Evaluating machine translation in a low-resource language combination: Spanish-Galician

María do Campo Bayón
Grup Tradumàtica
Dept. de Traducció i d’Interpretació i d’Estudis de l’Àsia Oriental
Universitat Autònoma de Barcelona
maria.docampo@e-campus.uab.cat

Pilar Sánchez-Gijón
Grup Tradumàtica
Dept. de Traducció i d’Interpretació i d’Estudis de l’Àsia Oriental
Universitat Autònoma de Barcelona
pilar.sanchez.gijon@uab.cat
Context

- This paper belongs to a bigger investigation for a Master's Dissertation (Official Master's Degree in Tradumatics: Translation Technologies)
- This work has been supported by the ProjecTA-U project, Grant Number FFI2016- 78612-R (MINECO/FEDER/UE)

Motivation
- Evaluating if the translation industry seeming preference for NMT can be applied to low-resource languages
- Defining the available possibilities in Spanish-Galician MT
Aim

• **Main aim:**
  • Determining which type of MT system (RBMT, PBMT or NMT) is perceived as more adequate in the context of a minoritized language such as Galician in a MT+PE workflow.

• **Specific aims:**
  • BLEU automatic evaluation.
  • Human evaluation (quality perception survey conducted among experienced professional post-editors)
  • Error analysis framework (MQM)
Background: MT evaluation

- Shterionov et al. (2018) show that a few translators see NMT as a booster of their productivity
- Some translators even see (N)MT as a handicap for their productivity while others perceive it the other way around (Sánchez-Gijón et al. 2019)
- Castilho et al. (2017) conclude that raw NMT segments may not be preferred by translators. Compared to PBMT, NMT represents a step forward but it implies also some limitations
- The same idea of strengths and weaknesses on NMT with respect to PBMT can be found in Popovic, 2017
Background: MT in Galician

- Low-resource language (technological and data resources)
- Enormous effort, mainly from the university community, to develop NLP resources and compile corpora (GalNet, SemCor, Freeling, LinguaKit...)
- There are also some MT systems specifically created for Galician (Apertium)
- Clear lack of NLP and DL tools
Methodology

• Object of study

• Perceived quality of a specialized text processed with the Google Translate, Apertium and ModernMT v. 2.5 engines evaluated using 3 different metrics:
  ▪ BLEU
  ▪ Professional ES>GL post-editors
  ▪ Error Analysis

Evaluating machine translation in a low-resource language combination: Spanish-Galician
Methodology

• Process

• Part I (preparation of materials):
  • Selection of the text of study from a professional context and definition of the human evaluation participants’ profile
    • Marketing text from a CPSL energy client
    • 15 professional post-editors from CPSL database
  • Customization of the Apertium engine
    • Addition of 30 lemmas from client’s terminology
  • Creation of a new PB engine in MMT
    • Client’s TM cleaning and extraction of bilingual corpora in Moses format (Olifant y Rainbow): 4315 units
    • Compilation of a general corpora Spanish-Galician: 6 million of words
  • Processing of the text in the three engines
  • Design of the survey on the perception of the quality (Ranking inspired by DQF + good/bad to post-edit binary response)
Methodology

• Process

• Part II (evaluation and comparison):
  • BLEU automatic evaluation: quantitative data
  • Human evaluation by means of a survey: quantitative data
  • Error analysis following the MQM main error categories: qualitative data
  • Separated analysis on short and long segments
Data analysis: BLEU score (I)

- Whole document evaluation: 30 segments

<table>
<thead>
<tr>
<th></th>
<th>NMT</th>
<th>PBMT</th>
<th>RBMT</th>
</tr>
</thead>
<tbody>
<tr>
<td>BLEU</td>
<td>43.71</td>
<td>75.27</td>
<td>76.39</td>
</tr>
</tbody>
</table>
Data analysis: BLEU score (II)

- Long segments evaluation:

