by the smallest data component, in both ATIS and TREC cases the query set sizes (p. 175)

Ablation procedures

- Destroy/add randomly accents
- Add distracting text to a collection of texts to summarize
- Add noise to speech signal
- Increase the collection with difficult texts including "false friends" or ambiguous words
- Use OCR'ed material

Adding noise as ablation

- Quinlan (1986) describes ID3, a method for induction of decision trees with an information-theory based evaluation function (to choose which attribute to select at each step)

  maximize gain (A) = I(p,n)-E(A)

  minimize $E(A)=\Sigma(pi+ni)/(p+n)I(pi,ni)$
- Adding noise:

1. Artificially corrupt the training set
   - to class information
   - to values of attributes
2. Add unknown attribute values

Quinlan (contd.) on analysis of decision trees

- *do not test attributes whose irrelevance cannot be rejected with a very high (99%) confidence level* (p. 93)
- *for higher noise levels, the performance of the correct DT on corrupted data was found to be inferior to that of an imperfect DT formed from data corrupted to a similar level* (p.96)
- in DT creation, so that unknown values can only decrease the information gain of an attribute, distribute the unknown values among the possible known values (according to these latter distribution)
- in DT application, create as many paths as possible when the value is unknown, and choose the class with higher value
- Test several of these measures and criteria through simulation

Quinlan (contd.) on analysis of decision trees

- but: the gain criterion tends to favor attributes with many values!
- possible solutions:

1. all tests have only two outcomes
   - but large increase in computation
   - but difficult to read by experts
2. add in information value of attribute, maximize gain (A) / IV (A) for those attributes with an average-or-better gain

IV (A) = - $\Sigma(pi+ni)/(p+n)\log_2(pi+ni)/(p+n)$

Information theoretic evaluation (Pearl 1979)

Entropic measures for decision making: are they appropriate?

- Decision problem: T tests or information sources, with C(t) costs associated, Z states with P(z) likelihood that ztZ will occur next, A actions, and a payoff matrix U (utility or benefit of joint a and z)
- Goal: design a plan of sequential testing followed by a terminal action that maximizes payoff minus the cost of testing
- measure of the uncertainty content of various entities (signals, probabilities, etc.)

OR

- measure of the effort necessary for removing uncertainty (given a uniform cost test space)

Garvey et al. (1981)

- Integrating knowledge from disparate sources
- how to effectively combine (sometimes contradictory) information from multiple knowledge sources to compensate for their individual deficiencies
- specifying that "nothing is known" is different from P=0.5!

Geographical IR

- Assigning geographical scopes to Web pages
- Extracting geographical information from Web pages
- Assigning geographical topicality to Web documents

- Different tasks,
- similar evaluations?

References

Amigó, Enrique, J. Giménez, Julio Gonzalo & Lluis Màrquez. "MT Evaluation: Human-like vs. Human Acceptable". In *Proceedings of the COLING/ACL 2006* (Sydney, Australia, July 2006), pp. 17-24.

Automatic Language Processing Advisory Committee. "Language and machines: computers in translation and linguistics", Division of behavioral sciences, National Research Council, National Academy of Sciences, Washington, 1966. (= the ALPAC Report)

Babych, Bogdan, Anthony Hartley & Erik Atwell. "Statistical modelling of MT output corpora for Information Extraction". In Dawn Archer, Paul Rayson, Andrew Wilson and Tony McEnery (eds.), *Proceedings of the Corpus Linguistics 2003 conference*. Lancaster University (UK), 28 - 31 March 2003, pp. 62-70.

Biber, Douglas. *Variation across speech and writing*. Cambridge University Press, 1988.

Black, E., S. Abney, D. Flickinger, C. Gdaniek, R. Grishman, P. Harrison, D. Hindle, R. Ingria, F. Jelinek, J. Klavans, M. Liberman, M. Marcus, S. Roukos, B. Santorini & T. Strzalkowski. "A procedure for quantitatively comparing the syntactic coverage of English grammars". In *Proceedings of the February 1991 DARPA Speech and Natural Language Workshop* (Pacific Grove, CA, February 1991), pp. 306-311.

