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!)

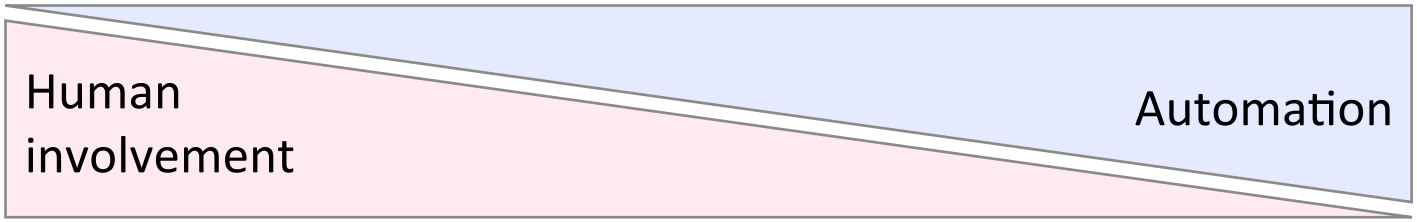
- Multilingual parallel corpus
  - Japanese (Ja) → English (En), Chinese (Zh), Korean (Ko)
  - Spoken language: travel (8,783 segs), hospital (1,676 segs)
  - 5-tuples for each target language
    - ***hyp***: generated by phrase-based SMT systems
    - ***ref***, ***pe***, and ***grade***: manually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

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

Background

Human-machine collaboration for translation

- A spectrum of translation methods [Hutchins+, 1992]



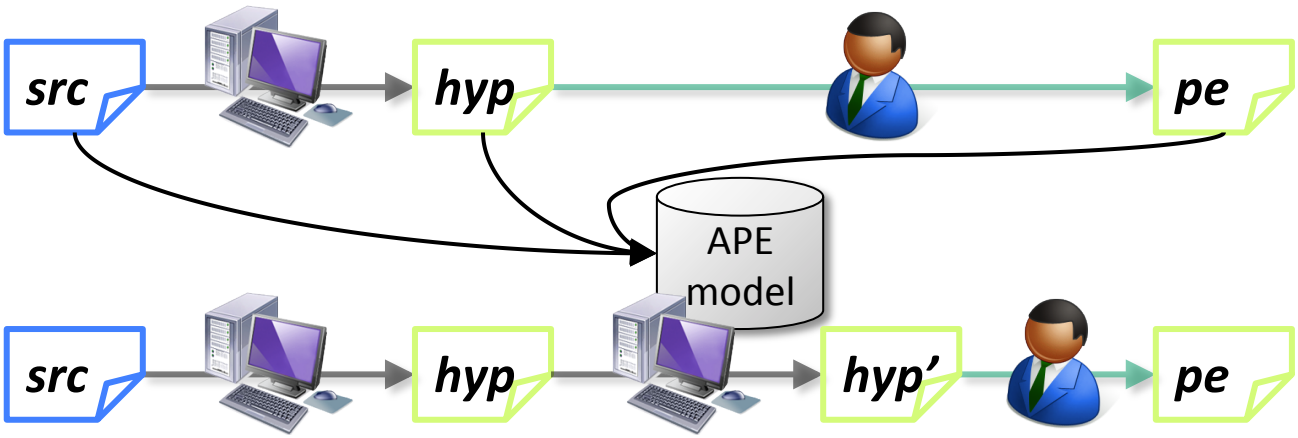
(a)	(b)	(c)	(d)
Human translation	Machine-aided Human translation	Human-aided MT	Fully automatic MT

- 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 (***src***, ***hyp***, ***pe***) triplets [Allen+, 00]

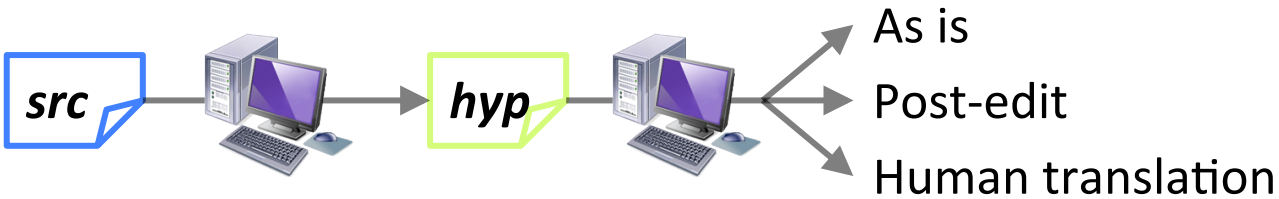


Prevalent approaches: Quality Estimation

- 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]

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

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



Limitations of existing datasets

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

Prior art	Translation direction	Domain	MT system	# of segments		
				train	dev	test
[Potet+, 12]	Fr → En	news	SMT	10,881	0	1,500
[Bojar+, 15]	En → Es	news	SMT	11,271	1,000	1,817
[Bojar+, 16]	En → De	IT	SMT	12,000	1,000	2,000
[Bojar+, 17]	En → De	IT	SMT	23,000	1,000	2,000
	De → En	pharmaceutical	SMT	25,000	1,000	2,000

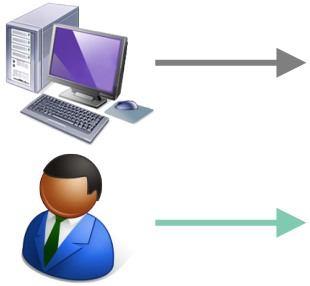
- Only PE-based QE
  - Word-level QE: OK/BAD tags generated from (***hyp***, ***pe***) pairs
  - Sentence-level QE: HTER computed from (***hyp***, ***pe***) pairs
- Only written texts

Can we rely on HTER to decide which ***hyp*** to be delivered?

