Evaluating an MT French / English System

Widad Mustafa El Hadi
Ismaïl Timimi
Université de Lille III

Marianne Dabbadie
LexiQuest - Paris

Evaluating an MT French / English System

Context : evaluation of a translation made by an MT System

- source text : 604.rtf : corpus INRA – corpus biotechnologique sur la reproduction chez l'animal
- Source language : French
- Target language : English

Available tools :

- French / English MT System
- French and English index of all specific words
- Original Indexes are not aligned

Object :

set a methodology for non interactive machine translation evaluation

Goal of translation :

simple understanding of original message (veille)

Evaluating an MT French / English System

Evaluation procedure:

Prequisites:

We cannot carry out verification because we do not have the system specifications

- *We can only carry out evaluation from user point of view*
- *We have no reference translation, no gold standard.*
- *French speaker intuitive correctness is our standard*

Evaluating an MT French / English System

Basis of the evaluation :

Gather numeric relevant data in order to identify
Problems and incorrections that can be analyzed later

On big corpora numeric automated evaluations are the only
Are the only efficient way to identify bugs and possible theoretical
Weaknesses of a system.

Numeric data:

- *Number of words in the source text = 562*
- *Number of unknown words for MT system = 35*
- *Number of NPs in source text and target text*
- *Number of VPs in source text and target text*

Question : should source and target number of NPs and VPs necessarily be equal ? Let us assume that this is the case and check

Evaluation criteria

Correction rate = grammatical correctness

As will be seen in further slides…

Informativeness (defined as characteristics of the translation process – output characteristics - quality of translation - quality of a text as a whole)

Is the text understandable ?

Basic criteria to work out informativeness rates:

- General language word level : corresponds to two categories (simple lexical morphemes or simple grammatical words.
- Polysemous words resolution : does the system suggest the right equivalent
- Segmentation problems
- Fluency problems (non idiomatic expressions – will be explained but no numeric data given because we assume that MT goal in this case is limited to information

Metrics

Informativeness

Rates should be calculated for each criterion

- For each criterion work out:
  - The total number of words corresponding to this criterion in source and target text and work out precision levels

Correction rate

Detailed in next slide along with results…

Calculate an average of precision rates assigned to all criteria to obtain general informativeness rate

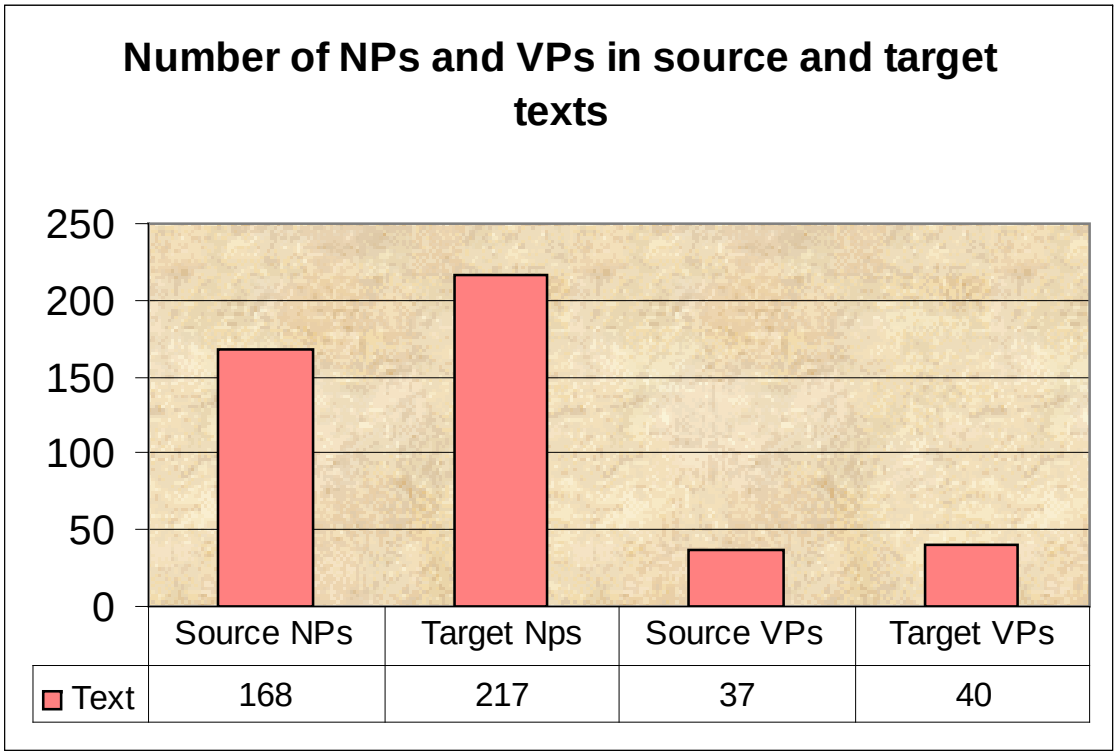
Adequacy will be an average of correction rate + informativeness