Breck, Eric J., John D. Burger, Lisa Ferro, Lynette Hirschman, David House, Marc Light & Inderjeet Mani. "How to evaluate your question answering system every day … and still get real work done". In Maria Gavriladou, George Carayannis, Stella Markantonatou, Stelios Piperidis and Gregory Stainhaouer (eds.), *Proceedings of the Second International Conference on Language Resources and Evaluation, LREC2000* (Athens, 31 May-2 June 2000), pp. 1495-1500.

Brewster, Christopher, Fabio Ciravegna & Yorick Wilks. "Background and Foreground Knowledge in Dynamic Ontology Construction: Viewing Text as Knowledge Maintenance", in Y. Ding, K. van Rijsbergen, I. Ounis & J. Jose (eds.), *Semantic Web, Workshop held the 26th Annual International ACM SIGIR Conference* (Toronto, Canada, July 28-August 1, 2003), no pp.

Brewster, Christopher, José Iria, Fabio Ciravegna & Yorick Wilks. "The Ontology: Chimaera or Pegasus?". In Nicholas Kushmerick, Fabio Ciravegna, AnHai Doan, Craig Knoblock and Steffen Staab (eds.), *Proceedings of the Dagstuhl seminar on Machine Learning for the Semantic Web* (Wadern, Germany, 13-18 February 2005).

Buitelaar, Paul. "CoreLex: An Ontology of Systematic Polysemous Classes". In *Formal Ontology in Information Systems. Proceedings of FOIS'98* (Trento, Italy, June 6-8, 1998). IOS Press, Amsterdam, 1998.

Cabral, Luís Miguel. "SUPeRB - **S**istema **U**niformizado de **Pe**squisa de **R**eferências **B**ibliográficas". MSc dissertation, Faculdade de Engenharia da Universidade do Porto, March 2007.

Carletta, Jean. "Assessing Agreement on Classification Tasks: The Kappa Statistic", *Computational Linguistics* **22** (1996), pp. 249-54.

Chinchor, Nancy. "The statistical significance of the MUC-4 results", *Proceedings of the Fourth Message Understanding Conference (MUC-4)*, Morgan Kaufmann Publ. San Mateo, CA, 1992, pp. 30-50.

Chinchor, Nancy. "MUC-7 Named Entity Task Definition. 1997. http://www.itl.nist.gov/iad/894.02/related_projects/muc/proceedings/ne_task.html

Chinchor, Nancy, Lynette Hirschmann & David D. Lewis. "Evaluating message understanding systems: an analysis of the Third Message Understanding Conference (MUC-3)", *Computational Linguistics* **19**, 1993, pp. 409-49.

Church, Kenneth W. & Eduard H. Hovy. "Good Applications for Crummy Machine Translation". *Machine Translation* **8**, 1993, pp. 239-258.

Cohen, Paul R. *Empirical Methods for Artificial Intelligence*. The MIT Press, 1995.

Cormack, Gordon V., Christopher R. Palmer & Charles L.A. Clarke. "Efficient construction of large test collections". In *Proceedings of the 21st Annual international ACM SIGIR Conference on Research and Development in information Retrieval* (Melbourne, Australia, August 24 - 28, 1998). SIGIR '98. ACM Press, New York, NY, pp. 282-289.

Craggs, Richard & Mary McGee Wood. "Evaluating Discourse and Dialogue Coding Schemes", *Computational Linguistics* **31**, No. 3, September 2005, pp. 289-296.

Cunningham, Hamish. "Information Extraction, Automatic". *Encyclopedia of Language and Linguistics*, 2nd Edition, Elsevier, 2006, vol. 5, pp. 665-677.

Di Eugenio, Barbara & Michael Glass. " The Kappa Statistic: A Second Look", *Computational Linguistics* **20** no. 1, 2004, pp. 95-101.

EAGLES. "Evaluation of Natural Language Processing Systems: FINAL REPORT". EAGLES document EAG-EWG-PR.2, Version of September 1995, http://www.issco.unige.ch/ewg95/ewg95.html.

Ellis, John M. *Language, Thought and Logi*c. Evanston IL: Northwestern University Press, 1993.

