Training a Super Model Look-Alike:

Featuring Edit Distance, N-Gram Occurrence, and One Reference Translation

Eva Forsbom, Uppsala University

`evafo@stp.ling.uu.se`

MT Summit IX

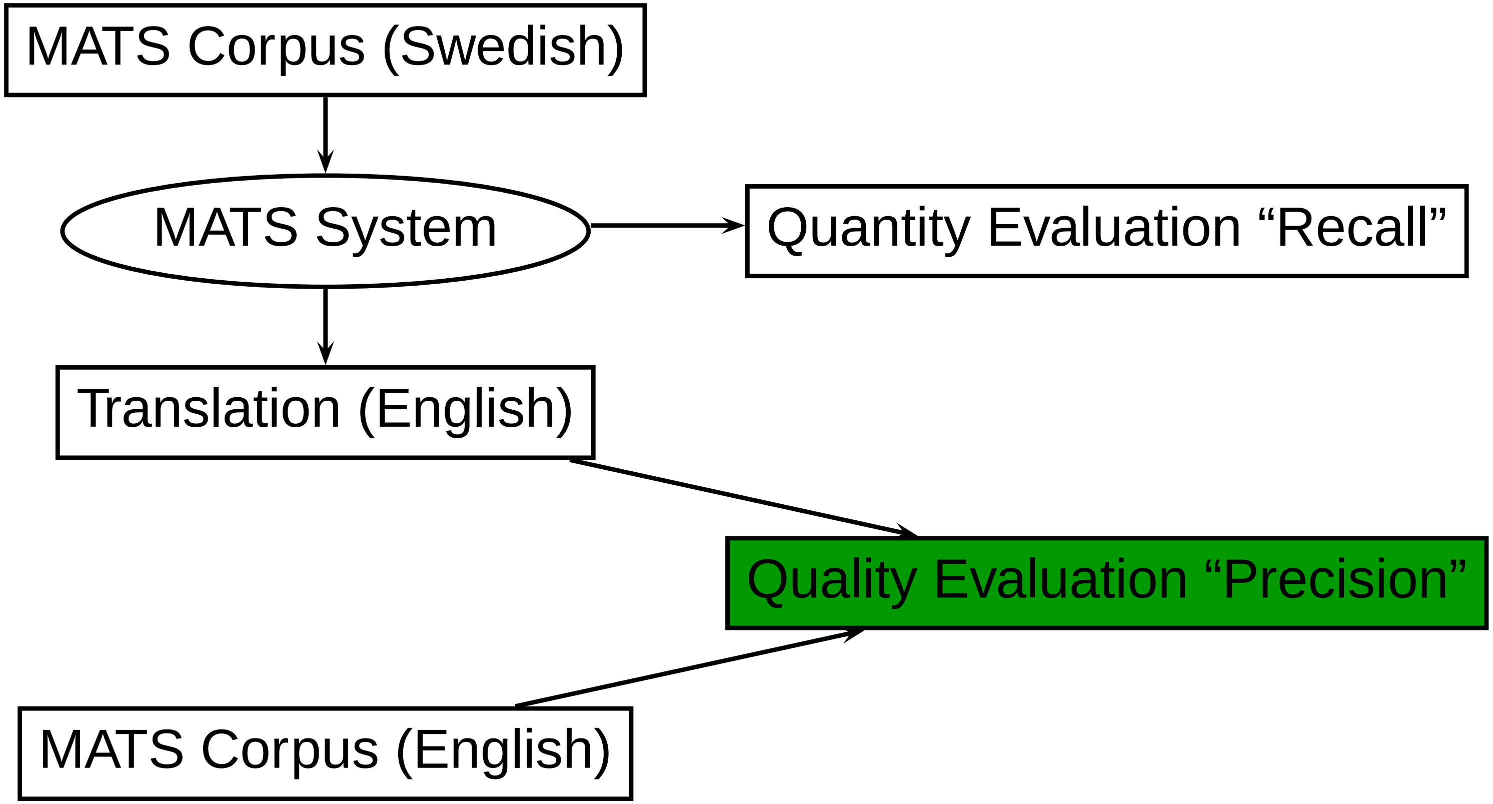
Workshop on Machine Translation Evaluation

Towards Systematizing MT Evaluation

Saturday, September 27, 2003

New Orleans, Louisiana, USA

Evaluation Context



MATS Corpus (Swedish)

MATS System

Quantity Evaluation “Recall”

Translation (English)

Quality Evaluation “Precision”

MATS Corpus (English)

WANTED: Translation Quality Measure!

- Be automatic.
- Work for various kinds of evaluations:
    - declarative,
    - progressive,
    - diagnostic.
- Work at various levels:
    - system,
    - document,
    - segment.
- Work for various text types (news/technical manuals).
- Work with one reference translation.
- Exist.

UPPSALA
UNIVERSITET

Applicants

- Edit distance: Word Accuracy (Alshawi et al. 1998)
- N-gram occurrence: BLEU (Papineni et al. 2001)
- N-gram occurrence: NIST (Doddington 2002)
- Possible redefinitions...

Heat 1: Edit Distance – Word Accuracy

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

where

$d =$ deletions

$s =$ substitutions

$i =$ insertions

$r =$ length of reference

WA Scoring Card

- Work for various kinds of evaluations:
  - declarative, √
  - progressive, √
  - diagnostic.
- Work at various levels:
  - system, √
  - document, √
  - segment.
- Work for various text types.
- Work with one reference translation. √

WA Scoring Card

- Work for various kinds of evaluations:
  - declarative, $\surd$
  - progressive, $\surd$
  - diagnostic.
- Work at various levels:
  - system, $\surd$
  - document, $\surd$
  - segment. Failed!
- Work for various text types.
- Work with one reference translation. $\surd$

WA Failed for Segments!

