Evaluation of Machine Translation Systems

Cristina Vertan

University of Hamburg • Informatics Department

Natural Language Systems Group

WWW: http://nats-www.informatik.uni-hamburg.de/~cri/

E-Mail: cri@informatik.uni-hamburg.de

Evaluation of MT-Systems

- In contrast to other software there is no “best solution” by human translators, which can be compared with the output of the system
- I.e., for one input sentence there are many different correct translations
- Quality measurement of an MT System depends on its purposes and on the requirements of potential users.
- Possible participants in evaluation :
  - Researchers
  - Research sponsors
  - Purchasers
  - Translators

Evaluation strategies

Black Box vs. Glass Box

- MT system is seen as a black box, whose operation is treated purely in terms of its input-output behaviour
- Should not be conducted by the developers
- Tests: functionality, volume of data handled, recovery situations

- Components of the system are inspected as well a their effect in the system
- Relevant to researchers and developers
- Static analysis: checking the system without running it (automatic syntax and type checking by a compiler, manual inspection of the system, symbolic execution, data flow analysis)
- Dynamic glass box requires running the program (e.g. trying the program on many logical paths and ensuring that every logical branch is executed at least once).

Evaluation strategies

Test Suite vs. Test corpus

- Carefully constructed set of examples, each testing a particular linguistic or translation problem (e.g. different lexical and structural differences)
- Problem: it is assumed that the behaviour of a system can be projected from carefully constructed examples to real texts
- Test suite evaluations are difficult to compare

- An adequate corpus (for the domain of the system) is used as input
- Problem: it does not test systematically all possible sources of incorrect translations, but considers the most frequent constructions
- It is difficult to estimate the behaviour of the system for other types of text

Evaluation - Linguistic Quality measures

- Intelligibility - measures the fluency and grammaticality of the TL text, with concern for whether it faithfully conveys the meaning of the SL
- Accuracy - indicates how the translated text preserves the content of the source text. (a high intelligible sentence may not convey the meaning of the source text because of incorrect disambiguation)
- Error analysis : e.g. count the number of words inserted, modified, deleted and moved by a post-editor. However, deciding hat is an acceptable translation is subjective.

Intelligibility scale 1-5

(Nagao et. Al. 1988)

1. The meaning of the sentence is clear and there are no questions. Grammar, word usage, and/or style are all apropriate, and no rewriting is needed
2. The meaning of the sentence is clear but there are some problems in grammar, word usage, and/or style, making the overall quality less than 1
3. The basic thrust of the sentence is clear, but you are not sure of some detailed pars because of grammar and word usage problems. You would need to look at the original SL sentence to clarify the meaning
4. The sentence contains many grammatical and word usage problems, and you can only guess at the meaning after careful study, if at all
5. The sentence cannot be understood at all.

Intelligibility -Problems

- Difficult to preserve objectivity: what someone classifies as 1 may be classified as 2 by someone else.
- Solution:
  - Combine scores of several evaluators and compute a statistical mean
  - Cloze tests: words in the translated text are masked and the evaluators are asked to guess the masked word. The correlation between the masked word and the guessed word (PoS, semantic category, etc.) is an indication of how interlligible the text is

Accuracy Scale 1-7

(Nagao et al. 1988)

