MT Eval – Hands-on exercise

Preliminary results and comments

Widad Mustafa El Hadi /Marianne Dabbadie
EVALING / University of Lille 3 – IDIST / CERSATES

Human evaluation: H4

- Georges Van Slype (1979)
  - It is a critical study of methods for evaluating the quality of Machine Translation. Final Report, Bureau Marcel van Dijk / European Commission, Brussels

Measuring intelligibility

- What is intelligibility?
- According to ISLE taxonomy:
  - Intelligibility / comprehensibility - how intelligible is the output under different conditions (e.g, are the sentence fragments translated while being entered). Comprehensibility reflects the degree to which a complete translation can be understood (whereas the intelligibility is based on the general clarity of translation, whether this is considered in its entirety or by segments out of context).

Widad Mustafa El Hadi - Marianne Dabbadie - LREC 2002

Measuring intelligibility

- It is a subjective measure
- It is a user oriented metric
- It gathers in fact two metrics :
  - Syntactic correctness
  - Semantic correctness
- Sentence level rating

Widad Mustafa El Hadi - Marianne Dabbadie - LREC 2002

Measuring intelligibility

- Measures intelligibility on a 0 to 3 scale
  - 3: Very intelligible: all the content of the message is comprehensible, even if there are errors of style and/or of spelling, and if certain words are missing,or are badly translated, but close to the target language
  - 2: Fairly intelligible:the major part of the message passes
  - 1: Basely intelligible:a part only of the content is understandable, representing less than 50% of the message
  - 0: Unintelligible: nothing or almost nothing of the message is comprehensible

Preliminary comments

- Chosen text : Taliban and women
  - Source text : French
  - Translation : English
  - 15 sentences in source text (including title and bibliographic information)

Preliminary comments

- Type of text: discourse
- Possible problem: coordination of long source sentences
- In reference translation:
  - sentences n°1 was split into 3 short sentences
  - Sentence n°3 was split into 2 short sentences

Each time this phenomenom occured in a translation i.e. : 1 source sentence=X target sentences these sentences were rated as one single entity

Widad Mustafa El Hadi - Marianne Dabbadie - LREC 2002

Preliminary comments

- Untranslated words were rated as unintelligible translations
- Examples : intelligibility - NP : l'ordre Taliban
- Reference translation : the Taliban order – either :
  - **Not translated**
    - translation n°1 : the Taliban
    - Translation n°12: The order
  - **Acceptable** : the Taliban rules (translation n°2)
  - **Unacceptable** :
    - the order Taliban (n°8/n°11)
    - The command Taliban (n°10)

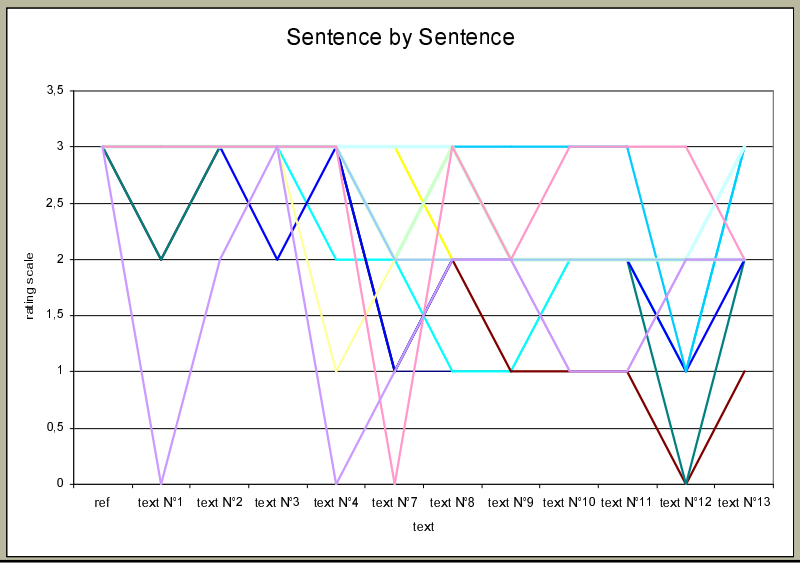
Widad Mustafa El Hadi - Marianne Dabbadie - LREC 2002

