Gathering empirical data to evaluate MT from English to Portuguese
Diana Santos, Belinda Maia & Luís Sarmento

Context for this paper
• Translation teaching at FLUP
• Linguateca’s evaluation contest initiative for Portuguese processing tools;
  – MT with emphasis on Portuguese

Objective in presenting CorTA here
• Rather than present a final corpus, we are more interested in the methodology used to create it

The problem we address
• Every one (laymen) judges translation quality
• But there is no consensual evaluation method
• Test-suites are artificial and examine syntactic structure with only a few lexical items
• Our approach is a small contribution to a difficult problem
• BUT it allows experimentation with a large body of judgements about (machine) translations
• In a nutshell, we want to make use of people’s judgements over translation of real text

Our (partial) solution is to:
• Try to collect a sizeable set of judgements
• Organize them systematically
• Use real text (the source text has not been created to evaluate the systems)
• Observe real MT systems in use
• Create a parallel corpus, CorTA, using a judgement collecting tool > TrAva

CorTA
(Corpus de Traduções Avaliadas)
• A parallel corpus annotated also with translation judgements
• It contains ~1000 sentences
• + 4 translations by 4 different MT systems
• Often a human translation as well
• Encoded in both IMS-CWB and XML
A screenshot of CorTA

The internal data of CorTA

- Each source sentence is associated with
  - The observation window (ex. N N A)
  - The number of errors
  - A human translation
  - The origin of the sentence
- Each target sentence (one per translation engine) is associated with
  - The kind(s) of error(s)
  - A comment

An example

BNC: The proposed new independent status would mean a split in the party.
Padrão POS Frase Original: AT0 AJ0 AJ0 AJ0 NN0
FT: O novo estado independente proposto quereria dizeria um fende no partido.
Sy: O status independente novo proposto significaria um split no partido.*
Am: Os estados independentes novos propostos significariam uma divisão na festa.
Número de traduções erradas: 2
Problemas Identificados na Tradução:
  Substantivos: Escolha Lexical (lexical choice of noun)
  Concordância: Número (agreement in number)

CorTA: a deeper characterization

- Kinds of errors
- Source sentence origin (Figure 2 in the paper)
- Sentence size
- Number of translation errors per sentence
- Kind of window
- Kind of errors per window

Distribution of the origins of English sentences

Kinds of errors

- Verbs: Lexical choice 357
- Nouns: Lexical choice 156
- Adjectives: Lexical choice 99
- Prepositions: Lexical choice 96
- Verbs: Simple tenses 78
- Ordering: Inside the NP 72
- Adverbs: Lexical choice 64
- Verbs: Compound tenses 60
- Ordering: Inside the clause 48
- POS disambiguation 47
- Determiners: Lexical choice 39
- Verbs: Verbal expressions 38
- Determiners: Articles 35 etc. out of X errors in 1147 sentences
Errors per sentence

1270

Kind of source window

<table>
<thead>
<tr>
<th>Kind of source window</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>VBB TO0</td>
<td>82</td>
</tr>
<tr>
<td>VM0</td>
<td>30</td>
</tr>
<tr>
<td>A10 A10 A10 NN0</td>
<td>18</td>
</tr>
<tr>
<td>NN1-VVG</td>
<td>16</td>
</tr>
<tr>
<td>VBB</td>
<td>11</td>
</tr>
<tr>
<td>A10 A10 NN0</td>
<td>10</td>
</tr>
<tr>
<td>A10-NN1</td>
<td>10</td>
</tr>
<tr>
<td>VM0 PNP VVI CIS</td>
<td>8</td>
</tr>
<tr>
<td>PNP VM0 AJC VVI</td>
<td>7</td>
</tr>
<tr>
<td>PRP PNI NN1</td>
<td>5</td>
</tr>
<tr>
<td>A10 A10 yet A10 NN0</td>
<td>4</td>
</tr>
</tbody>
</table>

Problems with the source window

- **PNP VVD PNQ A10 A10**
  - He remembered her veiled brown gaze. (COMPARA)

- **can assure PNP A10 NN1**
  - But I can assure you appropriate action will be taken when we have found the guilty party. (Web)

- **can afford TO0 VVD**
  - nobody can afford to get stuck in that way of thinking. (ASC)

Non-trivial search in CorTA

- What search engine performs best for NN?
- What kind of window displays least errors?
- Which errors were discovered for a particular kind of window?
- In which cases did all systems fail (or succeed)?
- Which windows displayed a particular error kind?

The underlying collecting tool: TrAva: Traduz e Avalia

- This system collected the judgements
- Collective gathering by students and interested researchers
- Four step decision
  - The window
  - How many errors
  - Error classification
  - Manual translation and/or comment

TrAva screenshot
Problems in using TrAva

- The people employed had no experience, either in systematization or evaluation
- There was no off-the-shelf taxonomy for E-to-P
- People tried to compare systems instead of assess them
- Errors were very hard to classify (source of the problem?, resulting problem?, ...)
- People kept assuming internal system behaviour
- Several versions of TrAva – but it is still developing and far from perfect

Preliminary results

- CorTA allows complex searches
- The data can (and should) still be improved (reclassifying / refining / adding / removing)
- Open system (anyone can use it for this language pair)
- We were able to do considerable training
- A cooperative endeavour

Future work

- Study inter-judgment agreement, by asking several subjects to reclassify sentences
- Mark also (whenever at all possible) the corresponding windows in translation
- Store linguistic variant of the judge (Brazil, Portugal, etc.), genre of the source text, ...
- Study in which cases the judgement is variant dependent (is it an error or not?)
- Machine-assisted generation of test-suites
- Comparing with COMPARA’s human translations
- Study the overlap between several machine translations
Please have a look and experiment with…

- [http://www.linguateca.pt/CorTA/](http://www.linguateca.pt/CorTA/)
- [http://www.linguateca.pt/TrAvA/](http://www.linguateca.pt/TrAvA/)

and there is also METRA, an meta-MT server which is very useful

and Boomerang! for entertainment