UPPSALA
UNIVERSITET

GSLT

Machine Translation Evaluation

Eva Forsbom

evafo@stp.ling.uu.se

Uppsala University

Evaluation Standardisation Efforts

```
           ISO                   LISA
         Software                 /\
         Quality                 /  \
            |                   /    \
          EAGLES             QA    Segmentation
         Framework
      /     |     \
Writing  Dialogue  ISLE          SAE
 Aids    Systems  Taxonomy        |
                                  |
                                J2450
```

Quality Attributes

ISO 8402: “The totality of features and characteristics of a product or service that bears on its ability to satisfy stated or implied needs”

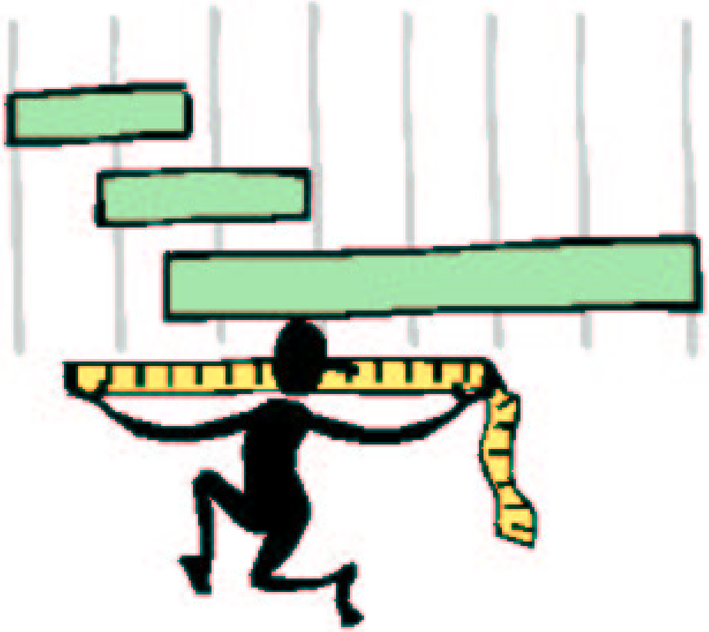
ISO/IEC 9126 series: Product quality

ISO/IEC 14598 series: Software product evaluation

- Functionality
- Reliability
- Usability
- Efficiency
- Maintainability
- Portability

UPPSALA
UNIVERSITET

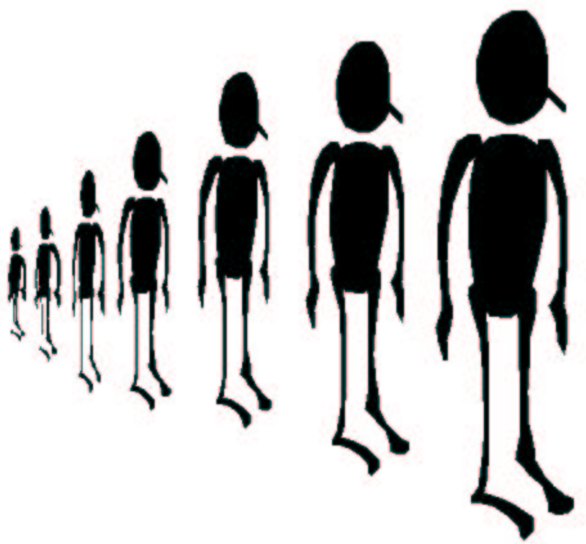
Evaluation Context



- For whom?
- Why?
- What?
- By whom?
- How?

For Whom?

Different users have different needs. The quality attributes should be picked and weighted accordingly.



- Consumer agency
- Manager
- Developer
- Experienced user
- Consumer
- ...

UPPSALA
UNIVERSITET

Why?

The purpose of the evaluation depends on the kind of user it is done for, and on the maturity of the product. There is a type of evaluation for each purpose... Some examples:

Type	Purpose
Feasability	See if the product is needed/worth developing
Diagnostic	Trace errors
Progressive	See changes between product versions
Adequacy	See if the product is adequate for a certain task
Performance	Compare different systems

What?