Word Accuracy can result in a score less than 0 if the length of the reference is shorter than the length of the candidate:

Src:	Tätningsring	
Cand:	Sealing ring	length = 2
Ref:	Seal	length = 1

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

WA – Redefinition

Word Accuracy For Translation:

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

where

$d =$ deletions

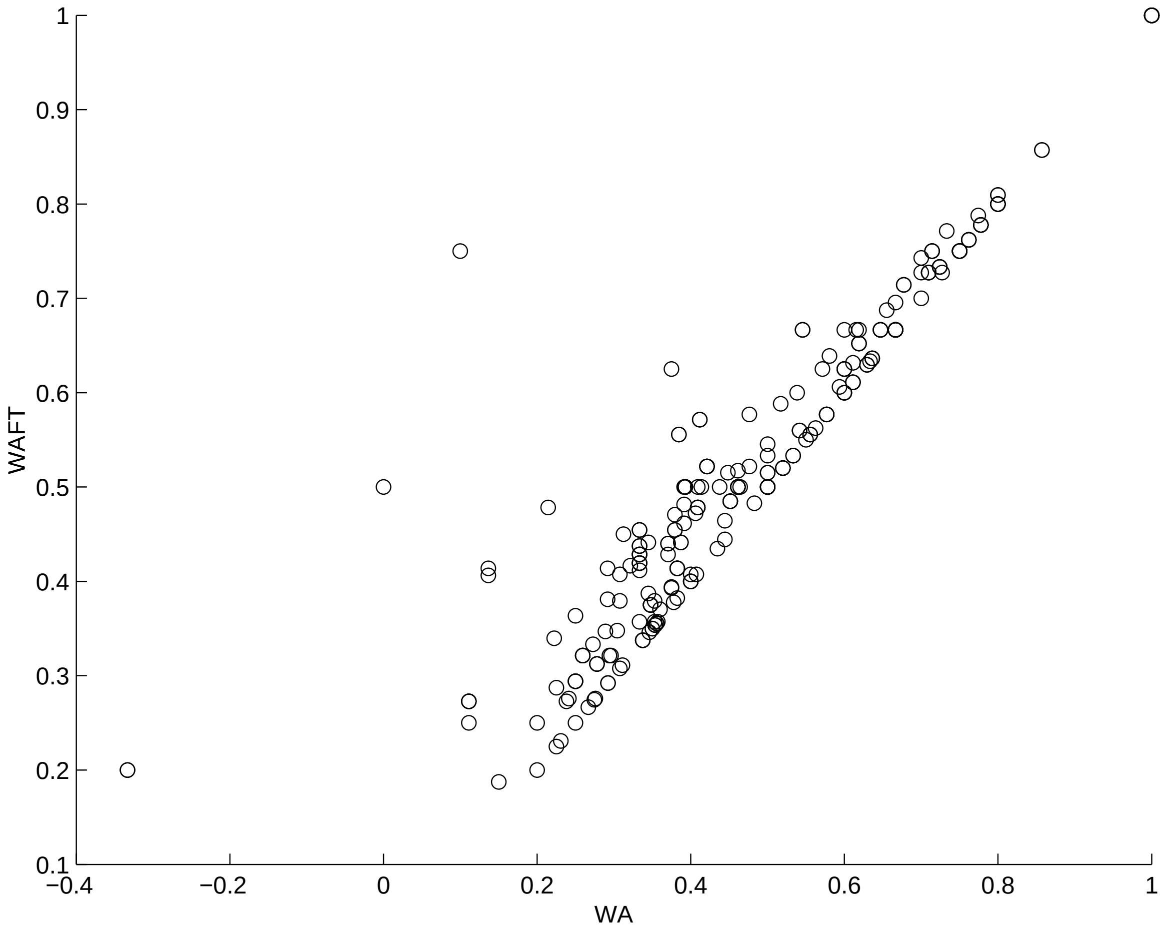
$s =$ substitutions

$i =$ insertions

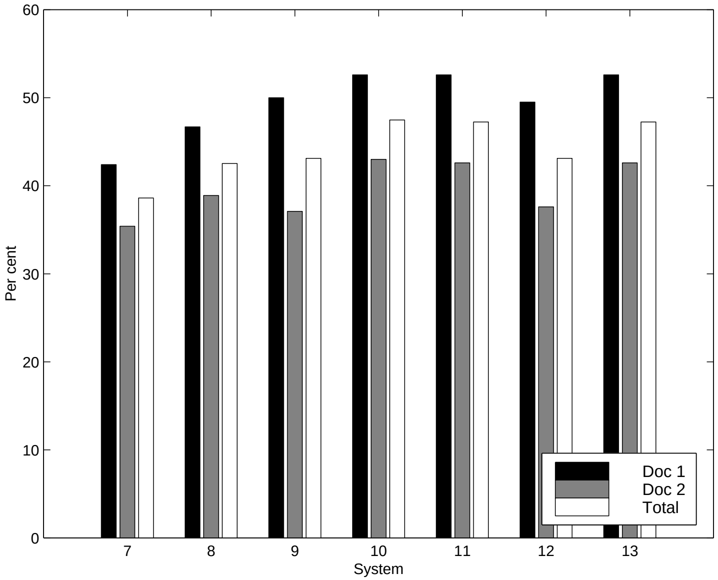
$r =$ length of reference

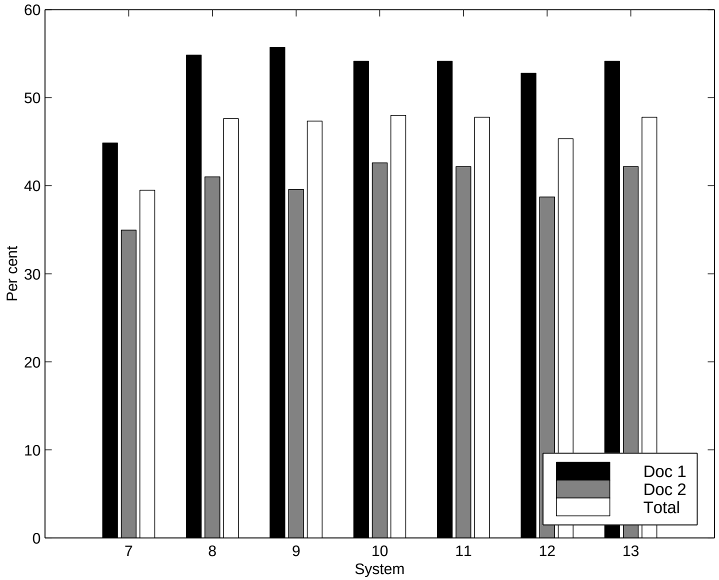
$c =$ length of candidate

WA vs. WAFT for Segments (LREC 6/4)



WA vs. WAFT: Docs & Sys (LREC 6/4)





- Ranking is the same (except for systems 8 and 9 on document 1).
- WAFT yields slightly higher scores than WA.

WAFT Scoring Card

- Work for various kinds of evaluations:
  - declarative, √
  - progressive, √
  - diagnostic.
- Work at various levels:
  - system, √
  - document, √
  - segment.
- Work for various text types.
- Work with one reference translation. √

WAFT Scoring Card

- Work for various kinds of evaluations:
  - declarative, √
  - progressive, √
  - diagnostic.
- Work at various levels:
  - system, √
  - document, √
  - segment. √
- Work for various text types.
- Work with one reference translation. √

WAFT Scoring Card

- Work for various kinds of evaluations:
  - declarative, √
  - progressive, √
  - diagnostic.
- Work at various levels:
  - system, √
  - document, √
  - segment. √
- Work for various text types. √
- Work with one reference translation. √

Heat 2: N-Gram Occurrence

BLEU

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

NIST

- Score $= [0, ?;$
- Compensates for difference in length by another brevity penalty;
- Applies different weights for the n-grams.

NIST Scoring Card

- Work for various kinds of evaluations:
  - declarative
  - progressive, $\surd$
  - diagnostic.
- Work at various levels:
  - system, $\surd$
  - document, $\surd$
  - segment.
- Work for various text types.
- Work with one reference translation.

NIST Scoring Card

- Work for various kinds of evaluations:
  - declarative Failed!
  - progressive, √
