Towards Fairer Evaluations of Commercial MT Systems on BTEC

Hervé Blanchon
GETA-CLIPS
herve.blanchon@imag.fr

Outline

- Why did we embark?
- Which Systems did we embark with?
- Chinese-English unlimited data track
- Japanese-English unlimited data track
- Competitive evaluation
- Some Q&A
- Final personal comments

Why did we embark?

- View actual data and associated evaluation results, to:
  - Follow the state of the art techniques
  - Clarify our ideas about some problems either of generaly proposed evaluation settings or of the communication around the results

Which systems, pairs and tracks?

- Widespread and available for the language pairs
  - Systran Web & Systran Prof. Premium V5
- Rule-based approach
- Unlimited data track
- Systran Web already been used a baseline system
- Runs

C-E	J-E	
C_1	J_1	Systran Web V5
C_2	J_2	Systran PP v5 with original dictionaries
C_3	J_3	Systran PP v5 with original and user dictionaries

Chinese-English

- Subjective evaluation (**C_3**)
  - non-native English > Fluency > disfluent English
  - Adequacy ≈ much of the meaning is expressed
- Objective evaluation

	BLEU	GMT	NIST	PER	WER
C_3	0.1620 1	0.5845 1	6.0061 1	0.5429 2	0.6581 2
C_1	0.1600 3	0.5802 3	5.9143 3	0.5423 1	0.6474 1
C_2	0.1620 1	0.5841 2	6.0039 2	0.5429 2	0.6581 2

- ***Same results for each version!***

Japanese-English (results)

- Subjective evaluation (**J_3**)
  - non-native English > Fluency > disfluent English
  - much > Adequacy > little
- Objective evaluation

	BLEU	GMT	NIST	PER	WER
J_3	0.1320 1	0.5687 1	5.6476 1	0.5978 1	0.7304 1
J_2	0.1311 2	0.5672 2	5.6096 2	0.6012 2	0.7349 2
J_1	0.0810 3	0.5116 3	4.1935 3	0.7179 3	0.8726 3

- ***Systems are ordered according to expectation***

Japanese-English (explanations)

- Bad translation when subject is omitted
  - ここ で 降り ます 。 **It gets** off here.
- Euphemistic utterance が translated by “but”
  - 両替 を し たい の です が 。It is to like to exchange **but**.
- Question word order
  - 入場 料 は いくら です か 。Is admission fee **how much**?
- Requests or invitations
  - 一緒 に 行き ましょ う 。**It will** go together.
- …

Competitive evaluation

- Observed results for Japanese-English
-

	BLEU	GMT	NIST	PER	WER
JE_1	0.6306 1	0.7967 2	10.7201 2	0.2333 1	0.2631 1
JE_3	0.6190 2	0.8243 1	11.2541 1	0.2492 2	0.3056 2
JE_4	0.3970 3	0.6722 3	7.8893 3	0.4202 3	0.4857 3
J_3	0.1320 4	0.5687 4	5.6476 4	0.5978 4	0.7304 4

- From rough scores Systran is 4th
- What does it means knowing that subjective evaluation is bad?
- Is this ranking relevant?

Competitive evaluation

- Observed results for Human Japanese-English

	BLEU		GMT		NIST		PER		WER	
JE_1	0.6306	1	0.7967	2	10.7201	2	0.2333	1	0.2631	1
JE_3	0.6190	2	0.8243	1	11.2541	1	0.2492	2	0.3056	2
J_4	0.4691	-	0.7777	-	9.9189	-	0.3236	-	0.3711	-
JE_4	0.3970	3	0.6722	3	7.8893	3	0.4202	3	0.4857	3
J_3	0.1320	4	0.5687	4	5.6476	4	0.5978	4	0.7304	4

- Perfect human minimal post-edition does not over-score MT
- What does it means knowing that subjective evaluation should be good?

Q&A (actual questions from IR people)

Q: Is-it NORMAL?

A: *YES of course! We are NOT evaluating translation QUALITY but SIMILARITY between candidate translations and references translations!*

Q: BUT, references are produced by humans!!!

A: *Yes, But … the post-editor may have produced translations having different style or wording compared with the references!*

Q&A (actual questions from IR people)

Q: Objective evaluation is said to be good because it results correlates with subjective evaluation?

R: *Correlation is still a hot topic! Sometimes the correlation is good, sometimes it is not the case.*

Q: Getting better, only mean that the system produces translations that resemble better to the references? Is that why post-edition is not ranked first?

R: *Yes.*

Concluding *personal* comments

- Systran, as it is, cannot be used as a baseline system for comparative, competitive evaluation, at least for English to Japanese on the BTEC corpus
  - Other language pairs have to be examined

Concluding *personal* comments

- Objective evaluation techniques do not evaluate translation quality, they evaluate the capacity of the system to mimic the reference
- then
  - good scores mean good mimicking
  - bad scores mean nothing on their own
- These techniques may be well suited for “systems that learn” from the data but not for others and the comparison is meaningless.

Comments

Questions