Ja to En/Zh/Ko Datasets for Translation Quality Estimation and Automatic Post-Editing

Atsushi Fujita  Eiichiro Sumita
ASTREC, NICT, Japan
firstname.lastname@nict.go.jp
NICT QE/APE Dataset (google it!)

- **Multilingual parallel corpus**
  - Japanese (Ja) → English (En), Chinese (Zh), Korean (Ko)
  - Spoken language: travel (8,783 segs), hospital (1,676 segs)
  - 5-tuples for each target language
    - **hyp**: generated by phrase-based SMT systems
    - **ref, pe, and grade**: manually created

<table>
<thead>
<tr>
<th>Source segment in Ja (src)</th>
<th>片道だけで買えますか。</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human translation (ref)</td>
<td>May I get it for one way?</td>
</tr>
<tr>
<td>MT output (hyp)</td>
<td>Can I buy just one way?</td>
</tr>
<tr>
<td>Manually post-edited hyp (pe)</td>
<td>Can I just buy a one way ticket?</td>
</tr>
<tr>
<td>Quality grade of hyp (grade)</td>
<td>B ∈ {S, A, B, C, D}</td>
</tr>
</tbody>
</table>
Background
Human-machine collaboration for translation

A spectrum of translation methods [Hutchins+, 1992]

- Human translation
- Machine-aided Human translation
- Human-aided MT
- Fully automatic MT

Questions

- (a) How often only human can translate?
- (b)(c) How to collaborate?
- (d) What is a realistic goal?
  - “Fully automatic MT is not achievable” [Bar-Hillel, 51]

The word “translation” consists of 11 letters.
Prevalent approaches: Post-Editing

- Post-editing MT outputs (PEMT)
  - Standard in industries in Europe [ISO18587, 17]

- Automatic PE (APE)
  - Trainable with \((\text{src}, \text{hyp}, \text{pe})\) triplets [Allen+, 00]
Prevalent approaches: Quality Estimation

- **Quality Estimation (QE)** [Blatz+, 04][Specia+, 10]
  - For the users (!= confidence score of MT systems)
  - Word-level QE
    - e.g., visualization of unreliable parts [Bach+, 11]

<table>
<thead>
<tr>
<th>Source</th>
<th>Human correction</th>
</tr>
</thead>
<tbody>
<tr>
<td>أنت مختلف تماماً عن زيد وعمرو فلا تحشر نفسك في سرداد التقليد والمحاكاة والذوبان</td>
<td>you are quite different from zaid amr, so do not cram yourself in the tunnel of simulation, imitation and assimilation.</td>
</tr>
<tr>
<td>MT output</td>
<td>We predict and visualize</td>
</tr>
<tr>
<td>you totally different from zaid amr, and not to deprive yourself in a basement of imitation and assimilation.</td>
<td>you totally different from zaid amr, and not to deprive yourself in a basement of imitation and assimilation.</td>
</tr>
</tbody>
</table>

- Sentence-level QE
  - e.g., whether **hyp** can be directly delivered [Soricut+, 10]
  - e.g., routing according to the expected PE labor [Specia, 11]

As is
- Post-edit
- Human translation

Figure 7: MT errors visualization based on confidence scores.

*Acknowledgements*

We would like to thank anonymous reviewers, Qin Gao, IBM machine translation team for their supports. Also, we would like to thank Christoph Tillmann and the MTSummit-XII Association for Machine Translation.
Limitations of existing datasets