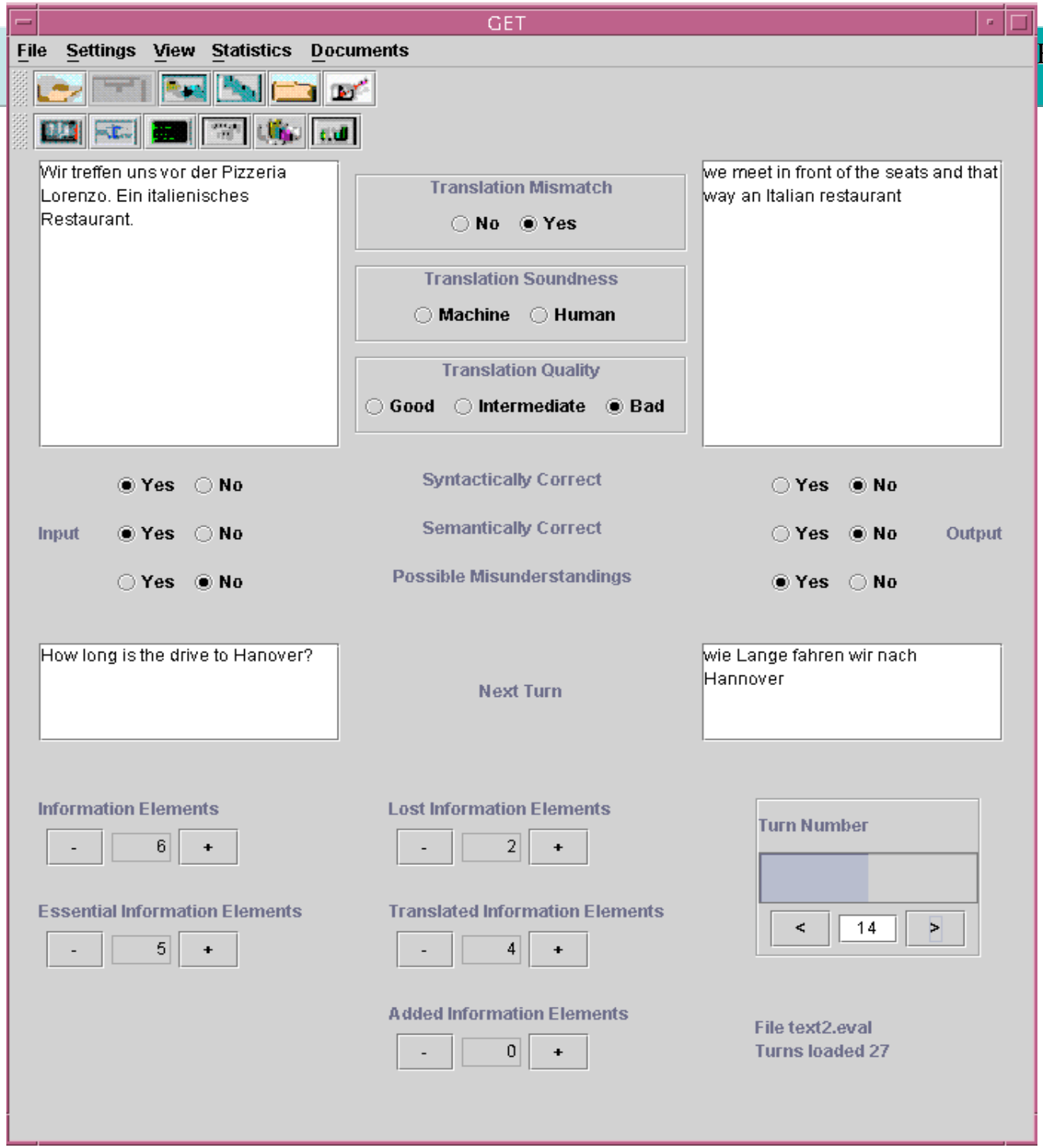
- Dependent on intelligibility. A text classified with 4 or 5 on intelligibility scale is unlikely to convey any content
- However a highly intelligible sentence may not convey the meaning of the spurce text because of incorrect disambiguation