Evert, Stefan & Brigitte Krenn. "Methods for the qualitative evaluation of Lexical Association Measures". In *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics* (Toulouse, 9-11 July 2001), 2001, pp. 188-195.

Evert, Stefan & Brigitte Krenn. "Computational approaches to collocations". Introductory course at the *European Summer School on Logic, Language, and Information (ESSLLI 2003)*, Vienna.

Gaizauskas, Robert. "Evaluating Language Processing Applications and Components". Invited lecture at *PROPOR'2003*, Faro, 27 June 2003.

Gale, William A., Kenneth W. Church & David Yarowsky. "One sense per Discourse", *Proceedings of 4th DARPA Workshop on Speech and Natural Language*, 1992, pp. 233-7.

Gale, William A., Kenneth W. Church & David Yarowsky. "Estimating Upper and Lower Bounds on the Performance of Word-Sense Disambiguation Programs". In *Proceedings, 30th Annual Meeting of the Association for Computational Linguistics* (Columbus, OH, 1992), pp. 249-256.

Garvey, Thomas D., John D. Lowrance & Martin A. Fischler. "An inference technique for integrating knowledge from disparate sources". *Proceedings of the Seventh IJCAI*, Vancouver, B.C., Canada: Morgan Kaufmann, 1981, pp. 319-25.

Grishman, Ralph & Beth Sundheim. "Message Understanding Conference - 6: A Brief History". *Proceedings of the 16th International Conference on Computational Linguistics (COLING-96)*, Copenhagen, 1996, pp. 466-71.

Halliday, M.A.K. *Computational and Quantitative Studies*, vol 7 in the Collected Works of MAK Halliday, edited by Jonathan J. Webster. London & New York: Continuum, 2005.

Hirschman, Lynette. "Comparing MUCK-II and MUC-3: assessing the difficulty of different tasks". In *Proceedings of the 3rd Conference on Message Understanding* (San Diego, California, May 21 - 23, 1991). Association for Computational Linguistics, Morristown, NJ, 1991, pp. 25-30.

Hirschman, Lynette. "The evolution of Evaluation: Lessons from the Message Understanding Conferences". *Computer Speech and Language* **12** (4), 1998, pp. 281-305.

House, Ernest R. *Evaluating with validity*. Sage Publications, Beverly Hills, CA, 1980.

Hovy, Eduard, Margaret King & Andrei Popescu-Belis. "Principles of Context-based Machine Translation Evaluation". *Machine Translation* **17**, no. 1, 2002, pp. 43-75.

Jarke, M., J.A. Turner, E.A. Stohr, Y. Vassiliou, N.H. White & K. Michielson. "A field evaluation of natural language for data retrieval", *IEEE Transactions on Software Engineering* **11**, no. 1, 1985, pp. 97-113.

Jijkoun, Valentin & Maarten de Rijke. "WiQA: Evaluating Multi-lingual Focused Access to Wikipedia". In *The First International Workshop on Evaluating Information Access (EVIA)* (May 15, 2007, Tokyo, Japan), 2007, no pp.

Katz, Graham & Fabrizio Arosio. "The Annotation of Temporal Information in Natural Language Sentences", *Proceedings of the Worskhop for Temporal and Spatial Information Processing* (Toulouse, July 7th 2001), EACL-ACL 2001, Toulouse, pp. 104-111.

Kazai, Gabriella & Mounia Lalmas. "eXtended Cumulated Gain Measures for the Evaluation of Content-Oriented XML Retrieval". *ACM Transactions on Information Systems* **24,** no 4, October 2006, pp. 503-542.

Keen, E. Michael. "Presenting results of experimental retrieval comparisons". *Information Processing and Management* **28** no. 4, 1992, pp. 491-502.

Kehler, Andrew, Douglas Appelt & John Bear. "The Need for Accurate Alignment in Natural Language System Evaluation". *Computational Linguistics* **27**, no. 2, June 2001, pp. 231-48.

Krippendorff, Klaus. *Content Analysis: an introduction to its Methodology*. Sage Publications, 2nd edition, 2004. First edition: 1980.