- Only for European languages
  - Those with 10k+ segments

<table>
<thead>
<tr>
<th>Prior art</th>
<th>Translation direction</th>
<th>Domain</th>
<th>MT system</th>
<th># of segments</th>
</tr>
</thead>
<tbody>
<tr>
<td>[Potet+, 12]</td>
<td>Fr → En</td>
<td>news</td>
<td>SMT</td>
<td>10,881 0 1,500</td>
</tr>
<tr>
<td>[Bojar+, 15]</td>
<td>En → Es</td>
<td>news</td>
<td>SMT</td>
<td>11,271 1,000 1,817</td>
</tr>
<tr>
<td>[Bojar+, 16]</td>
<td>En → De</td>
<td>IT</td>
<td>SMT</td>
<td>12,000 1,000 2,000</td>
</tr>
<tr>
<td>[Bojar+, 17]</td>
<td>En → De</td>
<td>IT</td>
<td>SMT</td>
<td>23,000 1,000 2,000</td>
</tr>
<tr>
<td></td>
<td>De → En</td>
<td>pharmaceutical</td>
<td>SMT</td>
<td>25,000 1,000 2,000</td>
</tr>
</tbody>
</table>

- Only PE-based QE
  - Word-level QE: OK/BAD tags generated from \((\text{hyp}, \text{pe})\) pairs
  - Sentence-level QE: HTER computed from \((\text{hyp}, \text{pe})\) pairs

- Only written texts

Can we rely on HTER to decide which \text{hyp} to be delivered?
Our 6-step procedure for corpus construction

- Cover Asian languages
- Cover spoken language
- Conduct also manual grading
Steps 1 & 2: Creation of parallel corpus
Step 1: Collecting Japanese utterances

- Spoken language in two domains
  - travel: transcribed utterances
    - Log data of our speech translation app, VoiceTral
  - hospital: written utterances
    - Manually created role-play dialogs

- De-duplication & manual cleaning
  - Non-understandable ones
  - Ungrammatical ones
  - Inappropriate ones wrt. social standards
Step 2: Manual translation

- Native speakers of each target language

- Instructions
  - Translate each segment
  - Assuming reasonable context
    - But without adding too much information
  - Only one ref, even if the src has more than one interpretation
    - e.g., filling zero-anaphor (he/she)
    - e.g., specifying sg./pl. (book/books)

  **src** หยิบ แค่ เดิน บ่ม มี หรือ
  (one way) (only) (ACC) (buy) (modesty) (INT) .

  **ref** May I get it for one way?
Steps 3 & 4: Creation of QE/APE dataset
Steps 3 & 4: MT and post-editing

- **Step 3**: $src \rightarrow hyp$ using phrase-based SMT systems

<table>
<thead>
<tr>
<th>Translation direction</th>
<th>MT system</th>
<th>Model</th>
<th># of sentences for training</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ja $\rightarrow$ En</td>
<td>SMT</td>
<td>2013/Oct.</td>
<td>736k</td>
</tr>
<tr>
<td>Ja $\rightarrow$ Zh</td>
<td>SMT</td>
<td>2016/Sep.</td>
<td>1.44M</td>
</tr>
<tr>
<td>Ja $\rightarrow$ Ko</td>
<td>SMT</td>
<td>2016/Sep.</td>
<td>1.40M</td>
</tr>
</tbody>
</table>

- **Step 4**: $src$, $hyp$, and $ref \rightarrow pe$
  
  - Correcting $hyp$ with minimum number of edits [Snover+, 06]
    - Deletion of a word (Del)
    - Insertion of a word (Ins)
    - Substitution of a word (Sub)
    - Shift of a word or a phrase (Shift)
  
  - A lenient criterion: $\text{TER}(hyp \rightarrow pe) \leq \text{TER}(hyp \rightarrow ref)$
    - Prohibited: copying $ref$ to $pe$
Detail of computing TER

- tercom-0.7.25 w/ OSS tokenizers
  - En: Moses’s tokenizer
  - Zh: Character
  - Ko: MeCab + mecab-ko-dic

- Asymmetric
  - For Step 4: TER(hyp → pe), TER(hyp → ref)
  - For sentence-level QE: TER(pe → hyp) (=HTER)

<table>
<thead>
<tr>
<th>src</th>
<th>あなたは22時までにチェックインする必要があります。</th>
</tr>
</thead>
<tbody>
<tr>
<td>hyp</td>
<td>You should check in by twenty two o'clock .</td>
</tr>
<tr>
<td>pe</td>
<td>You have to check in by 22 00 .</td>
</tr>
</tbody>
</table>