Depending on user and purpose, attributes at an appropriate level of specificity should be chosen for evaluation. Weighted results for specific attributes could be combined into a higher level attribute.

$$\begin{bmatrix} functionality : \begin{bmatrix} suitability : true \\ accuracy : 60\% \\ interoperability : xx \\ security : high \\ compliance : true \end{bmatrix} \\ reliability : 7 \\ usability : good \\ efficiency : basic \\ maintainability : xx \\ portability : yy \end{bmatrix}$$

By Whom?

The different types of evaluations requires different kinds of evaluators with different backgrounds. Some evaluations could be performed automatically, some not.

- Evaluation agency
- Business Manager
- Developer
- Domain expert
- Experienced user
- Bilingual user
- Consumer
- ...

How?

The evaluation process can be divided into three general stages:

1. Defining the quality requirements
   - requirements analysis
   - evaluation modelling
2. Preparing the evaluation
   - quality metrics selection
   - rating levels definition
   - assessment criteria definition
3. Proceeding with the evaluation
   - measurement
   - rating
   - assessment

MT Evaluation Smorgasbord

`http://www.issco.unige.ch/projects/isle/taxonomy2/`

Using ISLE’s MT Evaluation Taxonomy, evaluators can slide down a tree of increasingly specific quality attributes and find appropriate measures for evaluating them. It has two entry points, which are both mapped to the metrics.

1 Specifying user needs
- The purpose of evaluation
- The object of evaluation
- Characteristics of the translation task
  - Assimilation
  - Dissemination
  - Communication
- User characteristics
- Input characteristics (author and text)

2 System characteristics to be evaluated
- System internal characteristics
  - MT system-specific characteristics
  - Model of translation process
  - Linguistic resources and utilities
  - Characteristics of the intended mode
- System external characteristics
  - Functionality
  - Reliability
  - Usability
  - Efficiency
  - Maintainability
  - Portability
  - Cost

Blackbox Evaluation

In cases where the evaluator has no possibility to see output from the system components, or for high level quality attribute evaluation, a blackbox evaluation is appropriate. Then, only the input, possible settings, and output are known.

```
                         Input Overview
--------------------------------------------------------------------
Words                   Total: 11192          Unique: 2393 (21.38%)
Segments                Total:  1772          Unique: 1187 (66.99%)
--------------------------------------------------------------------
                         System Recall
--------------------------------------------------------------------
                               Words
Source Language Words   Total: 11025 (98.51%)  Unique: 2322 (97.03%)
--------------------------------------------------------------------
                             Segments
Fully Translated        Total:   594 (33.52%)  Unique:  210 (17.69%)
Translated              Total:   678 (38.26%)  Unique:  285 (24.01%)
--------------------------------------------------------------------
```

Glassbox Evaluation

In cases where the evaluator has possibility to see output from the system components, or for low level quality attribute evaluation, a glassbox evaluation is appropriate. Then, input, possible settings, and output to some or all components are known.

```
                  Error Reports
------------------------------------------------
                       Words
Source Language Words  Total:  167   Unique:  71
Translation Links      Total: 1795   Unique: 371
Target Language Words  Total:   18   Unique:   3
Target Language Code   Total:    7   Unique:   1
------------------------------------------------
                      Segments
Not Parsed             Total:  347   Unique: 304
Partially Parsed       Total:  712   Unique: 577
Not Transferred        Total:   15   Unique:   6
Not Generated          Total:   17   Unique:  12
------------------------------------------------
```

Evaluating Translation Quality

Translation quality is usually evaluated by comparison of the translated text to the source text (by bilingual evaluators) or to a reference translation (by monolingual evaluators). Some evaluations could be performed automatically.

- Fidelity (how close)
- Correctness (how correct)
- Adequacy (how adequate)
- Informativeness (how informative)
- Intelligibility (how understandable)
- Fluency (how fluent)

UPPSALA
UNIVERSITET

Manual Evaluation – Tests

- Grading
- Cloze test
- Comprehension test
- Task-based test
- Reading time
- Typing
- Post-editing

Example: Adequacy Scale

(Doyon, Taylor, and White, 1998)