  - diagnostic.
- Work at various levels:
  - system, √
  - document, √
  - segment.
- Work for various text types.
- Work with one reference translation.

NIST Failed for Declarative Evaluation!

NIST can yield different scores for equivalent objects of evaluations, due to its weighting method:

Src:	Antal	**Src:**	Beteckning
Cand:	Number	**Cand:**	Designation
Ref:	Number	**Ref:**	Designation

$\text{NIST} = 4.6267$ $\text{NIST} = 8.0311$

Heat 2: N-Gram Occurrence – BLEU

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

where

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

$r =$ length of reference

$c =$ length of candidate

$N = N_{max}$ (=4)

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

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

BLEU Scoring Card

- Work for various kinds of evaluations:
  - declarative, √
  - progressive, √
  - diagnostic.
- Work at various levels:
  - system, √
  - document, √
  - segment.
- Work for various text types.
- Work with one reference translation.

BLEU Scoring Card

- Work for various kinds of evaluations:
  - declarative, $\surd$
  - progressive, $\surd$
  - diagnostic.
- Work at various levels:
  - system, $\surd$
  - document, $\surd$
  - segment. Failed!
- Work for various text types.
- Work with one reference translation.

UPPSALA UNIVERSITET

BLEU Failed for Segments!

BLEU measure is not defined for segments with a length shorter than N_{max}:

Src:	Cylinder, underdel	
Cand:	Bottom cylinder	length $< N_{max}$
Ref:	Cylinder bottom	

$$\text{BLEU} = \text{BP} \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right) = undefined/0$$

BLEU – Redefinition

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

where

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

First Draft Still Failed for Segments!

First draft measure is not defined for segments lacking co-occurrence for at least 1 n-gram level:

Src:	Ledningsnät för bränslepump	
Cand:	Cable harness for fuel pump	no 3- or 4-grams
Ref:	Fuel pump cable harness	

$$\text{First draft} = \text{BP} \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right) = undefined/0$$