Krovetz, Robert. "Homonymy and Polysemy in Information Retrieval". In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and 8th Conference of the European Chapter of the Association for Computational*

Linguistics ACL 1997 (Madrid, Spain, 7-12 July 1997). Morgan Kaufmann Publishers, 1997, pp. 72-79.

Krovetz, Robert. "More than One Sense Per Discourse". 2000 http://www.itri.brighton.ac.uk/events/senseval/ARCHIVE/PROCEEDINGS/krov2.ps

Krovetz, Robert & W. Bruce Croft. "Lexical ambiguity and information retrieval". *ACM Trans. Inf. Syst.* **10** no. 2, April 1992, pp. 115-141.

Levenshtein, Vladimir. "Binary codes capable of correcting deletions, insertions, and reversals". *Soviet Physics Doklady* **10**, 1966, pp. 707–710.

Melamed, I. Dan. "Automatic Construction of Clean Broad-Coverage Translation Lexicons", *Proceedings of the 2nd Conference of the Association for Machine Translation in the Americas (AMTA)*, Montreal, PQ, 1996.

Melamed, I. Dan. "A word-to-word model of translational equivalence". In *Proceedings of the 35th Annual Meeting on Association For Computational Linguistics* (Madrid, Spain, July 07 - 12, 1997). Association for Computational Linguistics, Morristown, NJ, pp. 490-497. Corrected version from http://cs.nyu.edu/~melamed/ftp/papers/transmod.ps.gz

Melamed, I. Dan. "Annotation Style Guide for the Blinker Project". Version 1.0.4. IRCS Technical Report #98-06, Dept. of Computer and Information Science, University of Pennsylvania, February 6, 1998.

Melamed, I. Dan. "Manual Annotation of Translational Equivalence: The Blinker Project". IRCS Technical Report #98-07, Dept. of Computer and Information Science, University of Pennsylvania, 1998.

Melamed, I. Dan. “Models of Translational Equivalence among Words”. *Computational Linguistics* **26** no. 2, 2000, pp. 221-249.

Mitkov, Ruslan. "Towards a more consistent and comprehensive evaluation of anaphora resolution algorithms and systems". Keynote speech, *Proceedings of the Discourse Anaphora and Anaphora Resolution Colloquium (DAARC-2000)*, Lancaster, UK, pp. 96-107.

Moore, Robert C. "Towards a Simple and Accurate Statistical Approach to Learning Translation Relationships among Words". In *Proceedings of Workshop on Data-driven Machine Translation, 39th Annual Meeting and 10th Conference of the European Chapter of the Association for Computational Linguistics* (Toulouse, France, 2001), pp. 79-86.

Noreen, Eric W. *Computer Intensive Methods for Testing Hypotheses: An Introduction*. John Wiley and Sons, 1989.

Papineni, Kishore, Salim Roukos, Todd Ward & Wei-Jing Zhuw. “BLEU: a Method for Automatic Evaluation of Machine Translation”. Research Report, Computer Science IBM Research Division, T.J.Watson Research Center, RC22176 (W0109-022), 17 September 2001, http://domino.watson.ibm.com/library/CyberDig.nsf/Home (keyword=RC22176).

Pearl, Judea. "Entropy, information, and rational decisions". *Policy Analysis and Information Systems*, Special Issue on "Mathematical Foundations", **3** no. 1, July 1979, pp. 93-109.

Pedersen, Ted. "Fishing for Exactness". In *Proceedings of the South - Central SAS Users Group Conference (SCSUG-96)* (Austin, TX, Oct 27-29, 1996).

Quinlan, J.R "Induction of decision trees". *Machine Learning* **1**, 1986, pp. 81-106.

Raghavan, Vijay V., Petter Bollmann & Gwang S. Jung. "Retrieval system evaluation using recall and precision: problems and answers". In N. J. Belkin andfC. J. van Rijsbergen (eds.), *Proceedings of the 12th Annual international ACM SIGIR Conference on Research and Development in information Retrieval, SIGIR '89* (Cambridge, Massachusetts, United States, June 25 - 28, 1989). ACM Press, New York, NY, pp. 59-68.