Our 6-step procedure for corpus construction

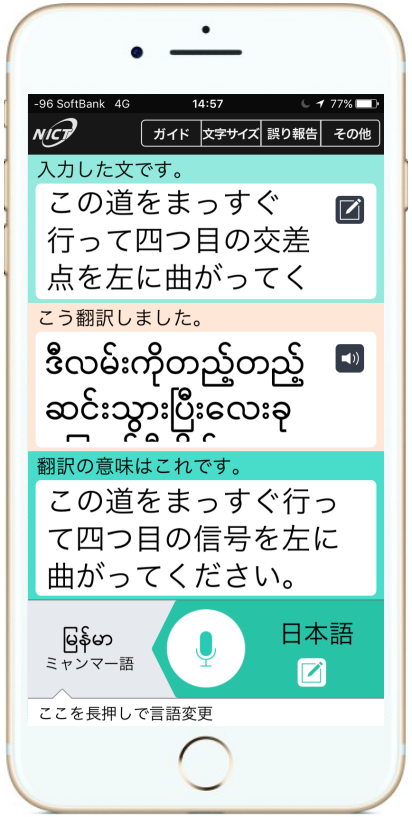
- Cover Asian languages
- Cover spoken language
- Conduct also manual grading

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, VoiceTra
  - 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

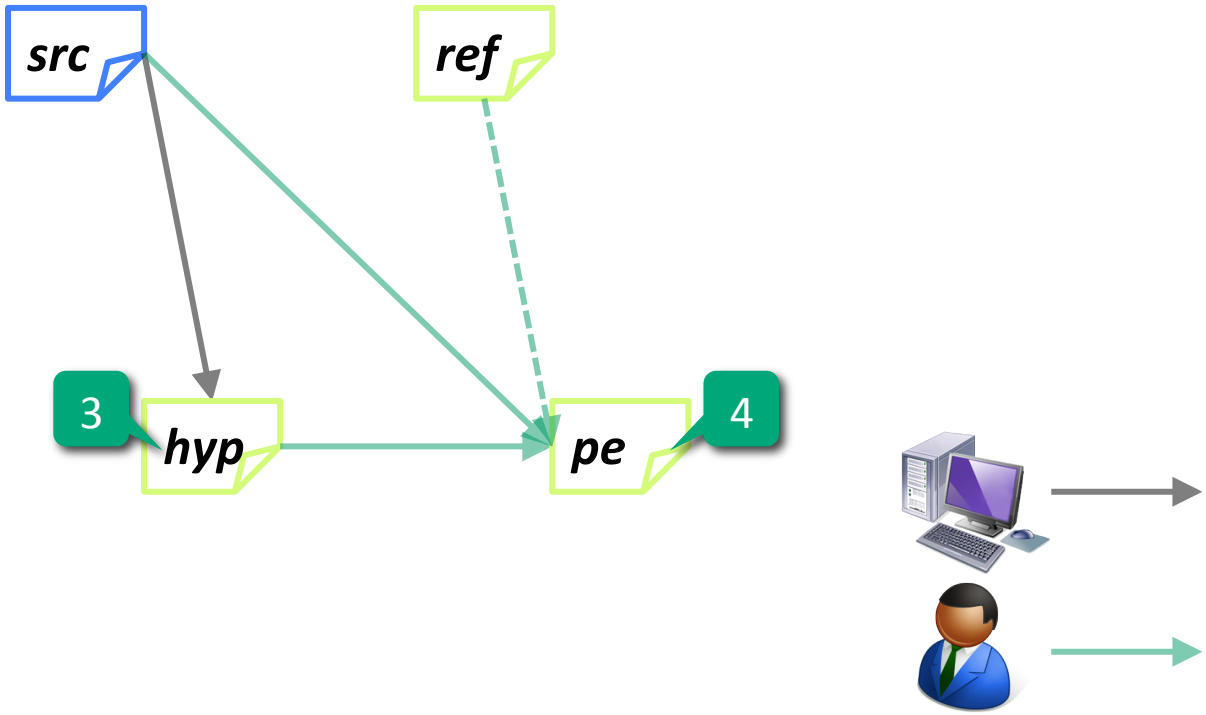
- Native speakers of each target language
- Instructions
  - Translate each segment
  - Assuming reasonable context
    - But without adding too much information
  - Only one ***ref***, even if the ***src*** has more than one interpretation
    - e.g., filling zero-anaphor (he/she)
    - e.g., specifying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