5 All meaning expressed in the source fragment appears in the translation fragment

4 Most of the source fragment meaning is expressed in the translation fragment

3 Much of the source fragment meaning is expressed in the translation fragment

2 Little of the source fragment meaning is expressed in the translation fragment

1 None of the meaning expressed in the source fragment is expressed in the translation fragment

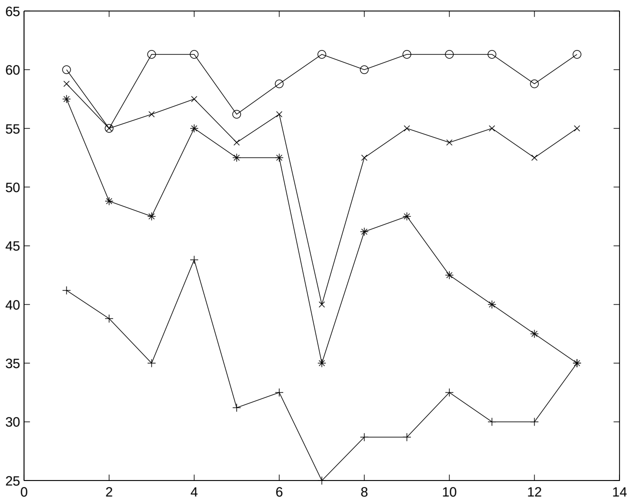
Example: Adequacy Test for LREC’02

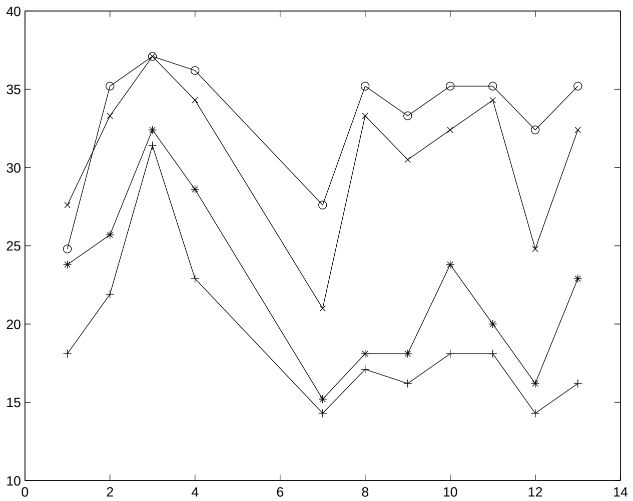
(`http://stp.ling.uu.se/˜evafo/lrec_eval/`)

1 2 3 4 5 **Source:** Prévenir ses enfants des problèmes de drogue

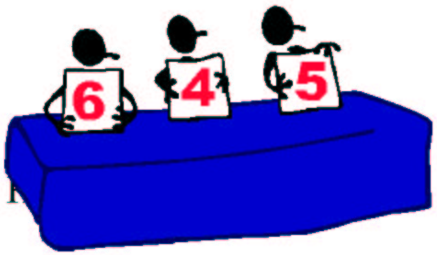
○ ○ ○ ○ ○ **Reference:** Prevent your children from having drug problems

Translation: Prevent your children from drug problems

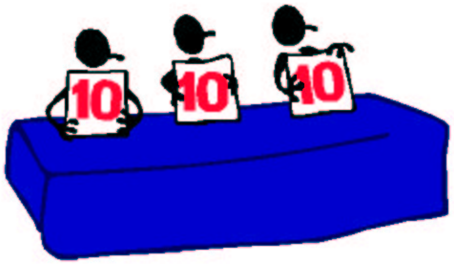




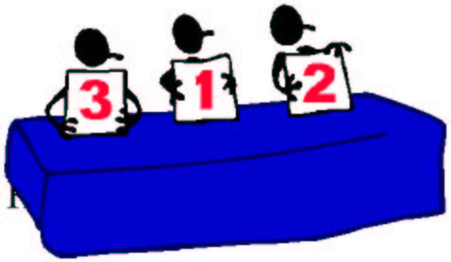
Manual Evaluation – Problems



The hat is fat.



The cat is fat.



The hat is fat.

Semi-Automatic Evaluation

Semi-automatic evaluation usually involves some form of manual mark-up, followed by automatic comparison and computation, e.g. by certain words, constructions, or information units.

- Named entity translation
- EvalTrans
- Syntactic correctness
- Domain terminology translation
- Information unit translation
- Test suite creation

Example: Named Entity Translation

(Reeder et al. 2001)

In this evaluation, some human annotators marks up named entities (NE) in a reference translation. All unique NE's from the reference translation are then searched in the translations, and all unique occurrences counted. Some normalisation processes could also be applied.