Roberts, Ian & Robert Gaizauskas. "Evaluating Passage Retrieval Approaches for Question Answering". Research Memorandum CS-03-06, University of Sheffield, 2003. ftp://ftp.dcs.shef.ac.uk/home/robertg/papers/dcs-tr-03-06.pdf

Sampson, Geoffrey & Anna Babarczy. "A test of the leaf–ancestor metric for parse accuracy". *Journal of Natural Language Engineering* **9**, 2003, pp. 365–80.

Sanderson, Mark & Hideo Joho. "Forming test collections with no system pooling". In *Proceedings of the 27th Annual international ACM SIGIR Conference on Research and Development in Information Retrieval SIGIR '04* (Sheffield, United Kingdom, July 25 - 29, 2004). ACM Press, New York, NY, pp. 33-40.

Sang, Erik F. Tjong Kim. "Introduction to the CoNLL-2002 Shared Task: Language-Independent Named Entity Recognition". In *Proceedings of CoNLL-2002* (Taipei, Taiwan, 2002), pp. 155-158.

Santos, Diana. "Português Computacional". In Inês Duarte & Isabel Leiria (orgs.), *Actas do Congresso Internacional sobre o Português, 1994, Volume III*, Lisboa, Edições Colibri / APL, Junho de 1996, pp. 167-84.

Santos, Diana, Luís Costa & Paulo Rocha. "Cooperatively evaluating Portuguese morphology". In Nuno J. Mamede, Jorge Baptista, Isabel Trancoso and Maria das Graças Volpe Nunes (eds.), *Computational Processing of the Portuguese Language, 6th International Workshop, PROPOR 2003, Faro, 26-27 June 2003, Proceedings*, Springer Verlag, 2003, pp. 259-266.

Santos, Diana, Nuno Seco, Nuno Cardoso & Rui Vilela. "HAREM: An Advanced NER Evaluation Contest for Portuguese". In Nicoletta Calzolari, Khalid Choukri, Aldo Gangemi, Bente Maegaard, Joseph Mariani, Jan Odjik and Daniel Tapias (eds.), *Proceedings of LREC 2006 (LREC'2006)* (Genoa, 22-28 May 2006), pp. 1986-1991.

Sazedj, Peyman & H. Sofia Pinto. "Time to evaluate: Targeting Annotation Tools". In *Proc. of Knowledge Markup and Semantic Annotation at ISWC 2005 (Semannot 2005)*, Nov. 2005, pp. 37-48.

Setzer, Andrea & Robert Gaizauskas. "A Pilot Study on Annotating Temporal Relations in Text". In *Proceedings of the Worskhop for Temporal and Spatial Information Processing* (Toulouse, July 7th 2001), EACL-ACL 2001, Toulouse, 2001, pp. 73-80.

Simard, Michel. "Translation spotting for translation memories". In *Proceedings of the HLT-NAACL 2003 Workshop on Building and Using Parallel Texts: Data Driven Machine Translation and Beyond - Volume 3* (Edmonton, Canada, May 31 - 31, 2003). Human Language Technology Conference. Association for Computational Linguistics, Morristown, NJ, 2003, pp. 65-72.

Simard, Michel, George Foster, Marie-Louise Hannan, Elliott Macklovitch & Pierre Plamondon. "Bilingual text alignment: where do we draw the line?". In Simon Philip Botley, Anthony Mark McEnery and Andrew Wilson (eds), *Multilingual Corpora in Teaching and Research*, Amsterdam – Atlanta, GA, Rodopi, 2000, pp. 38-64.

Simon, Julian L & Peter C. Bruce. *Resampling: A Better Way to Teach (and Do) Statistics* http://www.juliansimon.com/writings/Resampling_Statistics/Part_II/SHISTORY.txt

Sparck Jones, Karen. "Is question answering a rational task?". In R. Barnardi and M. Moortgat (eds.), *Questions and Answers: Theoretical and Applied Perspectives*, Second CoLogNET-ElsNET Symposium, (Utrecht Institute of Linguistics, Amsterdam, 2003), 2003, pp. 24-35.

Sparck-Jones, Karen & Julia R. Galliers. *Evaluating Natural Language Processing Systems: An Analysis and Review*. Springer Verlag, 1996. Extended version of Galliers & Sparck-Jones (1993).