<table>
<thead>
<tr>
<th></th>
<th>NMT</th>
<th>PBMT</th>
<th>RBMT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Segment 3</td>
<td>59.04</td>
<td>85.01</td>
<td>89.88</td>
</tr>
<tr>
<td>Segment 4</td>
<td>32.44</td>
<td>84.38</td>
<td>73.75</td>
</tr>
<tr>
<td>Segment 17</td>
<td>43.67</td>
<td>85.34</td>
<td>82.16</td>
</tr>
<tr>
<td>Segment 18</td>
<td>45.44</td>
<td>81.93</td>
<td>68.42</td>
</tr>
<tr>
<td>Segment 19</td>
<td>32.64</td>
<td>62.77</td>
<td>66.05</td>
</tr>
</tbody>
</table>
Data analysis: BLEU score (III)

- Short segments evaluation:

<table>
<thead>
<tr>
<th></th>
<th>NMT</th>
<th>PBMT</th>
<th>RBMT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Segment 5</td>
<td>21.02</td>
<td>35.36</td>
<td>35.36</td>
</tr>
<tr>
<td>Segment 6</td>
<td>35.36</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>Segment 7</td>
<td>37.99</td>
<td>45.18</td>
<td>45.18</td>
</tr>
<tr>
<td>Segment 20</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>Segment 21</td>
<td>63.89</td>
<td>63.89</td>
<td>63.89</td>
</tr>
<tr>
<td>Segment 22</td>
<td>64.32</td>
<td>15.21</td>
<td>64.32</td>
</tr>
<tr>
<td>Segment 23</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>Segment 24</td>
<td>66.87</td>
<td>66.87</td>
<td>66.87</td>
</tr>
<tr>
<td>Segment 25</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>Segment 30</td>
<td>60.65</td>
<td>60.65</td>
<td>60.65</td>
</tr>
</tbody>
</table>
Data analysis: survey (I)

- Quality perception
  1. Usable/not usable segments
  2. Ranking

Usable vs. non usable segments

Global ranking
Data analysis: survey (II)

• Long segments

Usable vs. not usable segments over 30 words

Long segments ranking

Evaluating machine translation in a low-resource language combination: Spanish-Galician
Data analysis: survey (III)

- Short segments

<table>
<thead>
<tr>
<th></th>
<th>RBMT</th>
<th>PBMT</th>
<th>NMT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Affirmative answers</td>
<td>93.3%</td>
<td>93.3%</td>
<td>60%</td>
</tr>
<tr>
<td>percentage</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Usable?</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
</tbody>
</table>

Short segment ranking

Evaluating machine translation in a low-resource language combination: Spanish-Galician
Evaluating machine translation in a low-resource language combination: Spanish-Galician

List of MQM error categories focused on:

- Accuracy
  - Addition
  - Mistranslation
  - Omission
  - Untranslated

- Fluency
  - Ambiguity
  - Unintelligible
  - Spelling
  - Typography
  - Grammar
    - Word-form
      - Part-of-speech
    - Agreement
    - Tense-mood-aspect
  - Word-order
  - Function-words

- Terminology
  - Inconsistent with termbase
  - Inconsistent with domain

- Style
  - Register
  - Awkward
  - Unidiomatic
Data analysis: error evaluation

Error Analysis

Error percentage / segment (%)

Whole document | Short segments | Long segments

RBMT | PBMT | NMT
Data analysis: error evaluation

Error Types

Evaluating machine translation in a low-resource language combination: Spanish-Galician
Conclusions

• Although NMT seems promising in frequent language combinations, especially if English is involved, it is not obtaining the desired results in low-resource languages such as the pair Spanish-Galician.

• NMT has not yet unseated RBMT and PBMT, performing, in fact, worse than these systems.

• There are enough resources for translators to use

• More tests should be done to replicate results and evaluate special needs to have a competitive NMT.

• Any future lines of investigation in MT and minoritized languages should be focused on searching and optimizing NLP and text resources.
Evaluating machine translation in a low-resource language combination: Spanish-Galician

María do Campo Bayón
Grup Tradumàtica
Dept. de Traducció i d’Interpretació i d’Estudis de l’Àsia Oriental
Universitat Autònoma de Barcelona
maria.docampo@e-campus.uab.cat
@Maria_doCampoB

Pilar Sánchez-Gijón
Grup Tradumàtica
Dept. de Traducció i d’Interpretació i d’Estudis de l’Àsia Oriental
Universitat Autònoma de Barcelona
pilar.sanchez.gijon@uab.cat
@psgijon