src 片道　だけ　で　買え　ます　か　。

(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

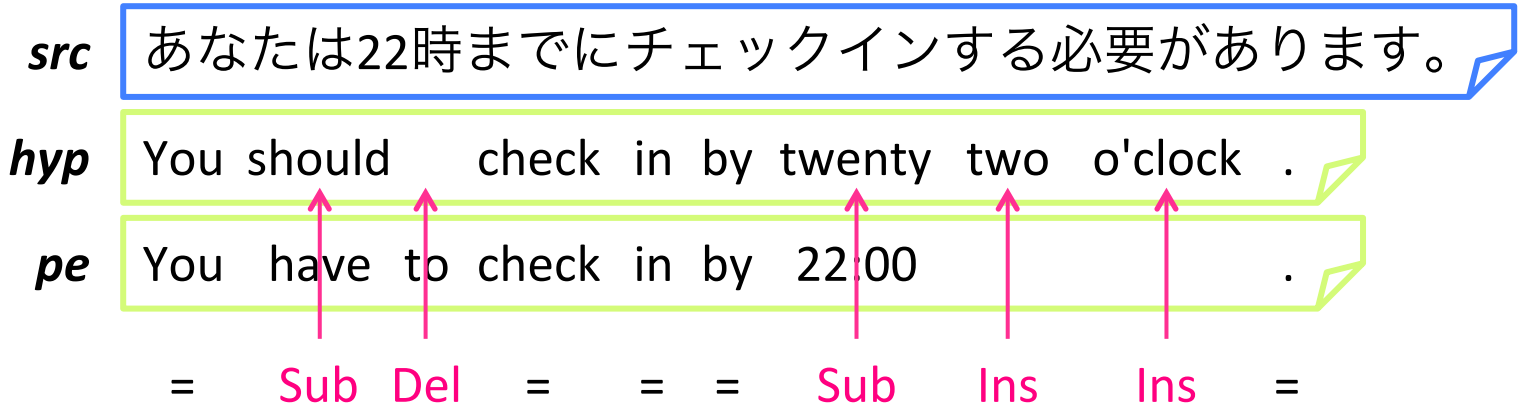
- Step 3: ***src*** → ***hyp*** using phrase-based SMT systems

Translation direction	MT system	Model	# of sentences for training
Ja → En	SMT	2013/Oct.	736k
Ja → Zh	SMT	2016/Sep.	1.44M
Ja → Ko	SMT	2016/Sep.	1.40M

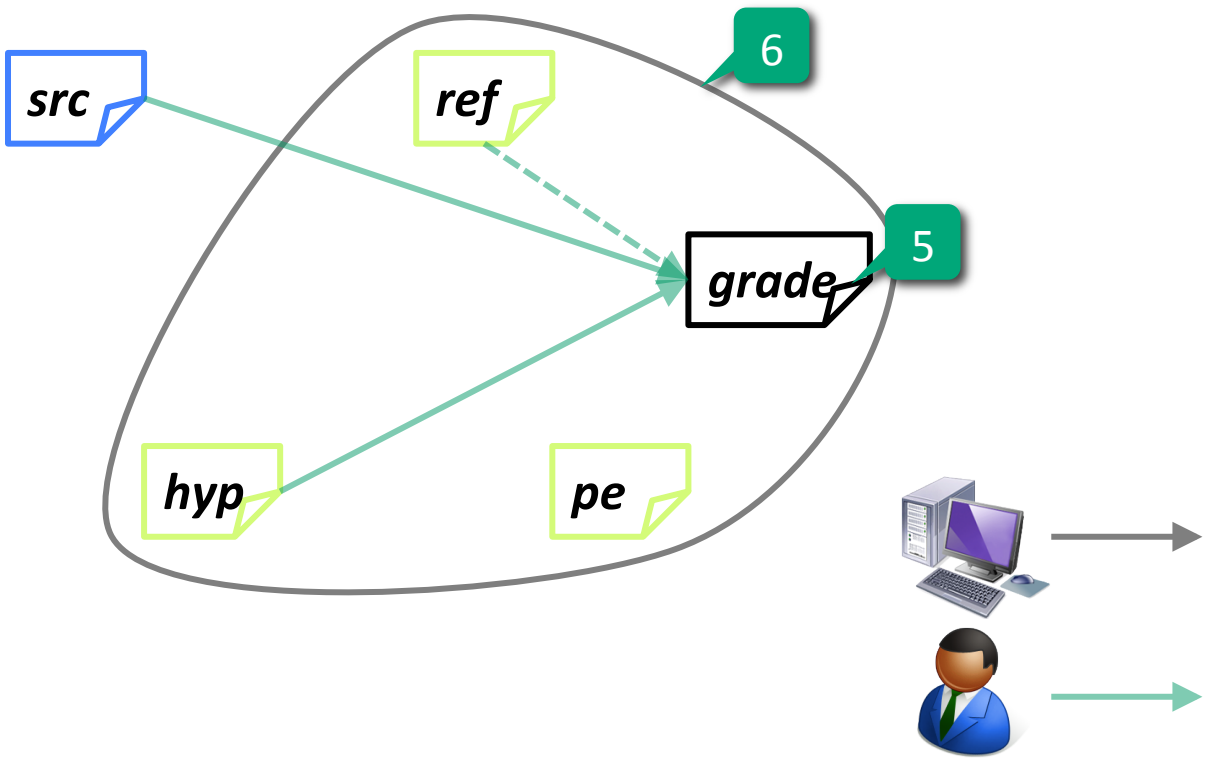
- Step 4: ***src***, ***hyp***, and ***ref*** → ***pe***
  - Correcting ***hyp*** with minimum number of edits [Snover+, 06]
    - Deletion of a word (Del)
    - Insertion of a word (Ins)
    - Substitution of a word (Sub)
    - Shift of a word or a phrase (Shift)
  - A lenient criterion: TER(***hyp*** → ***pe***) ≦ TER(***hyp*** → ***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)



