UPPSALA
UNIVERSITET

Maskinöversättning: Utvärdering

Eva Forsbom

evafo@stp.lingfil.uu.se

Översättarprogrammet

15 november 2005

UPPSALA UNIVERSITET

Strumpor eller översättningssystem?

märke & modell	Blue Forêt/ Wool & silk	Burlington/ Universal	Dressman/ Manbasic	H&M/ Wool mix socks	Hugo Boss/ Basic wool	KappAhl/ Body Zone
						
cirkapris	129 kr	119 kr	59:50 kr/2 par	39:50 kr	149 kr	39 kr
slittålig*	1 SLITS LÄTT	1	2	3	2	5 SLITSTARK
plus & minus	+ Mjuk, skön och smidig - Hård resår, blir hårdare efter tvätt - Tvätt i 30 grader	- Stor för de mindre storlekar den ska täcka. - Tvätt i 30 grader	- Stor för de mindre storlekar den ska täcka	- Hård resår	+ Tunn och smidig - Hård resår - Tvätt i 30 grader	- Liten i storleken. - Krymper en storlek efter tvätt.
material	60 % ull, 23 % silke, 17 % nylon	98 % ull, 2 % elastan SLITS LÄTT	60 % ull, 38 % nylon, 2 % elastan	70 % ull, 26 % nylon, 4 % elastan	70 % ull, 30 % nylon	60 % ull, 38 % nylon, 2 % elastan

Fotnot
* Betygsskala 1–5, där 5 är bäst

24

ICA-KURIREN 45/08

Attribut	System 1	System 2
Attribut 1	Värde 1.1	Värde 2.1
Attribut 2	Värde 1.2	Värde 2.2
Attribut 3	Värde 1.3	Värde 2.3

UPPSALA UNIVERSITET

Standardiseringsinitiativ

- ISO:s programvarukvalitet
  - EAGLES ramverk
    - skrivhjälpmedel
    - dialogsystem
    - ISLE:s taxonomi för maskinöversättning (FEMTI)
- LISA
  - kvalitetsgranskningsmodell
  - TMX, TBX, XLIFF, SRX, GIST
  - (branschöversikter)
- SAE
  - J2450

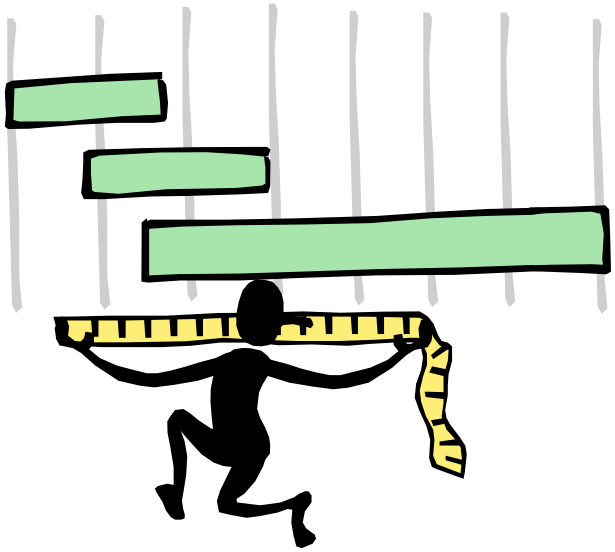
UPPSALA UNIVERSITET

ISO:s kvalitetsattribut

- 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:s 9126-serie: Produktkvalitet
- ISO/IEC:s 14598-serie: Utvärdering av programvaruprodukter
  - funktionalitet
  - pålitlighet
  - användbarhet
  - effektivitet
  - underhållbarhet
  - portabilitet

UPPSALA UNIVERSITET

Utvärderingskontexter

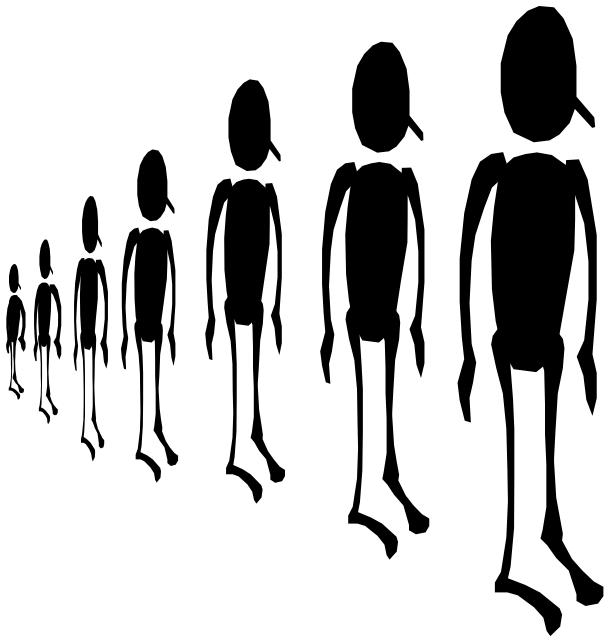


- För vem?
- Varför?
- Vad?
- Av vem?
- Hur?