UPPSALA UNIVERSITET

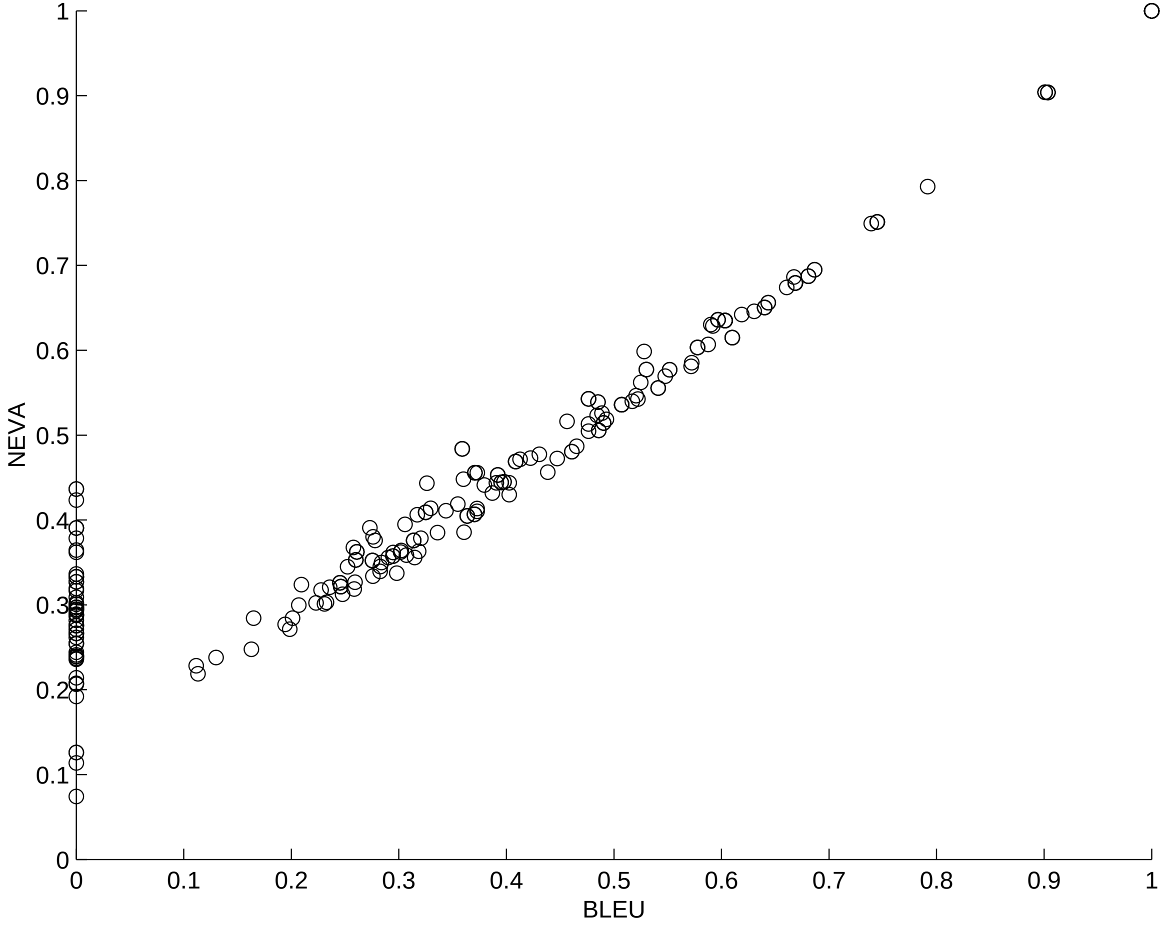
BLEU – Redefinition 2

N-gram EVAluation:

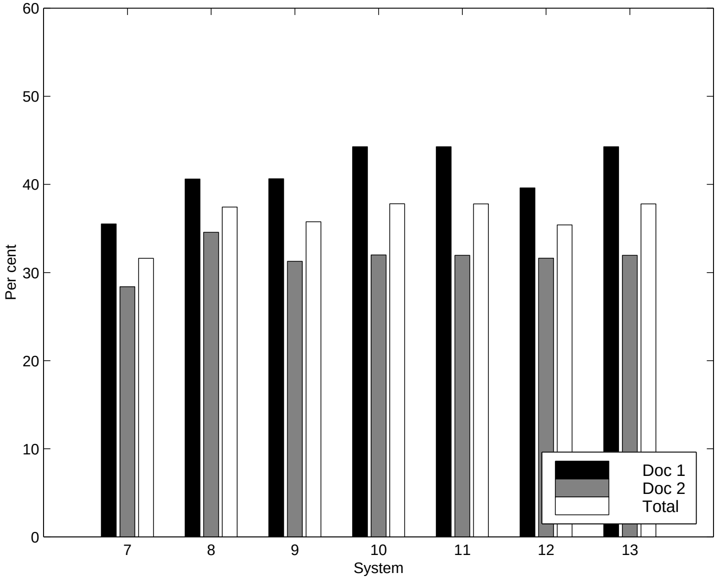
$$\text{NEVA} = \text{BP} \cdot \sum_{n=1}^{N} w_n p_n$$

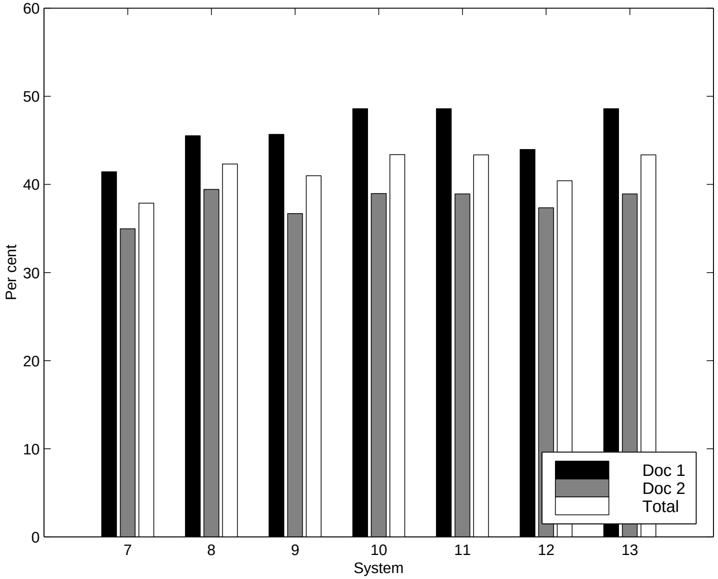
UPPSALA
UNIVERSITET

BLEU vs. NEVA for Segments (LREC 6/4)



BLEU vs. NEVA: Docs & Sys (LREC 6/4)





- Ranking is the same.
- NEVA yields slightly higher scores than BLEU.