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)
  - A: Good
  - B: Fair
  - C: Acceptable
  - D: Incorrect

Correct (goal of PE)

Some errors (need PE)

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.

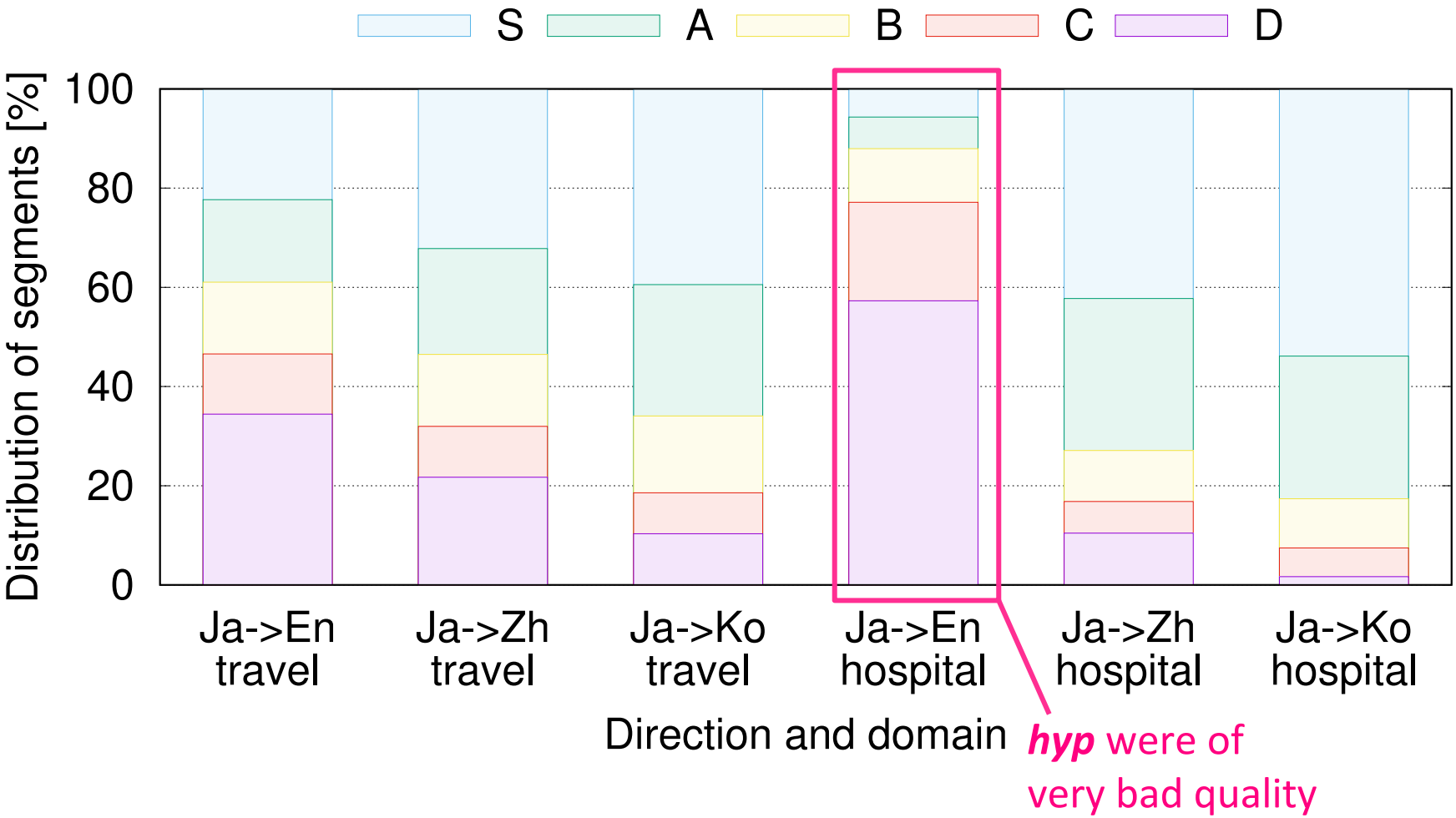
Step 6: Consistency check

- Resolve the discrepancies
  - If (***grade*** is “S” or “A”) && (***pe*** != ***hyp***)
    → then, retry both grading and post-editing
  - If (***grade*** is “B” or “C” or “D”) && (***pe*** == ***hyp***)
    → then, retry both grading and post-editing

