From Human to Automatic Error Classification for Machine Translation Output

German Research Center for Artificial Intelligence

Maja Popović, Aljoscha Burchardt

EAMT 2011 (Leuven, Belgium)

31 May 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

– human evaluation is resource-intensive and time-consuming

⇒ automatic methods are needed

Related work

Human error analysis and classification:

- Vilar & Xu+ 2006
  - classification scheme and detailed analysis of the obtained results are presented
  - the method has become widely used in recent years

Automatic error analysis:

- Lopez & Resnik 2005
  - analysis of POS sequences
- Popović & de Gispert+ 2006, Popović & Ney 2007
  - first steps towards the use of WER and PER for automatic error analysis
- Zhou & Wang+ 2008
  - analysis of parsed source and target sentences

Automatic method for error classification

Two main goals of the proposed automatic method:

- distribution of errors over the error classes within an output
- distribution of errors over translation outputs within a class

Five error classes based on the scheme (Vilar & Xu+ 2006):

- inflectional errors
- reordering errors
- missing words
- extra words
- incorrect lexical choice

using WER and PER decomposition method (Popović & Ney 2007)

Definition of error classes

After identifying actual words contributing to the:

- Levenshtein distance WER
- reference position-independent error rate RPER
- hypothesis position-independent error rate HPER

the error classes 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

Human error classification

- can be carried out in various ways
  - □ comparing with given reference translations
  - □ comparing with the source text
  - * strictly, flexibly, freely
- by no means unambiguous

In this work, two variants are carried out:

- strict comparison with the given reference
- flexible comparison with the given reference (natural way)
  - – syntactically correct differences in word order, synonyms or different expressions are not considered as errors

Analysed translation outputs

Six English translation outputs obtained by state-of-the-art statistical phrase-based systems:

- GALE texts
  - two Arabic-to-English and one Chinese-to-English translation outputs
  - strict human error classification
- WMT texts
  - three German-to-English translation outputs
    - translations of the same German source text
    - $\Rightarrow$ appropriate for comparing translation systems
  - flexible human error classification

Results of human and automatic classification

- raw error counts N_{hum}/N_{aut}
- Spearman's and Pearson's correlation coefficients ρ and r

GALE	inflection	order	missing	extra	lexical	ρ_{sys}	r_{sys}
ArEn1	20/23	39/66	79/63	127/137	135/147	0.90	0.96
ArEn2	22/24	30/41	97/102	73/76	140/131	1.00	0.99
CnEn	38/40	127/171	288/244	95/117	203/239	1.00	0.93

WMT	inflection	order	missing	extra	lexical	ρ_{sys}	r_{sys}
DeEn1	12/32	60/235	204/199	52/40	189/521	0.70	0.72
DeEn2	16/44	41/212	172/200	30/56	163/495	0.7	0.74
DeEn3	17/46	100/274	107/153	68/99	171/508	0.90	0.91
ρ_{class}	1.00	1.00	0.60	0.5	1.00		
r_{class}	0.90	0.99	0.90	0.62	0.96		

- high correlations (> 0.700)
  - across the error classes
  - across the translation outputs
- slightly lower for the flexible human classification (WMT)
- extra words class has the weakest correlation across the translation outputs

*note:
correlations with the results of (Vilar & Xu^{+}) are > 0.500
(free human error analysis, English and Spanish outputs)

Analysis of the differences

- automatic method can successfully substitute human error classification
- nevertheless, it would be useful to better understand certain differences:
  - large number of automatic reordering errors
  - disambiguation between lexical errors and missing/extra words
  - low correlations for extra words
  - are all errors detected by humans successfully covered?