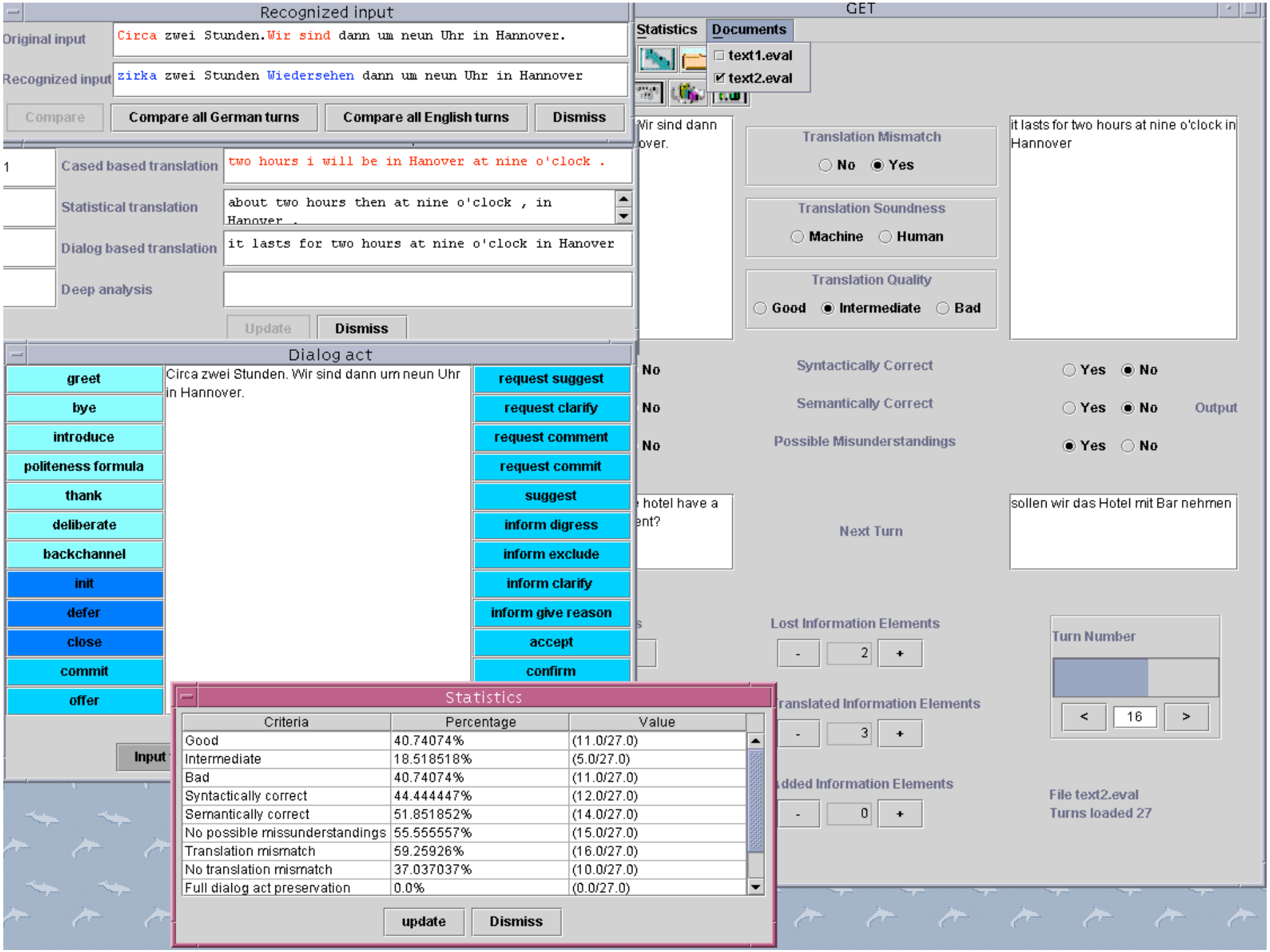
1. The content of SL sentence is faithfully conveyed to the TL sentence. The translated sentence is clear to a native speaker of the TL and no rewriting is needed
2. The content of the SL sentence is faithfully conveyed to the TL sentence, and can be clearly understood by a native speaker, but some rewriting is needed
3. The content of the SL sentence is faithfully conveyed in the TL sentence, but some changes are needed in word order
4. While the content of the SL sentence is generally conveyed faithfully in the TL sentence, there are some problems with things like relationships between phrases and expressions, and with tense, plurals, and the position of adverbs. There is some duplication of nouns in the sentence.

Accuracy Scale (cont.)

5. The content of the SL sentence is not adequately conveyed in the TL sentence. Some expressions are missing, and there are problems with the relationships between clauses, between phrases and clauses, or between sentence elements.
6. The content of the SL sentence is not conveyed in the TL sentence
7. The content of the SL sentence is not conveyed at all. The output is not a proper sentence; subjects and predicates are missing

- Again there is the problem of objectivity. In Interlingua Systems one solution could be to use also the back-translation


GET
File Settings View Statistics Documents
Wir treffen uns vor der Pizzeria Lorenzo. Ein italienisches Restaurant.
Translation Mismatch
No Yes
Translation Soundness
Machine Human
Translation Quality
Good Intermediate Bad
we meet in front of the seats and that way an Italian restaurant
Yes No
Input
Yes No
Yes No
Syntactically Correct
Semantically Correct
Possible Misunderstandings
Yes No
Yes No
Output
Yes No
How long is the drive to Hanover?
Next Turn
wie Lange fahren wir nach Hannover
Information Elements
- 6 +
Lost Information Elements
- 2 +
Turn Number
Essential Information Elements
- 5 +
Translated Information Elements
- 4 +
< 14 >
Added Information Elements
- 0 +
File text2.eval
Turns loaded 27


