Hjerson: An Open Source Tool for Automatic Error Classification

Maja Popović
MT Marathon 2011 (Trento, Italy)
06 September 2011
standard automatic evaluation metrics (BLEU, TER, METEOR) do not provide answers on questions such as:

- what is a particular strength/weakness of the system?
- what does a particular modification exactly improve?
- does a worse-ranked system outperform a better-ranked one in any aspect?

⇒ human error analysis and classification have become widely used in recent years for these purposes (e.g. Vilar^+ 06, Farrús^+ 09)
  - human evaluation is resource-intensive and time-consuming

⇒ automatic methods are needed
After identifying actual words contributing to the:
- Levenshtein distance \( \text{WER} \)
- reference position-independent error rate \( \text{RPER} \)
- hypothesis position-independent error rate \( \text{HPER} \) (Popović & Ney 07)

the following five error classes based on (Vilar et al. 06) are defined:

- inflectional error: full form is an \( \text{RPER} \) or \( \text{HPER} \) error, base form is correct
- reordering error: a \( \text{WER} \) error which is neither \( \text{RPER} \) nor \( \text{HPER} \) error
- missing word: a \( \text{WER} \) deletion which is also an \( \text{RPER} \) error
- extra word: a \( \text{WER} \) insertion which is also an \( \text{HPER} \) error
- lexical error: an error which is neither inflectional nor missing/extra word
Workflow

- Reference and additional information
- Reference base forms
- Translation reference(s)
- Translation hypothesis
- Hypothesis base forms
- Hypothesis additional information

Identifying erroneous words contributing to WER, RPER and HPER

Error classification (inflectional errors, reordering errors, missing words, extra words, lexical errors)

Overall (document level) raw error counts and error rates

Sentence level raw error counts and error rates

Reference and hypothesis with error class labels (text format)

Reference and hypothesis with error class colours (HTML format)
Correlation with human results

Tested on various language pairs:

- English to Spanish
- Spanish, German, Arabic and Chinese to English

- high correlations (> 0.500)
  - across the error classes
  - across the translation outputs
- high recall (> 50%)

- extra word class should be improved
  - low recall
  - the weakest correlation across the translation outputs
Usage – input options

- required:
  -R, --ref       translation reference
  -H, --hyp       translation hypothesis
  -B, --baseref   reference base forms
  -b, --basehyp   hypothesis base forms

- optional:
  -A, --addref    additional reference information
  -a, --addhyp    additional hypothesis information

- format:
  - raw text
  - one sentence per line
  - multiple references separated by #
<table>
<thead>
<tr>
<th>example.hyp</th>
<th>example.ref</th>
</tr>
</thead>
<tbody>
<tr>
<td>This time, the reason for the collapse on Wall Street. The proper functioning of the market and a price.</td>
<td>This time the fall in stocks on Wall Street is responsible for the drop. The proper functioning of the market environment and the decrease in prices.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>example.hyp.base</th>
<th>example.ref.base</th>
</tr>
</thead>
<tbody>
<tr>
<td>This time, the reason for the collapse on Wall Street. The proper functioning of the market and a price.</td>
<td>This time the fall in stock on Wall Street be responsible for the drop. The proper functioning of the market environment and the decrease in price.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>example.hyp.pos</th>
<th>example.ref.pos</th>
</tr>
</thead>
<tbody>
<tr>
<td>DT NN, DT NN IN DT NN IN NP NP SENT DT JJ NN IN DT NN CC DT NN SENT</td>
<td>DT NN DT NN IN NNS IN NP NP VBZ JJ IN DT NN SENT DT JJ NN IN DT NN NN CC DT NN IN NNS SENT</td>
</tr>
</tbody>
</table>
Usage – output options

- standard output:
  - overall (document level) raw counts and error rates

- optional outputs:
  - -s, --sent sentence_errors.txt
    sentence level raw counts and error rates
  - -c, --cats categories.txt
    labelled reference and hypothesis words in text format
  - -m, --html categories.html
    labelled reference and hypothesis words in HTML format
<table>
<thead>
<tr>
<th>Metric</th>
<th>Value 1</th>
<th>Value 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wer</td>
<td>15</td>
<td>53.57</td>
</tr>
<tr>
<td>Rper</td>
<td>11</td>
<td>39.29</td>
</tr>
<tr>
<td>Hper</td>
<td>5</td>
<td>22.73</td>
</tr>
<tr>
<td>rINFER</td>
<td>1</td>
<td>3.57</td>
</tr>
<tr>
<td>hINFER</td>
<td>1</td>
<td>4.55</td>
</tr>
<tr>
<td>rRER</td>
<td>2</td>
<td>7.14</td>
</tr>
<tr>
<td>hRER</td>
<td>2</td>
<td>9.09</td>
</tr>
<tr>
<td>MISER</td>
<td>6</td>
<td>21.43</td>
</tr>
<tr>
<td>EXTER</td>
<td>2</td>
<td>9.09</td>
</tr>
<tr>
<td>rLEXER</td>
<td>4</td>
<td>14.29</td>
</tr>
<tr>
<td>hLEXER</td>
<td>2</td>
<td>9.09</td>
</tr>
</tbody>
</table>

- **r** = reference
- **h** = hypothesis
- **b** = block
REF: This time the fall in stocks on Wall Street is responsible for the drop.

HYP: This time, the reason for the collapse on Wall Street.

REF: The proper functioning of the market environment and the decrease in prices.

HYP: The proper functioning of the market and a price.

- pink = inflectional errors
- green = reordering errors
- blue = missing/extra words
- red = lexical errors
1::ref-err-cats: This time the fall in stocks on Wall Street is miss responsible for the drop. 
1::hyp-err-cats: This time, the reason for the collapse on Wall Street. 
2::ref-err-cats: The proper functioning of the market environment and the miss decrease in prices. 
2::hyp-err-cats: The proper functioning of the market and a price increase.
REF: This time the fall in stocks on Wall Street is responsible for the drop.

HYP: This time, the reason for the collapse on Wall Street.

REF: The proper functioning of the market environment and the decrease in prices.

HYP: The proper functioning of the market and a price.
1::ref-err-cats:  This time the fall in stocks on Wall Street is miss responsible for reord the reord drop .

1::hyp-err-cats:  This time , ext the reason for reord the reord collapse on Wall Street .

2::ref-err-cats:  The proper functioning of the market environment and the miss decrease in prices .

2::hyp-err-cats:  The proper functioning of the market and a price .
a tool for systematic automatic error classification

- high correlations with human classification results
- high recall values

⇒ can replace (or facilitate) human error analysis

http://www.dfki.de/~mapo02/hjerson/
a number of possibilities for future work:

- synonym lists
- word position (especially for frequent words)
- using other types of alignments (Zeman+ 11)
- assigning multiple errors per word (with probabilities)

- currently being tested and further developed in the framework of the TARAXÜ project (http://taraxu.dfki.de)
D. Vilar, J. Xu, L.F. D’Haro, H. Ney: 
**Error Analysis of Machine Translation Output**
LREC 2006, Genoa, Italy

M. Popović, H. Ney: 
**Word Error Rates: Decomposition over POS classes and Applications for Error Analysis**
WMT 2007, Prague, Czech Republic

M. Popović, A. Burchardt: 
**From Human to Automatic Error Classification for Machine Translation Output**
EAMT 2011, Leuven, Belgium

M. Popović, H. Ney: 
**Towards Automatic Error Analysis of Machine Translation Output**
Computational Linguistics 37(4), December 2011