UPPSALA
UNIVERSITET

NEVA Scoring Card

- Work for various kinds of evaluations:
  - declarative, $\surd$
  - progressive, $\surd$
  - diagnostic.
- Work at various levels:
  - system, $\surd$
  - document, $\surd$
  - segment.
- Work for various text types.
- Work with one reference translation.

NEVA Scoring Card

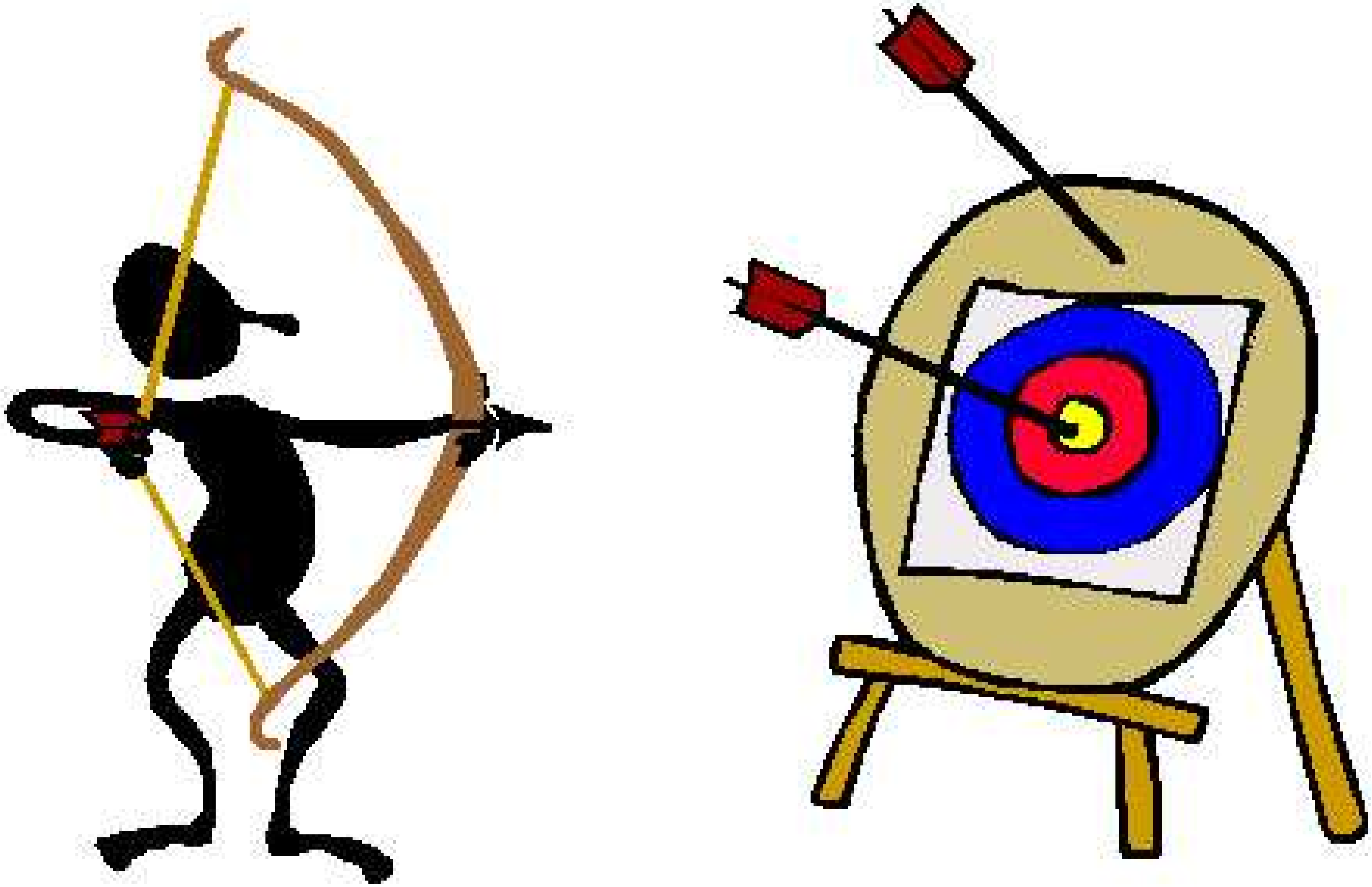
- Work for various kinds of evaluations:
  - declarative, √
  - progressive, √
  - diagnostic.
- Work at various levels:
  - system, √
  - document, √
  - segment. √
- Work for various text types.
- Work with one reference translation.

NEVA Scoring Card

- Work for various kinds of evaluations:
  - declarative, √
  - progressive, √
  - diagnostic.
- Work at various levels:
  - system, √
  - document, √
  - segment. √
- Work for various text types. √
- Work with one reference translation.

UPPSALA
UNIVERSITET

One Reference Translation?



6/4 vs. 1 Reference Translation

- Ranking is basically the same (except for 8 and 9 on document 1 for WAFT).
- Scores are higher for 6/4 references (much higher for NEVA).
- Scoring levels for document 1 and 2 are reversed.

