TOIN

Will Neural MT be a Breakthrough in English-to-Japanese Technical Translation?

Tsunao Mikasa and Nobuko Kasahara, TOIN Corporation

September 2017

Agenda

- **Question**: Is MT really usable for English-to-Japanese translation?
- **Pilot project** we carried out for assessing quality and productivity
  - Overview
  - MT engines examined
  - Methodology and assumptions
  - Results
- **Conclusion**

QUESTION

Is MT usable for English-to-Japanese (E2J) translation services where the required quality is at the same level as Human Translation (HT)?

- Until recently, the answer was **NO**; to obtain certain productivity gains in post-editing, quality of final translation needed to be compromised
- In other words, only “Light PE” was worth considering, and “real” translation was achievable only by human translators with no help of “machine translators”

QUESTION (CONTD.)

Is MT usable for English-to-Japanese (E2J) translation services where the required quality is at the same level as Human Translation (HT)?

- We claim that the answer will be **YES** if using the latest MT technologies, in particular **neural** engines (under some reasonable assumptions about content types)
- In other words, MT will enable most E2J translators to achieve the **same quality** without compromise at **higher productivity** (except for some special content types, such as marketing materials)

PILOT PROJECT

To examine our claim, we carried out a simple pilot project for accessing quality and productivity in Human Translation (HT) and Post-Editing (PE)

Key Assumptions:

- We focused on **Technical** documents, as this sector accounts for the largest portion of many language service providers in Japan
- PE quality was required to be **the same level as HT**, since our interest was in examining whether HT quality can be achieved by PE without any compromise in quality (not "Light PE")

MT ENGINES EXAMINED

We examined two engines which are recognized as ones of the best Neural and Statistical English-to-Japanese MT engines:

- **Google NMT**—Neural
- **NICT みんなの自動翻訳@TexTra®** —Statistical

  (NICT: National Institute of Information and Communications Technology 情報通信研究機構)

Note: NICT has recently also released its Neural engine

TOIN

METHODOLOGY AND ASSUMPTIONS

Content translated: A typical technical document, User Manual of a major PLM software product

- **Not too technical**, easy-to-understand for the average user (and for translators!)

Volume:

- **5k** words for PE/HT productivity evaluation
- Additional **10k** words for MT quality evaluation

TOIN

METHODOLOGY AND ASSUMPTIONS (CONTD.)

Sample segments:

			Translator KH			
Segm	Source Segment	MT Target Segment	MT Engine	PE-er	Post-edited Target	Translated Target
245	The action is only available when creating or editing a change task.	この操作は、変更タスクを作成または編集するときに使用できるようになります。	NICT	KH	この操作は、変更タスクを作成または編集するときにのみ使用できます。	
246	The action is only available when you access the Resulting Objects table from the change task information page.	操作は、変更タスクの情報ページの「結果オブジェクト」(Resulting Objects)テーブルにアクセスした場合にのみ使用できます。	NICT	KH		この操作は、変更タスク情報ページから「結果のオブジェクト」テーブルにアクセスするときにのみ使用できます。
247	Open a new window to edit the change task.	新しいウィンドウが開き、変更タスクを編集します。	NICT	KH	新しいウィンドウを開き、変更タスクを編集します。	
248	Set effectivity on an object.	オブジェクトのエフェクティビティを設定します。	NICT	KH		オブジェクトで有効性を設定します。
249	View effectivity on an object.	オブジェクトのエフェクティビティを表示します。	NICT	KH	オブジェクトの有効性を表示します。	

METHODOLOGY AND ASSUMPTIONS (CONTD.)

Resources—Linguists (Translators/Post-Editors) who worked in the pilot:

- **Four** senior-level linguists with 10+ year-experience in E2J technical translation
- Past experience in PE was **not** required (though two of them did have some PE experience)
- Each of them translated/PE'd the same 5,000-word sample document
- They focused on achieving sufficient (HT-level) quality in PE; never forced to use MT outputs or "hurry up" in PE

METHODOLOGY AND ASSUMPTIONS (CONTD.)

Linguistic reference:

Made linguistic reference as **simple** as possible to see the pure impact of MT on quality and productivity:

- No Translation Memory (TM)
- No Terminology Database (TD)
- No Style Guidelines (SG)

METHODOLOGY AND ASSUMPTIONS (CONTD.)