- Minimality of post-edits is assessed again (if necessary)
  - If TER(***hyp*** → ***pe***) > TER(***hyp*** → ***ref***)
    → then, retry post-editing

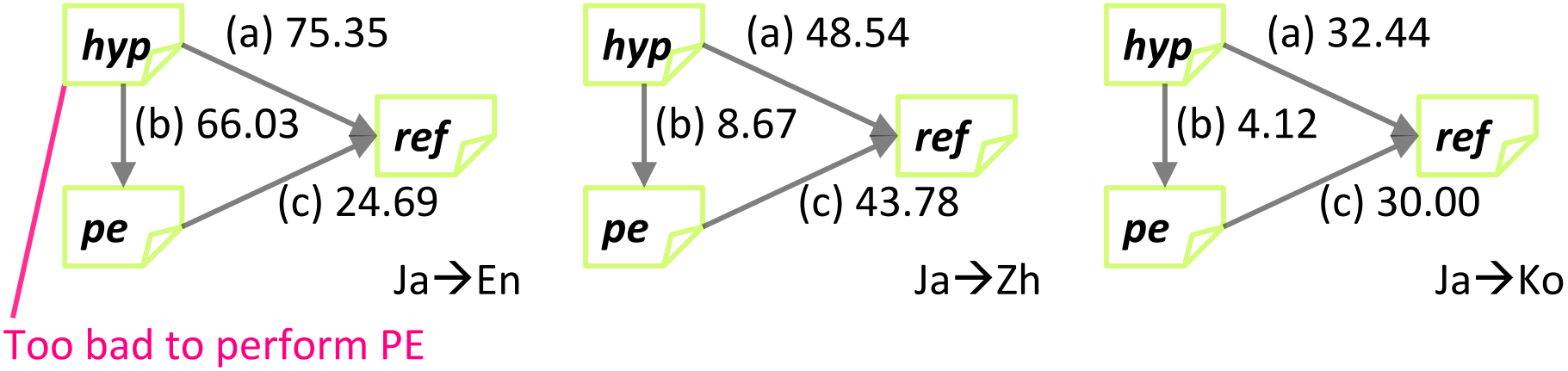
Analyses of the created datasets

Results of manual grading



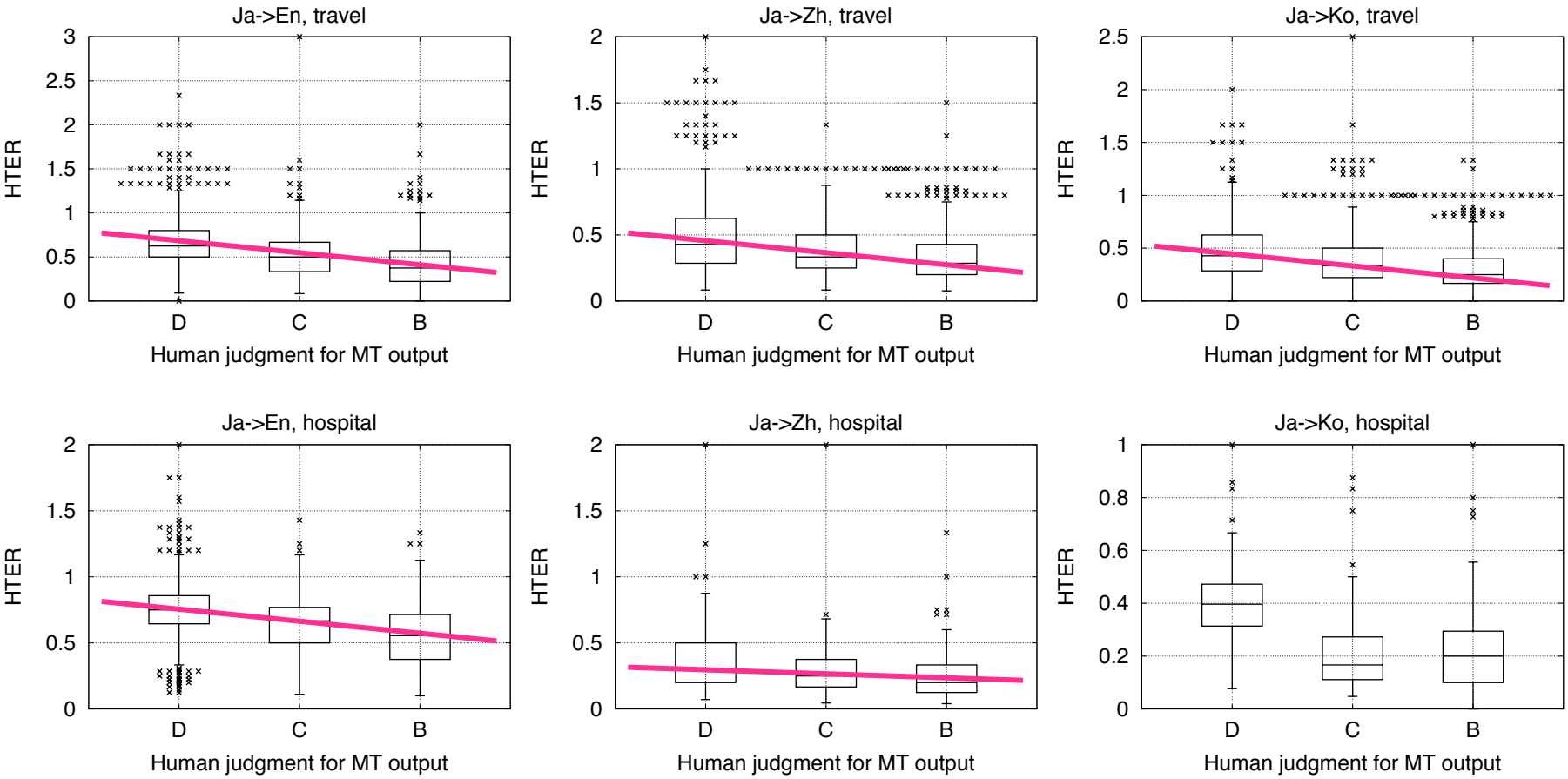
Proximity between translations

