Assessing Inter-Annotator Agreement for Translation Error Annotation

German Research Center for Artificial Intelligence

Arle Lommel, Maja Popović, Aljoscha Burchardt

MTE-LREC 2014
(Reykjavik, Iceland)

26 May 2014

Introduction

Typical ways of using human knowledge for assessing machine translation output:

- generating reference translations
- rating MT output based on quality
- post-editing MT output (implicit error markup)
- error classification (explicit error markup)

However:

- no single objectively correct translation of a given text
- no single correct error type for a number of translation errors

⇒ inter-annotator agreement (IAA)

* this work: error classification

IAA for error classification

Introduction

IAA for error classification

MQM error classification scheme

Annotated texts – first step

Annotated texts – filtering

Word level IAA

Span issues

Unclear error categorisation

Conclusions

Lessons learnt and ongoing/future work

Numerous possible interpretations:

- word level (strong effect of exact span)
  - □ F-score
    takes into account only absolute agreement
    $F = P(a) = \frac{\sum_k N(a_1{=}k, a_2{=}k)}{N(words)}$
  - □ Kappa coefficient
    takes into account agreement by chance
    $\kappa = \frac{P(a) - P(e)}{1 - P(e)}$
    $P(a) = \sum_k P(a_1 = k, a_2 = k)$ absolute agreement
    $P(e) = \sum_k P(a_1 = k) * P(a_2 = k)$ agreement by chance
- sentence level (no effect of exact span)
- analysis on different error class levels
  (error/no error, simpler error tags, detailed error tags)

MQM error classification scheme

- accuracy
  - terminology
  - mistranslation
  - omission
  - addition
  - untranslated
- fluency
  - style/register
  - spelling
    - capitalisation
  - typography
    - punctuation
  - grammar
    - morphology (word form)
    - part of speech
    - agreement
    - word order
    - function words
    - tense/aspect/mood
  - unintelligible

Annotated texts – first step

Starting point:

- WMT 2013 translation outputs
  - □ Spanish→English
  - □ English→Spanish
  - □ German→English
  - □ English→German

  produced by state-of-the-art MT systems
  - □ statistical (SMT)
  - □ rule-based (RBMT)
  - □ hybrid (HMT) (only for translation from English)

- only “native” source sentences were used
  - □ in order to evaluate human translations (HT) as well
  - □ in order to avoid influences of intermediate (human) translation

Annotated texts – filtering

Only high quality translations were annotated in order to minimise effects of overlapping errors :

- the sentences were filtered according to the following criterion:
  - □ rank 1: perfect output (no editing needed)
  - □ rank 2: near miss translations (up to three edits needed)
  - □ rank 3: bad (more than 3 edits needed)

- subsets of sentences with rank 2 were annotated:

# of sentences	es-en	en-es	de-en	en-de
SMT	60	40	60	40
RBMT	60	40	60	40
HMT	0	40	0	40
HT (references)	40	40	40	40
# of annotators	4	4	3	4

Word level IAA

Introduction
IAA for error classification
MQM error classification scheme
Annotated texts – first step
Annotated texts – filtering
Word level IAA
Span issues
Unclear error categorisation
Conclusions
Lessons learnt and ongoing/future work

Kappa coefficients:

κ	es-en	en-es	de-en	en-de
a1-a2	0.30	0.35	0.23	0.36
a1-a3	0.18	0.36	0.36	0.28
a2-a3	0.19	0.28	0.29	0.33
a1-a4	0.25	0.33	/	0.30
a2-a4	0.26	0.36	/	0.34
a3-a4	0.34	0.35	/	0.30
avg	0.25	0.34	0.29	0.32

For comparison:
Kappa coefficients for WMT ranking tasks:

	es-en	en-es	de-en	en-de
avg	0.40	0.32	0.38	0.40

Span issues

% of span disagreement	es-en	en-es	de-en	en-de
accuracy	0	0.1	0	0.4
addition	0.5	1.3	0.4	2.2
agreement	0.4	2.8	0.3	1.4
capitalisation	0	0.6	0.3	0.3
fluency	0	0	0	0
function words	9.2	10.1	4.1	1.9
grammar	3.0	0.3	0.1	9.5
mistranslation	6.4	6.9	4.4	8.0
morphology	0	0.1	1.0	0.1
POS	1.1	0.5	1.2	0
punctuation	2.0	0.7	1.2	1.5
spelling	0.4	0.6	0.1	0.2
style/register	7.1	7.4	3.8	6.3
tense/aspect/mood	1.6	4.4	0.5	2.3
terminology	6.3	14.2	8.9	2.8
typography	0	0.4	0	0
unintelligible	0.1	0	0.3	1.2
untranslated	0.3	0	0.3	0.5
word order	8.0	10.1	24.2	6.1

Unclear error categorisation

Introduction
IAA for error classification
MQM error classification scheme
Annotated texts – first step
Annotated texts – filtering
Word level IAA
Span issues
Unclear error categorisation
Conclusions
Lessons learnt and ongoing/future work

% on sentence level disagreement	es-en	en-es	de-en	en-de
accuracy	0.2	0.2	0	1.0
addition	2.1	4.8	4.0	3.5
agreement	6.2	7.3	3.6	4.7
capitalisation	0.3	2.9	1.1	1.2
fluency	0	3.0	0	0.2
function words	**30.4**	**21.9**	**18.9**	**7.6**
grammar	6.3	1.0	0.7	16.8
mistranslation	**23.6**	**22.8**	**27.1**	**24.1**
morphology	0.2	0.3	3.8	5.4
omission	5.3	6.6	7.6	5.2
POS	2.9	2.2	2.4	1.0
punctuation	4.0	4.5	9.1	9.3
spelling	0.8	2.2	1.1	0.9
style/register	16.3	9.1	3.3	11.0
tense/aspect/mood	3.9	11.3	3.1	7.1
terminology	**12.9**	**24.5**	**19.1**	**13.1**
typography	0.2	0.8	0.2	0
unintelligible	0.2	0	1.3	1.2
untranslated	0.9	0.9	0.9	0.5
word order	7.2	5.9	8.9	4.4
no error	**15.4**	**10.4**	**7.6**	**8.7**

Conclusions

Results have shown:

- very high confusion for Mistranslation and Terminology
- high confusion for FunctionWords
- high agreement for WordOrder on the sentence level but high confusion on the word level i.e. span
- relatively large confusion between Error and NoError

Causes of disagreements:

- confusion within the hierarchy
  - which level is the most appropriate?
- lack of clear decision tools
  - mismatch between annotators’ and MQM creators’ knowledge
- annotators’ personal opinion

Lessons learnt and ongoing/future work

- revising MQM hierarchy
  - □ merge Terminology with Mistranslation
  - □ merge Agreement, POS, Tense/Aspect/Mood into Morphology(WordForm)
  - □ split FunctionWords into Missing, Extra and Wrong
- create a formal decision tree
- improve guidelines
- deeper analysis and understanding of disagreements can provide insight into how humans perceive translation quality