- Only relevant when many NE's.
- Depends on the annotators' consistency.
- Depends on the reference translation quality.

Example: EvalTrans

(Nießen et al. 2000)

EvalTrans is a tool for semi-automatic evaluation of translations. Storing of previous evaluations makes the manual evaluations more consistent.

- Manual seeding of scores (SSER)
- Storing of evaluations (WER and SSER)
- Automatic comparison of new translations with old
- Extrapolation of SSER for new translations
- Highlighting of new translations (with mark-up of edit operations)
- Possibilty of splitting segments into information units

UPPSALA
UNIVERSITET

Automatic Evaluation

Automatic evaluation is usually some form of approximate string matching or a count of mark-ups. If there exist advanced linguistic resources for the languages under scrutiny, much mark-up could be done automatically.

- Edit distance
- N-gram occurrence
- Number of untranslated words
- (Named entity translation)
- (Syntactic correctness)
- (Domain terminology translation)
- (Information unit translation)
- (Test suite creation and evaluation)

Edit Distance – Dynamic Programming

		A	N	D	E	R	S
	0	1	2	3	4	5	6
A	1						
R	2						
N	3						
E	4						

Edit Distance – Dynamic Programming

		A	N	D	E	R	S
	0	1	2	3	4	5	6
A	1	0					
R	2						
N	3						
E	4						

Edit Distance – Dynamic Programming

		A	N	D	E	R	S
	0	1	2	3	4	5	6
A	1	0	1	2	3	4	5
R	2						
N	3						
E	4						

Edit Distance – Dynamic Programming

		A	N	D	E	R	S
	0	1	2	3	4	5	6
A	1	0	1	2	3	4	5
R	2	1	1	2	3	3	
N	3						
E	4						

Edit Distance – Dynamic Programming

		A	N	D	E	R	S
	0	1	2	3	4	5	6
A	1	0	1	2	3	4	5
R	2	1	1	2	3	3	4
N	3	2	1	2	3	4	4
E	4	3	2	2	2	3	4

Edit Distance – Aligning

		A	N	D	E	R	S
	0	1	2	3	4	5	6
A	1	0	1	2	3	4	5
R	2	1	1	2	3	3	4
N	3	2	1	2	3	4	4
E	4	3	2	2	2	3	4

*	*	*	*	*	*	S
*	*	*	*	*	*	*
*	*	*	*	*	*	i

Edit Distance – Aligning

		A	N	D	E	R	S
	0	1	2	3	4	5	
A	1	0	1	2	3	4	
R	2	1	1	2	3	3	
N	3	2	1	2	3	4	
E	4	3	2	2	2	3	

*	*	*	*	*	R	S
*	*	*	*	*	*	*
*	*	*	*	*	i	i

Edit Distance – Aligning

		A	N	D	E	R	S
	0	1	2	3	4		
A	1	0	1	2	3		
R	2	1	1	2	3		
N	3	2	1	2	3		
E	4	3	2	2	2		

*	*	*	*	E	R	S
*	*	*	*	E	*	*
*	*	*	*	c	i	i

Edit Distance – Aligning

		A	N	D	E	R	S
	0	1	2	3			
A	1	0	1	2			
R	2	1	1	2			
N	3	2	1	2			
E							

*	*	*	D	E	R	S
*	*	*	*	E	*	*
*	*	*	i	c	i	i

Edit Distance – Aligning

		A	N	D	E	R	S
	0	1	2				
A	1	0	1				
R	2	1	1				
N	3	2	1				
E							

*	*	N	D	E	R	S
*	*	N	*	E	*	*
*	*	c	i	c	i	i

Edit Distance – Aligning

		A	N	D	E	R	S
	0	1					
A	1	0					
R	2	1					
N							
E							

*	*	N	D	E	R	S
*	R	N	*	E	*	*
*	d	c	i	c	i	i

Edit Distance – Aligning

		A	N	D	E	R	S
	0	1					
A	1	0					
R							
N							
E							

A	*	N	D	E	R	S
A	R	N	*	E	*	*
c	d	c	i	c	i	i

Example: Word Accuracy

(Alshawi et al. 1998)

$$\text{WA} = \left(1 - \frac{d + s + i}{r}\right)$$

where

$d =$ deletions

$s =$ substitutions

$i =$ insertions

$r =$ length of reference

Word Accuracy Problem

The original word accuracy measure could result in a score less than 0, as in the following example:

Src:	Tätningsring
Cand:	Sealing ring
Ref:	Seal

$$\left(1 - \frac{1+1+0}{1}\right) = -1$$

UPPSALA
UNIVERSITET

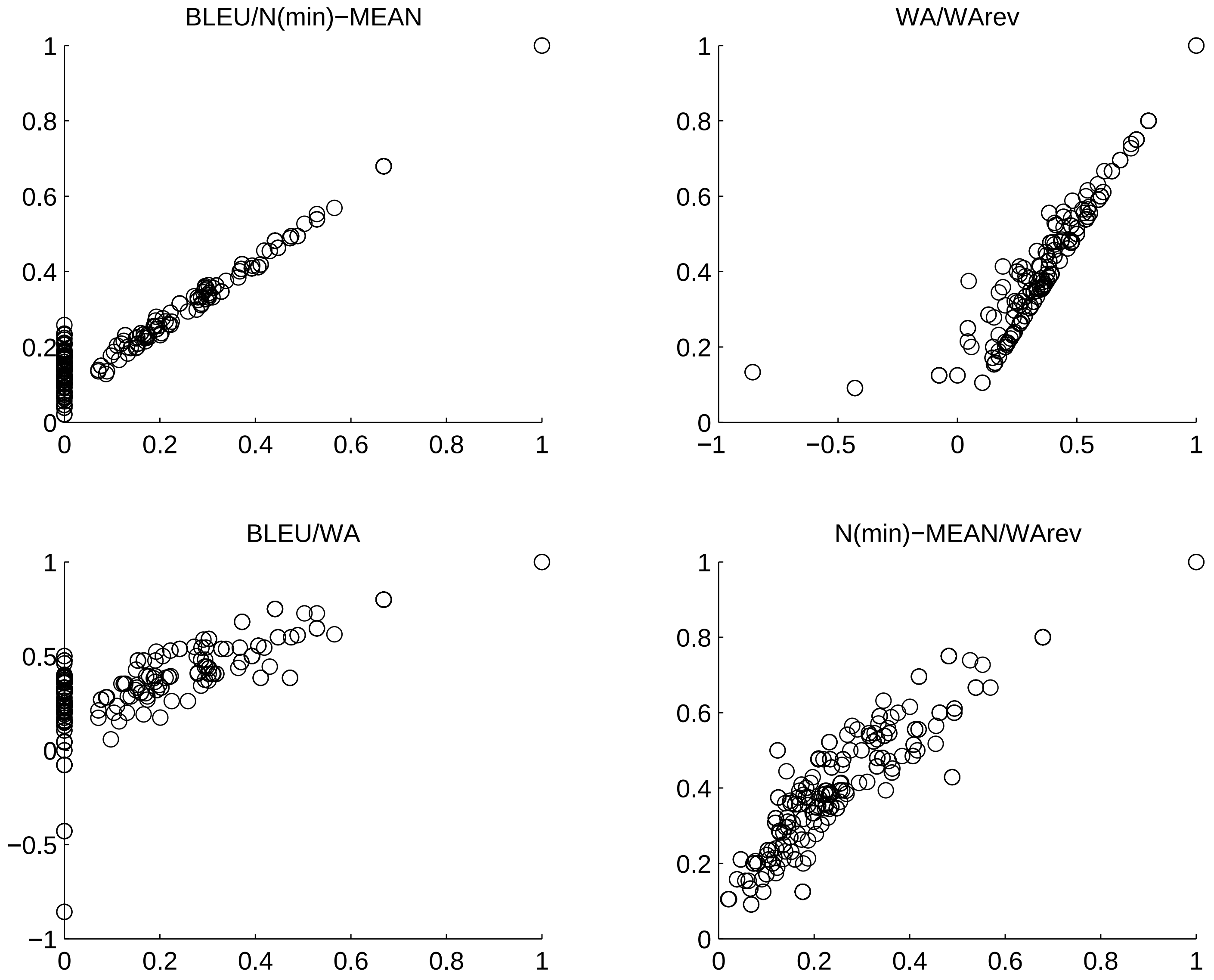
Revised Word Accuracy

$$\text{WArev} = \left(1 - \frac{d + s + i}{\max(r, c)}\right)$$

where

$$\begin{aligned} d &= \text{deletions} \\ s &= \text{substitutions} \\ i &= \text{insertions} \\ r &= \text{length of reference} \\ c &= \text{length of candidate} \end{aligned}$$