UPPSALA UNIVERSITET

För vem?

Olika användare har olika behov.
Kvalitetsattributen bör väljas och viktas därefter.



- Konsument-organisation
- Projektledare
- Utvecklare
- Avancerad användare
- Konsument
- ...

UPPSALA UNIVERSITET

Varför?

Syftet med utvärderingen beror på typ av användare och produktens mognad. Exempel:

Typ	Syfte
Genomförbarhet	Se om produkten behövs/är värd att utveckla.
Diagnostik	Spåra fel.
Progression	Se ändringar mellan olika versioner.
Adekvans	Se om produkten är adekvat för en viss uppgift.
Prestanda	Jämföra olika system.

UPPSALA UNIVERSITET

Vad?

```
[funktionalitet:
  [lämplighet:sant
   korrekthet:60%
   interoperab.:hög
   säkerhet:hög
   följer std: sant]
 pålitlighet:7
 användbarhet:bra
 effektivitet:std
 underhållbarhet:god
 portabilitet:Linux]
```

Typ av användare och syfte bör bestämma på vilken detaljnivå attribut ska väljas. Viktade värden för vissa attribut kan kombineras till ett attribut på högre nivå.

UPPSALA UNIVERSITET

Av vem?

- Utvärderingsorgan
- Affärschef
- Utvecklare
- Domänexpert
- Avancerad användare
- Tvåspråkig användare
- Konsument
- ...

Olika typer av utvärderingar kräver utvärderare med olika bakgrund. Vissa utvärderingar kan utföras automatiskt; andra inte.

UPPSALA UNIVERSITET

Hur?

1. Definition av kvalitetskraven
   - kravanalys
   - utvärderingsmodellering
2. Förberedelse av utvärderingen
   - val av kvalitetsmått
   - definition av graderingsnivåer
   - definition av bedömningskriterier
3. Utförande av utvärderingen
   - mätning
   - gradering
   - bedömning

UPPSALA UNIVERSITET

FEMTI - användarbehov

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

UPPSALA UNIVERSITET

FEMTI - systemegenskaper

1. 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 of use
   - System external characteristics
     - Functionality
     - Reliability
     - Usability
     - Efficiency
     - Maintainability
     - Portability
     - Cost

UPPSALA UNIVERSITET

Svarta lådan-utvärdering

Input Overview

Words				
Words	Total	Unique		
	44107	15.58%	6874	

Segments				
Segments	Total	Unique		
	7414	63.57%	4713	

System Recall

Words				
Source Language Words	Total		Unique	
	99.15%	43730	97.70%	6716

Segments				
Fully Translated	Total		Unique	
	39.26%	2911	24.89%	1173
Translated	Total		Unique	
	42.97%	3186	29.13%	1373

UPPSALA UNIVERSITET

Glaslådeutvärdering

Error Reports

Words		
Source Language Words	Total: 377	Unique: 158
Translation Links	Total: 6622	Unique: 750
Target Language Words	Total: 180	Unique: 13
Target Language Code	Total: 70	Unique: 8

Segments		
Not Parsed	Total: 161	Unique: 141
Partially Parsed	Total: 3658	Unique: 3055
Not Transferred	Total: 250	Unique: 14
Not Generated	Total: 159	Unique: 130

UPPSALA UNIVERSITET

Utvärdering av översättningskvalitet

- Jämförelse med källtext (tvåspråkig utvärderare/automatiskt)
- Jämförelse med referensöversättning (enspråkig utvärderare/automatiskt)

- Trogenhet (hur nära)
- Korrekthet (hur korrekt)
- Adekvans (hur adekvat)
- Informationsinnehåll (hur informativ)
- Förståelighet (hur förståelig)
- Flyt (hur flytande)

UPPSALA UNIVERSITET

Manuell utvärdering

- Gradering
- Lucktest
- Läsförståelsetest
- Uppgiftsbaserat test
- Lästid
- Inskrivningstid
- Efterredigering

UPPSALA UNIVERSITET

Gradering - trogenhet

- 5 = All betydelse från källfragmentet finns kvar i det översatta fragmentet .
- 4 = Det mesta av betydelsen från källfragmentet finns kvar i det översatta fragmentet.
- 3 = Mycket av betydelsen från källfragmentet finns kvar i det översatta fragmentet.
- 2 = Lite av betydelsen från källfragmentet i källfragmentet finns kvar i det översatta fragmentet.
- 1 = Ingen betydelse från källfragmentet finns kvar i det översatta fragmentet.