= Sub Del = = = Sub Ins Ins =
Steps 5 & 6: Additional human evaluation
Step 5: Manual grading

Evaluate the quality of hyp wrt. its src
- Optionally referring to ref

5-ary classification [Goto+, 13] rather than scoring
- S: Perfect (native)  Correct (goal of PE)
- A: Good
- B: Fair
- C: Acceptable  Some errors (need PE)
- D: Incorrect
Definition of acceptable classes

S: Perfect ("AA" in [Goto+, 13])
- Information of the source text has been completely translated. There are no grammatical errors in the target text. Lexical choice and phrasing are natural even from a native speaker point-of-view.

A: Good
- Information of the source text has been completely translated. There are no grammatical errors in the target text, but lexical choice and phrasing are slightly unnatural.
Definition of unacceptable classes

- **B: Fair**
  - There are *some minor errors* in the target text of less important textual information, but the *meaning of the source text can be easily understood*.

- **C: Acceptable**
  - Important parts of the source text are *omitted or incorrectly translated*, but the *meaning of the source text can be easily understood*.

- **D: Incorrect ("F" in [Goto+, 13])**
  - The meaning of the source text is *incomprehensible* from target text.
Step 6: Consistency check

- Resolve the discrepancies
  - If $(\text{grade} \text{ is } \text{“S” or “A”}) \land \land (\text{pe} \neq \text{hyp})$
    → then, retry both grading and post-editing
  - If $(\text{grade} \text{ is } \text{“B” or “C” or “D”}) \land \land (\text{pe} = \text{hyp})$
    → then, retry both grading and post-editing

- Minimality of post-edits is assessed again (if necessary)
  - If $\text{TER}(\text{hyp} \rightarrow \text{pe}) > \text{TER}(\text{hyp} \rightarrow \text{ref})$
    → then, retry post-editing
Analyses of the created datasets
Results of manual grading

Direction and domain

hyp were of very bad quality
Proximity between translations

Observations

- (a) **hyp → ref**: automatic MT evaluation
  - Correlates with the proportion of “S” and “A”
- (b) **hyp → pe**: amounts of post-edits
  - Significantly better than (a), except for Ja→En hospital
- (c) **pe → ref**: unrelated
  - Fallen between (a) and (b), except for Ja→En tasks

e.g., TER scores in the hospital domain

- Ja→En: hyp (a) 75.35, (b) 66.03, (c) 24.69
  - Too bad to perform PE
- Ja→Zh: hyp (a) 48.54, (b) 8.67, (c) 43.78
- Ja→Ko: hyp (a) 32.44, (b) 4.12, (c) 30.00
Distribution of sentence-wise HTER score

- Overall: some level of correlation
- But, lower HTER scores do not mean higher quality
HTER vs. *grade*

Ja->En, hospital

Human judgment for MT output

Critical errors were corrected with small numbers of edits

Too much edits
Ja ➔ En examples with HTER/grade mismatch

- **HTER=0.22 / “D” Drop of negation**
  - *src* 多額の現金は持ってこないでください。
  - *hyp* Please bring a lot of cash.
  - *pe* Please don’t bring a lot of cash.

- **HTER=0.13 / “D” Incorrect reference**
  - *src* 頭が痛くありませんか。
  - *hyp* Do you have pain in my neck?
  - *pe* Do you have pain in your neck?

- **HTER=0.78 / “B” Too much edits**
  - *src* 素晴らしい景色だね
  - *hyp* It’s beautiful scenery.
  - *pe* The scenery’s beautiful, isn’t it?