Word Accuracy vs. Revised Word Accuracy



UPPSALA
UNIVERSITET

Word Accuracy Weaknesses

- Sensitive to word order reversal
- Only evaluated against one reference translation at a time

Src:	Cylinder, underdel
Cand:	Bottom cylinder
Ref:	Cylinder bottom

Src:	Ledningsnät för bränslepump
Cand:	Cable harness for fuel pump
Ref:	Fuel pump cable harness

N-Gram Occurrence

N-gram occurrence is a way of measuring if words are correctly translated (1-grams) and if the translation is idiomatic ($n > 1$). It seems to correlate well with human evaluation of accuracy and fluency.

BLEU (Papineni et al. 2001)

- Grade $= [0, 1]$;
- Compensates for difference in length by a brevity penalty;
- Applies equal weights for all n-grams.

NIST (DARPA 2001(?))

- Grade $= [0, \infty)$;
- Compensates for difference in length by another brevity penalty;
- Applies different weights for the n-grams.

UPPSALA
UNIVERSITET

Example: BLEU

$$\text{BLEU} = \text{BP} \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right)$$

where

$$\text{BP} = \begin{cases} 1 & \text{if } c > r \\ \mathrm{e}^{\left(1-\frac{r}{c}\right)} & \text{if } c \leq r \end{cases}$$

$$r = \text{length of reference}$$

$$c = \text{length of candidate}$$

$$N = 4$$

$$w = \frac{1}{N}$$

$$p = \frac{\sum_{C\in\{Candidates\}} \sum_{n\in\{Candidates\}} Count_{clip}(n)}{\sum_{C\in\{Candidates\}} \sum_{n\in\{Candidates\}} Count(n)}$$

BLEU Problem

The original BLEU measure is not defined for all cases, as in the following examples:

Src:	Cylinder, underdel
Cand:	Bottom cylinder
Ref:	Cylinder bottom

Src:	Ledningsnät för bränslepump
Cand:	Cable harness for fuel pump
Ref:	Fuel pump cable harness