UPPSALA UNIVERSITET

Resultat - trogenhet

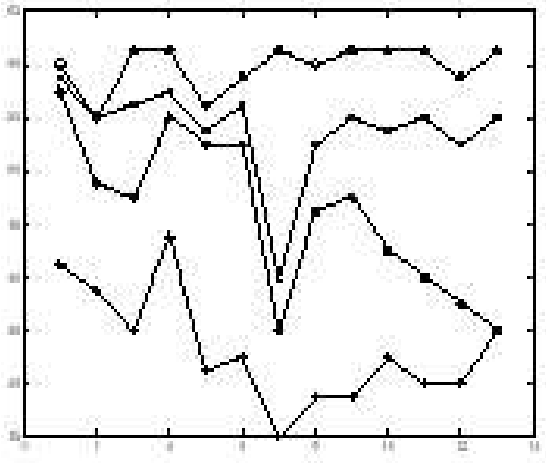
http://stp.lingfil.uu.se/~evafo/lrec_eval/

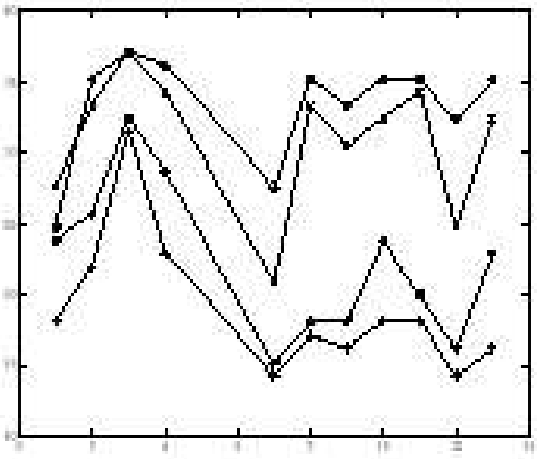
1 2 3 4 5

Source: Prévenir ses enfants des problèmes de drogue

Reference: Prevent your children from having drug problems

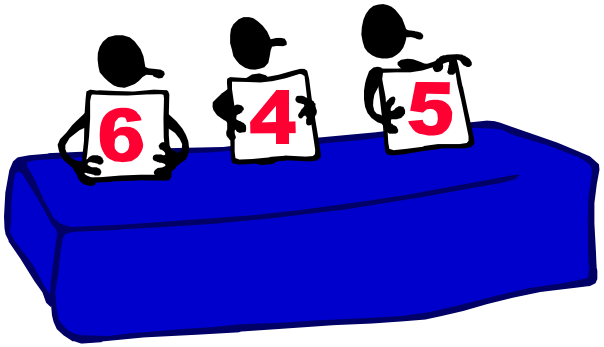
Translation: Prevent your children from drug problems



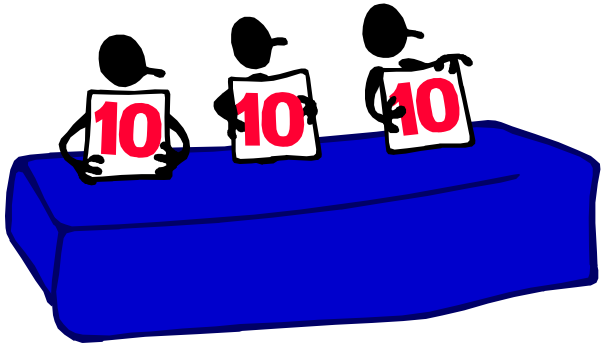


UPPSALA UNIVERSITET

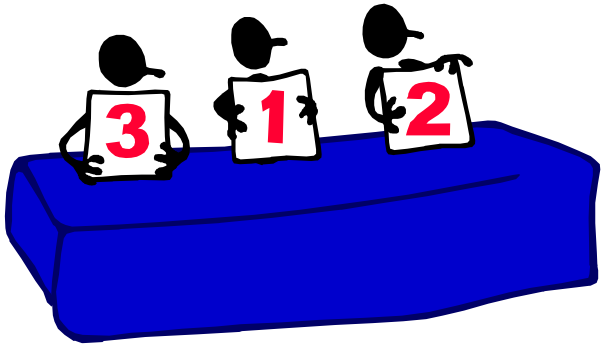
Manuell utvärdering - problem



The hat is fat.