Piot project for Productivity evaluation

- Each linguist produces a translation of each segment, either by
  - **HT**: translating the source segment without referring to any MT outputs, or
  - **PE**: editing MT output of the source segment
  - To do HT or PE is randomly chosen by the system so that the total # of HT/PE'd segments will be equal

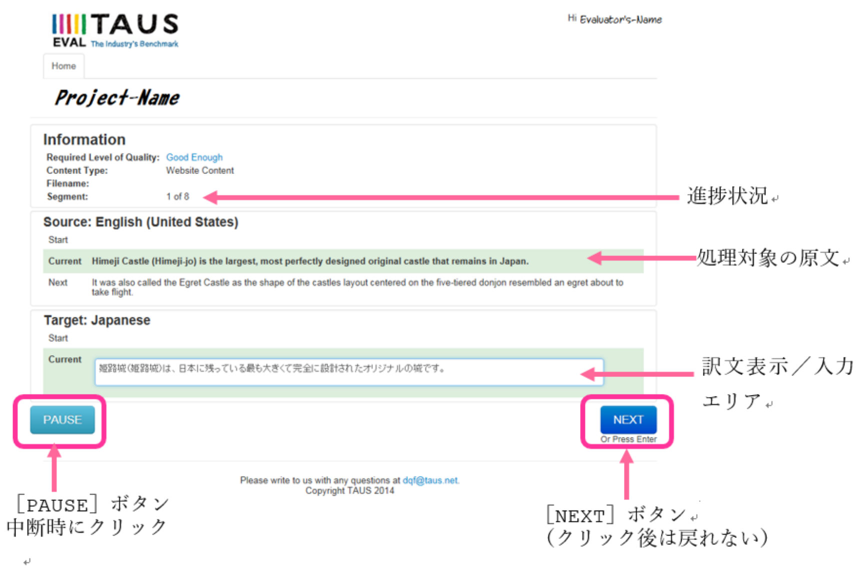
METHODOLOGY AND ASSUMPTIONS (CONTD.)

Piot project for Productivity evaluation (contd.)

- For **PE**, either **GNMT** or **NICT** engine applied
  - Randomly chosen by the system so that the total # of the segments from each engine will be equal
  - Not make it visible to the linguist which engine was used (to avoid any bias)
- We used TAUS DQF tools for productivity evaluation

Methodology and Assumptions (contd.)

Post-editing on TAUS DQF tools:



Methodology and Assumptions (contd.)

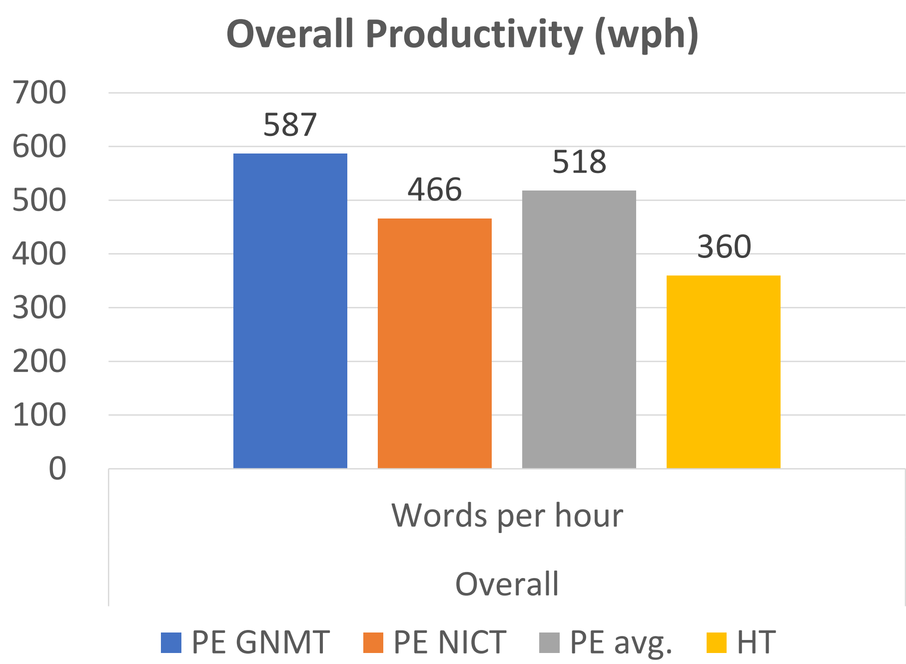
Quality evaluation of raw MT outputs

- Evaluated quality of raw MT outputs of **GNMT** and **NICT** engines
  - Randomly chosen by the system so that the total # of the segments from each engine will be equal
  - Not make it visible to the evaluator which engine was used (to avoid any bias)

