Hjerson: An Open Source Tool for Automatic Error Classification

German Research Center for Artificial Intelligence

Maja Popović

MT Marathon 2011 (Trento, Italy)

06 September 2011

Motivation

- 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

Algorithm

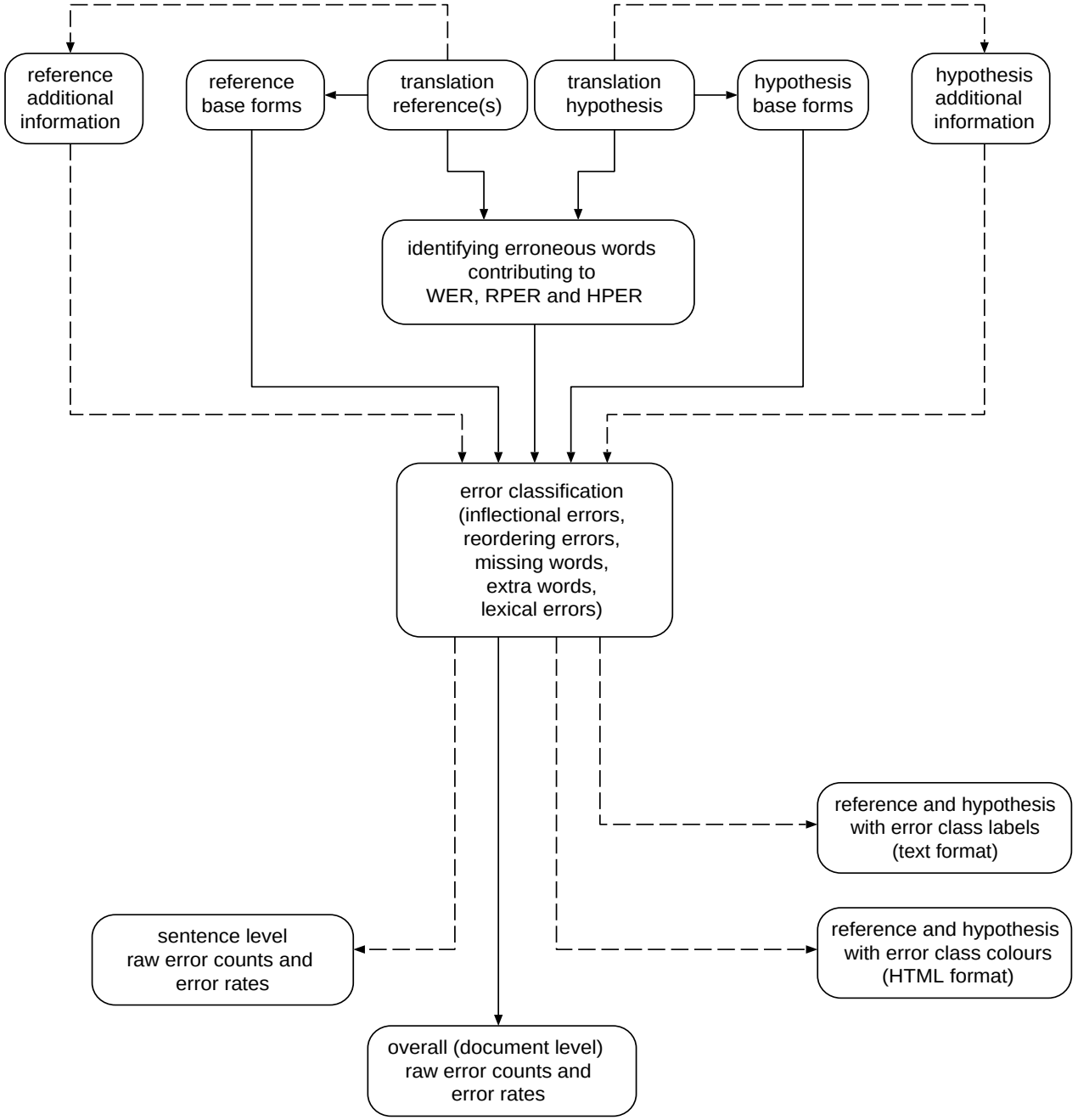
After identifying actual words contributing to the:

- Levenshtein distance WER
- reference position-independent error rate RPER
- hypothesis position-independent error rate HPER (Popović & Ney 07)

the following five error classes based on (Vilar+ 06) are defined:

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

Workflow


reference 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)
reference and hypothesis with error class labels (text format)
sentence level raw error counts and error rates
reference and hypothesis with error class colours (HTML format)
overall (document level) raw error counts and error rates

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 #

Input example

example.hyp	example.ref
This time , the reason for the collapse on Wall Street . The proper functioning of the market and a price .	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 .

example.hyp.base	example.ref.base
This time , the reason for the collapse on Wall Street. The proper functioning of the market and a price .	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 .

example.hyp.pos	example.ref.pos
DT NN , DT NN IN DT NN IN NP NP SENT DT JJ NN IN DT NN CC DT NN SENT	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

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

Standard output example

Wer:	15	53.57
Rper:	11	39.29
Hper:	5	22.73

rINFer:	1	3.57	brINFer:	1	3.57
hINFer:	1	4.55	bhINFer:	1	4.55
rRer:	2	7.14	brRer:	1	3.57
hRer:	2	9.09	bhRer:	1	4.55
MISer:	6	21.43	bMISer:	4	14.29
EXTer:	2	9.09	bEXTer:	2	9.09
rLEXer:	4	14.29	brLEXer:	2	7.14
hLEXer:	2	9.09	bhLEXer:	2	9.09

- r = reference
- h = hypothesis
- b = block

HTML example

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

Annotation example

1::ref-err-cats: This∼∼x time∼∼x the∼∼x fall∼∼lex in∼∼lex stocks∼∼lex on∼∼x Wall∼∼x Street∼∼x is∼∼miss responsible∼∼miss for∼∼reord the∼∼reord drop∼∼miss .∼∼x

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

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

2::hyp-err-cats: The∼∼x proper∼∼x functioning∼∼x of∼∼x the∼∼x market∼∼x and∼∼x a∼∼lex price∼∼infl .∼∼x

HTML with POS tags

REF: This#DT time#NN the#DT fall#NN in#IN stocks#NNS on#IN Wall#NP Street#NP is#VBZ responsible#JJ for#IN the#DT drop#NN .#SENT

HYP: This#DT time#NN ,#, the#DT reason#NN for#IN the#DT collapse#NN on#IN Wall#NP Street#NP .#SENT

REF: The#DT proper#JJ functioning#NN of#IN the#DT market#NN environment#NN and#CC the#DT decrease#NN in#IN prices#NNS .#SENT

HYP: The#DT proper#JJ functioning#NN of#IN the#DT market#NN and#CC a#DT price#NN .#SENT

Annotation with POS tags

1::ref-err-cats: This#DT∼∼x time#NN∼∼x the#DT∼∼x fall#NN∼∼lex in#IN∼∼lex stocks#NNS∼∼lex on#IN∼∼x Wall#NP∼∼x Street#NP∼∼x is#VBZ∼∼miss responsible#JJ∼∼miss for#IN∼∼reord the#DT∼∼reord drop#NN∼∼miss .#SENT∼∼x

1::hyp-err-cats: This#DT∼∼x time#NN∼∼x ,#,∼∼ext the#DT∼∼x reason#NN∼∼ext for#IN∼∼reord the#DT∼∼reord collapse#NN∼∼lex on#IN∼∼x Wall#NP∼∼x Street#NP∼∼x .#SENT∼∼x

2::ref-err-cats: The#DT∼∼x proper#JJ∼∼x functioning#NN∼∼x of#IN∼∼x the#DT∼∼x market#NN∼∼x environment#NN∼∼miss and#CC∼∼x the#DT∼∼miss decrease#NN∼∼miss in#IN∼∼lex prices#NNS∼∼infl .#SENT∼∼x

2::hyp-err-cats: The#DT∼∼x proper#JJ∼∼x functioning#NN∼∼x of#IN∼∼x the#DT∼∼x market#NN∼∼x and#CC∼∼x a#DT∼∼lex price#NN∼∼infl .#SENT∼∼x

Summary

- a tool for systematic automatic error classification
  - high correlations with human classification results
  - high recall values

$\Rightarrow$ can replace (or facilitate) human error analysis

`http://www.dfki.de/~mapo02/hjerson/`

Outlook

- 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`)

References

- 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