The cat is fat.



The hat is fat.

UPPSALA UNIVERSITET

Halvautomatisk utvärdering

- Manuell uppmärkning
- Automatisk jämförelse och beräkning

- Namngivna enheter
- Syntaktisk korrekthet
- Översättning av domänspecifik terminologi
- Översättning av informationsenheter
- Skapande av testsviter

UPPSALA UNIVERSITET

SAE J2450

Category *c*	serious $w_{c,s}$	minor $w_{c,m}$
Wrong term (WT)	5	2
Syntactic error (SE)	4	2
Omission (OM)	4	2
Word structure or agreement error (SA)	4	2
Misspelling (SP)	3	1
Punctuation (PE)	2	1
Miscellaneous error (ME)	3	1

$$\text{SAE J2450} = \frac{1}{N}\sum_{c}(s_c \cdot w_{c,s} + m_c \cdot w_{c,m})$$

s_c = number of serious errors in the category c
m_c = number of minor errors in the category c
N = number of words in the source text

UPPSALA UNIVERSITET

SAE J2450 – statistiskt system

WTs SEs OMs SAs SPs PEs MEs

Vi anser också att det i detta sammanhang bör fastläggas i gemenskapens regelverk att det godkännande som den nationella myndigheten årligen senast den 15 december skall meddela beträffande en producentorganisations verksamhetsprogram för nästkommande år skall vara bindande för den nationella myndigheten och inte senare kunna ändras om det kan klarläggas att samtliga de omständigheter som avser en åtgärd är redovisade vid detta tillfälle .

WTm SEm OMm SAm SPm PEm MEm

We also believes that it in this context should fastläggas in the Community rules that the approval of the national department annually later 15 December must Parliament´s rejected a producentorganisations vegetables for the next year shall be bindande for the national department and no later be changed if there may klarläggas all those circumstances which pay an action is drafts at this occasion .

UPPSALA UNIVERSITET

SAE J2450 – regelbaserat system

WTs SEs OMs SAs SPs PEs MEs

Vi anser också att det i detta sammanhang bör fastläggas i gemenskapens regelverk att det godkännande som den nationella myndigheten årligen senast den 15 december skall meddela beträffande en producentorganisations verksamhetsprogram för nästkommande år skall vara bindande för den nationella myndigheten och inte senare kunna ändras om det kan klarläggas att samtliga de omständigheter som avser en åtgärd är redovisade vid detta tillfälle .

WTm SEm OMm SAm SPm PEm MEm

We consider also to it in this context shall fastläggas in the community's legislation to the approval as the national authority annually late the 15 December will announce concerning a producer organisation's operational programmes for the following years will product bindande for the national authority and not late can change if it can klarläggas that all the circumstances that intend a measure are presented at this occasion.

UPPSALA UNIVERSITET

SAE J2450 – resultat

RBMT	evaluation 1			evaluation 2		
c	s_c	m_c	score	s_c	m_c	score
WT	365	0	1825	361	0	1805
SE	34	0	136	33	0	132
OM	6	6	36	2	8	24
SA	11	70	184	9	75	186
SP	0	2	2	0	5	5
PE	0	1	1	2	1	5
ME	6	10	28	4	7	19
SAE J2450	overall score		0.9559	overall score		0.9404
SMT	evaluation 1			evaluation 2		
c	s_c	m_c	score	s_c	m_c	score
WT	496	0	2480	512	0	2560
SE	35	0	140	26	0	104
OM	40	9	178	29	14	144
SA	7	45	118	9	41	118
SP	1	0	3	0	1	1
PE	1	2	4	6	2	14
ME	23	8	77	29	4	91
SAE J2450	overall score		1.2706	overall score		1.2853

UPPSALA UNIVERSITET

Automatisk utvärdering

- Ungefärlig strängmatchning
- Automatisk uppmärkning och beräkning

- Redigeringsavstånd
- N-gramsamförekomst
- Antal oöversatta ord
- Namngivna enheter
- Syntaktisk korrekthet
- Översättning av domänspecifik terminologi
- Översättning av informationsenheter
- Skapande av testsviter och utvärdering med dem

