On the Value of Machine Translation Adaptation

Stacey Bailey and Keith Miller

LREC Workshop: Automatic and Manual Metrics for Operational Translation Evaluation (MTE 2014)

May 26, 2014

MITRE

MT, DT, DTMT

- **Domain tuning (DT): Tailoring a machine translation (MT) system to (ideally) better translate the kinds of data users want to translate.**
  - Training MT using data that is representative of a topic's subject matter (e.g., scientific and technical literature)
  - Adding terminology that is relevant to that subject matter

- **Does DTMT improve the quality of MT output?**
  - *What do you mean by 'improve'?*
- **If so, what's the best way to go about DT?**
  - *What do you mean by 'best'?*
- **What are the trade-offs for different approaches to DT?**
  - *Oh, and is it worth the effort?*

What We Collected

- Chinese-English parallel texts in the cyber domain
- 10 abstracts (with a total of roughly 70 segments) were selected for a task-based evaluation.
- Three additional translations were produced for each test segment for a total of four reference translations.

What We Thought

- **Hypothesis 1**. DT improves the quality of machine translation output over baseline capabilities, as measured by automatic evaluation metrics.
- **Hypothesis 2.** Human translation time can be reduced by requiring human translators to post-edit the output of DTMT systems.
- **Hypothesis 3.** The linguistic quality of the target language document can be improved by requiring human translators to post-edit the output of DTMT systems.

What We Measured

- Automatic evaluation across all four domains
  - MT evaluation scores with and without DT
    - BLEU, METEOR, ROUGE, WER, TER
- Task-based evaluation for a subset of the Cyber data
  - MT evaluation scores for 10 experimental conditions (next slide)
    - BLEU, METEOR, ROUGE, WER, TER
    - HTER (Human-mediated translation error rate)
    - Relative translation rates – Total time to complete a translation (in minutes) relative to the length (number of characters) of the source Chinese text
    - Translator judgments – Questionnaire responses about MT utility and quality
    - Quality control (QC) judgments for translated segments
      - Ratings for meaning and grammaticality
      - Counts of missing and incorrectly translated terms