- 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

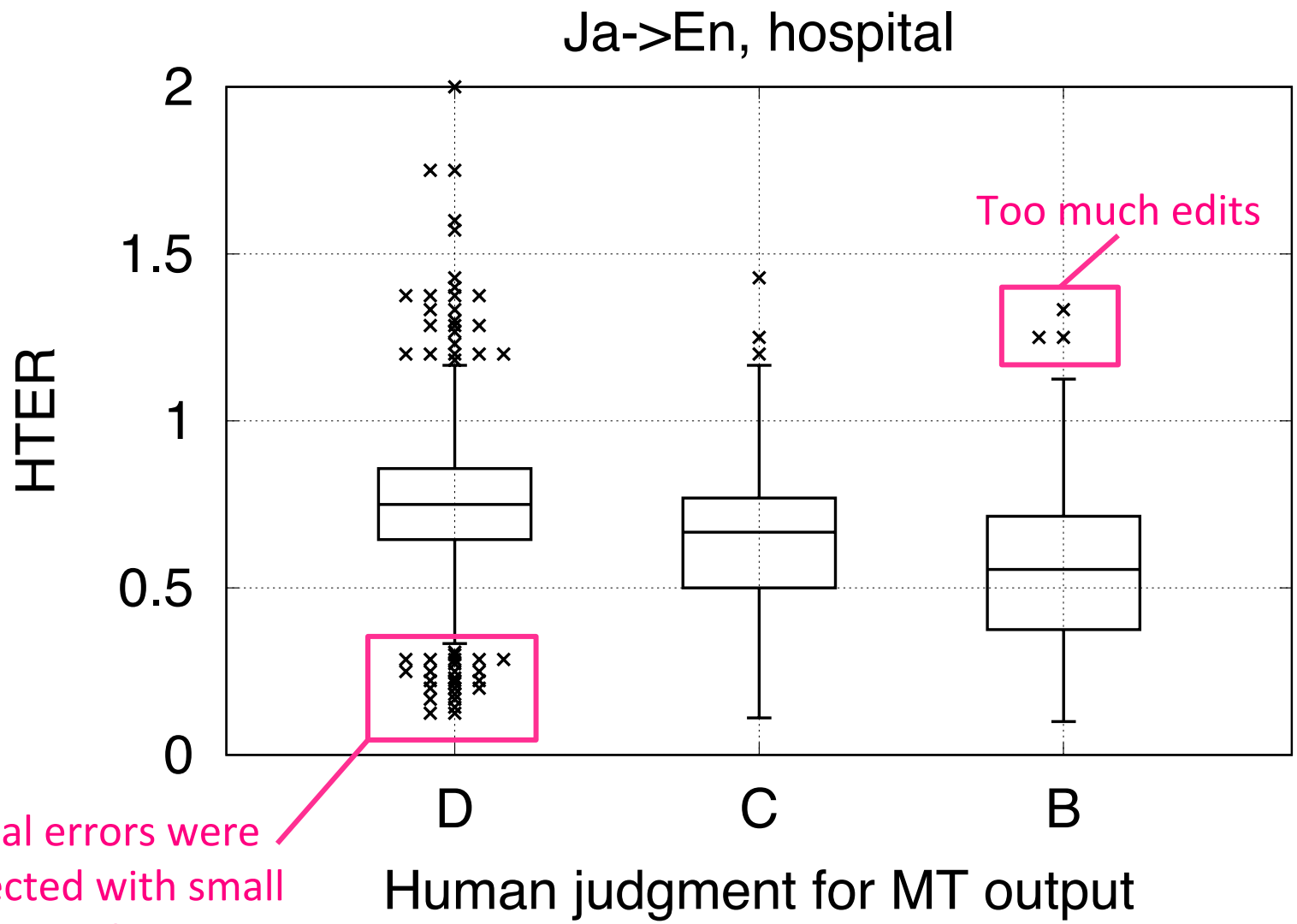


Distribution of sentence-wise HTER score

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



HTER vs. *grade*



Critical errors were corrected with small numbers of 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 首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