UPPSALA UNIVERSITET

Redigeringsavstånd - ordnoggrannhet

$$\mathrm{WA} = \left(1 - \frac{d+s+i}{l_r}\right), \mathrm{WAFT} = \left(1 - \frac{d+s+i}{\max(l_r, l_c)}\right)$$

l_r = length of reference
l_c = length of candidate translation

Src: Tätningsring

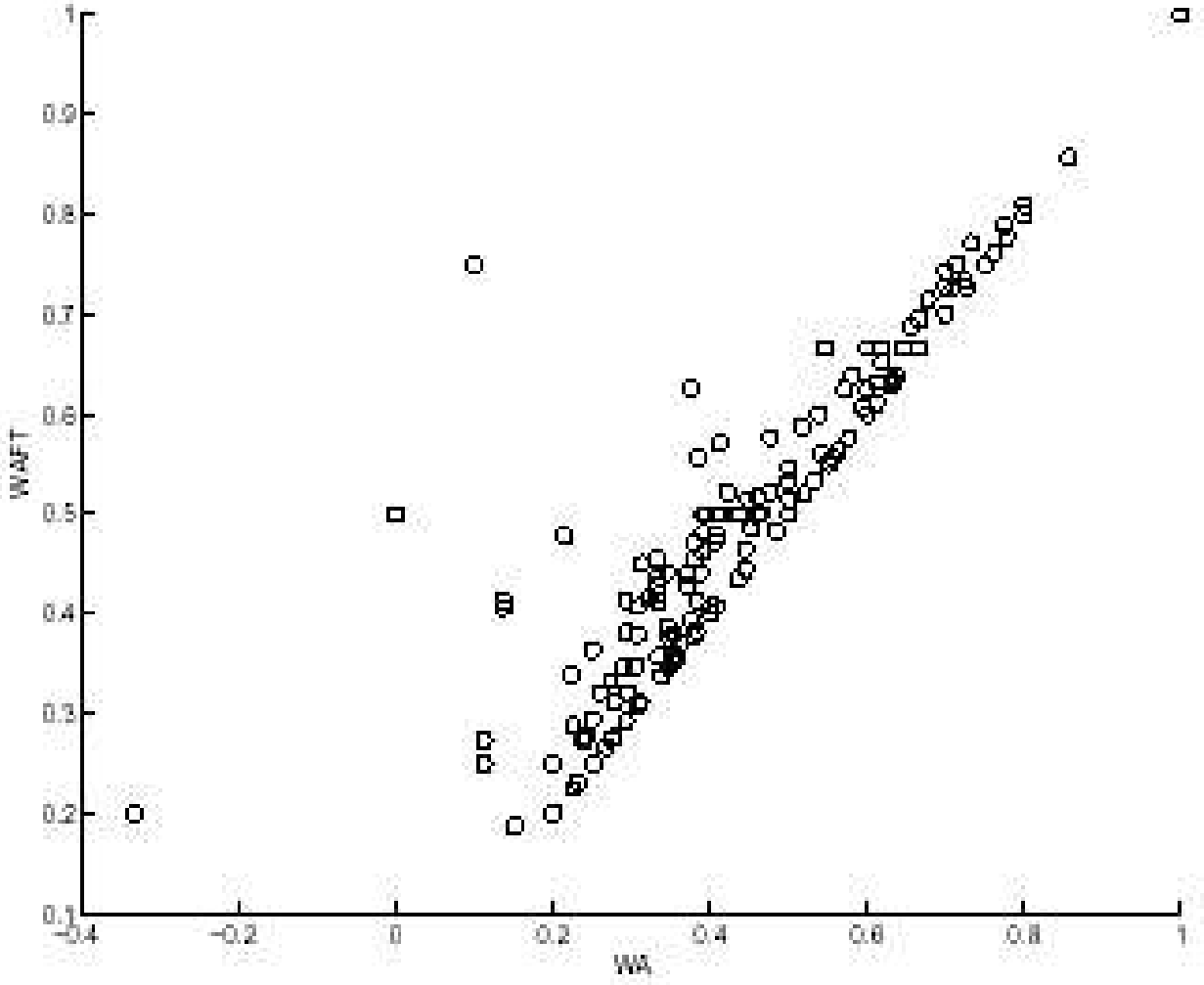
Cand: Sealing ring length = 2

Ref: Seal length = 1

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

UPPSALA UNIVERSITET

WA vs. WAFT för segment



UPPSALA UNIVERSITET

Fördelar och nackdelar

- Möjligt att länka operationer
  - *clip/clamp*
  - *tensioner/tensioners*
  - *in/into*
  - *the/∅*
  - */∅*
- Känslig för ordföljdsändringar

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

UPPSALA UNIVERSITET

N-gramsamförekomst

Kandidat: <It is a guide to action> <which> <ensures that the military> <always> obeys <the> <commands> <of the party .>

Referens1: <It is a guide to action> that <ensures that the military> will forever heed Party <commands> .