Generate numeric data to calculate correction rate

Sentence	Source NPs	Target Nps	Source VPs	Target VPs
1	8	10	3	4
2	11	10	1	3
3	5	9	2	3
4	9	9	3	3
5	9	9	1	1
6	7	6	1	1
7	9	17	2	2
8	7	9	2	2
9	7	10	1	2
10	9	18	1	2
11	10	11	2	1
12	9	13	1	1
13	9	10	3	2
14	6	6	2	2
15	6	13	1	1
16	6	9	2	2
17	15	15	3	3
18	10	13	2	2
19	6	6	2	1
20	10	14	2	2
	168	217	37	40
TOTAL		49		3

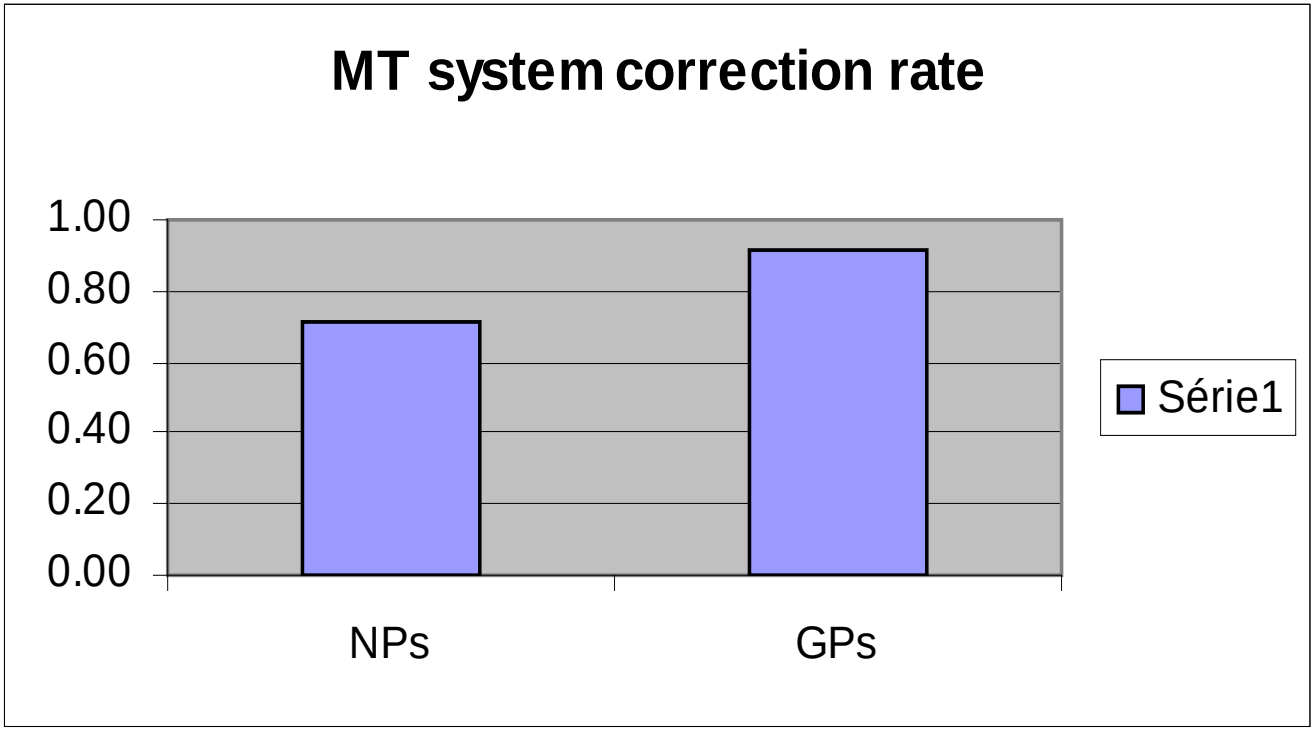
Input data to calculate MT system correction rate

After a previous source and target text tagging with LATL bilingual parser



Calculating MT system correction rate

$$1 - \frac{\text{(Number of target NP – source NPs)}}{\text{Number of source NPs}}$$



Further work (*when back to Paris…)*

Checking grammatical correctness

•Wherever there appears a difference between number of NPs and VPs

Finer grained analysis that includes adjectives

•Try to give a clear diagnostic of any problem (semantic or syntactic)

•Generate adequacy rates along with analysis

Thank you very much…

Now Widad will tell you how to generate informativeness data…