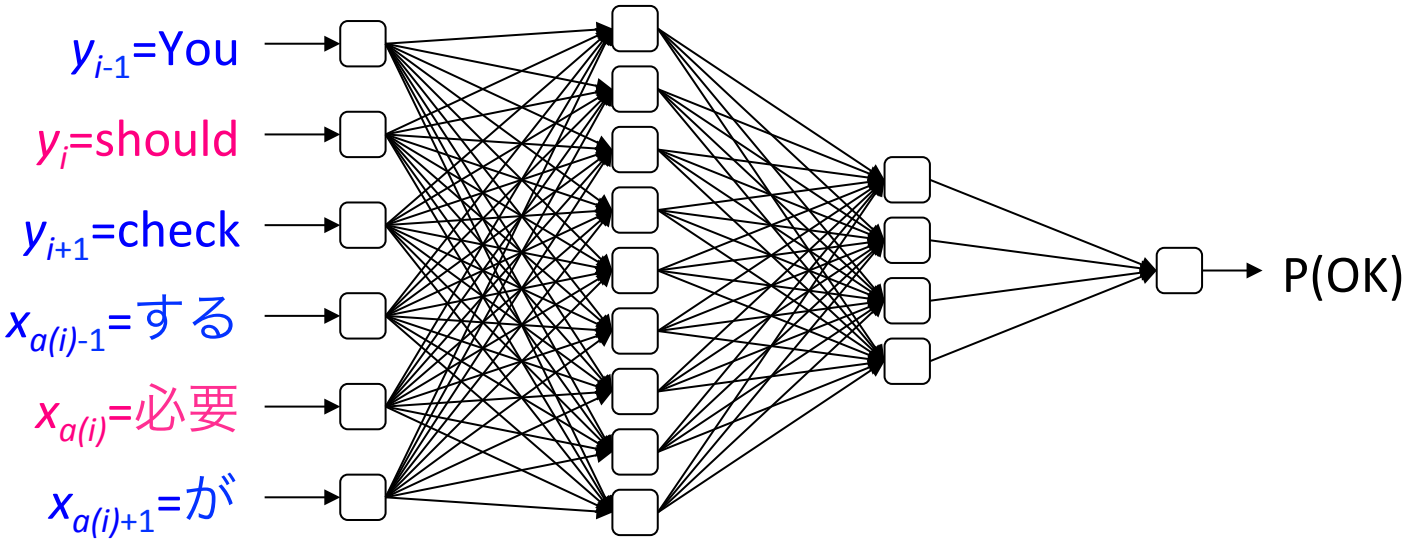
Partition	# of sentences	# of tokens
train	1.57M	20.1M-2.51M
dev	14k	179k-224k

Word-level QE: Methods

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

src あなた は 22時 まで に チェックイン する 必要 が あり ます 。

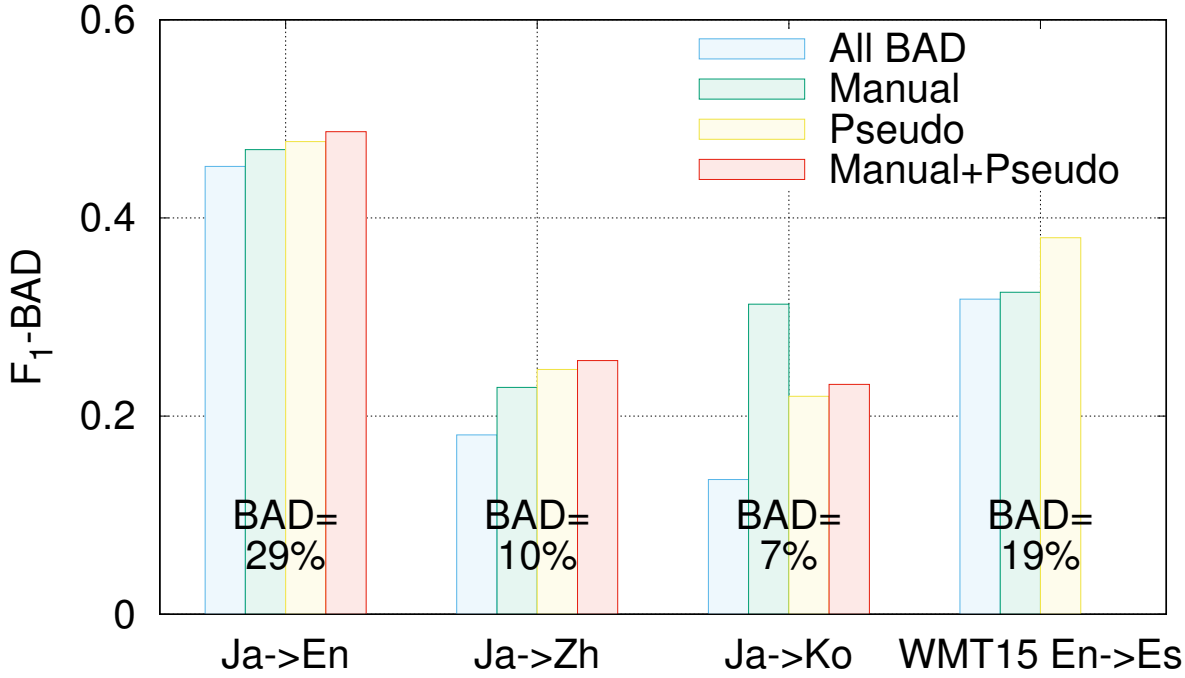
hyp You should check in by twenty two o'clock .