Referens2: It is the guiding principle <which> guarantees the military forces <always> being under <the> command of the Party .

Referens3: It is the practical guide for the army always to heed the directions <of the party .>

UPPSALA UNIVERSITET

BLEU (Bi-Lingual Evaluator Understudy)

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

where

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

$r =$ length of reference

$c =$ length of candidate

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

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

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

UPPSALA UNIVERSITET

BLEU – problem (⇒ NEVA)

Src: Cylinder, underdel

Cand: Bottom cylinder length $< N_{max}$

Ref: Cylinder bottom

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

Src: Ledningsnät för bränslepump

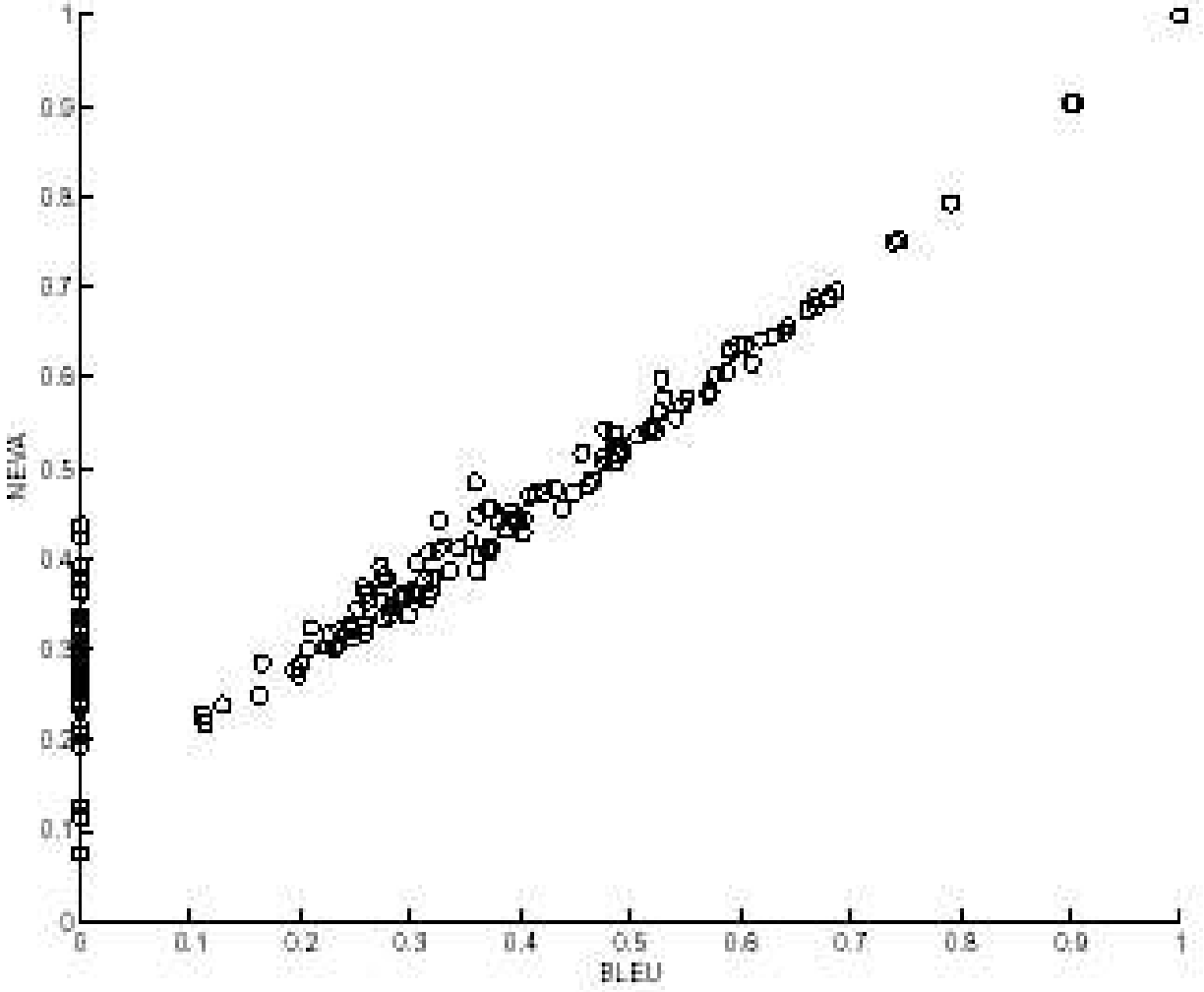
Cand: Cable harness for fuel pump no 3- or 4-grams