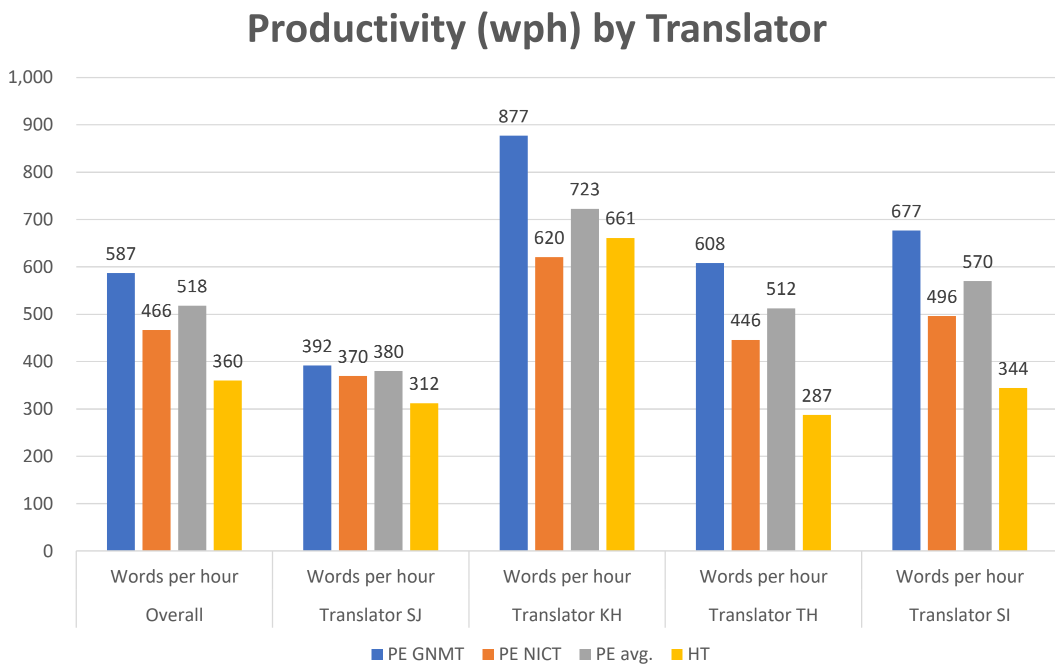
Methodology and Assumptions (contd.)

- We used TAUS DQF tools and their evaluation criteria for quality evaluation
  - **Fluency**
    - **Flawless** (4) —refers to a perfectly flowing text with no errors.
    - **Good** (3) —refers to a smoothly flowing text even when a number of minor errors are present.
    - **Disfluent** (2) —refers to a text that is poorly written and difficult to understand.
    - **Incomprehensible (1)** —refers to a very poorly written text that is impossible to understand.
  - **Adequacy**
    - **Everything** (4)—All the meaning in the source is contained in the translation, no more, no less.
    - **Most** (3)—Almost all the meaning in the source is contained in the translation.
    - **Little** (2)—Fragments of the meaning in the source are contained in the translation.
    - **None** (1)—None of the meaning in the source is contained in the translation.

Results—Productivity



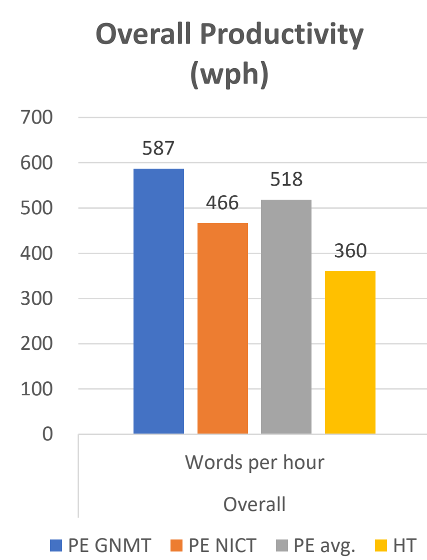
RESULTS—PRODUCTIVITY (CONTD.)