Tennant, Harry. "Experience with the Evaluation of Natural Language Question Answerers". In *Proceedings of the International Joint Conference on Artificial Intelligence, IJCAI 79*, 1979, pp. 874-876.

Tinsley, Howard E. A. & David J. Weiss. "Interrater Reliability and Agreement." In Howard E. A. Tinsley and Steven D. Brown (eds.), *Handbook of Applied Multivariate Statistics and Mathematical Modeling*, San Diego, CA, Academic Press, 2000, pp. 95-124.

Véronis, Jean. "Tagging guidelines for word alignment", Version 1.0, April 26, 1998, http://www.up.univ-mrs.fr/veronis/arcade/arcade1/2nd/word/guide/index.html.

Véronis, Jean & Philippe Langlais. "Evaluation of paralell text alignment systems: the ARCADE project". In Jean Véronis (ed.), *Parallel Text Processing*, Dordrecht: Kluwer Academic Publishers, 2000, pp. 369-388.

Vilain, Marc, John Burger, John Aberdeen, Dennis Connolly & Lynette Hirschman. "A model-theoretic coreference scoring scheme". *Proceedings of the Sixth Message Understanding Conference (MUC-6)*, Morgan Kaufmann, 1995, pp. 45-52.

Voorhees, Ellen M. "Philosophy of IR Evaluation". Presentation at CLEF, 2001, http://www.ercim.org/publication/ws-proceedings/CLEF2/vorhees.pdf

Voorhees, Ellen M. & Chris Buckley. "The effect of topic set size on retrieval experiment error", in *Proceedings of the 25th annual international ACM SIGIR conference on research and development in information retrieval* (Tampere, Finland, 2002), pp. 316-323.

Voorhees, Ellen M. & Donna Harman. "Common Evaluation measures". In *The Eleventh Text Retrieval Conference (TREC 2002)*, NIST Special Publication: SP 500-251, 2003, http://trec.nist.gov/pubs/trec11/appendices/MEASURES.pdf

Voorhees, Ellen M. & Donna Harman. "The Text REtrieval Conference", in Ellen M. Voorhees and Donna K. Harman (eds.), *TREC: Experiment and Evaluation in Information Retrieval*, MIT Press, 2003, pp. 3-20.

Voorhees, Ellen M. & Dawn M. Tice. "Building a Question Answering Test Collection", in *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (Athens, Greece, July 2000), pp. 200-207.

White, John S. & Theresa A. O'Connell. "Evaluation in the ARPA machine translation program: 1993 Methodology". In *Proceedings of the Human Language Technology Workshop*, San Francisco, Morgan Kaufmann, 1994, pp. 135-140.

Wilks, Yorick A. "Unhappy Bedfellows: The Relationship of AI and IR". In John I. Tait (ed.), *Charting a New Course: Natural Language Processing and Information Retrieval: Essays in Honour of Karen Spärck Jones*, Springer, 2005, pp. 255-282.

Yarowsky, David. "Word sense disambiguation using statistical models of Roget's categories trained on large corpora". In *Proceedings of COLING'92* (Nantes, 23-28 July 1992), 1992, pp. 454-60.

Yarowsky, David. "One sense per collocation". In *Proceedings of Human Language Technology, ARPA*. San Francisco, Calif., Morgan Kaufmann, 1993, pp. 266-271.

Yarowsky, David. "Unsupervised Word Sense Disambiguation Rivaling Supervised Methods." In *Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics*. Cambridge, MA, 1995, pp. 189-196.

Zobel, Justin. "How Reliable Are the Results of Large-Scale Information Retrieval Experiments?". In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, *SIGIR '98* (Melbourne, Australia, August 24-28, 1998). ACM, 1998, pp. 307-314.

Acknowledgments

This work was done in the scope of the Linguateca project jointly funded by the Portuguese Government and the European Union (FEDER and FSE) under contract ref. POSC/339/1.3/C/NAC.

I am indebted to many people from Linguateca and from SINTEF ICT, who have attended preliminary presentations of the material here and helped improve it significantly with comments and questions.

Oslo, 17 August 2007