HTER and grade define different objectives for sentence-level QE
Benchmarking

- How the data resemble WMT?
- Are existing methods useful?
Benchmarking experiments

Tasks

- Word-level QE
- Sentence-level QE: two variants
- Automatic Post-editing (APE)

External resource

- We reluctantly used our in-house parallel corpus
  - Daily life conversations
  - Reasonably large

<table>
<thead>
<tr>
<th>Partition</th>
<th># of sentences</th>
<th># of tokens</th>
</tr>
</thead>
<tbody>
<tr>
<td>train</td>
<td>1.57M</td>
<td>20.1M-2.51M</td>
</tr>
<tr>
<td>dev</td>
<td>14k</td>
<td>179k-224k</td>
</tr>
</tbody>
</table>
Word-level QE: Methods

- **Feed-forward NN model** [Kreutzer+, 15]
  - Relying on word alignment and context words in a window

  ![Diagram of feed-forward NN model](image)

  \[ \begin{align*}
  y_{i-1} &= \text{You} \\
  y_i &= \text{should} \\
  y_{i+1} &= \text{check} \\
  x_{a(i)-1} &= \text{する} \\
  x_{a(i)} &= \text{必要} \\
  x_{a(i)+1} &= \text{が}
  \end{align*} \]

  - Use of large pseudo training data
    - generated from \((\text{hyp}, \text{ref})\) pairs [Liu+, 17]
**Word-level QE: Results**

- On F<sub>1</sub> score for BAD tags (F<sub>1</sub>-BAD)
- The less the BAD tags were, the lower F<sub>1</sub>-BAD was
- Were pseudo training data useful?
  - ✔ Ja → En and Ja → Zh: as in En → Es [Liu+, 17]
  - ✗ Ja → Ko: we have not yet found the reason

![Bar chart showing F<sub>1</sub>-BAD for different conditions and languages](chart.png)
Sentence-level QE: Methods

- **Two variant tasks**
  - Prediction of HTER as regression [Bojar+, 14;15;16;17]
  - Binary classification
    - No PE is needed: “S” and “A”
    - PE is needed: “B”, “C”, “D”

- **SVM with heuristic features**
  - **QuEst17**: the 17 MT system-independent features [Specia+, 15]
    - # of tokens, LM prob., Avg. ambiguity based on phrase table, etc.
  - **SntEmb**: distributed representation for each of src & hyp
    - Avg. of 300-dim embeddings of constituent words [Shah+, 16]
Sentence-level QE: Results

- SntEmb improved the performance of QuEst features

- Prediction of HTER
  - No naïve baseline

- Binary classification
  - No WMT counterpart
  - No naïve baseline
  - On F1-mult (F_1-BAD * F_1-OK)
  - Two expected usages
    - To filter out truly BAD hyp
    - To filter truly OK hyp
Conclusion

NICT QE/APE Dataset (google it!)
- Japanese (Ja) $\rightarrow$ English (En), Chinese (Zh), Korean (Ko)
- Spoken language: travel (8,783 segs), hospital (1,676 segs)
- 5-tuples for each target language
  - $\textit{hyp}$: generated by phrase-based SMT systems
  - $\textit{ref}$, $\textit{pe}$, and $\textit{grade}$: manually created

<table>
<thead>
<tr>
<th>Source segment in Ja ($\textit{src}$)</th>
<th>片道だけで買えますか。</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human translation ($\textit{ref}$)</td>
<td>May I get it for one way?</td>
</tr>
<tr>
<td>MT output ($\textit{hyp}$)</td>
<td>Can I buy just one way?</td>
</tr>
<tr>
<td>Manually post-edited $\textit{hyp}$ ($\textit{pe}$)</td>
<td>Can I just buy a one way ticket?</td>
</tr>
<tr>
<td>Quality grade of $\textit{hyp}$ ($\textit{grade}$)</td>
<td>$B \in {S, A, B, C, D}$</td>
</tr>
</tbody>
</table>
Future work

(ongoing) Creating QE data for NMT outputs
- *grade* only (by Mar. 2018?)
  - PE is more difficult to perform
  - PE costs significantly more

Evaluation of recent NN-based methods
- e.g., Predictor-estimator NN model for QE [Kim+, 17]
- e.g., Multi-source NN-based APE [Junczys-Dowmunt+, 16]

Application to MT-related tasks
- e.g., Enhancing our translation services
- e.g., Filtering automatically harvested parallel sentences

A shared-task in the next WAT?