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 vs NEVA för segment



UPPSALA UNIVERSITET

Fördelar och nackdelar

- Alltid >0 om något är rätt
- Möjligt att få listor på saknade *n*-gram

- Känslig för ordfel (särskilt i mitten)

Src: Kontrollera backventilen.

Cand: Check the check valve.

Ref: Check the non-return valve.

UPPSALA UNIVERSITET

Antal referensöversättningar

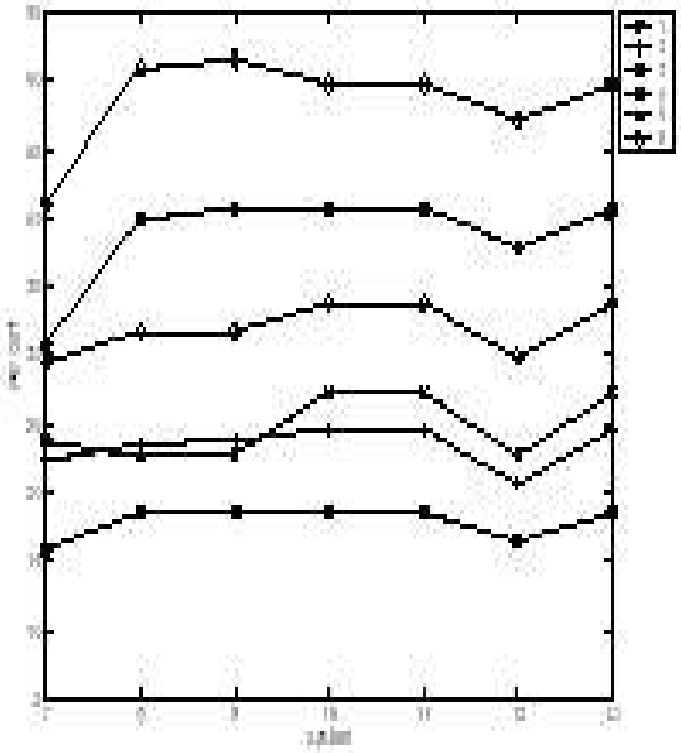
- Rankning är i princip densamma.
- Poängen ökar med antalet referensöversättningar.

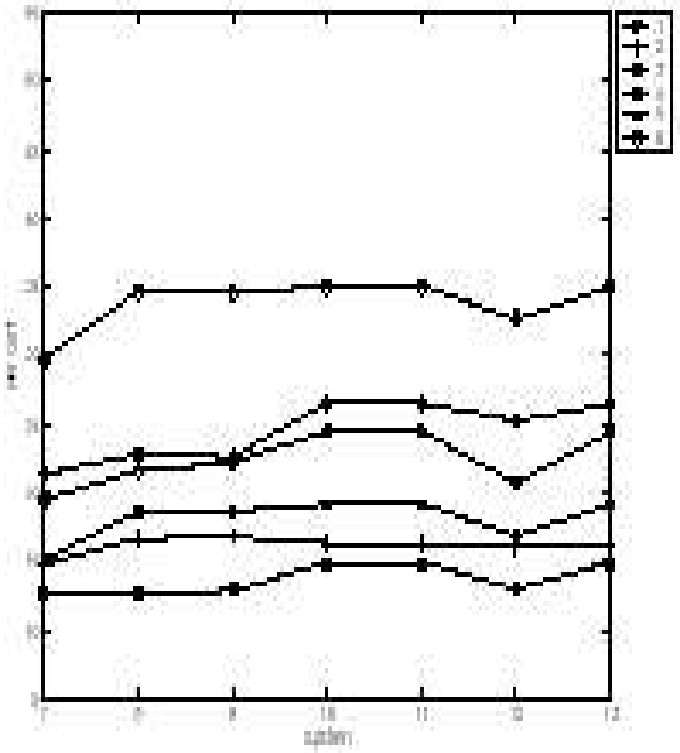
Level	WAFT	NEVA
System	0.8589	0.9857
Document 1	0.6854	0.9983
Document 2	0.9348	0.9632
Segment	0.6215	0.7274

UPPSALA UNIVERSITET

Kvalitet på referensöversättning

- Kvaliteten har betydelse för poängnivån.
- Kvaliteten har mindre betydelse för rankning.