Preliminary comments

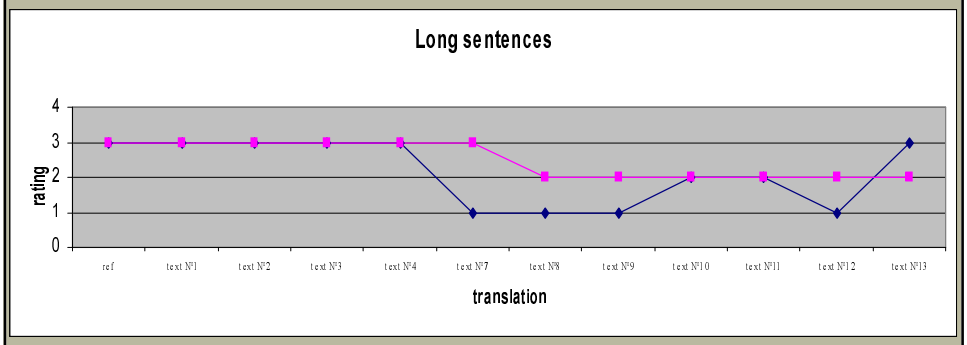
- Short sentences generally intelligible
- Translations n°7 and n°13:
  - word to word
  - Untranslated words
  - Lexical incorrectness
    - (e.g.: imposé:taxed / L'ensemble de la société : the body of the corporation)
- translations n°8 and n°9
  - Long sentences were not split
  - Several suggested translations
  - Probably MT

Widad Mustafa El Hadi - Marianne Dabbadie - LREC 2002

Preliminary results

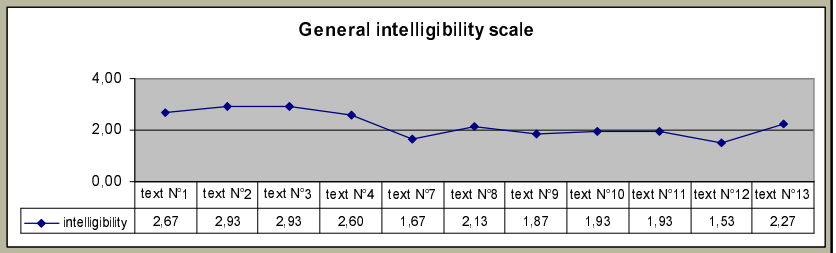


Preliminary results



Widad Mustafa El Hadi - Marianne Dabbadie - LREC 2002

Preliminary results



	text N°1	text N°2	text N°3	text N°4	text N°7	text N°8	text N°9	text N°10	text N°11	text N°12	text N°13
intelligibility	2,67	2,93	2,93	2,60	1,67	2,13	1,87	1,93	1,93	1,53	2,27

Widad Mustafa El Hadi - Marianne Dabbadie - LREC 2002

A6: any other ideas ?

- Yes – we had an idea

Aim: introduce the idea and consider it together

  - This exercise is the result of our previous studies at the hands-on machine translation workshops (2001 Geneva workshop (article for MT Summit – further work presented at the poster session during this conference)

Widad Mustafa El Hadi - Marianne Dabbadie - LREC 2002

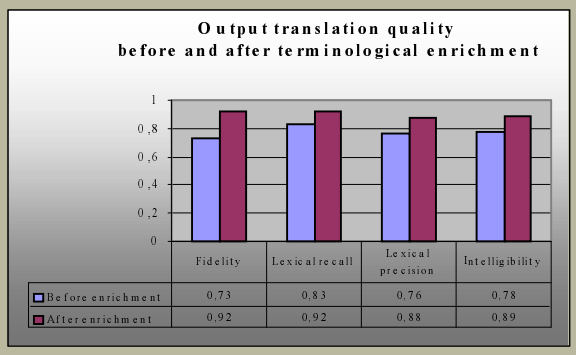
A6: Any other ideas

- Geneva 2001: work on a translation made by Reverso on the INRIA corpus on bio-technologies
- we had chosen to count the number of *NPs* (noun phrases) and *VPs* (verb phrases) in source text and target texts, a first indication being given by non parallel data

Widad Mustafa El Hadi - Marianne Dabbadie - LREC 2002

A6: any other ideas

- Another study, presented the results on the same corpus after terminological enrichment



Widad Mustafa El Hadi - Marianne Dabbadie - LREC 2002

A6: any other ideas?

- But this methodology is still imprecise and limited to a first indication of MT system's analysis failure, when a gap is observed on non parrallel data.

Widad Mustafa El Hadi - Marianne Dabbadie - LREC 2002

A6: any other ideas?

- Nevertheless, the use of finer grained criteria such as adjectives or prepositional phrases count could also be envisaged.
- A methodology including a test tool that would implement source and target transfer rules might probably prove more accurate and also apply to non isomorphic languages.

Any idea?

Widad Mustafa El Hadi - Marianne Dabbadie - LREC 2002


Thank you very much…
Widad Mustafa El Hadi - Marianne Dabbadie - LREC 2002