RESULTS—PRODUCTIVITY (CONTD.)

Key findings:

- PE w/ GNMT
  - Highest productivity
  - 63% faster than HT on average
- PE w/ NICT
  - 30% faster than HT on average

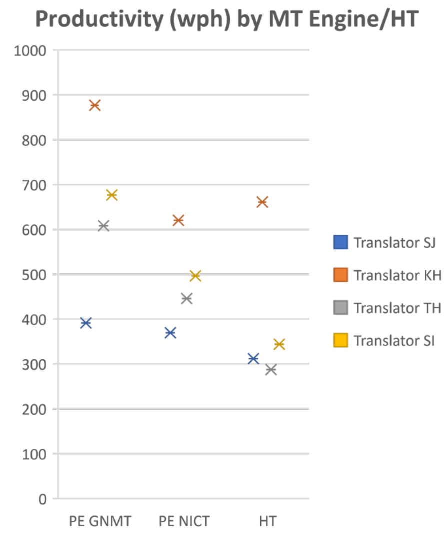


Results—Productivity (contd.)

- **PE GNMT > PE NICT > HT**
  —The same tendency observed almost **independent of the translator**
- **Correlation ratio** between **Productivity** and **MT Engine/HT:**
  η = **0.57**

$$\eta := \sqrt{\frac{\sum_{i=1} n_i (\overline{x_i} - \overline{x})^2}{\sum_{i=1} \sum_{j=1}^{n_i} (x_{ij} - \overline{x})^2}}$$

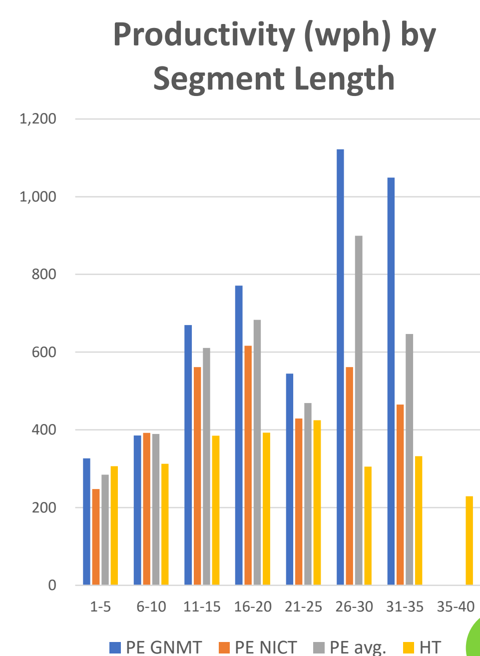
$$(0 \leq \eta \leq 1)$$



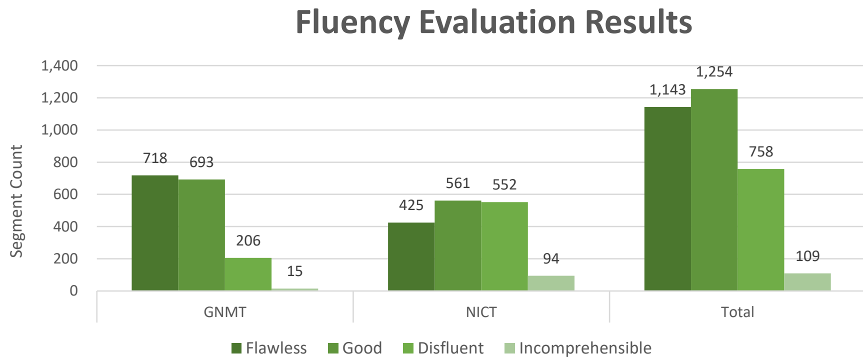
Results—Productivity (contd.)