What We Tested

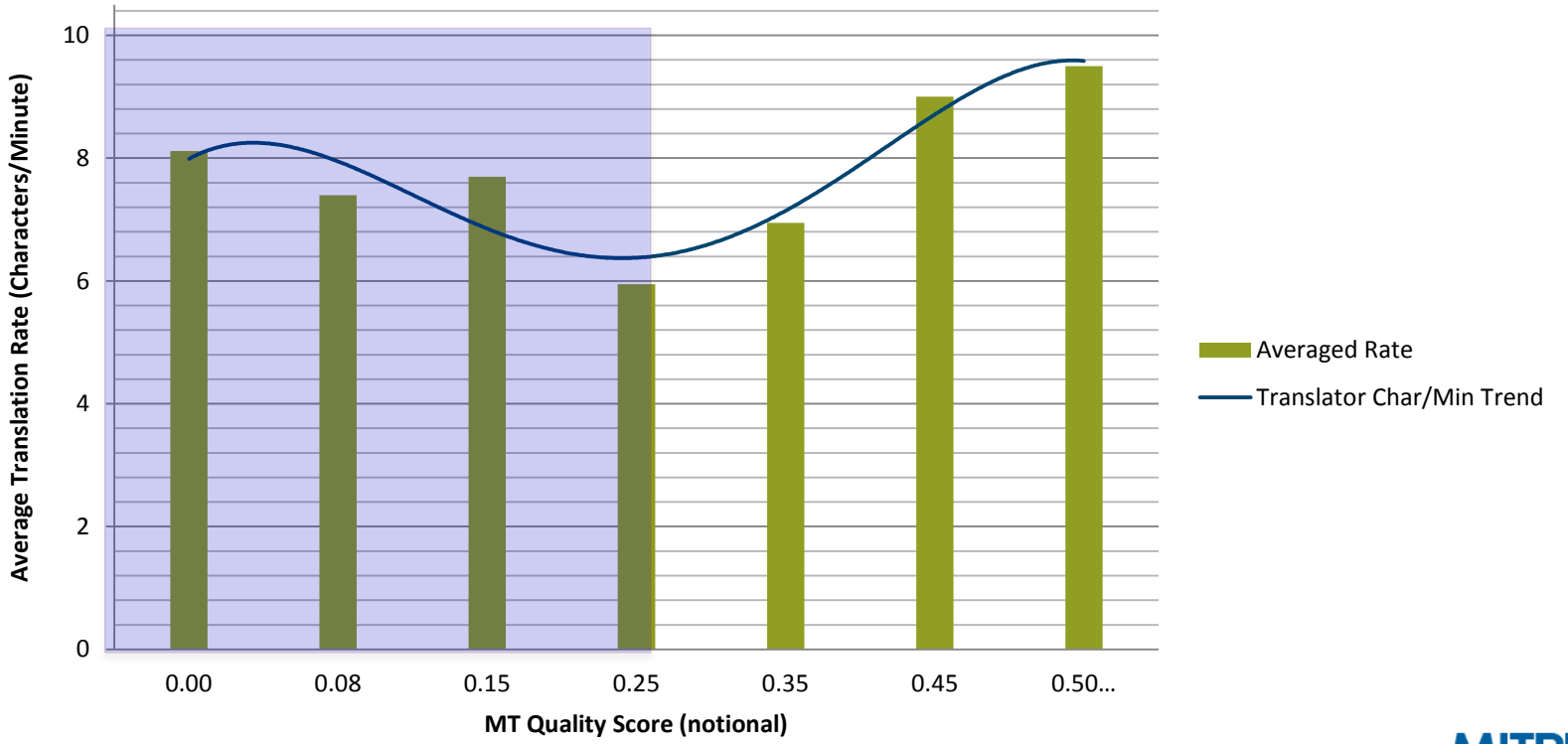
- Types of DT:
  - Statistical MT models: Training a custom engine using parallel, domain-specific data
  - Domain-specific glossaries: Lightweight adaptation with terms
- The MT systems: Two commercial off-the-shelf MT engines
- The experimental conditions:
  - For each MT system
    - **Baseline**. Translation using the MT engine without any DT.
    - **Lexicon**. Translation using the MT engine plus lightweight DT with a found domain-specific lexicon.
    - **Parallel Data**. Translation using the MT engine plus a statistically retrained engine based on the training data.
    - **Lexicon + Parallel Data**. Translation using the MT engine with a statistically retrained engine and a found lexicon.
  - Plus:
    - **Source Language + Term Highlighting**. Manual translation that does not use MT but does use a domain-specific lexicon for highlighting found terms with glosses
    - **Source Language Text Only**. Manual translation with no MT or term highlighting.

What We Found

- **Hypothesis 1**. DT improves the quality of machine translation output over baseline capabilities, as measured by automatic evaluation metrics.
  - DT did improve the automatic scores, but the BLEU scores never improved above the mid 20s.
- **Hypothesis 2.** Human translation time can be reduced by requiring human translators to post-edit the output of DTMT systems.
  - Results neither support or refute this hypothesis. Translation rates were faster for some MT conditions and slower for others.
  - Translators were slowest on output from the two MT conditions with the most involved DT (i.e., DT combining an MT engine trained on domain-specific data and a domain-specific glossary).
  - Translation times for source texts enhanced with term glosses were about the same or slower than those for translating the source texts from scratch.
- **Hypothesis 3.** The linguistic quality of the target language document can be improved by requiring human translators to post-edit the output of DTMT systems.
  - There was a clear pattern of improving (i.e., decreasing) HTER scores with customization for one MT engine, but a mixed pattern with the other MT engine. But this does not speak to the final translation quality after post-editing.
  - Translators' own opinions of the quality of the resulting translations were largely that MT had no effect.
  - Additional adequacy, fluency and error data have been collected but not yet analyzed.

What We Suspected

- Baseline MT might be of such low quality that the meaning may simply not be recoverable from the output provided.
- Translators have no option but to start from scratch, even when asked to post-edit.
- DT might improve the quality of the MT output but only enough to cause translators to take more time to consider the output rather than discarding it out of hand.



MITRE

What We Did Next

- The work raised more questions than it answered about the relationship between automated metrics, MT quality and the usefulness of MT for operationally focused tasks, so we…
  - proposed a workshop!
- (We also began follow-on work with a different language pair and data that might put us on a different starting point on the quality-productivity curve.)