UPPSALA UNIVERSITET

Supermodell via kloning

1) Trouble shooting
2) The fluid is cleaned by passing through a filter.
3) Failure to follow this instruction can...

1) Trouble shooting
2) The fluid is cleaned via a filter.
3) Failure to follow this instruction can...

1) Troubleshooting
2) The oil is cleaned via a filter.
3) Failure to follow this instruction can...

1) Troubleshooting
2) The oil is cleaned via a filter.
3) It can...

UPPSALA UNIVERSITET

Referenser

- Alshawi et al., 1998. Automatic acquisition of hierarchical transduction models for machine translation. I *Proceedings of ACL'98*, s. 41-47, Montreal. URL: http://acl.ldc.upenn.edu/P/P98/p98-1006.pdf.
- Doddington, 2002. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. I *Proceedings of HLT 2002*, s. 128-132, San Diego. URL: http://www.nist.gov/speech/tests/mt/doc/ngram-study.pdf.
- Doyon et al., 1998. The DARPA machine translation evaluation methodology: Past and present. I *Proceedings of AMTA'98*, Philadelphia.
- EAGLES (Expert Advisory Group on Language Engineering Standards). URL: http://issco-www.unige.ch/projects/eagles/.
- FEMTI (a Framework for the Evaluation of Machine Translation in ISLE). URL: http://www.issco.unige.ch/projects/isle/femti/.

UPPSALA UNIVERSITET

Referenser 2

- Forsbom, 2003. Training a Super Model Look-Alike: Featuring Edit Distance, N-Gram Occurrence, and One Reference Translation. I *Proceedings of the Workshop on Machine Translation Evaluation: Towards Systemizing MT Evaluation (MT SUMMIT IX)*, s. 29-36, New Orleans. URL: http://stp.lingfil.uu.se/~evafo/Papers/mtsummitixeval.pdf.
- Hovy et al. 2002. Principles of Context-Based Machine Translation Evaluation. *Machine Translation* 17(1), s. 43-75.
- ISLE (International Standards for Language Engineering). URL: http://www.issco.unige.ch/projects/isle.
- ISO (International Organization for Standardization). URL: http://www.iso.org/.
- LISA (Localization Industry Standards Association). URL: http://www.lisa.org/.
- Papineni et al., 2001. *BLEU: a method for automatic evaluation of machine translation*. Teknisk rapport IBM RC22176 (W0109-022), IBM Research Division, T. J. Watson Research Center. URL: http://domino.watson.ibm.com/library/cyberdig.nsf/ (nyckelord: RC22176.pdf).

UPPSALA UNIVERSITET

Referenser 3

- Popescu-Belis, 2003. An experiment in comparative evaluation: humans vs . computers. I *Proceedings of MT SUMMIT IX*, s. 307-314. New Orleans. URL: http://www.amtaweb.org/summit/MTSummit/FinalPapers/60-Popescu-final.pdf.
- Reeder et al., 2001. The naming of things and the confusion of tongues: an MT metric. I *Proceedings of the MT Evaluation Workshop: Who Did What To Whom (MT Summit VIII)*, s. 55-59, Santiago de Compostela. URL: http://issco-www.unige.ch/projects/isle/papersMTS/reeder-1.pdf.
- SAE (Society of Automotive Engineers). URL: http://www.sae.org/.
- Sågvall Hein et al., 2003. MATS - A Glass Box Machine Translation System. I *Proceedings of MT SUMMIT IX*, s. 491-493. New Orleans. URL: http://www.amtaweb.org/summit/MTSummit/FinalPapers/50-Sagvall-final.pdf.
- Weijnitz et al., 2004. MT goes farming: Comparing two machine translation approaches on a new domain. I *Proceedings of LREC'04*, volym VI, s. 2043-2046. Lissabon. URL: http://stp.ling.uu.se/~evafo/Papers/smtest_lrec04.pdf.