Level	WAFT	NEVA
System	0.8589	0.9857
Document 1	0.6854	0.9983
Document 2	0.9348	0.9632
Segment	0.6215	0.7274

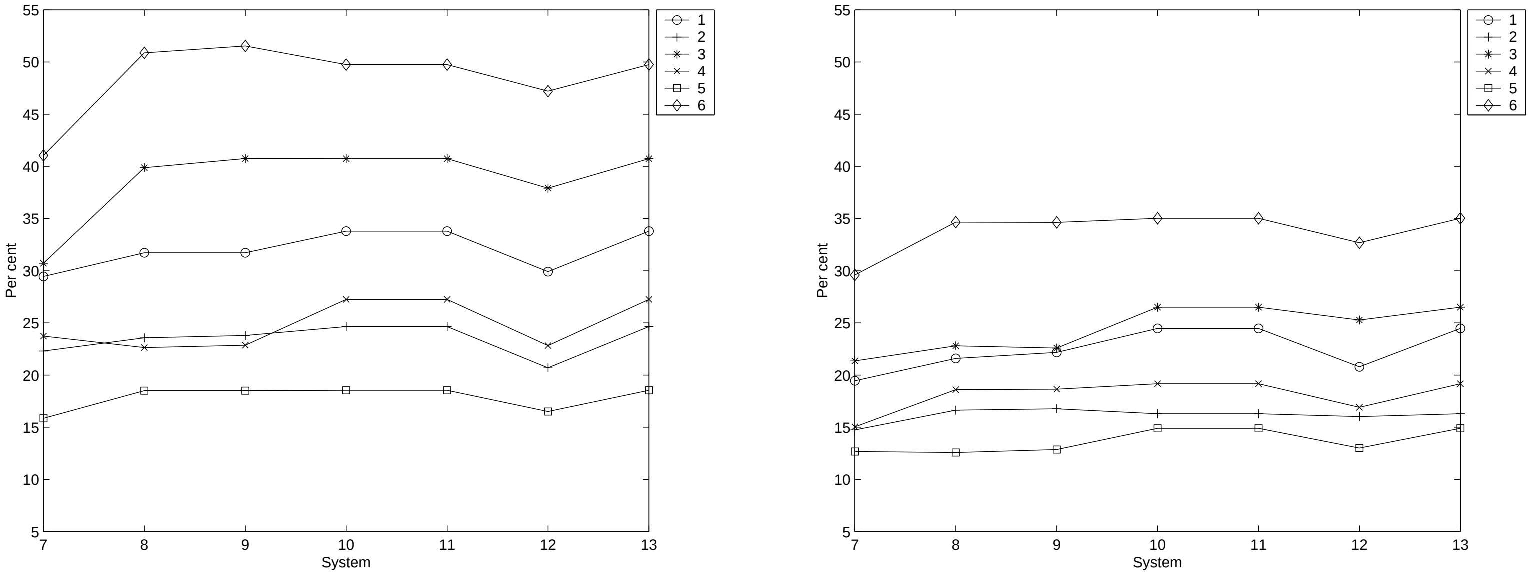
NEVA Scoring Card

- Work for various kinds of evaluations:
  - declarative, √
  - progressive, √
  - diagnostic.
- Work at various levels: √
- Work for various text types. √
- Work with one reference translation.

NEVA Scoring Card

- Work for various kinds of evaluations:
  - declarative, √
  - progressive, √
  - diagnostic.
- Work at various levels: √
- Work for various text types. √
- Work with one reference translation. √

Quality of Reference Translation

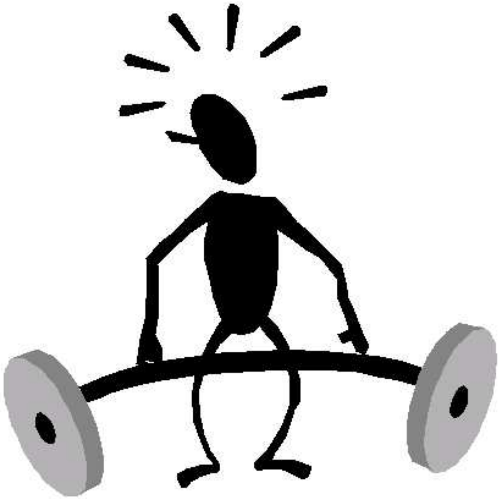


- Quality of reference translation matters for scoring.
- Quality of reference translation does not matter much for ranking.

The Super Model: Requirements

Free from

- errors (spelling, grammar, etc.);
- inconsistencies (variant forms, unwanted synonyms, etc.); and
- interpretations, additions, deletions, etc.