- Use of large pseudo training data
  - generated from (***hyp***, ***ref***) pairs [Liu+, 17]

Word-level QE: Results

- On F_1 score for BAD tags (F_1-BAD)
- The less the BAD tags were, the lower F_1-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



From [Liu+, 17]

Sentence-level QE: Methods

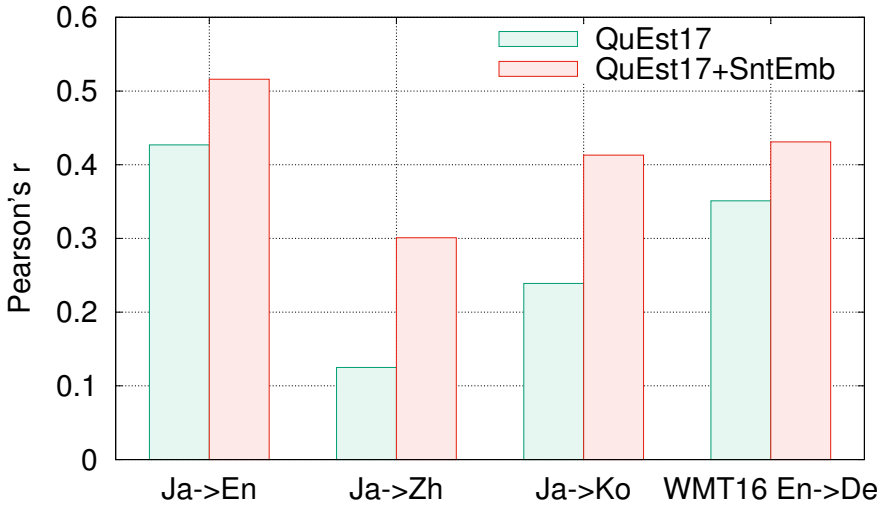
- Two variant tasks
  - Prediction of HTER as regression [Bojar+, 14;15;16;17]
  - Binary classification
    - No PE is needed: “S” and “A”
    - PE is needed: “B”, “C”, “D”

- 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]

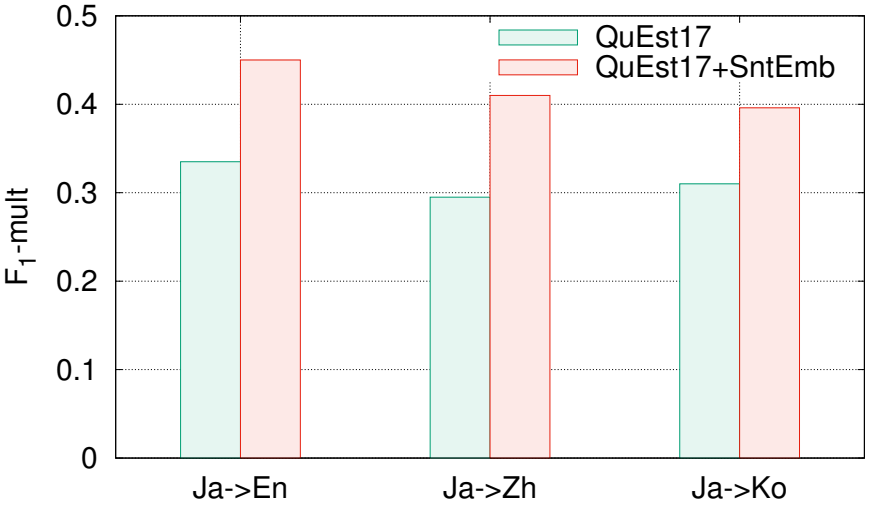
Sentence-level QE: Results

- SntEmb improved the performance of QuEst features

  - Prediction of HTER
    - No naïve baseline
    - On Pearson's cor. Coeff.



  - 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) → 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

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

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?