Recall and precision, GALE texts

ArEn1 ref	inflection	order	missing	lexical	x
inflection	**78.9/78.9**	/	2.2/10.5	0.8/5.3	0.1/5.3
order	/	**92.5/51.4**	8.8/11.1	3.2/5.6	1.8/**31.9**
missing	/	/	**53.8/81.7**	4.8/10.0	0.4/8.3
lexical	15.8/2.1	2.5/0.7	29.7/19.3	**85.5/75.7**	0.2/2.1
x	5.3/0.1	5.0/0.2	5.5/0.4	5.6/0.5	**97.5/98.8**

ArEn1 hyp	inflection	order	extra	lexical	x
inflection	**81.0/89.5**	/	0.7/5.3	/	0.1/5.3
order	4.8/1.4	**90.2/51.4**	3.6/6.9	5.8/12.5	1.5/**27.8**
extra	4.8/1.0	/	**53.3/72.3**	15.4/23.8	0.2/3.0
lexical	4.8/0.8	2.4/0.8	15.3/15.9	**64.1/75.8**	0.8/6.8
x	4.8/0.1	7.3/0.2	27.0/2.8	14.7/1.7	**97.5/95.2**

- missing and extra words: lowest recall
  - confusion with lexical errors
  - frequent extra words not detected
- reordering errors: lowest precision
  - correct frequent words tagged as errors

Examples, GALE texts

reference: hypothesis:	... of local party committees . **Secretaries** of the Commission ... ... of local party committees of the **provincial** Commission ...
errors:	Secretaries – missing(hum,aut) provincial – extra(hum,aut)

reference: hypothesis:	... , although the **Japanese friendly feelings for** China added an increase , ... , although China **can feel** the **Japanese** increase , ...
errors:	Japanese – order(hum,aut) friendly – missing(hum,aut) feelings for – missing(hum)/lexical(aut) can feel – extra(hum)/lexical(aut)

Recall and precision, WMT texts

DeEn1 ref	inflection	order	missing	lexical	x
inflection	**92.3**/37.5	1.6/3.1	2.0/12.5	1.6/9.4	1.1/37.5
order	/	**61.3**/15.3	5.9/4.8	2.6/2.0	17.3/**77.8**
missing	/	6.5/2.1	45.8/48.4	16.6/16.7	5.7/32.8
lexical	7.7/0.2	11.3/1.4	42.9/17.5	**78.2**/30.3	22.6/**50.6**
x	/	19.4/1.9	3.4/1.1	1.0/0.3	**53.4**/**96.6**

DeEn1 hyp	inflection	order	extra	lexical	x
inflection	**92.3**/37.5	5.4/12.5	/	2.6/12.5	1.1/37.5
order	/	**51.4**/15.3	14.8/3.2	4.5/2.8	17.8/**78.6**
extra	/	1.4/3.2	16.7/29.0	3.2/16.1	1.5/**51.6**
lexical	7.7/0.2	24.3/4.0	**57.4**/6.9	**85.8**/29.6	24.4/**59.3**
x	/	17.6/2.1	11.1/1.0	3.9/1.0	**55.3**/**96.0**

- lower precisions $\leftrightarrow$ lower recall of correct words
  - especially for reordering and lexical errors
- larger confusion “missing/extra words $\rightarrow$ lexical errors”
- a number of frequent extra words is
  - not detected
  - tagged as reordering errors

Examples, WMT texts

reference:	**Passengers can get** coffee and newspapers **when** boarding .
hypothesis:	Coffee and newspapers **can passengers in** boarding .
errors:	Passengers can – order(hum,aut) get – missing(hum,aut) when – lexical(hum,aut) in – lexical(hum,aut)

reference:	The **famous journalist Gustav Chalupa** , **born in České Budějovice** , **also confirms this** .
hypothesis:	The **also confirms the famous Austrian journalist Gustav Chalupa** , **from Budweis Lamborghini** .
errors:	famous journalist Gustav Chalupa – order(aut) born in České Budějovice – lex(aut) also confirms – order(hum,aut) this – missing(hum)/lexical(aut) the – extra(hum)/lexical(aut) Austrian – extra(hum,aut) from Budweis – lexical(aut) Lamborghini – extra(hum)/lexical(aut)

Summary

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

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

*except extra words

$\rightarrow$ not particularly stable and reliable at this stage

Outlook

- a number of possibilities for future work:
  - synonym lists
  - word position (especially for frequent words)
  - assigning multiple errors per word (with probabilities)

- currently being tested and further developed in the framework of the TARAXÜ project
  `http://taraxu.dfki.de`