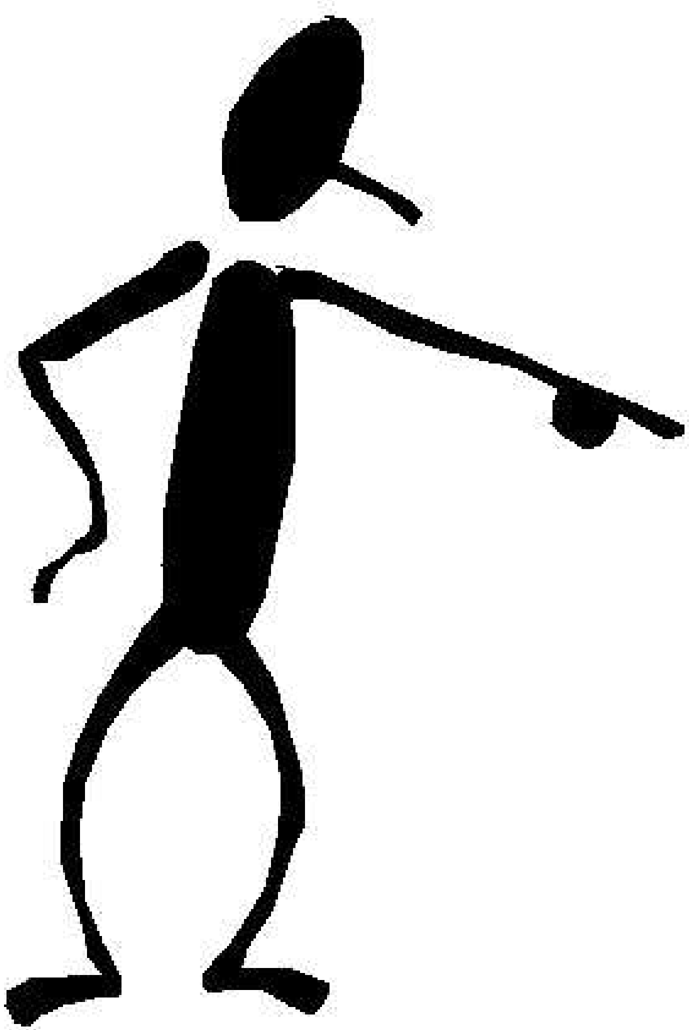


The Super Model: Cloning

Original	**Clone 1: Errors**	**Clone 2: Inconsistencies**	**Clone 3: Interpretations**
1. Trouble shooting	1. Trouble shooting	1. Troubleshooting	1. Troubleshooting
2. The fluid is cleaned by passing through a filter.	2. The fluid is cleaned via a filter.	2. The oil is cleaned via a filter.	2. The oil is cleaned via a filter.
3. Failure to follow this instruction can ...	3. Failure to follow this instruction can ...	3. Failure to follow this instruction can ...	3. It can ...

UPPSALA
UNIVERSITET

Diagnostic Evaluation?



Diagnostic Evaluation: Edit Distance

Weakness:

- Sensitive to word order reversal.

Src: Cylinder, underdel

Cand: Bottom cylinder

Ref: Cylinder bottom

Advantages:

- Possibility to align edit operations, and to find
  - variant forms and synonyms (*clip/clamp*);
  - inflectional errors (*tensioner/tensioners*);
  - word errors (*in/into*);
  - differences in definiteness (*the*);
  - specifications or generalisations (nominal modifiers);

WAFT Scoring Card

- Work for various kinds of evaluations:
  - declarative, $\surd$
  - progressive, $\surd$
  - diagnostic.
- Work at various levels: $\surd$
- Work for various text types. $\surd$
- Work with one reference translation. $\surd$

WAFT Scoring Card

- Work for various kinds of evaluations:
    - declarative, √
    - progressive, √
    - diagnostic. √
- Work at various levels: √
- Work for various text types. √
- Work with one reference translation. √

Diagnostic Evaluation: N-Gram Occurrence

Weakness:

- Sensitive to word errors (particularly mid-segment)

Src:	Kontrollera backventilen.
Cand:	Check the check valve.
Ref:	Check the non-return valve.

Advantages:

- If something is right, it always yields a score above 0.
- It would be possible to report all n-grams not found.

NEVA Scoring Card

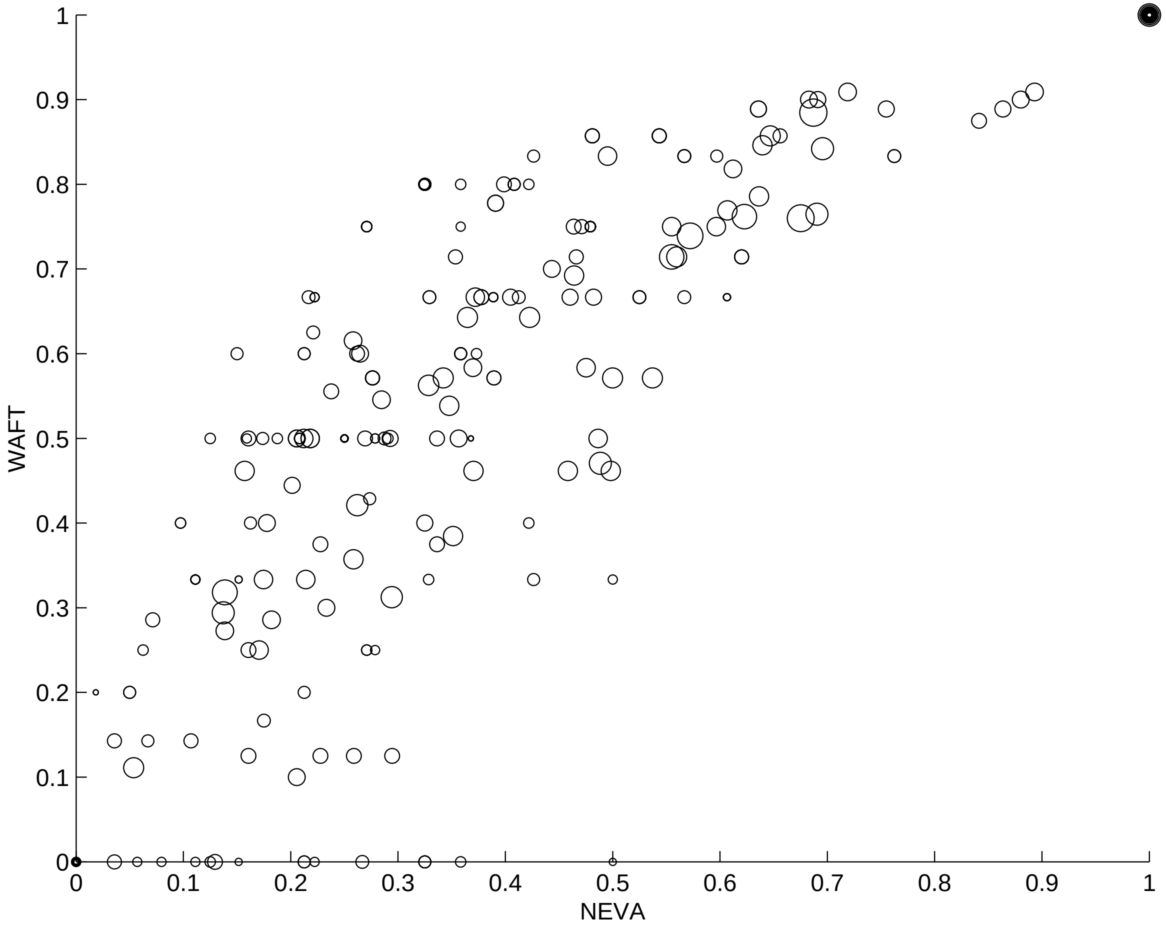
- Work for various kinds of evaluations:
  - declarative, √
  - progressive, √
  - diagnostic.
- Work at various levels: √
- Work for various text types. √
- Work with one reference translation. √