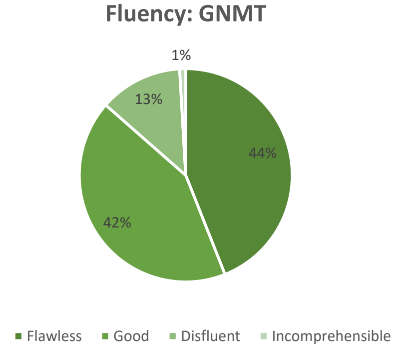
Other observations:

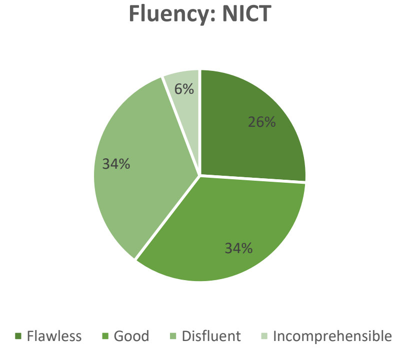
- **No** apparent correlation observed between **Productivity** and **Segment Length** (word count of each segment)
- In particular, in **HT**, SL does not seem to affect Productivity at all
- **GNMT** seems to show a slight tendency that the **longer SL**, the ***higher* productivity**, but it's not significant



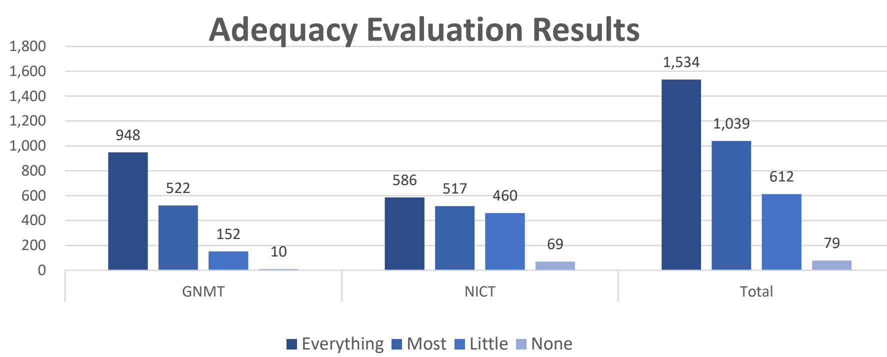
RESULTS—QUALITY: FLUENCY

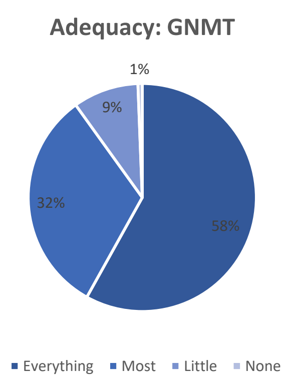


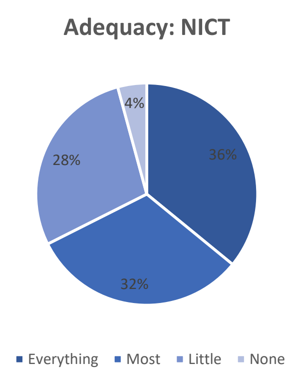




RESULTS—QUALITY: ADEQUACY



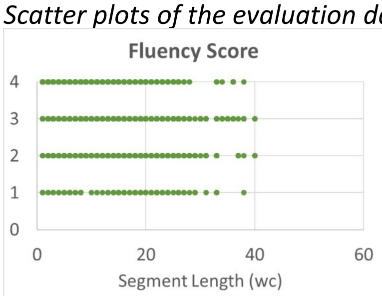


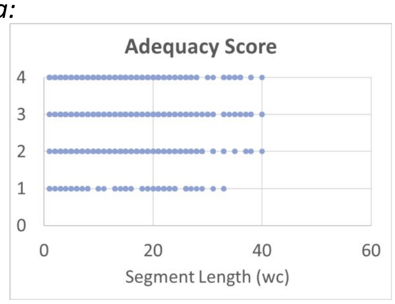


Results—Quality

- **Key findings**
  - **GNMT** had better scores overall, where
    - **+85%** segments had ***Flowless* (4) or *Good* (3) fluency**
    - **+90%** segments had ***Everything* (4) or *Most* (3) adequacy**
- **Other observations**
  - Almost **no** correlation observed between **Segment Length** and **Quality** of MT outputs in our pilot:

Scatter plots of the evaluation data:





Conclusion

Productivity gains

- We observed **63%** average productivity gains in **PE w/ GNMT** as well as **30%** gains in **PE w/ NICT.**
- This strongly suggests that significant improvement in efficiency can be achieved in most E2J technical localization projects by utilizing the latest MT engines, in particular, GNMT, in the translation process.

Other findings

- In our pilot, we didn't observe the tendency "Longer sentences give worse MT outputs, thus result in lower PE productivity", which may be a myth.