N-MEAN – Revised BLEU

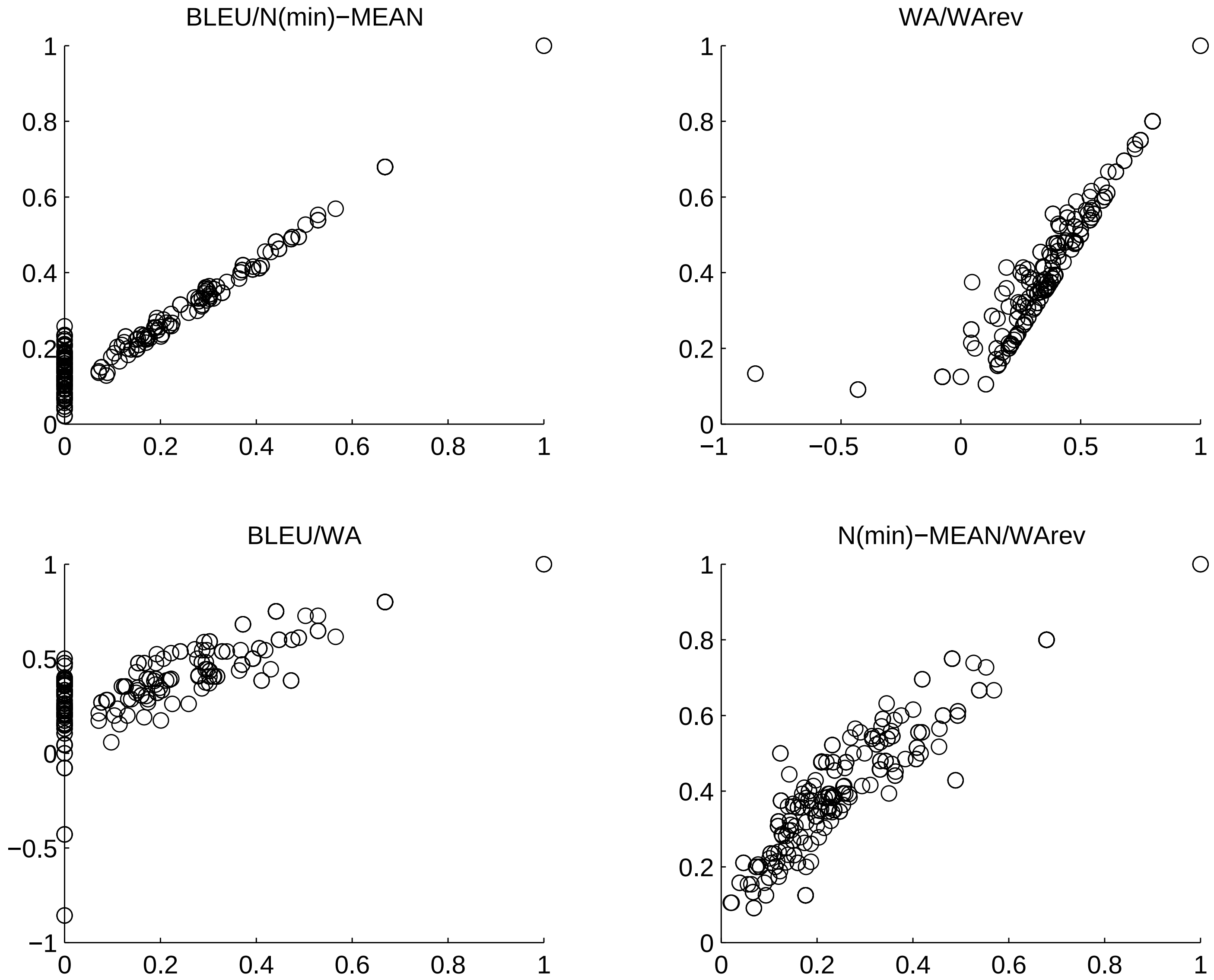
$$\text{N-MEAN} = \text{BP} \cdot \sum_{n=1}^{N} w_n p_n$$

where

$$\text{N} = \begin{cases} N_{max} & \text{if } c \geq N_{max} \\ c & \text{if } c < N_{max} \end{cases}$$

UPPSALA UNIVERSITET

BLEU vs. N-MEAN



N-Gram Occurrence Weakness

- Sensitive to word errors (particularly mid-segment)

Cand:	The cats is fat
Ref:	The cat is fat

Ongoing and Future Work

- Applying these automatic measures on another text type
- Applying these automatic measures on another domain
- Applying these automatic measures on another language pair
- Applying these automatic measures with only one reference translation
- Using other automatic measures
- Using more linguistic measures

References

- Alshawi et al. Automatic acquisition of hierarchical transduction models for machine translation. In *Proceedings of the ACL’98*, pp. 41–47, Montreal, Canada, 1998.
- DARPA. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics, 2001(?).
- Doyon et al. The DARPA machine translation evaluation methodology: Past and present. In *Proceedings of AMTA’98*, Philadelphia, PA, 1998.
- EAGLES (Expert Advisory Group on Language Engineering Standards) `http://issco-www.unige.ch/projects/eagles/`

References...

- ISLE (International Standards for Language Engineering) `http://www.issco.unige.ch/projects/isle`
- ISO (International Organization for Standardization) `http://www.iso.org`
- LISA (Localization Industry Standards Association) `http://www.lisa.org`
- Nießen et al. An evaluation tool for machine translation: Fast evaluation for MT research. In *Proceedings of LREC'00*, pp 39–45, Athens, Greece, 2000.

References...

- Papineni et al. BLEU: a method for automatic evaluation of machine translation. IBM RC22176 (W0109-022), IBM Research Division, T. J. Watson Research Center, 2001.
- Reeder et al. The naming of things and the confusion of tongues: an MT metric. In *Proceedings of the MT Evaluation Workshop: Who Did What To Whom, MT Summit VIII*, pp. 55–59, Santiago de Compostela, Spain, 2001.
- SAE (Society of Automotive Engineers). `http://www.sae.org/`