Recognized input
Original input
Circa zwei Stunden.Wir sind dann um neun Uhr in Hannover.
Recognized input
zirka zwei Stunden Wiedersehen dann um neun Uhr in Hannover
Compare
Compare all German turns
Compare all English turns
Dismiss
1
Cased based translation
two hours i will be in Hanover at nine o'clock .
Statistical translation
about two hours then at nine o'clock , in Hanover .
Dialog based translation
it lasts for two hours at nine o'clock in Hanover
Deep analysis
Update
Dismiss
Dialog act
greet
bye
introduce
politeness formula
thank
deliberate
backchannel
init
defer
close
commit
offer
Circa zwei Stunden. Wir sind dann um neun Uhr in Hannover.
request suggest
request clarify
request comment
request commit
suggest
inform digress
inform exclude
inform clarify
inform give reason
accept
confirm
Input
Statistics
Criteria
Percentage
Value
Good
40.74074%
(11.0/27.0)
Intermediate
18.518518%
(5.0/27.0)
Bad
40.74074%
(11.0/27.0)
Syntactically correct
44.444447%
(12.0/27.0)
Semantically correct
51.851852%
(14.0/27.0)
No possible missunderstandings
55.555557%
(15.0/27.0)
Translation mismatch
59.25926%
(16.0/27.0)
No translation mismatch
37.037037%
(10.0/27.0)
Full dialog act preservation
0.0%
(0.0/27.0)
update
Dismiss
GET
Statistics
Documents
text1.eval
text2.eval
Wir sind dann over.
Translation Mismatch
No
Yes
Translation Soundness
Machine
Human
Translation Quality
Good
Intermediate
Bad
it lasts for two hours at nine o'clock in Hannover
No
No
No
Syntactically Correct
Semantically Correct
Possible Misunderstandings
Yes
No
Yes
No
Yes
No
Output
hotel have a ent?
Next Turn
sollen wir das Hotel mit Bar nehmen
Lost Information Elements
-
2
+
Translated Information Elements
-
3
+
dded Information Elements
-
0
+
Turn Number
<
16
>
File text2.eval
Turns loaded 27

Evaluation - automatic measures

- Try to reduce the evaluation costs and to make them objective
- Compare the system‘s output with reference translations
- Different statistical measures are use to evaluate the distance between the output string and the reference translation: e.g. number of identical words, number of insertion/deletions, total number of words
- Most well-known measures: NIST, BLEU, developed at IBM

String based metrics

Two Methods both based on string manipulation

- Simple String Accuracy,

SSA = 1 – (I + D + S / |Reference Sentence|)

- Generation String Accuracy

GSA = 1 – (M + I + D + S/ |Reference Sentence|)

No evaluation of relationships between words

BLEU

•BiLingual Evaluation Understudy
•Bleu is an IBM-developed metric and is probably the best known and most used in the Machine Translation community
N-gram = sequence of n words (usually one counts uni- bi- maximum trigrams)

• It is based on two factors:
1. An n-gram count between Result and References
2. A brevity penalty if the |Result| < |References|

$$p_n = \frac{\sum_{c \in Candidate} \sum_{n\text{-}gram \in c} Count_{clip}(n\text{-}gram)}{\sum_{c \in Candidate} \sum_{n\text{-}gram \in c} Count(n\text{-}gram)}$$

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

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

Automatic measures - Problems

- Disadvantages:
  - Require preprocessing steps which can modify the results for some languages (e.g. every word is lower-cased)
  - Difficult to apply for spoken or spoken-like input, because normally there is no reference translation-corpus
  - Cannot evaluate translations which preserve the meaning but concise for example the expression.
  - What to do with synonym words ?

Evaluation: Linguistisc Quality (general for all NLP Systems)

- Coverage
  - Lexikon
  - Syntax
  - Semantik
- Pragmatics
- Compatibility
- Data formats
- Languages
- Domains

Evaluation - Software criteria

- Functionality - determines the degree to which it fulfils the stated or implied needs of a user
- Reliability - if the system maintains its level of performance under specified conditions and for a specified period of time
- Usability - indicates the effort needed to use the software by a stated or implied set of users
- Efficiency - relationship between the level of performance of the software and the amount of resources used to achieve that level of performance under specified conditions
- Maintainability - effort needed to make specified modifications to the software
- Portability - indicates the ability of the software to be transferred from one environment to another.