NEVA Scoring Card

- Work for various kinds of evaluations:
  - declarative, √
  - progressive, √
  - diagnostic. √
- Work at various levels: √
- Work for various text types. √
- Work with one reference translation. √

Diagnostic Evaluation: Correlation

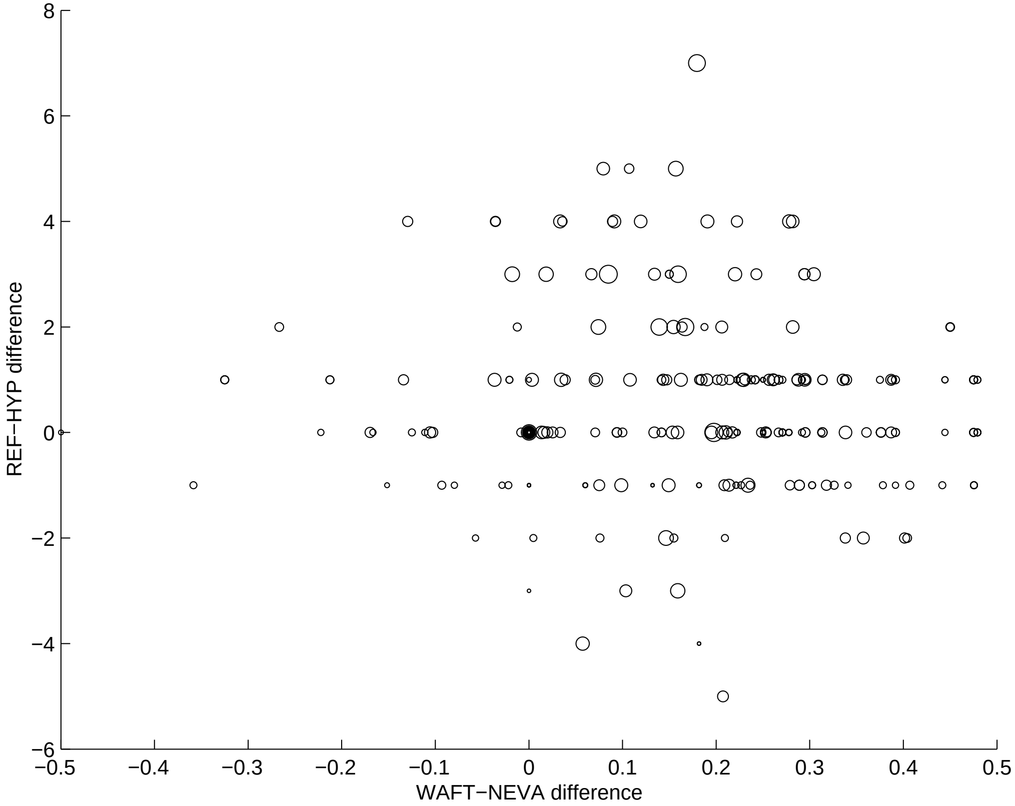
All segments in MATS where NEVA scores were greater than WAFT scores displayed a reversed word order problem:



UPPSALA UNIVERSITET

Diagnostic Evaluation: Function?

Diagnostic score is possibly a function involving difference in WAFT and NEVA scores and difference in candidate and reference length.



Conclusions 1

- Edit distance and n-gram co-occurrence measures are applicable
  - for declarative, progressive, and diagnostic evaluations;
  - at the system, document, and segment level;
  - for news text and technical manuals; and
  - for use with a single reference translation.
- The existing measures (WA and BLEU) needed redefinition to be applicable at the segment level.
- The redefined measures (WAFT and NEVA) gave higher scores, but kept the ranking.
- The measures gave higher scores when used with several reference translations, but kept the ranking.

Conclusions 2

- WAFT gave higher scores than NEVA, except for major word-order reversals.
- NEVA was more sensitive to word-level errors.
- The differences could be used to single out certain error types in diagnostic evaluation.
- The differences could be used to find inconsistences in a single reference translation.

References 1

- Alshawi et al. Automatic acquisition of hierarchical transduction models for machine translation. In *Proceedings of ACL'98*, pp. 41–47, Montreal, Canada, 1998.
- Doddington. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In *Proceedings of HLT 2002*, pp. 128–132, San Diego, USA, 2002.
- 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.

References 2

- Popescu-Belis. Meta-evaluation of evaluation metrics. tentative synthesis on the evaluation exercise. Talk presented at *Workshop on Machine Translation Evaluation: Human Evaluators Meet Automated Metrics (LREC'02)*, Las Palmas de Gran Canaria, Spain, 2002.
- Sågvall Hein et al. Scaling up an MT prototype for industrial use – databases and data flow. In *Proceedings from LREC'02*, pp 1759–1766, Las Palmas de Gran Canaria, Spain, 2002